Normative Acceptance of Power Abuse*

Leonard Hoeft[†] **Wladislaw Mill**[‡§] **Alexander Vostroknutov**[¶]

August 2022

Abstract

Abuse of institutional power is an issue that plagues economic efficiency even in developed countries. We hypothesize that the stability of this phenomenon hinges on its normative acceptance by all parties involved, even those disadvantaged by it. In a laboratory experiment, we create an environment conducive to unfair exploitation and study the normative perceptions of participants placed in a position of power and those who suffer from its abuse. We find that participants exposed to abuse start to believe that free-riding and punishment on the part of a powerful participant are socially acceptable. The participants who abuse their power also indicate that free-riding, while forcing others to cooperate, is not inappropriate. We find that the observed change in normative beliefs operates on the level of social norms, but not personal norms that remain unchanged by the experience of abuse. Thus, we find that power abuse may lead to pluralistic ignorance. Our study suggests that the human capacity to exculpate abusive behavior can be behind the stability of corrupt institutions.

JEL classifications: C91, C92, K42, H41, D73.
Keywords: power abuse, social norms, personal norms, public goods, punishment, experiments.

*We would like to thank Benedikt Werner, Christoph Engel, Eugenio Verrina, Nan Zhang, Angelo Romano, Yan Chen, Nora Szech, Ro'i Zulta, Ben Greiner, Simeon Schudy, Menusch Khadjavi, Oliver Kirchkamp, Martin Kocher and the participants of the Behavioral and Experimental Economics Network seminar in Rome (Sep 2018), the Center for Empirical Research in Economics and Behavioral Science at the Erfurt University, the Kiel Behavioral Economics Seminar, the MPI Bonn Research Seminar and participants of the EEA in Manchester, THEEM in Konstanz, IMEBESS in Utrecht, MBEES in Maastricht, Meeting of the standing field committee for social sciences of the German Economic Association for invaluable comments. We also would like to thank the anonymous Associate Editor, and three reviewers for their insightful comments, which helped to substantially improve this paper. We gratefully acknowledge funding from the Max-Planck Society and the IMPRS-Uncertainty. All mistakes are our own.

[†]Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Straße 10, 53113 Bonn Germany. e-mail: hoeft@coll.mpg.de

[‡]Department of Economics, University of Mannheim, L7 3-5, 68131 Mannheim, Germany. e-mail: mill@uni-mannheim.de

[§]Corresponding author.

[¶]School of Business and Economics, Maastricht University, Tongersestraat 53, 6211 LM Maastricht, Netherlands. e-mail: a.vostroknutov@maastrichtuniversity.nl

1 Introduction

Recent developments in various social sciences including sociology (Bicchieri, 2016), human evolutionary biology (Henrich, 2017), and more recently economics (Fehr and Schurtenberger, 2018), have led to a broad consensus that, contrary to the predictions of standard models with selfish preferences, people largely act in a pro-social manner (Schroeder and Graziano, 2015), which is backed by a fairly uncontroversial norm proscribing selfishness (Cubitt et al., 2011; Krupka and Weber, 2013a).

Nevertheless, unjust conditions and behavior are pervasive and hard to eradicate. Even developed countries with functioning legal and social systems witness high inequality and unfair distribution of power (Acemoglu et al., 2015; Rose-Ackerman and Palifka, 2016). Indeed, much of the policy debate involves arguing that some part of society is disproportionately favored, thus failing to contribute to the community: they essentially "play a rigged game" (Acemoglu and Robinson, 2008; Dal Bó et al., 2009). This is true despite the fact that most modern societies feature institutions that encourage prosocial behavior. Examples of unjust behavior include politicians using their position to obtain irregular benefits or, on a smaller scale, managers forcing their coworkers to invest in shared projects that they themselves skimp on.

The mismatch between individual prosociality and the persistence of certain kinds of corruption among the powerful may originate from the differences in the normative perception of wrongful acts. People, who are abused by the powerful, may internalize such experiences in the form of norms that they believe are in place. People who abuse their power may also believe that their behavior is not inappropriate (due to self-serving bias, motivated reasoning, or something else). Either way, no one in the abusive relationship may believe that something is not right. This idea finds support in a recent World Bank report (World Bank Group, 2017), which claims that top-down attempts at fighting corruption fail due to social norms that support it on all levels of hierarchy. Thus, in this paper we are interested in investigating the influence that the experience of power abuse may have on the normative beliefs of people who experience it and discuss what it can imply in general for the efforts to change inefficient institutions.

We tackle this question by experimentally investigating one specific instance of what we define as "power abuse" and its causal relationship with normative perceptions by participants experiencing it. We implement a repeated Public Goods game (PGG) that allows one powerful participant (punisher), who fulfills the role of a sanctioning authority, to dictate contribution norms, while being free to exempt himself from them (Hoeft and Mill, 2017a). This implies that the punisher can "abuse power" by punishing others for low contributions while not contributing anything himself (and not abuse power by contributing the same amounts as others). Thereby, our design maintains the inherent ambiguity of power abuse as the powerful may still play a

vital role in society (by increasing contributions of others) even though they exploit their power (by not contributing themselves).[1]

In order to understand how the powerful and the powerless perceive power abuse we elicit their beliefs about the prevalent norms of behavior *in their own reference group*. This means that punishers are only asked about the norms they believe exist among other punishers and symmetrically subjects who are not punishers are only asked about norms in the group of other such subjects. This is an important detail, as punishers and other subjects have different roles and different experiences in the game.

We find that participants who have experienced abuse start to believe that it is more socially appropriate for punishers to free-ride and to punish others than participants whose punisher contributed the same or more than them. We also find that punishers who undercontribute perceive this behavior to be more socially appropriate than punishers who do not. Both findings confirm our initial intuition that abusive behavior may be stable due to norms that surround it.

In the follow-up experiment, we have determined that the effect of changing normative beliefs among people who experience abuse works on the level of *social norms*, but not on the level of *personal norms* (Bašić and Verrina, 2020). This means that abusive experiences (when the abusive person is not punished) make people believe that others accept such abuse, though people themselves may personally disagree with it. This points towards a mechanism responsible for the emergence of pluralistic ignorance (Bicchieri, 2016) and, in general, situations where social norms exist that few people personally support.

Our findings, which contribute to the growing literature on the formation and robustness of social norms, draw a rather grim picture in which the powerful abuse their position, believing that they have done nothing wrong, while the powerless suffer from the abuse but consider their situation normatively appropriate. If our results generalize to environments outside the laboratory, they can explain a relative stability of corrupt institutions, since no party involved feels that anyone is doing anything inappropriate. It would also explain why removing those who abuse power is not enough, as others who fill their roles may act similarly even if they were disadvantaged by that same behavior before. This happens because both parties follow what they believe is a social norm.

[1] In our experiment, the powerful (punishers) are chosen randomly, whereas in the real world people with power are often chosen through some measure of merit or some form of voting. Whether the powerful are elected or assigned randomly might have a substantial influence on normative beliefs of others involved in the PGG. We acknowledge that this is so but would like to emphasize that the acceptance of power abuse through adjusting normative beliefs is probably stronger in case a powerful official was elected than when she was chosen randomly. This is because if an official has been approved by the society one way or another (election), then all her actions will be viewed as more legitimate, given the trust bestowed on her, than the actions of some randomly chosen official (as in our experiment). Thus, if we manage to detect an effect on normative beliefs in our environment, it would make it even more likely that the acceptance of power abuse is going on in the real world where officials are chosen through elections.

Even though this mechanism of stability of abusive behavior may sustain inefficient institutions for a prolonged periods of time, there is a chance for successful policy as personal norms do not seem to be influenced by negative experiences as much as social norms do. This implies that campaigns directed at clarifying the situation with normative beliefs (explaining people that few of them personally support the abusive social norm) can change the norm in a positive direction and as a result reduce abusive behavior.

2 Literature

Our study speaks to different literatures in economics. From the macro perspective, the form of power abuse that we study is specific to modern societies with strong institutions. Typically, such societies prevent the direct and forceful subjugation or mistreatment of others, which is considered extremely morally inappropriate, and few such problematic behaviors are widespread. Even authoritarian states avoid clear moral violations and choose to veil subjugation of their subjects behind normative reasons (Beetham, 2013). Only a small percentage of the population openly violate fundamental norms of fairness and respect for basic human rights in direct interactions with others. Those who steal or harm others are quickly ostracized and are often considered to be antisocial or dangerous.

However, institutions that promote public welfare regularly create unfair opportunities for their functionaries at the expense of the general population. People with power recurrently exploit their positions in questionable ways (Kipnis, 1972; Maner and Mead, 2010). Such behaviors often take the form of hypocritical enforcement of institutional rules that the enforcers do not adhere to themselves. Examples include politicians using their influence to attain atypical benefits (Grant and Keohane, 2005; Olken and Pande, 2012), police officers using illegal violence (Wong, 1998), doctors using their connections for special treatment (p. 71-73 Klitzman, 2007), and managers forcing their coworkers to invest in shared projects that they themselves skimp on (Xu et al., 2015; Vredenburgh and Brender, 1998). The ubiquity of these kind of practices and the ostensible perception that they are less deleterious than direct harm may be explained by their indirect consequences and the dilution of norms determining appropriate behavior in complex institutions, which leads to the spread of normative uncertainty (Tremewan and Vostroknutov, 2020). Moreover, it is easy to make excuses on the grounds that, even though an individual with power might use his position for personal benefit, he still provides an important social service.

In spite of all this evidence that shows how detrimental corruption can be, the relationship between abuse of institutional power and its associated normative perceptions remains unclear for the most part. Does everybody agree on the norms regarding indirect harm and abuse of power? Do the abusers simply use their advantageous position out of selfishness, or justify their behavior as socially appropriate? Do participants exposed to the abuse stay true to their moral

convictions or assimilate bad norms after being exposed to corrupt institutions? Causal evidence is hard to come by in real world, which is why we turn to the laboratory. Unlike the established designs, where all players have the means to punish others (e.g., Fehr and Gächter, 2000), our game models the ambivalence of indirect abuse of power: not contributing while forcing others to do so is unfair, but enforcing high contribution norms is beneficial, even if the punisher does not himself comply. This allows us to study the phenomenon in laboratory conditions.

From the micro perspective, we contribute to the literature on cooperative behavior, punishment, and their normative underpinnings. In laboratory tasks, normative beliefs were shown to influence a wide range of behavior, from dictator game giving to cooperation, trust, discrimination, and corruption (Barr et al., 2018; Bicchieri et al., 2021; Gächter et al., 2017; Krupka and Weber, 2013b; Banerjee, 2016; Bicchieri and Xiao, 2009a). Normative expectations are closely tied to descriptive norms, as the concepts of "common" and "moral" are strongly associated (Eriksson et al., 2015). Patterns of behavioral regularities together with corresponding empirical and normative expectations can constitute social norms, which in turn drive behavior (Bicchieri, 2016; Fehr and Schurtenberger, 2018). Our design sheds some light on the relationship between abuse of institutional power and its associated normative perceptions, which contributes to the growing literature on the formation and robustness of social norms. This allows us to see if there are differences in normative perceptions of the same situation generated by either being assigned to the position of power or experiencing the effects of presence/absence of power abuse. More importantly, our design is able to answer whether and how randomly experiencing abuse causally changes normative perceptions.

Our findings can be understood from the perspective of the literature on the role of norms in social learning. Having experienced power abuse or the absence of it in the repeated PGG updates subjects' *empirical expectations* (Bicchieri, 2016), which in their turn shift *normative expectations*—that we elicit in the experiment—towards the observed behavior. In other words, whatever is observed or experienced becomes more socially appropriate. Similar effects have been directly documented in previous studies (e.g., Bicchieri and Xiao, 2009b; Panizza et al., 2020). In other related studies it has been found that: witnessing the behavior of others can erode norm compliance (Lindström et al., 2018; Bicchieri et al., 2022; Fischbacher et al., 2001; Merguei et al., 2022; Bicchieri and Xiao, 2009a), yet social proximity is key to prevent the erosion of pro-social behavior (Bicchieri et al., 2022). The theoretical framework of Tremewan and Vostroknutov (2020) suggests that this kind of influence of descriptive information on normative beliefs can take place in the presence of substantial *normative uncertainty* that has been detected in some studies (d'Adda et al., 2020; Merguei et al., 2020). In such environments, rational Bayesian agents optimally interpret descriptive information as revealing injunctive norms, which is what seems to be happening in our experiment.

Our results are also related to the diverse literature on "unfair treatment." There is substantial evidence that people victimized by an unfair treatment may be influenced by and become more accepting of it: experiencing unfair behavior makes the punishment of similar acts in the future less likely (Herz and Taubinsky, 2017); witnessing social norm violations leads to less trust (Banerjee, 2016); experimental subjects from countries with high corruption indices are more likely to lie (Gächter and Schulz, 2016). The reaction to observed norm violations can be "contagious": criminal behavior is often spatially correlated (Glaeser et al., 1996; Zenou, 2003), which, according to the proponents of the "broken windows" hypothesis, is due to norm violations signaling a lack of commitment of a society to follow norms (Wilson and Kelling, 1982). Similarly, Fisman and Miguel (2007) observed that diplomats from corrupt countries committed more parking violations. On the institutional level, Tabellini (2008, 2010) shows that normative values in the regions that experienced the rule of despotic institutions in the past are less likely to be consistent with "generalized morality," or the norms of good conduct, than those in the countries that did not endure such rule.[2]

Our experiment also adds to the literature on peer effects. We show how peer effects impact not just behavior, but the corresponding norms themselves. What we observe goes beyond a mere behavioral spillover, that can be independent of a change in normative evaluations. Acting selfishly after witnessing or being a victim of norm violations may be due to moral licensing (Blanken et al., 2015), a feeling of entitlement(Hoffman et al., 1994), inequality aversion (Fehr and Schmidt, 1999), or being conditionally pro-social (Fischbacher et al., 2001). Moreover, we show this influence even for those who can not engage in the respective behavior and are disadvantaged by it. Participants internalize the selfish behavior of those in a privileged position, even if they harm them.

We also shed some light on the normative complexity and normative disagreement in laboratory experiments. Our design implements a normatively ambiguous scenario where the punisher may abuse his power but still fulfill an essential role in his group. On the one hand, undercontributing while imposing high contributions is frowned upon and generally viewed as inappropriate, especially if the punisher acts unilaterally and can profit from his punishment (Kimbrough and Vostroknutov, 2016; Faillo et al., 2013). It has been shown that cooperation is higher when punishment is consensual (Casari and Luini, 2009), or when participants vote on whom to punish (Ertan et al., 2009). Participants are also more likely to contribute if a single punisher is selected by group members then by randomness (Baldassarri and Grossman, 2011). Further, evidence suggests that in case punishers can profit from punishment, it's effectiveness is reduced (Xiao, 2013) or it may fail altogether (Gneezy and Rustichini, 2000; Fehr and Rocken-

[2]See also Becker et al. (2015). It should be mentioned that the opposite process has also been documented: Lowes et al. (2017) report the results of a field experiment showing that strong institutions in the past crowd out rule-following behavior today.

bach, 2003; Houser et al., 2008). Finally, Fuster and Meier (2010) show that private incentives can weaken norm enforcement mechanisms.

On the other hand, our punisher can enforce higher contributions by those without power and thereby raise the group earnings. This gives room for normative disagreement. Previous research has shown that in complex environments moral disagreement is pervasive (Reuben and Riedl, 2013); people are reluctant to harm others in a personal and direct way, while harming them as a side effect seems more permissible (Greene et al., 2009); there is a tendency to justify one's questionable actions with self-serving beliefs about the behavior of others (Di Tella et al., 2015). This may be especially true for our punishers. As power is associated with disinhibition, powerful people may be predisposed to violating norms and are punished less (van Kleef et al., 2015; Kassas and Palma, 2021; Kimbrough and Vostroknutov, 2020). But people in power may also feel entitled to selfish behavior due to self-serving biases (Kassas and Palma, 2019) and simply believe abusive behavior is justified in their position.

The aforementioned effects can lead to the development of bad social norms, and our study contributes to this literature as well. The effects similar to ours have been found in many studies. For example, Abbink et al. (2017) finds that peer punishment can sustain destructive norms; allowing for counter punishment can lead to destructive punishment cycles that reduce welfare (Nikiforakis, 2008) and the probability of such feuds is enhanced with normative complexity (Nikiforakis et al., 2012). Further, bad norms can persist because of pluralistic ignorance (Smerdon et al., 2020) and once a bad behavioral norm is established it may be difficult to escape due to a preference for imitation, a "conformity trap" (Andreoni et al., 2017). It was also shown that changing an established norm can require reaching a "tipping point" (Centola et al., 2018) as people fear the cost associated with transitioning to a new norm (Andreoni et al., 2021), and finally, punishment can be ineffective if it is not combined with information on norms (Bicchieri et al., 2021). Our study points out an additional mechanism by which bad norms can be sustained: the internalization of abusive behavior even among those who fall victim to it.

Our study is also related to the literature on leadership. For example, it was found that groups perform better with leaders who are cooperatively inclined and who contribute a lot not only due to favorable expectations about the cooperation of others, but also due to other social motivations (Gächter et al., 2012). Compensating a leader who moves first in a public good game can increase contributions, but only to a certain extent as high compensation attracts free-riders (Cappelen et al., 2016). A leader who moves first also shapes the followers' initial beliefs and contributions in the first rounds (Gächter and Renner, 2018). The decision to punish and reward others can increase the reputation of a leader, although rewards are more effective in this way if noise is introduced (de Kwaadsteniet et al., 2019). Acting in groups can make antisocial behavior normatively more acceptable (Behnk et al., 2022). Public goods that are not established yet are better promoted by those who engage in the contribution behavior themselves (Kraft-

Todd et al., 2018). Our design differs from the literature on leadership, where leaders typically have a dissimilar action space at the contribution stage that confers soft power. We implement a strong power asymmetry between the punisher and non-punishers and keep the contribution stage similar for punishers and non-punishers. Our results show that cooperative punishers (who behave like leaders) extract higher contributions as may be expected given the literature on leadership.

Finally, we also speak to the literature on corruption, as our design also investigates a type of antisocial behavior. Abbink et al. (2002) implemented a bribery game in which a first mover (briber) can pay a small fee to transfer an amount of money (bribe) to a second mover (public official), who can accept or reject the bribe. Afterwards, the second mover can decide between a option he slightly prefers and an option favorable to the first mover. In treatments, the latter option also inflicts costs on others (harm to the public). Participants bribe even if their preferred option includes harmful externalities. This effect can be mitigated if the public officials are rotated (Abbink, 2004), although loaded instructions do not reduce corruption (Abbink and Hennig-Schmidt, 2006). Higher salaries do not automatically diminish bribery (Abbink, 2006) if they are exogenously determined, yet can do so if they are chosen by a principal (Jacquemet, 2005). Our design differs from the research on corruption as in the latter the antisocial behavior is reciprocal and collusive, while power abuse does not require the participation of the powerless.

3 Experimental Design

The general idea of our design is to model a social interaction where a minority can abuse institutionalized power. Our goal is to estimate how norms change after participants experience different levels of power abuse. Thus, to study the abusive behavior and the normative perception of power, we conducted a two-part experiment. The first part, aimed at creating a situation of power abuse, is very similar to the design used in Hoeft and Mill (2017a) and in Hoeft and Mill (2017b). In particular, a standard Public Goods game (the PGG) is implemented for 15 rounds with one subject randomly assigned to the additional role of punisher throughout the game. Note that this is a rather conservative implementation to explore power abuse as in other experiments and in reality most people in power achieve it by some kind of merit or voting procedures.[3] The second part utilizes the design of Krupka and Weber (2013a) to elicit subjects' normative perceptions of different actions in the game (norm elicitation task). More specifically,

[3] Such a design feature, while an interesting follow-up, would have introduced more moving parts to an already complicated and lengthy experiment.

subjects in power, and subjects not in power are asked to provide normative evaluations of several situations that could take place in the PGG.[4,5]

3.1 Public Goods Game

All participants are randomly assigned a fixed role, either *punisher* or *non-punisher*, and appointed to a group of four, in which they remain for the 15 rounds of the PGG (partner matching). Each round of the PGG consists of three stages.

Stage 1. Contribution to the Public Good. The first stage is a standard PGG (see Figure C.1 for a screenshot). Each of the four participants is endowed with 20 tokens and is asked to allocate this endowment between private and group accounts (1 token = 20 Euro cents). Tokens allocated to the private account are the subject's to keep. Tokens allocated to the group account (c_i) have a marginal per-capita return (MPCR) of 0.5, so that each group member receives 0.5 times the total contribution. The payoff π_i of participant $i \in \{A, .., D\}$ is defined as

$$\pi_i = 20 - c_i + 0.5 \cdot \sum_{j \in \{A,..,D\}} c_j \tag{1}$$

Stage 2. Punishment. In the second stage (see Figure C.2 for a screenshot), the punishment decisions are made. While the three non-punishing group members (participants A, B, and C) are just shown a blank screen asking them to wait for the decision of the punisher, the punisher (participant D) is shown the contributions and current payoffs of all group members in an anonymized way. To rule out reputation effects and to reduce the possibility of a punisher spitefully targeting individual non-punishers, the information about non-punishing participants is presented to the punisher in random order in an anonymized way in each round (Fehr and Gächter, 2000). Specifically, in each round, non-punishers were randomly assigned the labels 1, 2, and 3 in this stage.

[4]Subjects only learned the nature of the task in the second part after the first part was concluded.

[5]We deliberately put the norm elicitation task only at the end of the experiment, instead of having it before and after the PGG (which seems reasonable if one wants to detect changes in normative beliefs). There are several reasons for this choice. First, it has been shown that repeating the same task, and specifically answering *the same questions*, activates a drive for consistency and, thus, may dilute studied effects (Johansson-Stenman and Svedsäter, 2008). Second, a repeated norm elicitation task might further dilute the effects as the study would take considerably longer. Third, a repeated norm elicitation task asking the same questions might be prone to demand effects as subjects may reason that the study is about changes in normative perceptions. Fourth, asking for the normative evaluation of situations before the PGG might frame punishers to behave less abusively, which might substantially reduce the participants' experience of power abuse. All these issues related to internal validity contributed to our choice of measuring normative beliefs only after the PGG. However, we still cannot exclude the possibility of demand effects. Also, a drive for consistency might still motivate participants as we do ask for norms in multiple (but different) situations.

At this stage, the punisher is asked to indicate how many tokens he would like to deduct from the payoff of subject i (the amount deducted is denoted by σ_i, $i \neq D$).[6] The overall maximal possible deduction in every round is restricted to 30 tokens, which is enough to deter every participant from free-riding.[7] The punishment is costless for D and unused punishment tokens are forfeited.[8] Thus, the punisher could reduce the payoff of the non-punishers by 30 tokens at most, but his payoff would not be directly influenced by punishing (as punishment is costless) or not punishing (as unused tokens are forfeited). This is to ensure that the contributions of the punisher can be directly compared to the contributions of others.

The payoff π_i of a non-punisher $i \neq D$ is given by

$$\pi_i = 20 - c_i + 0.5 \cdot \sum_{j \in \{A,..,D\}} c_j \; - \; \sigma_i. \tag{2}$$

The payoff of the punisher is described by equation (1). In Appendix B we show that with selfish players the unique SPNE of this game is for the punisher to mete out maximum punishment of 10 tokens to each other player who does not contribute 20 tokens in any period, in which case all other players contribute optimally 20 tokens in each period and the punisher contributes zero in each period.[9]

Stage 3. Feedback. The third stage provides feedback to the participants (see Figure C.3 for a screenshot). More specifically, they are informed about their own contribution to the private and group accounts, their own punishment (reduction), and their resulting payoff. Further, they are also informed about the contributions of all other group members labeled as players A, B, C, and D throughout all rounds. Importantly, subjects are able to track the contribution behavior of the punisher (as well as all other group members), which is common knowledge. This feedback

[6] To avoid framing and demand effects, we referred to the act as "reducing the payoff" and not as "punishment."

[7] Note that the highest individual benefit from free-riding when the other two non-punishers contribute 20 tokens, is 10 tokens. If a punisher was confronted with three free-riders and utilized all 30 punishment tokens, he could make every free-rider indifferent between free-riding and fully contributing by subtracting 10 tokens from each of them. As soon as one subject contributes more than zero, the punisher can already make contributing a preferential option. Hence, 30 tokens are sufficient to ensure punishment to be a deterrent.

[8] Making punishment costly would change the budget constraint of the punisher, thus making his contribution decisions incomparable to the contribution decision of the non-punishers. In the alternative case of not forfeiting punishment tokens, the punisher could contribute more in stage one, anticipating extra gains in the second stage, which again would make the contribution decisions of punishers and non-punishers incomparable.

[9] Note that a different way to create "power" would be to have a sequential stage game, similar to Gächter and Renner (2018). Specifically, the punisher could also be a first-mover. However, we decided against it, as it would mean an additional difference between the punisher and non-punishers. Our goal was to make power abuse as straightforward as possible. Thus, we wanted the punisher to differ from non-punishers solely in their power to punish, while they should be comparable in the contribution and feedback stage. Hence, any differences in the contribution stage would have made power abuse less clear, which is why we decided to induce power solely through a punishment opportunity.

ensures that group members can witness if there is abuse of power. Non-punishers are not informed about the punishments meted out to others.

The design choice concerning feedback was driven by two objectives. On the one hand, we wanted to ensure that participants could spot and witness abuse of power, so that we can measure how this experience might affect future normative perceptions. Thus, to investigate how subjects react to experiences of power abuse some feedback is essential. Therefore, all participants were informed about the contribution of punishers as well as all other group members. Further, participants were informed about the punishment they themselves received (which again was common knowledge).[10]

On the other hand, we wanted to reduce the complexity and increase the chances of observing abusive behavior. Thus, we decided against the full information approach (i.e., non-punishers observe the punishment of others on top of their own punishment). We did so for three reasons: 1) to not overload participants with information (as the experiment and the feedback are already rather complex), 2) to allow the punisher to try different punishment strategies (in particular in the beginning) without being constantly monitored and 3) we wanted to have sufficient scope for punishers to behave abusively, which might have been reduced if they would have been more closely monitored by non-punishers.

3.2 Norm Elicitation Task

To elicit normative perceptions, we utilize the norm elicitation task by Krupka and Weber (2013a). More specifically, subjects have to indicate how socially appropriate they find a certain action (five actions are assessed) in a certain situation (three situations are assessed). Thus, the norm elicitation task measures the injunctive norm. In order to be paid, participants are asked to indicate the *modal* appropriateness estimation of a specific group of other participants. If their assessment of the social appropriateness of a specific action in a specific situation in a specific group was identical to the modal response of other participants in this group, they are paid € 8, otherwise they are paid € 0.[11]

[10]Specifically, the feedback about contributions was necessary so that the non-punishers could directly observe the punisher using his power. If we were to give no feedback at all, there would be no experience of abuse as the hypocrisy of punishers would not be observed. If we were to only report averages, as opposed to the individual contributions of all participants, this would obfuscate the actions of the punisher and we would have to control for non-punishers' beliefs as they may think the punisher actually contributes a lot, and only non-punishers refuse to contribute.

[11]The Krupka-Weber task (KW task) is essentially a coordination game where subjects use what they believe is a social norm as a focal point to coordinate on the normative valence of each action in the game they evaluate. Given this, it may be argued that various other focal points (from the outside world or created during the PGG) can influence the evaluations in the KW task, which can potentially distort the measurements. We agree that this is possible. However, in the study by Fallucchi and Nosenzo (2021) this possibility is tested directly. The authors find no significant distortions in measurements in the KW task due to the presence of other focal points. Given this, we do not expect other focal points to distort our measurements much. In addition, our main results hinge on the

There are two reasons why participants may deem actions of the punisher inappropriate: 1) undercontribution is inappropriate in general, or 2) punishing while undercontributing is inappropriate. To disentangle these possibilities we elicit the appropriateness of contributions by the punisher in two situations: full (FC-Q) and medium (MC-Q) contributions by the other group members. Additionally we ask how appropriate it is to punish given different contributions by the punisher when the group contributes halfway (Pun-Q). The three situations, with the corresponding five actions to be normatively assessed, are as follows:

Full Contributing Question (FC-Q) Suppose the others (A, B, C) contributed 20 tokens each to the group account in the previous round. How socially appropriate are the following decisions by D?
D contributes 0, 5, 10, 15, 20 tokens to the group account.

Medium Contributing Question (MC-Q) Suppose the others (A, B, C) contributed 10 tokens each to the group account in the previous round. How socially appropriate are the following decisions by D?
D contributes 0, 5, 10, 15, 20 tokens to the group account.

Punishment Question (Pun-Q) Suppose the others (A, B, C) contributed 10 tokens each to the group account in the previous round. How socially appropriate is it for D to reduce the payoff of A, B, or C, if he contributed the following amounts?
D contributes 0, 5, 10, 15, 20 tokens to the group account and reduces the payoff of A, B, or C.

In each of the three situations, subjects rate the social appropriateness of each action (contribution by D of 0, 5, 10, 15, 20). For each action, the appropriateness is chosen on a seven-point Likert scale: very socially inappropriate, socially inappropriate, somewhat socially inappropriate, neither appropriate nor inappropriate, somewhat socially appropriate, socially appropriate, very socially appropriate.[12] To assess the social appropriateness of these situations, punishers indicate what level of appropriateness they think the mode of other punishers in the current session would choose (punishers' own reference group). Similarly, players A, B, and C (non-

comparisons of normative valences between treatments, which should mitigate the effect of focal points common to both treatments. With regard to (what may seem like) focal points that appear during the PGG (for example, some non-punishers observe power abuse, and some do not), we do expect these experiences to change normative beliefs because this is exactly the main hypothesis that we test in our experiment. We do not think of the experience in the PGG as learning about focal points, but rather as the experience from which people learn what the social norm in this game might be.

[12]We chose seven instead of five statements as originally used by Krupka and Weber (2013a) (see Tables C.1, C.2, and C.3 for further details). The main reason to do so was to ensure sufficient variation in the data and to eliminate a very clear focal point (which might result in a possible demand effect). Specifically, having only five appropriateness statements for each of the five actions would make it salient and likely that participants would answer diagonally, i.e., choosing different appropriateness levels for each of the five actions. Such a design decision could potentially reduce variation and bias results by providing a very salient artificial focal point. Using seven instead of five statements reduces this issue.

punishers) indicate the level of appropriateness that they think the mode of other such players in the current session would choose (ABCs' own reference group).[13]

3.3 Personal Norm Elicitation

While the norm elicitation task by Krupka and Weber (2013a) measures beliefs about social norms, personal norms—that were found to play a role in pro-sociality (Bašić and Verrina, 2020)—might also be affected by the experience of power abuse. Social norms are socially contingent and therefore especially prone to being distorted by socially shared experiences. Yet, the effect of experiencing power abuse may run even deeper and modify personal norms as well, which would make real world interventions targeted at a reversal of normative attitudes harder. Thus, in a follow-up experiment, a new set of participants was asked to state their personal appropriateness perception of the same situations as presented above. Specifically, participants in the roles of punishers and participants in the role of non-punishers were presented with the same situations and same actions as above. The only difference to the social norm elicitation task was that participants were not incentivized to match the modal response of others. They were rather asked to state what they believe the modal response of others in the relevant situations was. Importantly, as participants were not incentivized but rather paid a flat fee for this task, they had no incentive to focus on the social norm but rather on their personal norm.

We have also used an additional behavioral measure to estimate the personal norms. Specifically, participants were asked, as part of the follow-up experiment, to indicate how much additional money (between € 0 and € 10) they would give to punishers in a different session. This decision was costless for the participants to ensure that no wealth effects drive our results. Thus, punishers (in other sessions) could only earn additional money, while it had no costs for the decision-makers.[14] Similarly to the norm elicitation, we asked for this dictator decision conditional on five levels of punishers' average contributions. Thus, participants had to indicate how much money to give to a stranger in the role of a punisher in a different session, based on the average contribution of this stranger. Also, following the norm elicitation task, participants

[13] In the experiment, punishers/non-punishers were also asked to evaluate the levels of appropriateness chosen by the mode of the non-punishers/punishers in the current session. After that, both punishers and non-punishers evaluated the levels of appropriateness expressed by the mode of a third group of people. This group consisted of independent outsiders who did not participate in Part 1 of the experiment (the PGG), but were given the same instructions as punishers and non-punishers. These subjects simply had to indicate the appropriateness levels that they thought the mode of punishers, non-punishers, and other independent outsiders in their session have chosen. We do not discuss these data in this paper.

[14] We designed this task to be a mere money-giving task in order to have it as a surprise stage by the end of the experiment. This choice ensures that it does not confound behavior by explaining to participants that their payoff might be reduced based on their behavior. If we were to use a task where the final payment of participants might be reduced, they, in anticipation of this wealth effect, might have changed their behavior in the PGG. Further, to keep the wealth of decision makers (in particular, non-punishers) comparable between the original and the follow-up study, we decided not to use a dictator game but rather a game where the decision maker can make a costless but still incentivized decision.

were asked how much money to give to a punisher in the same three scenarios as in the norm elicitation task.[15] We present the results related to this measure in Section F.3.

3.4 Payment

At the end of the experiment, subjects in the original experiment were paid for both tasks: the PGG and the appropriateness evaluation.[16] Subjects in the role of punishers and non-punishers were paid for one randomly chosen round of the PGG. One random action from one random situation of Part 2 was drawn to determine the payment. In case a subject evaluated the payoff-relevant action in the payoff-relevant situation as the mode of other subjects in the same role, she obtained € 8, and zero otherwise. Subjects in the follow-up experiment were paid for one randomly chosen round of the PGG and they received a fix flat-fee of € 2.5 for personal norm elicitation task.

Overall, the average payoff for punishers and non-punishers in the original study was € 16.50 (including a show-up fee of € 5). The average payoff for punishers and non-punishers in the follow-up study was € 13.60 (including a show-up fee of € 5).

3.5 Subjects

Subjects were randomly assigned to computer cubicles. They received written instructions separately and were given an opportunity to ask questions for each task in the experiment.[17] After taking part in the PGG subjects were given on-screen instructions for the norm elicitation task and made their decisions in this task. After that, they filled in socio-demographic information and then were presented with their payoff information and received their payoff privately. The experiment lasted 1.5 hours (including seating, instructions, payoff, etc.). All measurements were computerized with the experimental software z-Tree (Fischbacher, 2007).

Original study: The original experiment was conducted in May 2018 at the Bonn DecisionLab and consisted of 7 sessions that were conducted with subjects in the roles of punishers and non-punishers (4 sessions with 32 subjects and 3 sessions with 28 subjects). Overall 212 subjects (60% female) were recruited with the online registration software Hroot (Bock et al., 2014). The

[15] To make this task incentive compatible, we explained that the scenario and situation which is the closest to the average behavior of the punisher would be implemented.

[16] Following the arguments of Charness et al. (2016) and Azrieli et al. (2018), we decided not to pay for all decisions in the experiment to reduce hedging. At the same time, we did not want to dilute the incentives, in particular in the norm elicitation task. Thus, to find the right balance, we incentivized the three tasks separately (for an overview of other papers using such an approach see Charness et al., 2016). We believe that this did not create any problems with hedging as there was no feedback between the tasks. Moreover, before the experiment subjects were only informed that there would be three tasks (without knowing whether and how the upcoming tasks will be incentivized).

[17] The instructions as well as an English version of the handout can be found in Appendix G.

subjects' age ranged from 17 to 72 years (median = 21). Most were bachelor students (semester median = 3).

Follow-up study: The follow-up experiment was conducted in March 2022 at the BEELab at Maastricht University and consisted of 9 sessions. All nine sessions were conducted with subjects in the roles of punishers and non-punishers. Overall 164 subjects (52% female) were recruited with the online registration software ORSEE (Greiner, 2015). The subjects' age ranged from 17 to 45 years (median = 20). Again, most were bachelor students (semester median = 3).

4 Hypotheses

Let us call subjects who played in role *D* in the PGG *punishers*, and subjects who played in roles *A*, *B*, and *C* *non-punishers*. Since the primary focus of this study is on the influence of abusive behavior on the normative perceptions of non-punishers, we need to divide our data by the behavior of punishers. We will call punishers who make below median contributions (among other punishers) to the public good in the first period of PGG *uncooperative punishers*. Notice that in this paper we use this term exclusively in this sense and we do not attach any additional meaning to it beyond the definition just given. Similarly and only in this paper, we will call the punishers whose initial contribution is in the upper half of the distribution (i.e., above, or equal to the median initial contribution of all punishers) *cooperative punishers*.[18] Respectively, we will use the following definitions for the purpose of naming subjects in specific groups in this paper: *non-punishers assigned an uncooperative punisher* are subjects in a group with an uncooperative punisher, and *non-punishers assigned a cooperative punisher* are those in a group with a cooperative punisher. We will refer to these groups as *cooperative punisher-groups* and *uncooperative punisher-groups* (the former contain a cooperative and the latter an uncooperative punisher).

Our null hypothesis is that subjects have robust and common beliefs about social appropriateness of actions in the PGG, and that they are not distorted by any experiences in the game. Hence, under the null hypothesis cooperative and uncooperative punishers are expected to have the same social appropriateness evaluations (normative valences). Similarly, non-punishers assigned an uncooperative and a cooperative punisher are expected to have the same normative valences. Some non-punishers experience power abuse while others do not, but under the null hypothesis they all agree on how socially appropriate the actions in the PGG are. Thus, the normative valences of punishers and non-punishers in both cooperative and uncooperative punisher groups are expected to be identical.

[18] In the results section below we will see that the initial contribution of punishers is predictive of their future contribution behavior, and that punishers who contribute relatively little in the first round are indeed the "uncooperative" punishers (i.e., in the sense that they, on average, contribute less than the non-punishers).

Hypothesis S0 *All types of subjects have the same beliefs about social appropriateness of actions in PGG that are not modulated by experience.*

The alternative hypothesis is that subjects report social norms aligned with the behavior in their PGG group. For non-punishers, we hypothesize that experiencing power abuse would change their normative perception. Since non-punishers are randomly assorted to different punishers, their pre-game perception of social norms should be similar on average. But punishers may be more or less inclined to behave abusively. One idea is that non-punishers exposed to abuse may become especially sensitive to the norm violations as they experience the negative effects first hand and become *disapproving* of such behavior. This is expressed in the following hypothesis about the behavior of non-punishers.

Hypothesis S1 *Non-punishers assigned an uncooperative punisher think that it is less socially appropriate for punishers to contribute less than them as compared to non-punishers assigned a cooperative punisher.*

Another idea comes from empirical research: societies that experience corruption and abuse of power may normatively internalize this behavior among their members, thus leading to the general *approval* of abusive behavior (World Bank Group, 2017). This would imply that participants experiencing abuse would change their perception of the social norm and find it more appropriate for the punisher to undercontribute. We formulate this as a hypothesis.

Hypothesis S2 *Non-punishers assigned an uncooperative punisher think that it is more socially appropriate for punishers to contribute less than them as compared to non-punishers assigned a cooperative punisher.*

The effect of experience in the PGG on social norms of punishers is not so obvious, as they themselves influence the environment in the game. We nonetheless hypothesize that punishers should report social norms that are in line with their own contributions and punishment (they should find their own actions more appropriate than other actions). There may be different mechanisms that lead punishers to report social norms in this way. For example, they may try to rationalize their own behavior (Murphy, 2012), form self-serving beliefs (Ploner and Regner, 2013; Grossman and van der Weele, 2017), or indicate a higher appropriateness out of social image concerns (Kim and Kim, 2019; Kassas and Palma, 2019; Bursztyn and Jensen, 2016). We summarize this in the following hypothesis about the behavior of punishers.

Hypothesis S3 *Uncooperative punishers consider it socially appropriate to contribute less than non-punishers, while cooperative punishers find it inappropriate.*

The hypotheses above make statements pertaining to social norms. It is also possible to reformulate them with personal norms in mind. One may expect personal norms to be especially resistant to social influence as they may stem from moral considerations and values that are in-

dependent of how others think and behave. Nonetheless, it is unclear whether the experience of power abuse only influences social norms or goes deeper and also changes personal norms. We test the following hypotheses that mirror hypotheses S0, S1, and S2 above.[19]

Hypothesis P0 *Non-punishers personal norms are not modulated by experience.*

Hypothesis P1 *Non-punishers assigned a uncooperative punisher personally think that it is less appropriate for punishers to contribute less than them as compared to non-punishers assigned a cooperative punisher.*

Hypothesis P2 *Non-punishers assigned a uncooperative punisher personally think that it is more appropriate for punishers to contribute less than them as compared to non-punishers assigned a cooperative punisher.*

5 Results

The structure of this section is the following. In Section 5.1 we analyze the behavior of punishers and non-punishers in the PGG to gain an understanding of the experience they had during the game. Then, we move to the main result (Section 5.2), where we present the analysis of the normative perception of non-punishers. Finally, we present the analysis of normative perception of punishers in Section 5.3.

To classify the experience of non-punishers, we divide groups of PGG participants into *cooperative* and *uncooperative* defined by the median split of contributions of punishers in the first period. This means that if a punisher has contributed more than or equal to the median among all punishers in the first period, then his group is classified as cooperative and if he contributed below the median then his group is classified as uncooperative.[20] This classification makes sure that the group definition is not influenced by the choices of non-punishers in any way. In Appendix A, we discuss the choice of this classification of groups in more detail and mention the alternative classifications that we use later in the analysis.

5.1 Behavior in the PGG

In this section we present some summary statistics for the PGG. Specifically, we will analyze how punishers contribute and punish, how non-punishers contribute, what drives their contribution, and how the average payoffs differ between groups.

[19] We do not provide hypotheses on how punishers personal norm change, as personal norms (as compared to social norms) should vary across subjects and punishers are less likely to be influenced by a social experience they can largely shape according to their personal beliefs.

[20] We use the median split explicitly for illustrative purposes. In Appendix F we replicate all our results using a continuous scale of punishers' initial contributions. Thus, the results below do not hinge on the median split of the groups.

First, we look at the dynamics of contributions in cooperative punisher-groups and uncooperative punisher-groups, defined by the median split of contributions of punishers in the first period. Figure 1 shows the average contributions of cooperative and uncooperative punishers and non-punishers. The left panel shows the behavior in the original study, while the right panel depicts the follow-up study. The graphical interpretations are supported by the corresponding regressions displayed in Table F.1. Per construction, cooperative and uncooperative punishers differ in their contributions in the first period. We, however, can also see that cooperative and uncooperative punishers still differ in their contribution throughout the experiment.[21] Specifically, uncooperative punishers contribute about 62% of the contribution of cooperative punishers in the last 10 rounds in the original study and 85% in the follow-up study.

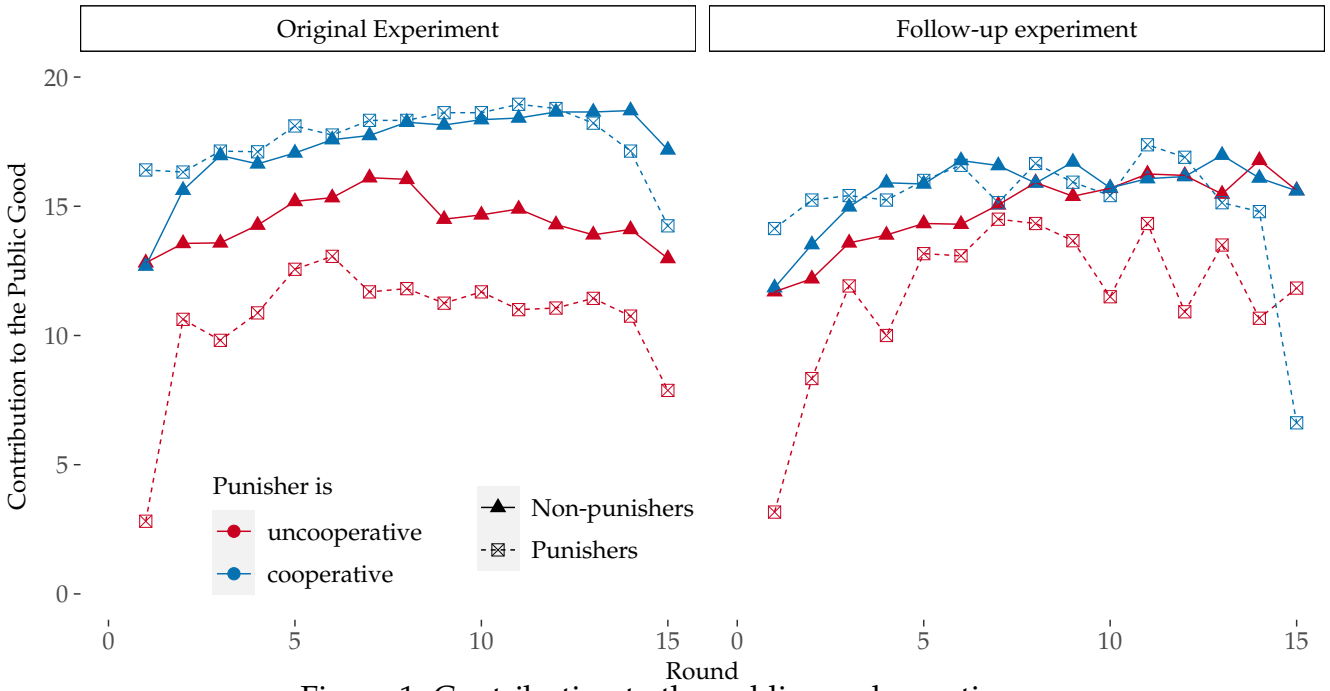


Figure 1: Contribution to the public good over time.

Blue lines represent the contribution of cooperative punisher-groups, while red lines represent the contributions of uncooperative punisher-groups. The punishers are represented by dashed lines with crossed cubes, while the non-punishers are shown with solid lines and solid triangles. The left panel depicts the behavior in the original study, while the right panel depicts the behavior in the follow-up study.

As the construction of cooperative and uncooperative punisher-groups solely hinges on the contribution of punishers in the first round, we would not expect any initial differences between non-punishers assigned an uncooperative or a cooperative punisher, since the assignment is random. Indeed, on both panels of Figure 1 we can see that in the first period the contributions are identical for non-punishers assigned an uncooperative and non-punishers assigned a coop-

[21] Note that this difference is not an artifact of the group-construction. Cooperative and uncooperative punishers are split by their initial contribution. The contributions could have converged over time, resulting in an identical average contributions in the second half of the experiment, which apparently is not the case.

erative punisher. However, the contribution behavior of non-punishers gradually diverges in the original study. In the follow-up study the contribution behavior of non-punishers assigned an uncooperative punisher differs for some time, but eventually converges back.

Most importantly, however, Figure 1 shows that the difference between punishers' and non-punishers' contributions depends on the initial contribution of punishers. Cooperative punishers act pro-socially and contribute on average about the same or even more than non-punishers in their groups (in both experiments). To the contrary, uncooperative punishers in both experiments contribute substantially less than their non-punishers. All these observations are confirmed by a mixed-effects regression in Table F.1 (Column 1).

We also analyze the punishment decisions by punishers. Figure E.2 shows that more punishment is used by uncooperative than by cooperative punishers in the original study, though, taken period-by-period or together, the amounts subtracted are not significantly different. From Table F.1 (Columns 10-12) we also see that cooperative and uncooperative punishers, on average, seem not to differ in their punishment behavior. However, in Table F.2 we see that non-punishers are adjusting their contributions as a consequence of punishment: they contribute more when punisher contributes more in the previous round and they contribute more when the punisher subtracts money from them in the previous round.

The differences in punishers' behavior, specifically their own contribution, also strongly affects the payoff received by the end of the game by non-punishers. Figure 2 shows the average payoffs of non-punishers and punishers in cooperative punisher-groups and uncooperative punisher-groups. While the punishers always earn more than non-punishers, cooperative punishers earn just slightly more than non-punishers in their group relative to the uncooperative punishers. More importantly, non-punishers assigned an uncooperative punisher earn substantially less than non-punishers assigned a cooperative punisher.[22] This is not the case for punishers. The punishers receive the same average payoff in cooperative punisher-groups and in uncooperative punisher-groups.[23] The reason for this potentially counter intuitive finding is that punishers in cooperative punisher-groups let non-punishers participate in their earnings by contributing to the public good game. Uncooperative punishers, on the other hand, extract a substantially larger part of the pie, but the pie itself is smaller as both punishers and non-punishers contribute less. Thus, while the overall payoff of punishers is identical between cooperative punisher-groups and uncooperative punisher-groups, non-punishers' payoff is influenced by their punishers being cooperative or uncooperative.

[22] This difference is significant in the original study $t(60)$= -5.4, $p \leq 0.001$ and goes in the same direction, without being significant, in the follow-up study $t(71.5)$= -1.2, p= 0.25.

[23] This is true both in the original study $t(19.5)$= -0.5, p= 0.591 and the follow-up study $t(14.5)$= 0.3, p= 0.77.

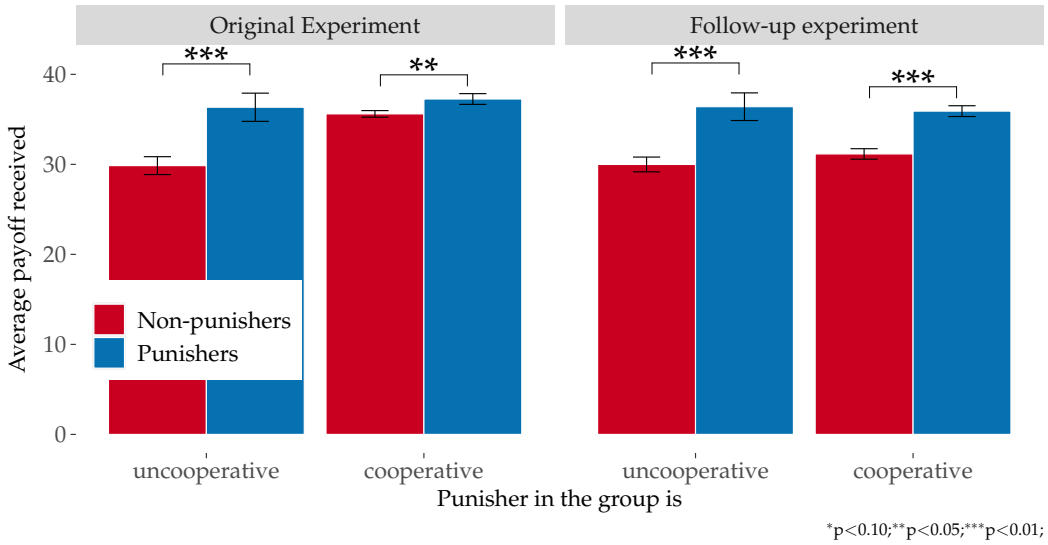


Figure 2: Average payoffs in the PGG.

The graph shows the payoffs obtained in different conditions averaged over all 15 rounds. The left panel shows the original experiment and the right panel shows the follow-up experiment. The left two bars present the averages in the cooperative punisher-groups, while the two right bars present uncooperative punisher-groups. Blue bars depict punishers' payoff, while red bars depict non-punishers' payoffs. Error bars denote standard errors.

5.2 Non-punishers' Normative Perception

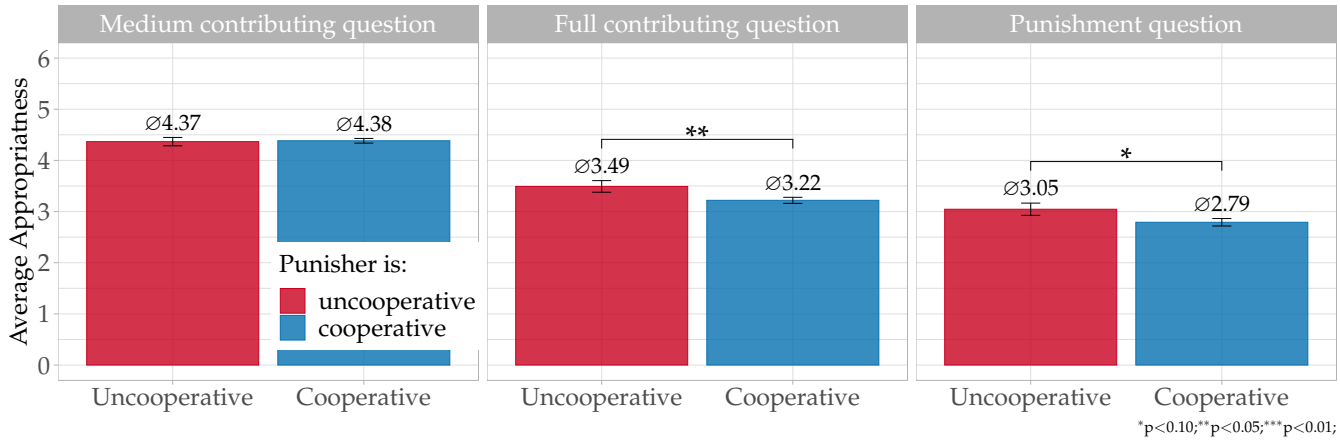
Before we get to our main result related to the normative perceptions of non-punishers, notice that there are five elicited normative valences for each question expressed by the participants in our experiment. Specifically, in each question, participants report their perceived normative valences for five levels of hypothetical contributions by a punisher (see Figure E.3 and Figure E.4). In order to relate these evaluations to contributions and punishment levels in the PGG and to ease interpretation, we transform these into a single number. We do so by considering *average normative valences*. The average normative valence is just the average of the five normative valences expressed by a participant in a given question. For example, for the normative valences shown in Figure 3a the average for each subject and each question (e.g., the *full contributing question*) would be taken over five levels of hypothetical contributions: 0, 5, 10, 15, and 20.

The interpretation of the average normative valence differs slightly between each of the three questions. For the *full contributing question*, the average normative valence describes how socially acceptable undercontribution by the punisher is. For the *medium contributing question*, the average normative valence describes how socially acceptable undercontribution by the punisher is if the non-punishers also undercontributed. For the *punishment question*, the average normative valence describes how socially acceptable the punishment of undercontributing non-punishers is if the punisher themselves undercontributed. Conceptually, the average normative valences

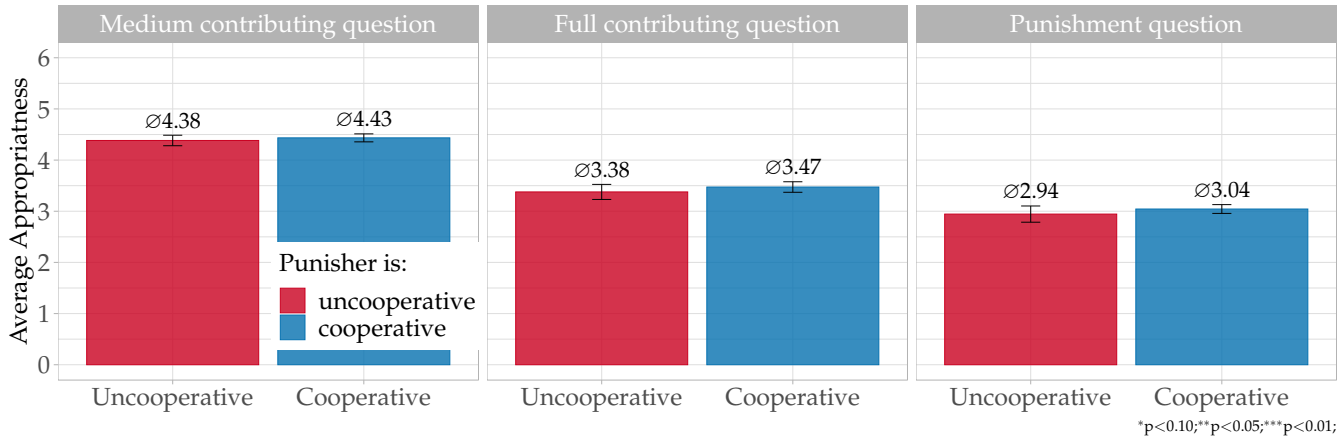
of all these questions can be interpreted as how socially acceptable participants consider power abuse.[24]

Notice that non-punishers were assigned randomly to cooperative punisher-groups and uncooperative punisher-groups. Therefore, any differences in norms that we detect between non-punishers assigned an uncooperative or a cooperative punisher must be due to the experience that they had during the PGG. In fact, in the first round of the PGG the contributions of non-punishers assigned an uncooperative or a cooperative punisher are statistically identical: they do not differ in their mean, median, minimum, or maximum contribution. Hence, all results for non-punishers can be *causally* attributed to the behavior of their punishers and the subsequent experience in the game. Specifically, from experiencing either a cooperative or an uncooperative punisher. This gives us an opportunity to see how the abusive behavior of uncooperative punishers and the cooperative behavior of cooperative punishers changes the non-punishers' perception of the appropriateness of the punishers' actions.

[24]In Appendix D we provide some argumentation for why this is a legitimate way to measure normative perceptions. In Appendix F.6 we further investigate the slopes of the normative perceptions. As the slopes are non-linear we use a GAM-model, which shows that all the results shown in the main text are also reflected in the slopes. Further, we also analyse the data by distinguishing between contributions of zero (free-riding) and non-zero in Appendix F.7. Essentially we find that most of our results are driven by the non-zero cases. This is particularly true as most participants agree on zero contributions being socially unacceptable.



(a) Average social norm perceptions of non-punishers.



(b) Average personal norm perceptions of non-punishers.

Figure 3: Normative valences reported by non-punishers.

The figure depicts the normative valences reported by non-punishers. The top panel shows the average social norms, the bottom panel shows the average personal norms. Left panels represent the normative valences for the *medium contributing question*, the mid-panels for the *full contributing question*, and the right panels for the *punishment question*. Blue bars present the averages normative valences in the cooperative punisher-groups (i.e. the punisher contributed above the median in the first round of the PGG), while red bars present the average normative valences in uncooperative punisher-groups. Error bars denote standard errors.

Figure 3a shows the non-punishers' average social norms elicited in the standard KW task (Figure E.3 also shows the same normative valences as functions). The answers to the *full contributing question* tell us what non-punishers believe is the common attitude among the non-punishers towards the *punishers'* free-riding. We see that non-punishers assigned an uncooperative punisher consider it significantly *more* appropriate than non-punishers assigned a cooperative punisher. This result is in support of Hypothesis S2: non-punishers assigned an uncooperative punisher justify the low contributions of punishers by believing that this is socially appropriate.[25]

[25] It may be argued non-punishers assigned an uncooperative punisher change their normative beliefs in comparison to non-punishers assigned a cooperative punisher because they observe the choices by the punishers that are consistent with the SPNE of the repeated PGG with selfish players described in Appendix B (punishers contribute

Figure 3a shows that non-punishers assigned an uncooperative punisher also consider it significantly more appropriate than non-punishers assigned a cooperative punisher when punishers subtract money from them. Column 6 of Table F.3 demonstrates the same point with a regression that treat punishers' initial contribution as a continuous variable. Importantly, unlike punishers, *the non-punishers are not those who punish, but those who receive the punishment*. Therefore, non-punishers assigned an uncooperative punisher — instead of seeing the hypocritical punishment, which comes from a person who contributes less than them, as "unfair" and thus inappropriate — start to believe that it is actually justified (Hypothesis S2).

Further, we can focus on regressions in Table F.3 to see how more exposure to hypocritical behavior of punishers affects the non-punishers' average normative valences. First, we see that focusing on the average contribution of punishers in the first two, first five and all rounds result in similar conclusions. More interestingly, we see that the level of the effect *grows* in magnitude and in significance. Thus, we find the strongest effect of non-punishers assigned an uncooperative punisher considering punishers' free-riding significantly more appropriate than non-punishers assigned a cooperative punisher if we focus on the average punishers' contributions over the whole game.[26]

Our final supporting evidence for Hypothesis S2 is presented in Figure 2. It shows the average earnings of non-punishers assigned a cooperative or an uncooperative punisher in the original experiment and in the follow-up experiment. We see that non-punishers assigned an uncooperative punisher earn on average 16% and 4% *less* money than non-punishers assigned a cooperative punisher in the original experiment and in the follow-up experiment, respectively. This provides strong evidence that non-punishers justify the behavior of punishers in accordance with Hypothesis S2. If anything, earning smaller amounts should make non-punishers assigned an uncooperative punisher realize that punishers are not doing something right. Nonetheless,

nothing, punish any deviation from full contribution, and all non-punishers contribute full amounts). However, we believe this to be rather unlikely for two reasons: 1) most of experimental economics suggests that people are not selfish utility maximizers, thus it is unlikely that our subjects believe that everyone is playing the game as if everyone is selfish (which would justify the focus on the mentioned SPNE); 2) from the norm elicitations in the PGG (e.g., Kimbrough and Vostroknutov, 2016) we know that people do not see the SPNE of the PGG with selfish players (zero contributions) as a norm, rather they believe that full contributions are the most appropriate actions. Thus, it seems unlikely that players change their beliefs about the norm just due to observation of equilibrium play in a selfish version of the PGG. Nevertheless, an interesting extension of our research would be to check if our results hold in games where abusive behavior of punishers is not an equilibrium of the game with selfish players.

[26] In Appendix F.5 in Table F.6 we also see that how often non-punishers have been punished by the punisher affects their normative valences. Non-punishers who have been punished more often consider undercontribution and punishment again *more* socially appropriate. Similarly, we see from Table F.7 that the number of times a punisher contributes less than their non-punishers contribution in the previous round also changes the normative valences. Specifically, non-punishers who experienced a punisher who undercontributed more often consider it, again, *more* appropriate to punish and undercontribute. At the same time, we need to point out that focusing on the average punishers' contribution over the whole game, or the number of times non-punishers have been punished, as well as the number of times punishers undercontribute might have issues with endogeneity as punishers might, in part, react to non-punishers' behavior. Thus, we obtain the cleanest (and clearly most unconfounded) results, if we focus on the punishers' initial contributions.

we observe the opposite trend. This finding together with the difference in normative perceptions reported above demonstrates an astounding effect that negative experiences can have on the perception of social appropriateness.

Result 1. *The non-punishers' perception of social norms are in line with Hypothesis S2. Non-punishers assigned an uncooperative punisher see low contributions of the punishers and the punishment that they receive as more appropriate than non-punishers assigned a cooperative punisher, even though they earn less.*

While, we have seen that experiencing an uncooperative punisher leads non-punishers to state higher social norm valences for free riding it is unclear whether non-punishers *personally* believe this to be the right action, or whether they believe that it is the socially acceptable action, without necessarily agreeing with it on a personal level. To answer this question we can focus on the unincentivized elicitations obtained from the follow-up study.

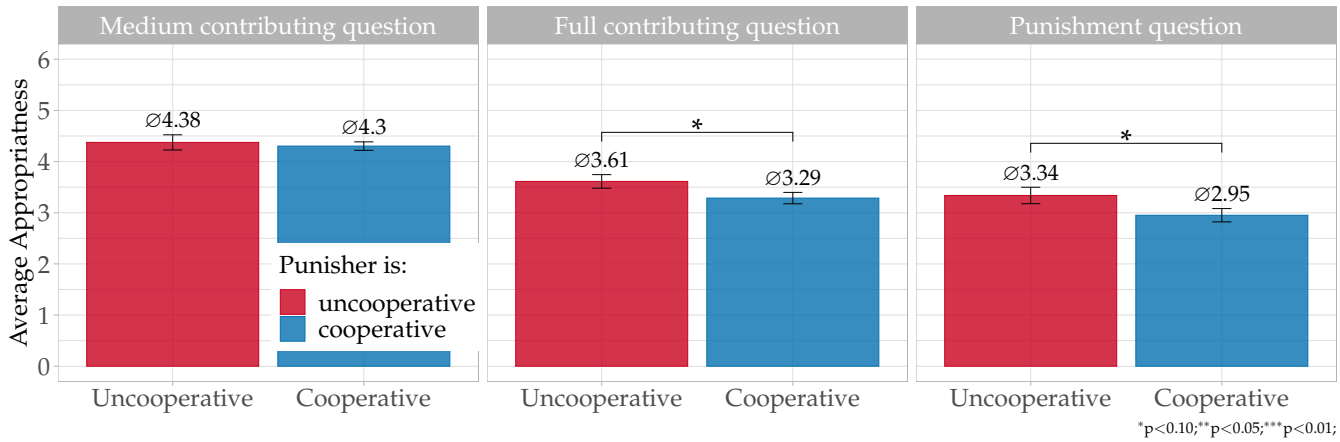
Figure 3b shows the personal norms of non-punishers for the three questions we considered above. We clearly see that experiencing an uncooperative punisher does not affect the personal norm of non-punishers, confirming Hypothesis P0. If anything, the personal norm changes to be less accepting of free riding and punishing, even though this effect is very small in magnitude.

Result 2. *The non-punishers' perceptions of personal norms are independent of their experiences in the PGG in line with Hypothesis P0.*

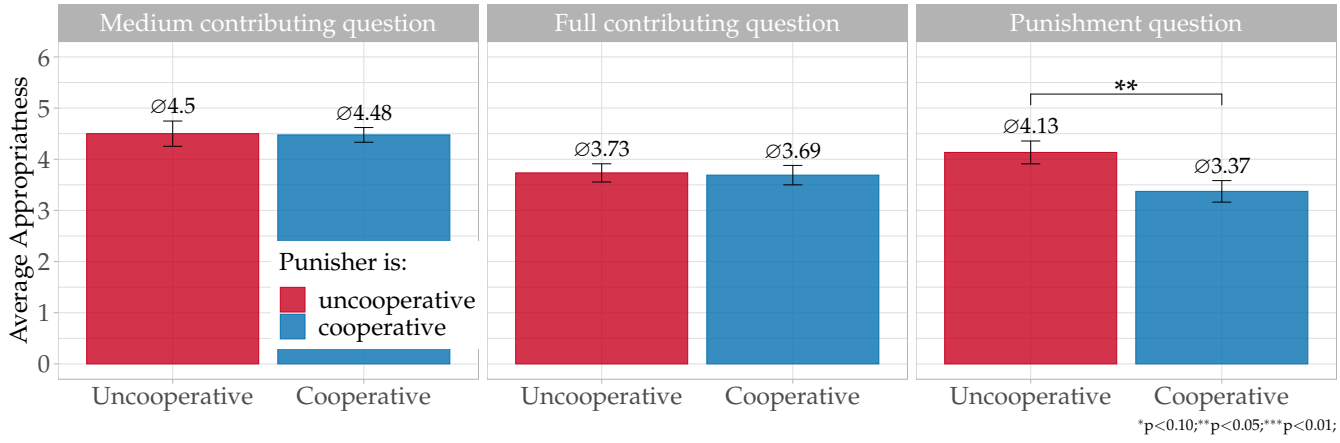
Bringing all these insights together leads to the following picture: when non-punishers are experiencing a punisher who initially undercontributed they believe that undercontribution is socially more acceptable than non-punishers who experienced a punisher who initially overcontributed. If non-punishers experience a punisher who undercontributed throughout the whole game they believe that undercontribution is even more socially acceptable than non-punishers who experienced a punisher who overcontributed throughout the whole game. However, non-punishers do not personally believe these actions to be socially more acceptable, but rather believe that this is what others believe on average. Hence, experiencing an abusive punisher does not change people's core beliefs about what is right or wrong but does change their beliefs about what others believe is right or wrong. This demonstrates one mechanism through which *pluralistic ignorance* (Bicchieri, 2016) may emerge that, in its turn, can lead to the social acceptance of corrupt institutions (because everyone wrongly believes that everyone thinks that they are legitimate).

5.3 Punishers' Normative Perception

When interpreting the normative valences of punishers it should be noted that punishers, by constructions, self select into either being a cooperative or uncooperative punishers. Therefore, the difference between cooperative and uncooperative punisher's normative valences should not be interpreted as causal evidence of their experience in the PGG. Rather we should interpret any insights from normative valences of punishers as correlational evidence of their motives. One idea would be that punishers who consider undercontribution as normatively acceptable are also the ones who undercontribute in the first round. It also could be that these punishers justify their behavior by stating that undercontribution is normatively acceptable (for example, due to self or social image concerns)



(a) Average social norm perceptions of punishers.



(b) Average personal norm perceptions of punishers.

Figure 4: Normative valences reported by punishers.

The figure depicts the normative valences reported by punishers. The top panel depicts the average social norms while the bottom panel depicts the average personal norms. Left panels denote the normative valences for the *medium contributing question*, the mid-panels denote the normative valences for the *full contributing question*, while the right panels denote the normative valences for the *punishment question*. Blue bars present the averages normative valences in the cooperative punisher-groups (i.e. the punisher contributed above the median in the first round of the PGG), while red bars present the average normative valences in uncooperative punisher-groups. Error bars denote standard errors.

Figure 4a shows the punishers' average perceptions of social norms. Figure 4b shows the punishers' average personal norms. We see that there is a significant difference in the average normative valences between cooperative and uncooperative punishers for the *full contributing question*. Uncooperative punishers consider it more appropriate than cooperative punishers to free-ride after others have contributed the full amount. A similar difference can be observed for the *punishment question*. Uncooperative punishers consider it more appropriate than cooperative punishers to punish a non-punisher's contribution of 10 tokens while contributing small amounts themselves. For the *medium contributing question*, we see neither a significant nor meaningful difference in the normative valences between cooperative and uncooperative punishers. The results for the personal norms trail the social norms except for the *full contributing question*, where we do not find a significant difference between cooperative and uncooperative punishers. Table F.3 reports same results as regressions with punishers' initial contribution treated as a continuous variable. Overall, these findings support Hypothesis S3. Specifically, uncooperative punishers indicate that contributing little and punishing non-punishers is not that bad from the moral perspective.[27]

Result 3. *Punishers' normative valences are in line with Hypothesis S3. Uncooperative punishers free-ride and indicate that this behavior is socially appropriate. Cooperative punishers contribute a lot and indicate that doing otherwise is inappropriate.*

Finally, in Appendix F.3 we discuss an additional behavioral measure that could shed some light on the changes of normative beliefs in our experiment. Specifically, we ask subjects to give money to punishers in other sessions conditional on their contributions (see Section 3.3 for more design details). We find that the results from this measure seem to be in line with both punishers' and non-punishers' personal norms and do not provide substantially relevant additional insight.

[27] As mentioned above, one possible reason for this correlation is that punishers who have such beliefs are also the ones who contribute little in the first round of the game. It might, however, also be that punishers change their beliefs (consciously or unconsciously) to be aligned with their behavior.

6 Conclusion

We study normative perceptions of power abuse in an experiment where only one randomly chosen player in a repeated Public Goods game (punisher) has the power to punish others conditional on their contributions. After the Public Goods game, we measure normative beliefs of all subjects about the appropriateness of the punisher's actions. We use the norm elicitation task by Krupka and Weber (2013a) to elicit social norms and an unincentivized version for eliciting personal norms. We hypothesize that the normative beliefs of subjects who did not have an opportunity to punish, but only could endure the consequences of punishment by others, are influenced by the experience of power abuse.

We find that subjects who experience the actions of the powerful, i.e. their abuse, change their beliefs about *social norms* in the direction of normative acceptance of these abusive actions. In other words, subjects who experienced abuse start believing that such state of affairs is a social norm (they believe that everyone believes that this is so). However, at the same time these subjects' *personal norms* seem not to change due to their experience in the Public Goods game. This means that even though our subjects believe that social norm has changed after they experienced abuse, they personally do not support abusive behavior and believe that it is not very appropriate as do subjects who did not experience abuse.

We also find that punishers who abuse their power by contributing little and forcing others to contribute a lot (uncooperative punishers) hold beliefs that this behavior is more appropriate than punishers who contribute the same or more than others (cooperative punishers). While these results are only correlational and cannot be used to determine what influences the normative beliefs of punishers, they still suggest that people who abuse power may do it because they believe that such abuse is not socially inappropriate.

Our results unveil a mechanism that might be responsible for many failed attempts to fight corruption on domestic and international levels, and point toward a reason why inefficient institutions endure. On the one hand, people in power may abuse it because they do not find anything wrong with such behavior. On the other hand, people who are being abused start believing that this is a social norm and may not voice their concerns. This result can contribute to the stability of corrupt institutions. Our findings are likely to underreport the extent of the problem: In our experiment, the powerful (punishers) are chosen randomly, whereas in the real world people with power are often chosen through some measure of merit or some form of voting. This may substantially legitimize their actions and make it more plausible to start assuming abusive behavior is backed by a social norm.

Despite such a grim picture, our experiment suggests that even though subjects, who experience abuse, change their views on social norms, they nonetheless retain their personal views on morality (personal norms) and the wrongness of power abuse. Thus, experiencing power abuse

may lead to the situations characterized by pluralistic ignorance where "bad" norms (legitimizing power abuse) are being maintained despite no one personally believing in such norms. Such situations can be improved if the information about actual low support for social norm is spread in the population (Schroeder and Prentice, 1998), thus suggesting one way to fight corruption.

To address the normalization of corruption and power abuse in the real world, it might be important to understand why those in power believe their actions to be in line with social norms. In our experiment, punishers self-selected into abusive behavior, which is why we can't make causal claims in this regard. However, punishers, too, could be influenced by shared experiences. They may believe that their behavior is inappropriate, but indicate otherwise to keep a positive social image (Kim and Kim, 2019; Kassas and Palma, 2019; Bursztyn and Jensen, 2016). Similarly, self-image concerns might lead the powerful to actually belief that their behavior is appropriate to keep a positive image of themselves (Ploner and Regner, 2013; Grossman and van der Weele, 2017). Future studies will have to extend these findings to understand whether unchecked power starts to erode even the social norms of those in power.

References

Abbink, K. (2004). Staff rotation as an anti-corruption policy: an experimental study. *European Journal of Political Economy*, 20(4):887–906.

Abbink, K. (2006). Fair salaries and the moral costs of corruption. In Kokinov, B., editor, *Advances in Cognitive Economics*. New Bulgarian University Press.

Abbink, K., Gangadharan, L., Handfield, T., and Thrasher, J. (2017). Peer punishment promotes enforcement of bad social norms. *Nature Communications*, 8(1):8:609.

Abbink, K. and Hennig-Schmidt, H. (2006). Neutral versus loaded instructions in a bribery experiment. *Experimental Economics*, 9(2):103–121.

Abbink, K., Irlenbusch, B., and Renner, E. (2002). An Experimental Bribery Game. *The Journal of Law, Economics, and Organization*, 18(2):428–454.

Acemoglu, D., Naidu, S., Restrepo, P., and Robinson, J. A. (2015). Democracy, redistribution, and inequality. In *Handbook of income distribution*, volume 2, pages 1885–1966. Elsevier.

Acemoglu, D. and Robinson, J. A. (2008). Persistence of power, elites, and institutions. *American Economic Review*, 98(1):267–93.

Andreoni, J., Nikiforakis, N., and Siegenthaler, S. (2017). Social change and the conformity trap. *Working paper*, page 34.

Andreoni, J., Nikiforakis, N., and Siegenthaler, S. (2021). Predicting social tipping and norm change in controlled experiments. *Proceedings of the National Academy of Sciences*, 118(16):e2014893118. Publisher: Proceedings of the National Academy of Sciences.

Azrieli, Y., Chambers, C. P., and Healy, P. J. (2018). Incentives in experiments: A theoretical analysis. *Journal of Political Economy*, 126(4):1472–1503.

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

Banerjee, R. (2016). Corruption, norm violation and decay in social capital. *Journal of Public Economics*, 137:14–27.

Barr, A., Lane, T., and Nosenzo, D. (2018). On the social inappropriateness of discrimination. *Journal of Public Economics*, 164:153–164.

Bašić, Z. and Verrina, E. (2020). Personal norms — and not only social norms — shape economic behavior. *SSRN Electronic Journal*.

Becker, S. O., Boeckh, K., Hainz, C., and Woessmann, L. (2015). The empire is dead, long live the empire! Long-run persistence of trust and corruption in the bureaucracy. *The Economic Journal*, 126(590):40–74.

Beetham, D. (2013). *The legitimation of power*. Macmillan International Higher Education.

Behnk, S., Hao, L., and Reuben, E. (2022). Shifting normative beliefs: On why groups behave more antisocially than individuals. *European Economic Review*, 145:104116.

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

Bicchieri, C., Dimant, E., Gächter, S., and Nosenzo, D. (2022). Social proximity and the erosion of norm compliance. *Games and Economic Behavior*, 132:59–72.

Bicchieri, C., Dimant, E., and Xiao, E. (2021). Deviant or wrong? The effects of norm information on the efficacy of punishment. *Journal of Economic Behavior & Organization*, 188:209–235.

Bicchieri, C. and Xiao, E. (2009a). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22(2):191–208. _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/bdm.621.

Bicchieri, C. and Xiao, E. (2009b). Do the right thing: but only if others do so. *Journal of Behavioral Decision Making*, 22(2):191–208.

Blanken, I., van de Ven, N., and Zeelenberg, M. (2015). A meta-analytic review of moral licensing. *Personality and Social Psychology Bulletin*, 41(4):540–558.

Bock, O., Baetge, I., and Nicklisch, A. (2014). Hroot: Hamburg Registration and Organization Online Tool. *European Economic Review*, 71(C):117–120.

Bursztyn, L. and Jensen, R. (2016). Social Image and Economic Behavior in the Field: Identifying, Understanding and Shaping Social Pressure.

Cappelen, A. W., Reme, B.-A., Sørensen, E. ø., and Tungodden, B. (2016). Leadership and Incentives. *Management Science*, 62(7):1944–1953.

Casari, M. and Luini, L. (2009). Cooperation under alternative punishment institutions: An experiment. *Journal of Economic Behavior & Organization*, 71(2):273–282.

Centola, D., Becker, J., Brackbill, D., and Baronchelli, A. (2018). Experimental evidence for tipping points in social convention. *Science*, 360(6393):1116–1119. Publisher: American Association for the Advancement of Science.

Charness, G., Gneezy, U., and Halladay, B. (2016). Experimental methods: Pay one or pay all. *Journal of Economic Behavior & Organization*, 131:141–150.

Cubitt, R. P., Drouvelis, M., Gächter, S., and Kabalin, R. (2011). Moral judgments in social dilemmas: How bad is free riding? *Journal of Public Economics*, 95(3?4):253–264.

d'Adda, G., Dufwenberg, M., Passarelli, F., and Tabellini, G. (2020). Social norms with private values: Theory and experiments. *Games and Economic Behavior*, 124:288–304.

Dal Bó, E., Dal Bó, P., and Snyder, J. (2009). Political dynasties. *The Review of Economic Studies*, 76(1):115–142.

de Kwaadsteniet, E. W., Kiyonari, T., Molenmaker, W. E., and van Dijk, E. (2019). Do people prefer leaders who enforce norms? Reputational effects of reward and punishment decisions in noisy social dilemmas. *Journal of Experimental Social Psychology*, 84:103800.

Di Tella, R., Perez-Truglia, R., Babino, A., and Sigman, M. (2015). Conveniently upset: avoiding altruism by distorting beliefs about others' altruism. *American Economic Review*, 105(11):3416–3442.

Eriksson, K., Strimling, P., and Coultas, J. C. (2015). Bidirectional associations between descriptive and injunctive norms. *Organizational Behavior and Human Decision Processes*, 129:59–69.

Ertan, A., Page, T., and Putterman, L. (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, 53(5):495–511.

Faillo, M., Grieco, D., and Zarri, L. (2013). Legitimate punishment, feedback, and the enforcement of cooperation. *Games and Economic Behavior*, 77(1):271–283.

Fallucchi, F. and Nosenzo, D. (2021). The coordinating power of social norms. *Experimental Economics*, 25(1):1–25.

Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90(4):980–994.

Fehr, E. and Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, 422(6928):137–140. Number: 6928 Publisher: Nature Publishing Group.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, 114(3):817–868.

Fehr, E. and Schurtenberger, I. (2018). Normative foundations of human cooperation. *Nature Human Behaviour*, 2(7):458–468.

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

Fischbacher, U., Gächter, S., and Fehr, E. (2001). Are people conditionally cooperative? evidence from a public goods experiment. *Economics Letters*, 71(3):397–404.

Fisman, R. and Miguel, E. (2007). Corruption, Norms, and Legal Enforcement: Evidence from Diplomatic Parking Tickets. *Journal of Political Economy*, 115(6):1020–1048.

Fuster, A. and Meier, S. (2010). Another hidden cost of incentives: The detrimental effect on norm enforcement. *Management Science*, 56(1):57–70.

Gächter, S., Gerhards, L., and Nosenzo, D. (2017). The importance of peers for compliance with norms of fair sharing. *European Economic Review*, 97:72–86.

Gächter, S., Nosenzo, D., Renner, E., and Sefton, M. (2012). WHO MAKES a GOOD LEADER? COOPERATIVENESS, OPTIMISM, AND LEADING-BY-EXAMPLE. *Economic Inquiry*, 50(4):953–967.

Gächter, S. and Renner, E. (2018). Leaders as role models and 'belief managers' in social dilemmas. *Journal of Economic Behavior & Organization*, 154:321–334.

Gächter, S. and Schulz, J. F. (2016). Intrinsic honesty and the prevalence of rule violations across societies. *Nature*, 531(7595):496–499.

Glaeser, E. L., Sacerdote, B., and Scheinkman, J. A. (1996). Crime and social interactions. *The Quarterly Journal of Economics*, 111(2):507–548.

Gneezy, U. and Rustichini, A. (2000). A fine is a price. *The Journal of Legal Studies*, 29:1–17.

Grant, R. W. and Keohane, R. O. (2005). Accountability and abuses of power in world politics. *American Political Science Review*, 99(1):29–43.

Greene, J. D., Cushman, F. A., Stewart, L. E., Lowenberg, K., Nystrom, L. E., and Cohen, J. D. (2009). Pushing moral buttons: The interaction between personal force and intention in moral judgment. *Cognition*, 111(3):364–371.

Greiner, B. (2015). Subject pool recruitment procedures: organizing experiments with orsee. *Journal of the Economic Science Association*, pages 1–12.

Grossman, Z. and van der Weele, J. J. (2017). Self-Image and Willful Ignorance in Social Decisions. *Journal of the European Economic Association*, 15(1):173–217.

Henrich, J. (2017). *The secret of our success: how culture is driving human evolution, domesticating our species, and making us smarter*. Princeton University Press.

Herz, H. and Taubinsky, D. (2017). What Makes a Price Fair? An Experimental Study of Transaction Experience and Endogenous Fairness Views. *Journal of the European Economic Association*, pages 1–37.

Hoeft, L. and Mill, W. (2017a). Abuse of power–an experimental investigation of the effects of power and transparency on centralized punishment. mimeo, University of Mannheim and MPI Bonn.

Hoeft, L. and Mill, W. (2017b). Selfish punishers. *Economics Letters*, 157:41–44.

Hoffman, E., McCabe, K., Shachat, K., and Smith, V. (1994). Preferences, property rights, and anonymity in bargaining games. *Games and Economic Behavior*, 7(3):346–380.

Houser, D., Xiao, E., McCabe, K., and Smith, V. (2008). When punishment fails: Research on sanctions, intentions and non-cooperation. *Games and Economic Behavior*, 62(2):509–532.

Jacquemet, N. (2005). Corruption as Betrayal: Experimental Evidence on Corruption under Delegation.

Johansson-Stenman, O. and Svedsäter, H. (2008). Measuring hypothetical bias in choice experiments: The importance of cognitive consistency. *The B.E. Journal of Economic Analysis & Policy*, 8(1).

Kassas, B. and Palma, M. A. (2019). Self-serving biases in social norm compliance. *Journal of Economic Behavior & Organization*, 159:388–408.

Kassas, B. and Palma, M. A. (2021). Social Norms: First Impulse or Last Resort?

Kim, C. and Kim, S.-H. (2019). Social image or social Norm?: Re-examining the audience effect in dictator game Experiments. *Journal of Behavioral and Experimental Economics*, 79:70–78.

Kimbrough, E. O. and Vostroknutov, A. (2016). Norms make preferences social. *Journal of the European Economic Association*, 14(3):608–638.

Kimbrough, E. O. and Vostroknutov, A. (2020). A Theory of Injunctive Norms. *SSRN Electronic Journal*.

Kipnis, D. (1972). Does power corrupt? *Journal of Personality and Social Psychology*, 24(1):33–41.

Klitzman, R. (2007). *When Doctors Become Patients*. Oxford University Press Inc.

Kraft-Todd, G. T., Bollinger, B., Gillingham, K., Lamp, S., and Rand, D. G. (2018). Credibility-enhancing displays promote the provision of non-normative public goods. *Nature*, 563(7730):245–248. Number: 7730 Publisher: Nature Publishing Group.

Krupka, E. L. and Weber, R. A. (2013a). Identifying social norms using coordination games: why does dictator game sharing vary? *Journal of European Economic Association*, 11(3):495–524.

Krupka, E. L. and Weber, R. A. (2013b). Identifying Social Norms Using Coordination Games: Why Does Dictator Game Sharing Vary? *Journal of the European Economic Association*, 11(3):495–524.

Lindström, B., Jangard, S., Selbing, I., and Olsson, A. (2018). The role of a "common is moral" heuristic in the stability and change of moral norms. *Journal of Experimental Psychology: General*, 147(2):228–242. Place: US Publisher: American Psychological Association.

Lowes, S., Nunn, N., Robinson, J. A., and Weigel, J. L. (2017). The Evolution of Culture and Institutions: Evidence From the Kuba Kingdom. *Econometrica*, 85(4):1065–1091.

Maner, J. K. and Mead, N. L. (2010). The essential tension between leadership and power: When leaders sacrifice group goals for the sake of self-interest. *Journal of Personality and Social Psychology*, 99(3):482–497.

Merguei, N., Strobel, M., and Vostroknutov, A. (2020). Moral opportunism and excess in punishment decisions. mimeo, Maastricht University.

Merguei, N., Strobel, M., and Vostroknutov, A. (2022). Moral opportunism as a consequence of decision making under uncertainty. *Journal of Economic Behavior & Organization*, 197:624–642.

Murphy, P. R. (2012). Attitude, machiavellianism and the rationalization of misreporting. *Accounting, Organizations and Society*, 37(4):242–259.

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

Nikiforakis, N., Noussair, C. N., and Wilkening, T. (2012). Normative conflict and feuds: The limits of self-enforcement. *Journal of Public Economics*, 96(9–10):797–807.

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

Panizza, F., Vostroknutov, A., and Coricelli, G. (2020). Norm conformity leads to extreme social behavior. mimeo, University of Trento, Maastricht University, University of Southern California.

Ploner, M. and Regner, T. (2013). Self-image and moral balancing: An experimental analysis. *Journal of Economic Behavior & Organization*, 93:374–383.

Reuben, E. and Riedl, A. (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior*, 77(1):122–137.

Rose-Ackerman, S. and Palifka, B. J. (2016). *Corruption and government: Causes, consequences, and reform*. Cambridge university press.

Schroeder, C. M. and Prentice, D. A. (1998). Exposing pluralistic ignorance to reduce alcohol use among college students1. *J. Appl. Soc. Psychol.*, 28(23):2150–2180.

Schroeder, D. A. and Graziano, W. G. (2015). *The Oxford Handbook of Prosocial Behavior*. Oxford Library of Psychology, Oxford ; NewYork.

Smerdon, D., Offerman, T., and Gneezy, U. (2020). 'Everybody's doing it': on the persistence of bad social norms. *Experimental Economics*, 23(2):392–420.

Tabellini, G. (2008). Institutions and culture. *Journal of the European Economic Association*, 6(2-3):255–294.

Tabellini, G. (2010). Culture and Institutions: Economic Development in the Regions of Europe. *Journal of the European Economic Association*, 8(4):677–716.

Tremewan, J. and Vostroknutov, A. (2020). *A Research Agenda in Experimental Economics*, chapter An Informational Framework for Studying Social Norms. Edward Elgar Publishers.

van Kleef, G. A., Wanders, F., Stamkou, E., and Homan, A. C. (2015). The social dynamics of breaking the rules: antecedents and consequences of norm-violating behavior. *Current Opinion in Psychology*, 6:25–31.

Vredenburgh, D. and Brender, Y. (1998). The hierarchical abuse of power in work organizations. *Journal of Business Ethics*, 17(12):1337–1347.

Wilson, J. Q. and Kelling, G. L. (1982). Broken windows. *The Atlantic Monthly*, 249(3):29–38.

Wong, K. C. (1998). A reflection on police abuse of power in the people's republic of china. *Police Quarterly*, 1(2):87–112.

World Bank Group (2017). *World Development Report 2017 : Governance and the Law*. Washington, DC: World Bank.

Xiao, E. (2013). Profit-seeking punishment corrupts norm obedience. *Games and Economic Behavior*, 77(1):321–344.

Xu, A. J., Loi, R., and Lam, L. W. (2015). The bad boss takes it all: How abusive supervision and leader–member exchange interact to influence employee silence. *The Leadership Quarterly*, 26(5):763–774.

Zenou, Y. (2003). The Spatial Aspects of Crime. *Journal of the European Economic Association*, 1(2-3):459–467.

Appendix (for online publication)

A Cooperative punisher-groups and Uncooperative punisher-groups

To classify the experience of non-punishers, we divide groups of PGG participants into cooperative and uncooperative by the initial contribution of their punishers (in the first period of PGG). Specifically, we take the initial contribution of each punisher and label her group as *uncooperative*, if her initial contribution is in the lower half (i.e., below the median initial contribution of all punishers), and as *cooperative*, if it is in the upper half (i.e., above, or equal to the median initial contribution of all punishers, see Figure E.1 for the histogram).

There is one important advantage of using the initial contribution as opposed to other possible classifications. It is that all our results for non-punishers divided into cooperative and uncooperative groups can be interpreted causally. Specifically, since in the first round of PGG the punishers have not yet seen any behavior by the non-punishers, it is impossible for non-punishers to influence the behavior of the punishers when they make their first choice in the game. Thus, should we detect any differences in non-punishers' norms between the cooperative and the uncooperative groups, they should be driven by the behavior of the randomly assigned punisher.

While in the main text we use this classification based on the median split of initial contributions of punishers, in the analysis reported in Appendix F.2 we also use other classifications that, in our opinion, deserve consideration. The reason is that there are downsides to using the initial contribution of the punisher to classify the experiences of non-punishers in the whole repeated PGG (a measure we want in order to understand the influence of different experiences on norms). One of them is that the behavior of the punisher in the first round might change later in the game. For example, a punisher may contribute nothing in the first round and contribute everything in all later rounds. Non-punishers in such groups might then have an overall positive experience with their punisher, which will not be reflected in his initial behavior. Similarly, some non-punishers might experience a high contribution from the punisher in the first round, and zero contributions later, which would have a reverse effect. So, the problem is that initial contributions of the punishers do not fully represent the experience of non-punishers throughout the whole game.

Thus—while in the main text we use the "clean" classification where non-punishers have not interacted with punishers at all—in Appendix F.2, we use multiple additional classifications involving measures that are more indicative of the experience of the non-punishers. Specifically, we use the mean contribution of the punishers averaged over the first two, five, or all rounds to classify punishers as cooperative and uncooperative. These classifications are increasingly indicative of the experience that non-punishers have throughout the game. However, they come with a possible downside of endogeneity. If we use all 15 rounds to classify a punisher as cooperative or uncooperative, this characterization might be confounded or even driven by the non-punishers' behavior. Similarly, but to a lesser extent, the average of the first five rounds might be affected by the non-punishers. The average contribution in the first two rounds, however, is not subject to this endogeneity problem. Specifically, the contribution of punishers in the second round *might* be driven by the behavior of non-punishers in the first round. But, we have seen that the behavior of non-punishers in the first round is identical between cooperative punisher-groups and uncooperative punisher-groups. The potential endogeneity problem kicks in in the third round, where non-punishers might react differently to different first-round contributions by the punisher in their second round. Thus, punishers in the third round might make their behavior dependent on non-punishers.

Further, we also use two additional alternative classifications of cooperative punisher-groups and uncooperative punisher-groups, which, however, are also prone to the potential endogeneity problem. The

first alternative measure is to focus on how often non-punishers have been punished by the punisher. This experience obviously might be driven by the non-punishers' contribution behavior; however, at the same time reflects how much the punisher used his power. The results of this measure are shown in Appendix F.5. The other alternative measure is how often the punisher contributed less than the average contribution of non-punishers in the previous round. This measure also captures how "abusive" the punisher is. However, it also is prone to the potential endogeneity problem. The results of this measure can also be found in Appendix F.5.

B Equilibrium of the Public Goods Game

In this section we describe a unique Subgame Perfect Nash equilibrium of the repeated Public Goods game with one punisher. To make things more tractable assume that before the game the punisher publicly announces the punishment strategy of the form: in period t if any player other than myself contributes less than $r_t \in [0, 20]$ then I will punish this player by 10 tokens. Here r_t can be potentially dependent on previous history in any way. Suppose in period t a non-punisher i contributes c_{it}. Then, if $c_{it} \geq r_t$ this player gets $20 - 0.5c_{it} + C_t$, where C_t is the contributions of all other players times 0.5. If $c_{it} < r_t$, the player gets $10 - 0.5c_{it} + C_t$. Without punishment the best payoff that i can get is $20 - 0.5r_t + C_t$. With punishment the best payoff that i can get is $10 + C_t$. The fact that $r_t \leq 20$ implies that $20 - 0.5r_t \geq 10$. Thus, player i strictly prefers to contribute r_t if $r_t < 20$ and is indifferent between full contribution and contribution of zero when $r_t = 20$.

It is clear that in any equilibrium punisher will choose to contribute zero, since otherwise he can always profitably deviate by contributing less. It is also clear that punisher's payoffs increase in r_t, given the best responses of the non-punishers described above. Thus, punisher will announce the highest r_t possible for all t. This is $r_t = 20$ for all t and all histories. So, one SPNE is to set $r_t = 20$ for all t. Punisher contributes zero tokens in all periods after any history, and all non-punishers contribute 20 tokens in all periods after any history.

Since in case $r_t = 20$ for all t the non-punishers are indifferent between contributing 20 or 0, it needs to be checked that in this case zero contributions by non-punishers in all periods is not an equilibrium. Indeed, there is a profitable deviation by the punisher who can announce before the game that $r_t = 20 - \varepsilon$ for small $\varepsilon > 0$. In this case the non-punishers optimally choose to contribute $r_t - \varepsilon$, which gives the punisher higher payoff than when they contribute zero tokens.

Therefore, the only SPNE of this game is for the punisher to announce $r_t = 20$ for all t, non-punishers contribute 20 tokens each, and the punisher contributes zero in all periods.

C Details of the Design

Suppose the others (A, B, C) contributed **20** tokens each into the group account in the previous decision.
How socially appropriate are the following decisions by D?

	Very socially inappropriate	Socially inappropriate	Somewhat socially inappropriate	Neither appropriate nor inappropriate	Somewhat socially appropriate	Socially appropriate	Very socially appropriate
D contributes 0 tokens to the Group account	✓						
D contributes 5 tokens to the Group account		✓					
D contributes 10 tokens to the Group account		✓					
D contributes 15 tokens to the Group account		✓					
D contributes 20 tokens to the Group account						✓	

Table C.1: Example of norm elicitation, *full contributing question*.

Suppose the others (A, B, C) contributed **10** tokens each into the group account in the previous decision.
How socially appropriate are the following decisions by D?

	Very socially inappropriate	Socially inappropriate	Somewhat socially inappropriate	Neither appropriate nor inappropriate	Somewhat socially appropriate	Socially appropriate	Very socially appropriate
D contributes 0 tokens to the Group account	✓						
D contributes 5 tokens to the Group account		✓					
D contributes 10 tokens to the Group account						✓	
D contributes 15 tokens to the Group account							✓
D contributes 20 tokens to the Group account							✓

Table C.2: Example of norm elicitation, *medium contributing question*.

Suppose the others (A, B, C) contributed **10** tokens each into the group account in the previous decision.
How socially appropriate is it for D **to reduce the payoff of A, B, or C** if he contributed the following amounts?

	Very socially inappropriate	Socially inappropriate	Somewhat socially inappropriate	Neither appropriate nor inappropriate	Somewhat socially appropriate	Socially appropriate	Very socially appropriate
D contributes 0 tokens to the Group account and reduces the payoff of A, B, or C.	✓						
D contributes 5 tokens to the Group account and reduces the payoff of A, B, or C.		✓					
D contributes 10 tokens to the Group account and reduces the payoff of A, B, or C.				✓			
D contributes 15 tokens to the Group account and reduces the payoff of A, B, or C.					✓		
D contributes 20 tokens to the Group account and reduces the payoff of A, B, or C.						✓	

Table C.3: Example of norm elicitation, *punishment question*.

Figure C.1: Contribution decision in the first stage.

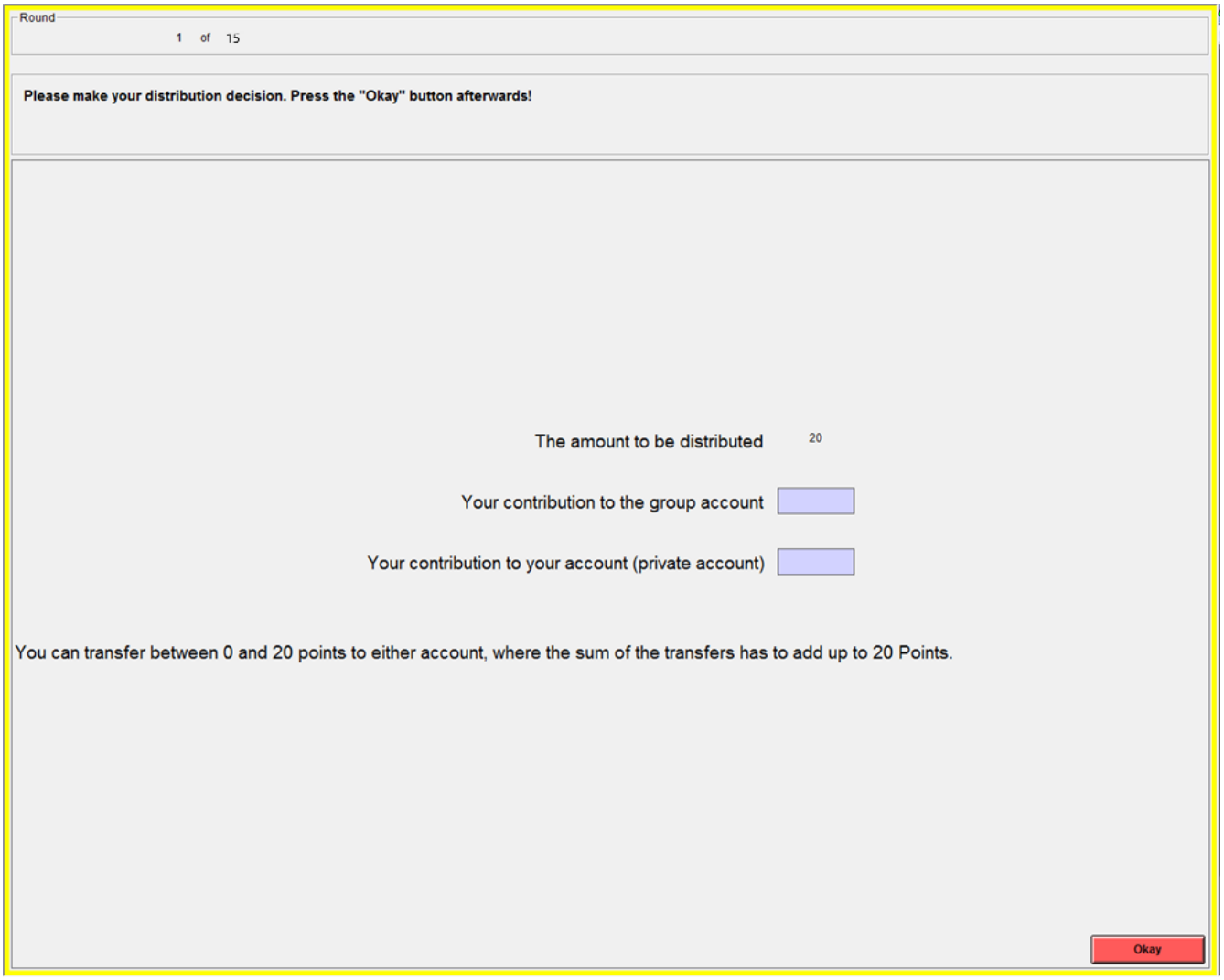


Figure C.2: Punishment decision in the second stage.

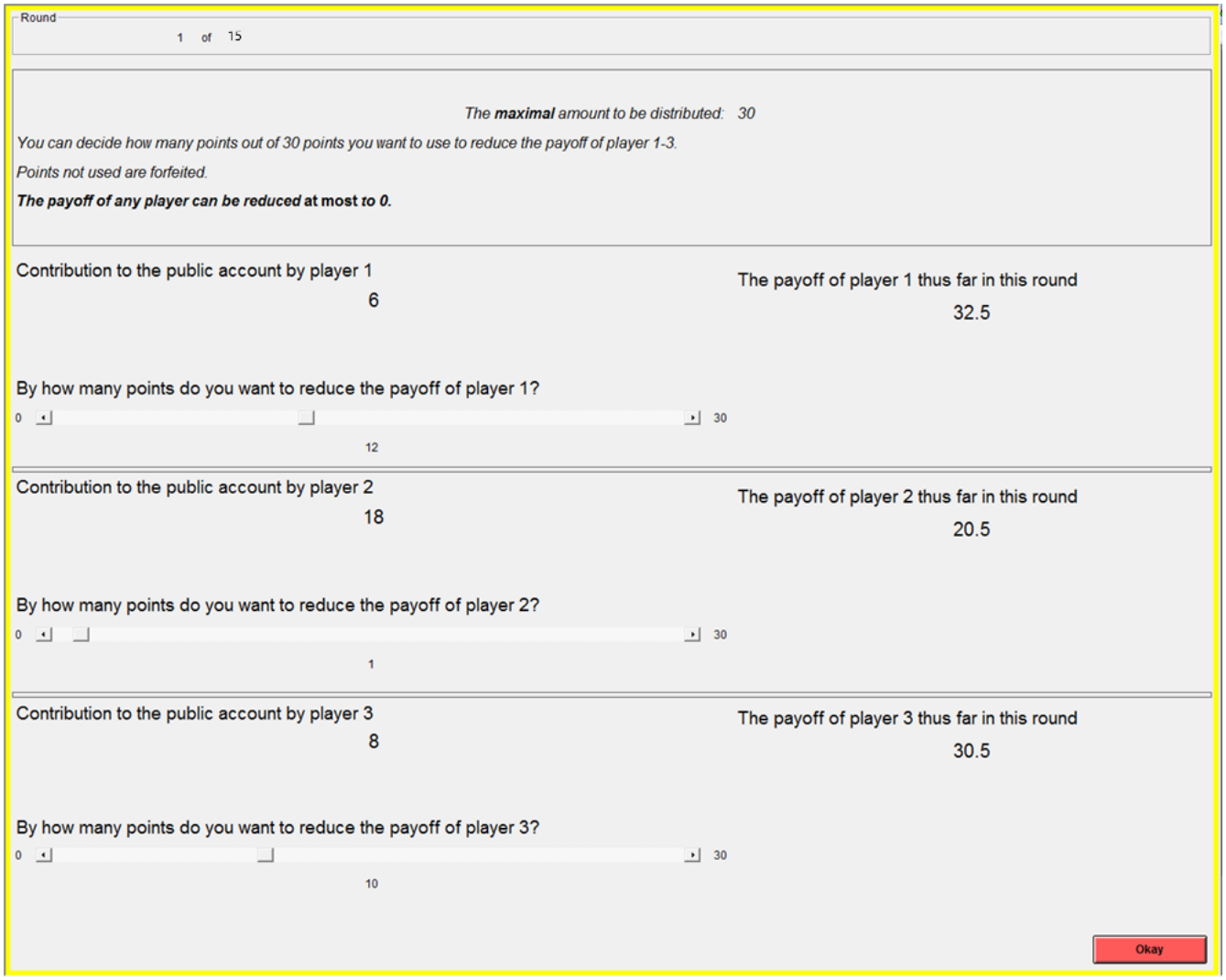
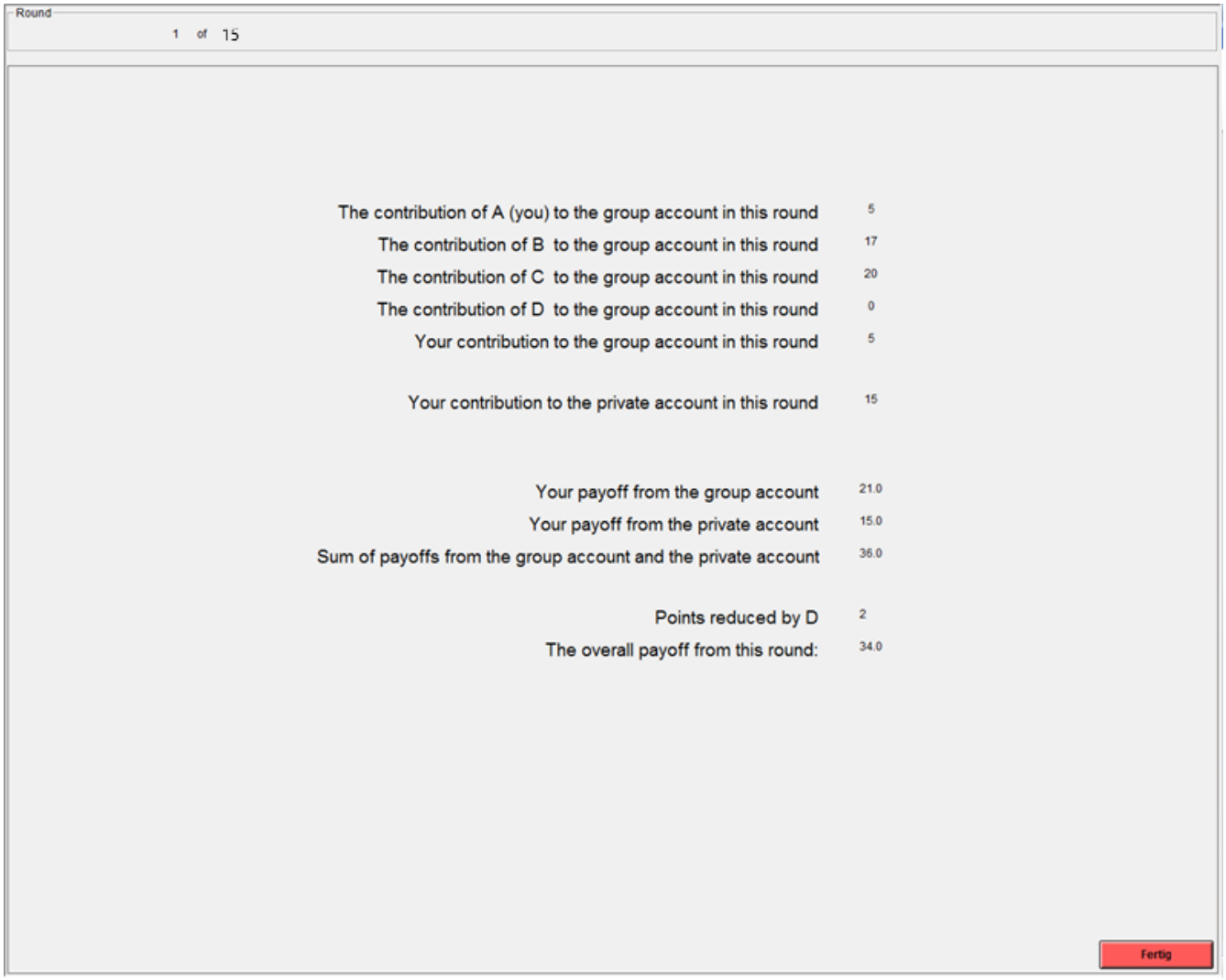


Figure C.3: Feedback in the third stage (non-punisher).



D Average Normative Valences and Comparison of Endpoints

In our analysis we compare normative valences within and between subjects. In particular, for each subject, each question, and each reference group we compute the *average normative valence* with the average taken over five levels of potential contributions of a punisher. Suppose we choose to compare the normative valences between two groups of subjects. For the *full contributing question*, *if the normative valences in these two groups are the same at the endpoints* (hypothetical punisher's contributions of 0 and 20), then the average normative valence becomes a measure of convexity of the normative valence function, or, in other words, the measure of steepness of the derivative in the vicinity of full contribution. For example, in the middle panel of Figure E.3 (i.e. the *full contributing question*), the average normative valence in the cooperative punisher-group is smaller than the average normative valence in the uncooperative punisher-group. With the assumption that the endpoints are the same, this implies that a lower average normative valence is equivalent to having steeper derivative close to full contribution. This means that, if a player is trading-off between maximizing normative valence and personal payoff, she will choose the contribution closer to full (20 tokens) when her average normative valence is lower. A similar argument holds for the *medium contributing question*. For the *punishment question*the logic is slightly different: punishers do not incur costs when they choose how much to punish, so in this case a lower average normative valence should automatically imply less punishment.

In order to meaningfully compare average normative valences in this way, we need to show that for Questions 20 and 10 it is indeed the case that the normative valences at the endpoints are the same for all groups of subjects that we consider. This Appendix provides the details of the statistical comparison of endpoints for the groups of subjects that we are interested in: cooperative/uncooperative punishers, and non-punishers assigned a cooperative or an uncooperative punisher. With few exceptions, which do not undermine our arguments, we show that there are no reasons to suspect that the endpoints in our groups of interest are different. Therefore, it is legitimate to conduct all analyses using average normative valences. However, we also focus on slopes directly in Appendix F.6. The results there provide the same insights and conclusions as the average normative valences reported in the main part of the paper.

We test the hypotheses that the normative valences elicited for the punisher's contributions 0 and 20, the endpoints, are the same across all types of subjects and across all reference groups in a given wave of the experiment. For each Question we compare normative valences in own reference group for punisher's contribution 0/20 in four groups: cooperative punishers, uncooperative punishers, non-punishers assigned a cooperative punisher, and non-punishers assigned an uncooperative punisher. We use Kruskal-Wallis tests for the comparison between these four groups. For all questions we run two sets of tests, one for the punisher's contribution 0 and another for the punisher's contribution 20. Further, we perform pairwise comparisons between cooperative punisher-groups and uncooperative punisher-groups for each questions. All the average responses, the Kruskal-Wallis tests and also the pairwise tests are reported in Table D.1. The Kruskal-Wallis tests show a significant difference in the endpoint 20 for the social norms in the *medium contributing question*. For the personal norms we find for both start and endpoints a significant difference of means in the *punishment question*. Focusing on the more relevant pairwise comparison of means between cooperative punisher-groups and uncooperative punisher-groups we find only two significant differences, as shown in Table D.1.

Therefore, overall, few exceptions, we cannot reject the hypotheses that the normative valences at the endpoints are different for any relevant comparisons and, thus, our method of comparing average norms is valid.

EndPoint: 0		Social Norms		Personal Norms	
Groups:		Non-punisher	Punisher	Non-punisher	Punisher
FC-Q	Cooperative	1.04	1.08	1.38]**	1.45
	Uncooperative	1.04	1.06	1.08	1.83
MC-Q	Cooperative	1.22	1.27	1.61	1.59
	Uncooperative	1.12	1.31	1.47	2.08
Pun-Q	Cooperative	1.04	1.03	1.32 */	1.31
	Uncooperative	1.02	1.44	1.39	1.75

EndPoint: 20		Social Norms		Personal Norms	
Groups:		Non-punisher	Punisher	Non-punisher	Punisher
FC-Q	Cooperative	6.78	6.95	6.57	6.86
	Uncooperative	6.88	6.75	6.72	6.08
MC-Q	Cooperative	6.58 */	6.59]*	6.33	6.48
	Uncooperative	6.4	5.81	5.94	5.67
Pun-Q	Cooperative	5.07	5.57	5.07 **/	5.9
	Uncooperative	5.29	5.5	5.06	6.33
Notes:			*p<0.10;**p<0.05;***p<0.01;		

Table D.1: Normative valences of non-punishers and punishers in cooperative punisher-groups and uncooperative punisher-groups in the two endpoints.

Pairwise tests between cooperative punisher-groups and uncooperative punisher-groups are performed using t-tests. Significant results are denoted by]*. Kruskal wallis tests of differences in endpoint between the four groups (uncooperative/cooperative punishers and non-punishers assigned a cooperative or an uncooperative punisher) are reported (in case of significance) with */. The top panel denotes the endpoint of punisher's contribution of zero while the lower panel denotes the endpoint of punisher's contribution of 20.

E Additional Figures

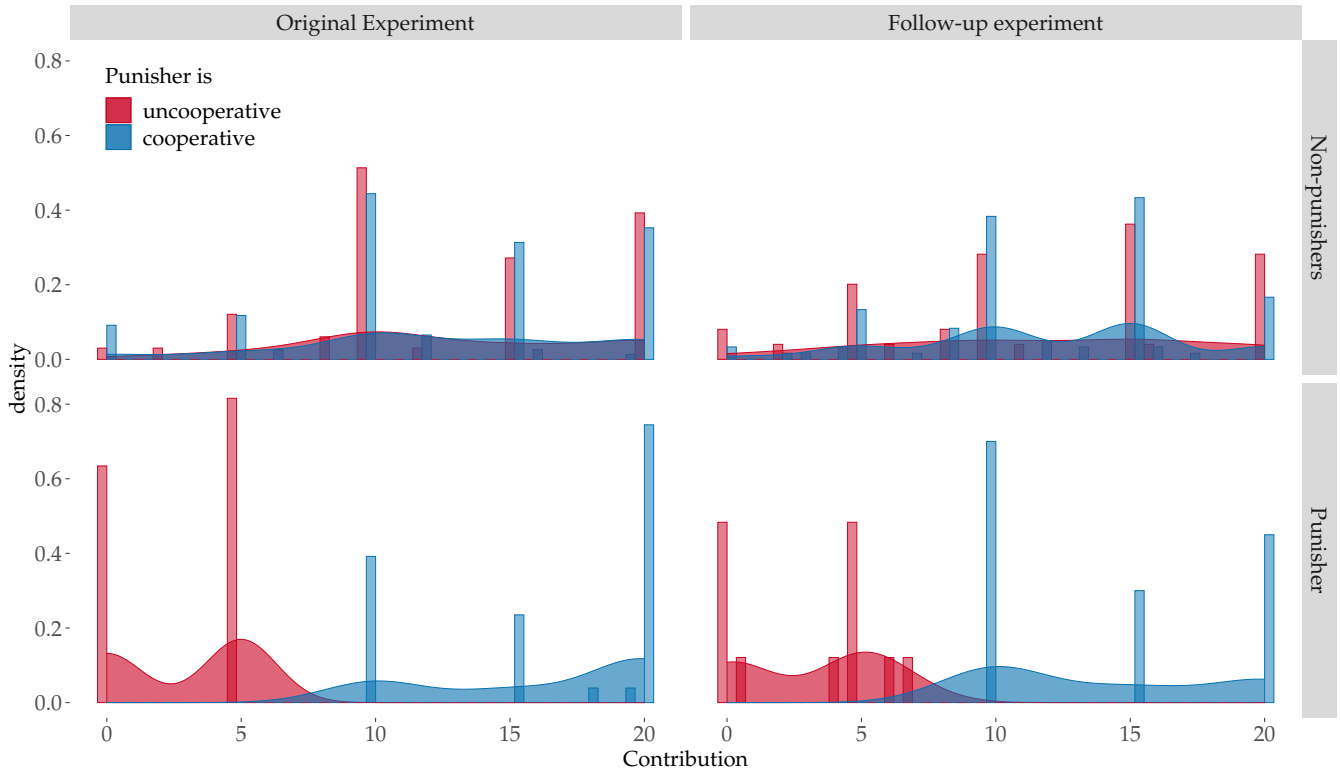


Figure E.1: Histogram of the initial PGG contribution of participant.s

The figure depicts the distribution of the first round contribution behavior. The top panel depicts the contribution behavior of non-punishers, while the bottom panel depicts the contribution behavior of punishers. The left panel depicts the original experiment while the right panel depicts the follow-up experiment. Blue bars present the contribution behavior in the cooperative punisher-groups (i.e. the punisher contributed above the median in the first round of the PGG), while red bars present the contribution behavior in uncooperative punisher-groups.

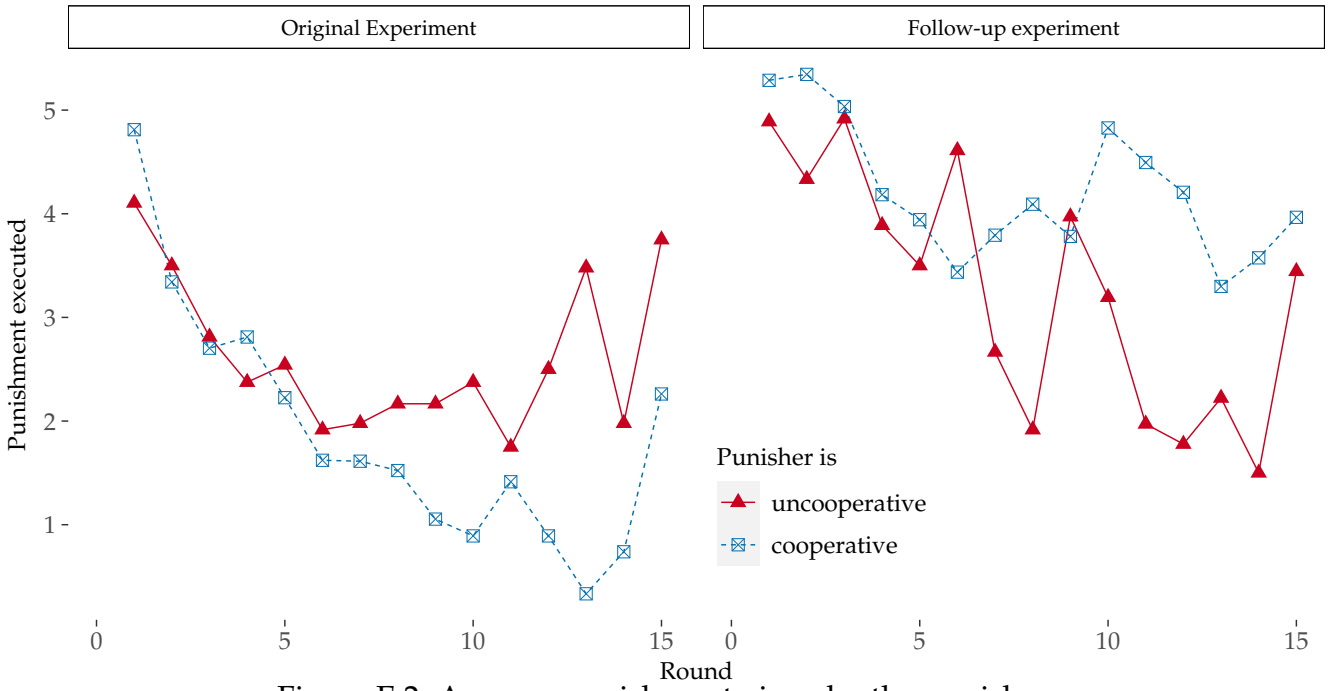


Figure E.2: Average punishment given by the punisher.

Blue dashed lines with crossed cubes represent the cooperative punisher's punishment behavior, while red solid lines with solid triangles represent the uncooperative punishers. The left panel depicts the behavior in the original study, while the right panel depicts the behavior in the follow-up study.

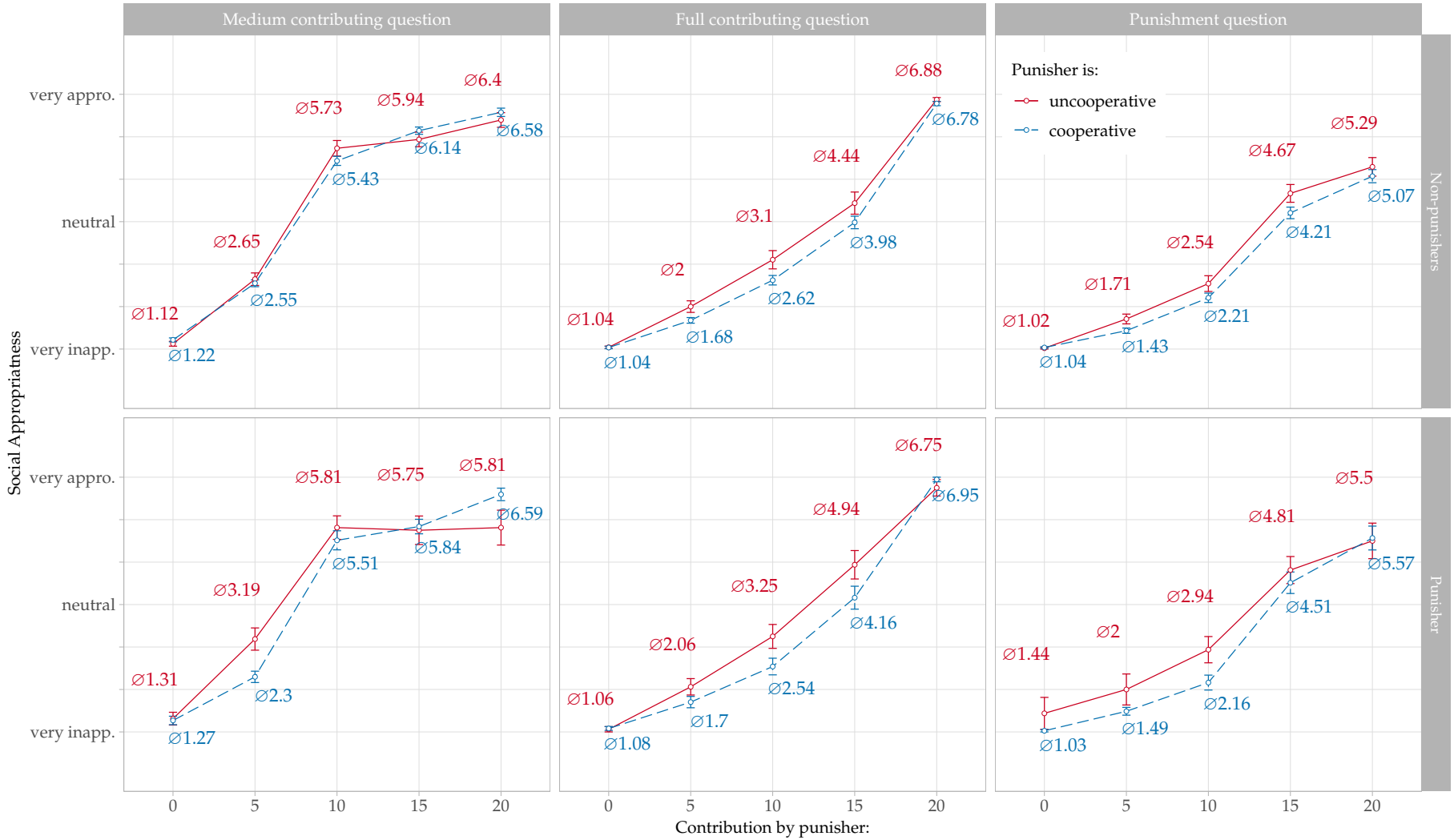


Figure E.3: Social norm as a function of the punishers' contribution.

The figure depicts the social normative valences reported by participants as a function of the punishers' contribution. The top panel depicts the normative valences reported by punishers while the bottom panel depicts the normative valences reported by non-punishers. Left panels denote the normative valences for the *medium contributing question*, the mid-panels denote the normative valences for the *full contributing question*, while the right panels denote the normative valences for the *punishment question*. Blue bars present the averages normative valences in the cooperative punisher-groups (i.e. the punisher contributed above the median in the first round of the PGG), while red bars present the average normative valences in uncooperative punisher-groups. Error bars denote standard errors.

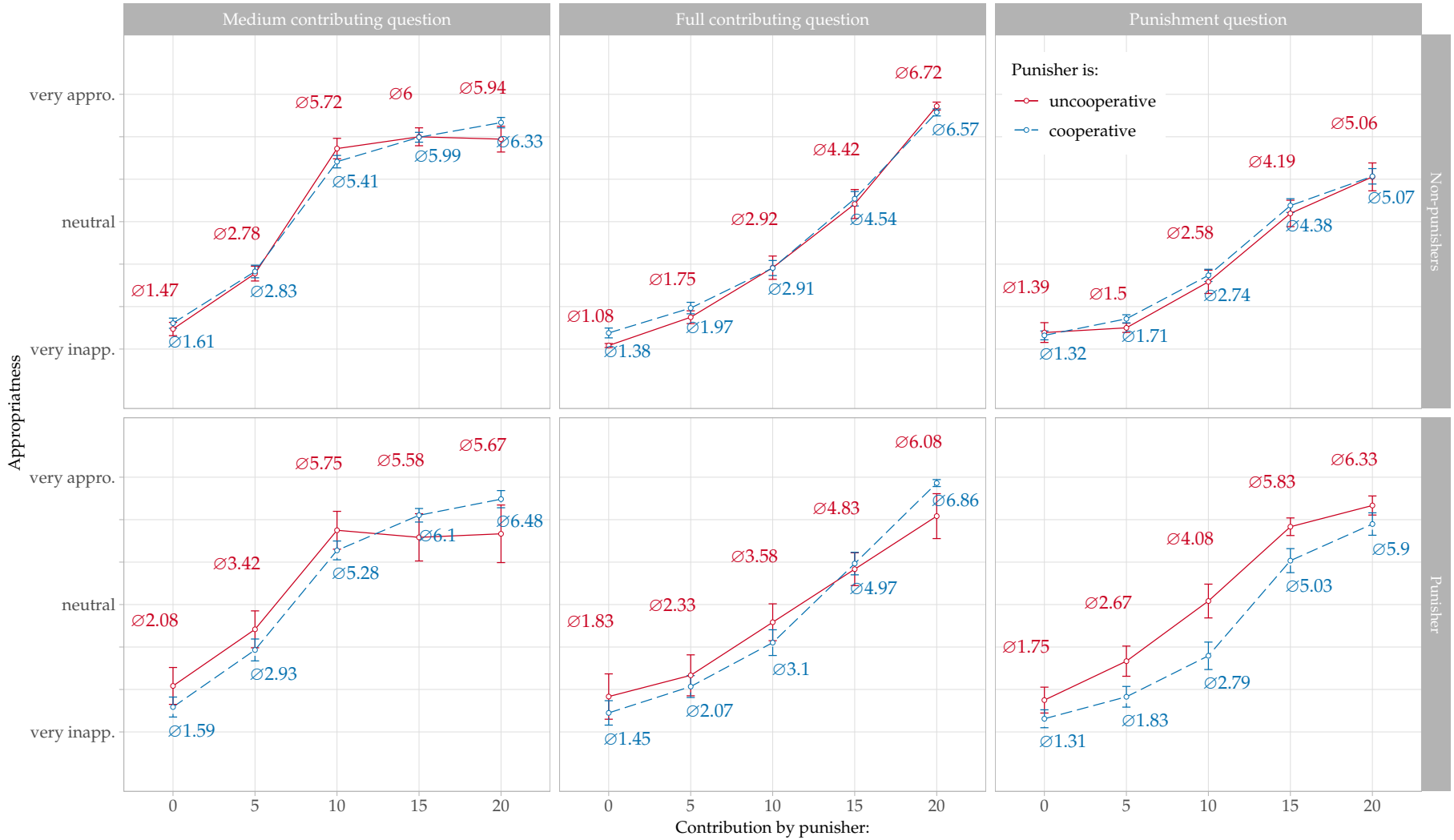


Figure E.4: Personal norm as a function of the punishers' contribution.

The figure depicts the personal normative valences reported by participants as a function of the punishers' contribution. The top panel depicts the normative valences reported by punishers while the bottom panel depicts the normative valences reported by non-punishers. Left panels denote the normative valences for the *medium contributing question*, the mid-panels denote the normative valences for the *full contributing question*, while the right panels denote the normative valences for the *punishment question*. Blue bars present the averages normative valences in the cooperative punisher-groups (i.e. the punisher contributed above the median in the first round of the PGG), while red bars present the average normative valences in uncooperative punisher-groups. Error bars denote standard errors.

F Additional Analyses

In this section we present multiple additional analyses. In section F.2 we present the main regression table. In section F.3 we present the results of the behavioral measure of personal norms. In section F.4 we account for the average contribution of non-punishers in the normative valences. In section F.5 we present multiple alternative classifications of uncooperative punisher-groups and cooperative punisher-groups. In section F.6 we focus on the functional form of the normative valences as a function of the hypothetical contribution of the punisher. In section F.7 we reanalyzes our results by distinguishing between a hypothetical contribution of zero and a hypothetical contribution of more than zero.

F.1 Regressions Supporting Figure 1

	Contribution behavior									Punishment behavior		
	Overall			Non-punishers			Punishers			Punishers		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Constant	12.96***	12.81***	13.16***	12.73***	12.63***	12.87***	10.09***	9.48***	10.89***	4.22***	4.14***	4.32***
	(0.68)	(0.93)	(0.94)	(0.68)	(0.95)	(0.93)	(0.87)	(1.13)	(1.31)	(0.55)	(0.63)	(0.88)
Cooperative Punisher-Group	2.02**	2.96***	0.82	2.02**	2.96***	0.82	5.48***	7.05***	3.44**	0.01	−0.74	0.96
	(0.80)	(1.10)	(1.09)	(0.80)	(1.12)	(1.08)	(1.00)	(1.31)	(1.48)	(0.63)	(0.73)	(1.01)
Punisher	−3.56***	−3.86***	−3.16***									
	(0.56)	(0.72)	(0.88)									
Punisher x Cooperative Punisher-Group	3.45***	4.09***	2.62**									
	(0.66)	(0.86)	(1.04)									
Period	0.23***	0.23***	0.23***	0.25***	0.24***	0.26***	0.17***	0.17***	0.16***	−0.18***	−0.21***	−0.14***
	(0.01)	(0.02)	(0.02)	(0.02)	(0.02)	(0.03)	(0.03)	(0.04)	(0.05)	(0.02)	(0.02)	(0.03)
Last Period	−3.08***	−3.00***	−3.19***	−2.13***	−2.44***	−1.73***	−5.94***	−4.68***	−7.57***	1.65***	2.21***	0.93*
	(0.25)	(0.30)	(0.42)	(0.28)	(0.34)	(0.46)	(0.54)	(0.61)	(0.95)	(0.32)	(0.39)	(0.54)
Contribution Behavior?	✓	✓	✓	✓	✓	✓	✓	✓	✓	×	×	×
Punishment Behavior?	×	×	×	×	×	×	×	×	×	✓	✓	✓
Original Experiment/Follow-up experiment	✓/✓	✓/×	×/✓	✓/✓	✓/×	×/✓	✓/✓	✓/×	×/✓	✓/✓	✓/×	×/✓
Sbj specific effects	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓	✓
Group specific effects	✓	✓	✓	✓	✓	✓	×	×	×	×	×	×
Log Likelihood	-16523.63	-8995.15	-7460.47	-12223.14	-6710.58	-5480.48	-4263.83	-2282.81	-1940.93	-3543.4	-1913.8	-1606.59
Observations	5,640	3,180	2,460	4,230	2,385	1,845	1,410	795	615	1,410	795	615

Notes: *p<0.10;**p<0.05;***p<0.01;

Table F.1: Contribution and punishment behavior.

This table evaluates the contribution and punishment behavior depicted in Figures 1 and E.2. *Cooperative Punisher-Group* denotes a dummy variable with value one if the punisher of the group contributed above the median in the first round of the game. *Punisher* denotes a dummy variable with value one for participants assigned the role of the punisher and zero otherwise. *Period* denotes the round of the PGG. *Last Period* denotes the last period to account for end-game effects. Heterogeneity on the subject level is accounted for by subject-specific random-intercept effects. Heterogeneity on the group level is accounted for by group-specific random-intercept effects. The first 9 models regress the contribution to the public good game as the dependent variable. The last three models regress the punishment meted out by punishers as the dependent variable.

	Contribution behavior		
	Pooled	Original	Follow-up
	(1)	(2)	(3)
Constant	6.87^{***}	6.76^{***}	7.03^{***}
	(0.42)	(0.53)	(0.66)
Punishment received$_{t-1}$	0.12^{***}	0.12^{***}	0.11^{***}
	(0.02)	(0.02)	(0.02)
Punisher's contribution$_{t-1}$	0.16^{***}	0.16^{***}	0.15^{***}
	(0.01)	(0.02)	(0.02)
Own contribution$_{t-1}$	0.40^{***}	0.43^{***}	0.36^{***}
	(0.02)	(0.02)	(0.03)
Period	0.04^{***}	0.03	0.06^{**}
	(0.02)	(0.02)	(0.03)
Last Period	-0.97^{***}	-1.28^{***}	-0.58
	(0.26)	(0.30)	(0.44)
Sbj specific effects	✓	✓	✓
Group specific effects	✓	✓	✓
Log Likelihood	-10990.12	-5942.86	-4996.13
Observations	3,948	2,226	1,722
Notes:	*p<0.10;**p<0.05;***p<0.01;		

Table F.2: Non-punishers' contribution behavior.

This table evaluates the contribution behavior of non-punishers. *Punishment received*$_{t-1}$ denotes the punishment received in the previous round. *Punisher's contribution*$_{t-1}$ and *Own contribution*$_{t-1}$ denote the punisher's and the own contribution in the previous round. *Period* denotes the round of the PGG. *Last Period* denotes the last period to account for end-game effects. Heterogeneity on the subject level is accounted for by subject-specific random-intercept effects. Heterogeneity on the group level is accounted for by group-specific random-intercept effects.

F.2 Main regression table

Table F.3 below shows the regression analyses supporting the main results of the study. In regressions (1-6) we use the classification of punishers as in the main text (the groups are defined by the median split of punishers' contributions in the first round). In regressions (7-12) we use a different classification that splits groups by the median of average contribution of punishers in the first two rounds of the PGG. Regressions (13-18) do the same for five rounds, and regressions (19-24) use the median split of average punishers' contributions in the whole game.

Panel A: Original experiment (Social Norms)

	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q
	Punishers			Non-punishers			Punishers			Non-punishers			Punishers			Non-punishers			Punishers			Non-punishers		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
Constant	3.65***	4.44***	3.21***	3.44***	4.40***	3.09***	3.88***	4.52***	3.33***	3.53***	4.39***	3.15***	4.04***	4.62***	3.43***	3.74***	4.36***	3.20***	3.98***	4.63***	3.51***	3.82***	4.44***	3.37***
	(0.17)	(0.14)	(0.21)	(0.13)	(0.08)	(0.12)	(0.21)	(0.18)	(0.26)	(0.16)	(0.11)	(0.16)	(0.24)	(0.20)	(0.30)	(0.18)	(0.12)	(0.18)	(0.26)	(0.22)	(0.32)	(0.19)	(0.13)	(0.19)
Cont.Pun$_{t\in\{1\}}$	−0.02*	−0.01	−0.01	−0.01	−0.002	−0.02**																		
	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)	(0.01)																		
Cont.Pun$_{t\in\{1,2\}}$							−0.04**	−0.01	−0.02	−0.02	−0.0005	−0.02**												
							(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)												
Cont.Pun$_{t\in\{1,...,5\}}$													−0.04***	−0.02	−0.02	−0.03***	0.001	−0.02**						
													(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)						
Cont.Pun$_{t\in\{1,...,15\}}$																			−0.04**	−0.02	−0.03	−0.03***	−0.004	−0.03***
																			(0.02)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)
Group specific effects	×	×	×	✓	✓	✓	×	×	×	✓	✓	✓	×	×	×	✓	✓	✓	×	×	×	✓	✓	✓
Log Likelihood	-50.07	-40.26	-60.2	-166.52	-123.65	-191.74	-48.37	-39.99	-59.9	-165.86	-123.5	-191.68	-47.46	-39.49	-59.69	-163.6	-123.42	-191.63	-48.82	-39.59	-59.41	-162.97	-123.28	-189.71
Observations	53	53	53	159	159	159	53	53	53	159	159	159	53	53	53	159	159	159	53	53	53	159	159	159

Panel B: Follow-up experiment (Personal Norms)

	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q
	Punishers			Non-punishers			Punishers			Non-punishers			Punishers			Non-punishers			Punishers			Non-punishers		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)	(13)	(14)	(15)	(16)	(17)	(18)	(19)	(20)	(21)	(22)	(23)	(24)
Constant	3.99***	4.78***	4.01***	3.18***	4.31***	2.93***	4.28***	4.78***	4.53***	3.20***	4.29***	2.88***	4.43***	4.79***	4.88***	3.36***	4.25***	2.81***	4.98***	5.14***	5.78***	3.44***	4.41***	2.83***
	(0.28)	(0.24)	(0.33)	(0.17)	(0.12)	(0.15)	(0.35)	(0.31)	(0.41)	(0.21)	(0.16)	(0.19)	(0.43)	(0.39)	(0.49)	(0.27)	(0.20)	(0.24)	(0.43)	(0.39)	(0.43)	(0.28)	(0.21)	(0.25)
Cont.Pun$_{t\in\{1\}}$	−0.03	−0.03	−0.04	0.02*	0.01	0.01																		
	(0.02)	(0.02)	(0.03)	(0.01)	(0.01)	(0.01)																		
Cont.Pun$_{t\in\{1,2\}}$							−0.05*	−0.02	−0.08**	0.02	0.01	0.01												
							(0.03)	(0.02)	(0.03)	(0.02)	(0.01)	(0.01)												
Cont.Pun$_{t\in\{1,...,5\}}$													−0.05*	−0.02	−0.10***	0.01	0.01	0.01						
													(0.03)	(0.03)	(0.03)	(0.02)	(0.01)	(0.02)						
Cont.Pun$_{t\in\{1,...,15\}}$																			−0.09***	−0.05*	−0.15***	0.0004	0.0003	0.01
																			(0.03)	(0.03)	(0.03)	(0.02)	(0.01)	(0.02)
Group specific effects	×	×	×	✓	✓	✓	×	×	×	✓	✓	✓	×	×	×	✓	✓	✓	×	×	×	✓	✓	✓
Log Likelihood	-53.27	-46.84	-60.14	-169.16	-134.38	-158.34	-52.43	-47.31	-58.13	-169.86	-134.29	-158.03	-52.4	-47.49	-57.53	-170.44	-134.15	-157.81	-49.36	-46.33	-50.09	-170.49	-134.55	-157.89
Observations	41	41	41	123	123	123	41	41	41	123	123	123	41	41	41	123	123	123	41	41	41	123	123	123

Notes: *p<0.10;**p<0.05;***p<0.01;

Table F.3: Estimation of the average normative valence by punishers' public good contributions.

The table depicts the average normative valence as a function of different measures of punishers' public good contributions. The top panel depicts the estimations of the original study were social norms where elicited, while the bottom panel depicts the estimations of the follow-up study where personal norms were elicited. FC-Q, MC-Q, and Pun-Q denote the average normative valence in the *full contributing question*, *medium contributing question* and the *punishment question*, respectively. *Cont.Pun*$_{t\in\{1\}}$ denotes the punishers contribution in the first round. *Cont.Pun*$_{t\in\{1,...,15\}}$ denotes the average punishers contribution in the first fifteen rounds. *Cont.Pun*$_{t\in\{1,2\}}$ and *Cont.Pun*$_{t\in\{1,...,5\}}$ are defined accordingly. Heterogeneity on the group level is accounted for by group-specific random-intercept effects.

F.3 Punishing Punishers

So far, we have focused on comparing the norms explicitly elicited from subjects. In this section, we test an indirect behavioral measure of personal norms to validate our findings. Specifically, participants were asked, as part of the follow-up experiment, to indicate how much additional money (between 0 and 10 euros) they would give to punishers in a different session.

Figure F.1 depicts the responses of punishers and non-punishers, both in cooperative punisher-groups and uncooperative punisher-groups. Table F.4 reports upon the regression of money given to punishers averaged over the five situations as a function of the punisher's initial public good contribution. First, we see that non-punishers assign substantially less money to punishers than punishers do. Second, we see that punishers who initially contributed little to the public-good game tend to reward positive contributions of other punishers more generously than punishers who initially contributed rather much. The difference between cooperative and uncooperative punishers is particularly pronounced in the *punishment question*, where uncooperative punishers give more money to other punishers if they punish non-punishers who contribute ten tokens while themselves contributing ten or more. These insights are in line with the personal and social norm measures reported by punishers. However, on average cooperative and uncooperative punishers do not differ significantly in the amount of money given to other punishers. As before, these insights are correlational and might merely reflect self-selection or image concerns.

Focusing on non-punishers, we find that both non-punishers assigned a cooperative or an uncooperative punisher assign very little money to abusive punishers (i.e., punishers contributing less than 10) and are more generous with their decision the more the punisher contributes to the public good. In line with the personal norm, we also see that non-punishers assigned a cooperative or an uncooperative punisher do not differ in their giving behavior. However, we see that the giving behavior of non-punishers follows rather the observations of our social norm measure. Specifically, non-punishers who experienced a punisher who abused their power reward a cooperative punisher more. This difference, however, is not significant for most specifications.

Overall, we conclude that the behavioral measure reflecting the personal norm follows the pattern of the personal norm rather closely. Uncooperative punishers reward power-abusing punishers more. Non-punishers seem to give rather little money to punishers who abuse their power – but they do so rather independently of their own experience during the PGG.

	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q
	Punishers			Non-punishers			Punishers			Non-punishers		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Constant	5.54***	6.14***	5.38***	2.46***	3.97***	2.02***	6.68***	7.45***	6.47***	1.82***	3.17***	1.89***
	(1.09)	(0.88)	(1.18)	(0.39)	(0.38)	(0.40)	(1.79)	(1.44)	(1.93)	(0.64)	(0.64)	(0.68)
Cont.Pun$_{t\in\{1\}}$	−0.02	−0.01	−0.03	−0.03	−0.06*	−0.03						
	(0.09)	(0.07)	(0.09)	(0.03)	(0.03)	(0.03)						
Cont.Pun$_{t\in\{1,\ldots,15\}}$							−0.10	−0.10	−0.10	0.02	0.01	−0.02
							(0.12)	(0.10)	(0.13)	(0.04)	(0.04)	(0.05)
Group specific effects	×	×	×	✓	✓	✓	×	×	×	✓	✓	✓
Log Likelihood	-108.7	-99.85	-111.71	-271.19	-269.5	-270.17	-108.38	-99.3	-111.46	-271.26	-270.83	-270.33
Observations	41	41	41	123	123	123	41	41	41	123	123	123
Notes:												*p<0.10;**p<0.05;***p<0.01;

Table F.4: Estimation of the average amount of money given to punishers.

The table depicts the average amount of money given to punishers as a function of different measures of punishers' public good contributions. FC-Q, MC-Q, and Pun-Q denote the average amount of money given in the *full contributing question*, *medium contributing question* and the *punishment question*, respectively. *Cont.Pun*$_{t\in\{1\}}$ denotes the punishers contribution in the first round. *Cont.Pun*$_{t\in\{1,\ldots,15\}}$ denotes the average punishers contribution in the first fifteen rounds. Heterogeneity on the group level is accounted for by group-specific random-intercept effects.

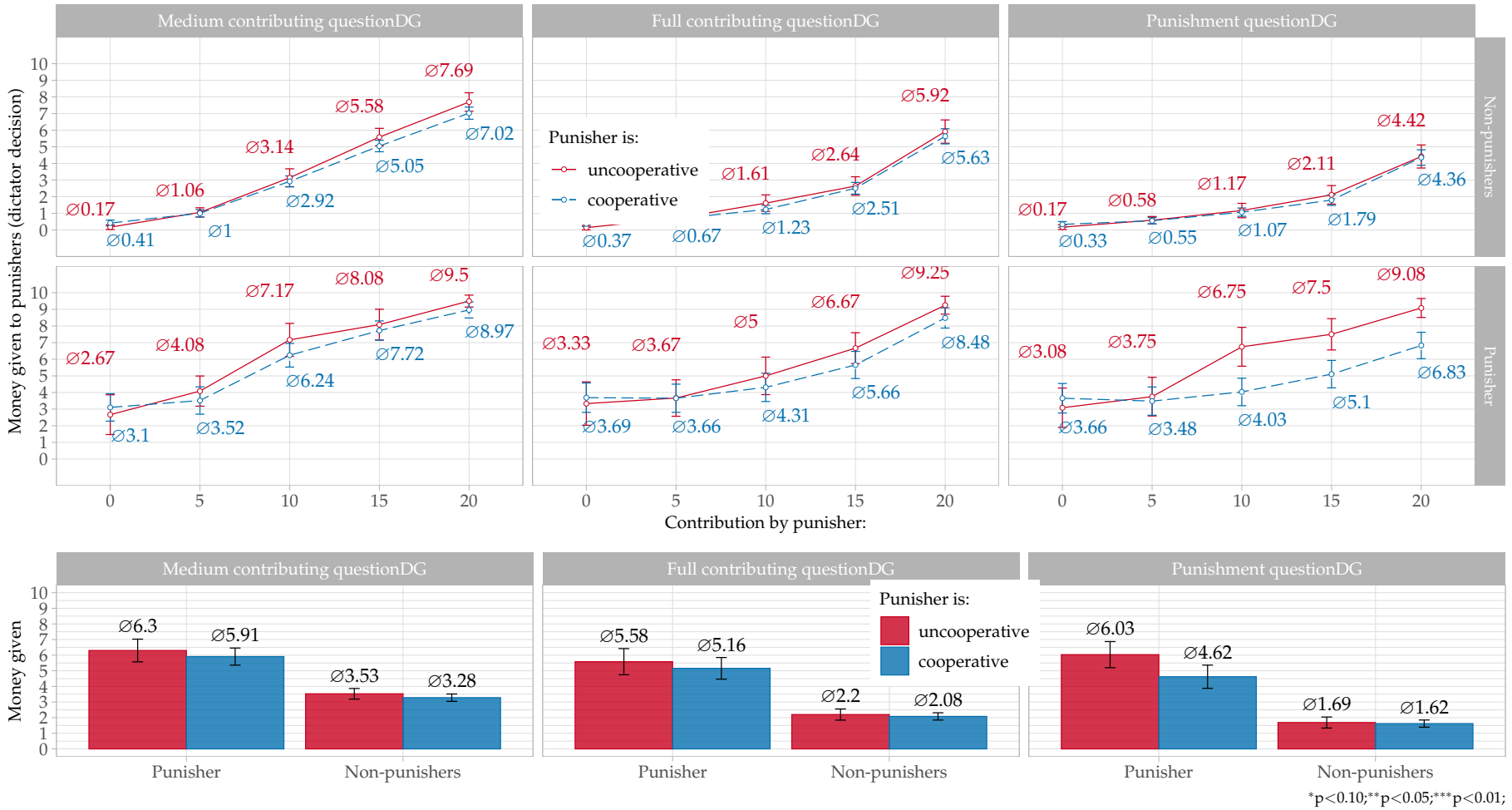


Figure F.1: Money given to punishers.

The figure depicts the how much money should be given to a randomly chosen punisher as a function of the punishers' contribution. The top panel depicts the money given to punishers by punishers while the middle panel depicts the money given to punishers by non-punishers. The bottom panel depicts the average amount of money given to punishers. Blue bars present the amount given in the cooperative punisher-groups (i.e. the punisher contributed above the median in the first round of the PGG), while red bars present the amount given in uncooperative punisher-groups. Left panels denote the amount given in the *medium contributing question*, the mid-panels denote the amount given in the *full contributing question*, while the right panels denote the amount given in the *punishment question*. Error bars denote standard errors.

F.4 Accounting for the non-punishers' Average Contributions

In this section we present the analysis that accounts for the average contribution of the non-punishers in the regressions. The results are reported in Table F.5. We see that the average contribution of non-punishers is highly predictive of the normative valences of non-punishers. This relationship is very similar to the main results of the paper: the more non-punishers contribute the less appropriate they consider undercontribution and punishment. However, we find no significant relationship between the punishers' contributions (neither in the first round nor averaged over all rounds) and the normative valences of non-punishers. One possible reason for this finding is that the non-punishers' contributions are a function of the initial (and average) punishers' contributions as we have shown in Table F.3 and as can be seen in Figure 1. Thus, the average contribution of non-punishers masks variation from the punishers contribution and therefore, the results seem to be affected. Thus, the estimation has an issue with multicollinearity, as the average and the initial contribution of punishers is highly correlated with the average contribution of non-punishers (r= 0.685, $p \leq 0.001$; r= 0.685, $p \leq 0.001$). Thus, this high multicollinearity, and the potential issue of endogeneity, should make the reader cautions in interpreting the results from Table F.5.

Panel A: Original experiment (Social Norms)

	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q
	Punishers			Non-punishers			Punishers			Non-punishers		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Constant	3.68***	4.41***	3.32***	4.22***	4.78***	3.75***	3.65***	4.40***	3.31***	4.22***	4.78***	3.73***
	(0.38)	(0.32)	(0.46)	(0.25)	(0.18)	(0.26)	(0.37)	(0.31)	(0.45)	(0.25)	(0.17)	(0.26)
Cont.Pun$_{t\in\{1\}}$	−0.02	−0.01	−0.01	−0.0003	0.003	−0.01						
	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)						
Cont.Pun$_{t\in\{1,...,15\}}$							−0.06**	−0.03*	−0.04	−0.01	0.02	−0.01
							(0.02)	(0.02)	(0.03)	(0.01)	(0.01)	(0.02)
Cont.Non-punishers$_{t\in\{1,...,15\}}$	−0.002	0.002	−0.01	−0.06***	−0.03**	−0.05***	0.04	0.03	0.02	−0.05**	−0.04***	−0.04*
	(0.02)	(0.02)	(0.03)	(0.02)	(0.01)	(0.02)	(0.03)	(0.02)	(0.04)	(0.02)	(0.01)	(0.02)
Group specific effects	×	×	×	✓	✓	✓	×	×	×	✓	✓	✓
Log Likelihood	-50.06	-40.25	-60.16	-164.25	-124.51	-191.06	-47.98	-39.02	-59.19	-163.47	-122.98	-190.73
Observations	53	53	53	159	159	159	53	53	53	159	159	159

Panel B: Follow-up experiment (Personal Norms)

	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q
	Punishers			Non-punishers			Punishers			Non-punishers		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Constant	4.72***	5.96***	6.51***	3.52***	4.53***	2.43***	4.66***	5.78***	6.43***	3.72***	4.61***	2.49***
	(0.75)	(0.62)	(0.78)	(0.44)	(0.33)	(0.40)	(0.67)	(0.61)	(0.67)	(0.44)	(0.32)	(0.39)
Cont.Pun$_{t\in\{1\}}$	−0.02	−0.02	−0.03	0.03*	0.01	0.01						
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)						
Cont.Pun$_{t\in\{1,...,15\}}$							−0.10***	−0.02	−0.13***	0.01	0.01	0.0002
							(0.03)	(0.03)	(0.04)	(0.02)	(0.02)	(0.02)
Cont.Non-punishers$_{t\in\{1,...,15\}}$	−0.05	−0.08**	−0.17***	−0.02	−0.01	0.03	0.03	−0.06	−0.07	−0.03	−0.02	0.03
	(0.05)	(0.04)	(0.05)	(0.03)	(0.02)	(0.02)	(0.05)	(0.05)	(0.05)	(0.03)	(0.02)	(0.03)
Group specific effects	×	×	×	✓	✓	✓	×	×	×	✓	✓	✓
Log Likelihood	-52.7	-44.66	-54.56	-171.5	-137.1	-160.22	-49.15	-45.32	-49.22	-172.61	-137.02	-159.83
Observations	41	41	41	123	123	123	41	41	41	123	123	123

Notes: *p<0.10;**p<0.05;***p<0.01;

Table F.5: Average normative valence and non-punishers' contributions.

The table depicts the average normative valence as a function of different measures of punishers' public good contributions while controlling for the non-punishers' average contributions. The top panel depicts the estimations of the original study were social norms where elicited, while the bottom panel depicts the estimations of the follow-up study where personal norms were elicited. FC-Q, MC-Q, and Pun-Q denote the average normative valence in the *full contributing question*, *medium contributing question* and the *punishment question*, respectively. *Cont.Pun*$_{t\in\{1\}}$ denotes the punishers' contribution in the first round. *Cont.Pun*$_{t\in\{1,...,15\}}$ denotes the average punishers' contribution in the first fifteen rounds. *Cont.Non-punishers*$_{t\in\{1,...,15\}}$ denotes the average non-punishers' contribution in the public good game. Heterogeneity on the group level is accounted for by group-specific random-intercept effects.

F.5 Alternative measures of cooperative and uncooperative punishers

In this section we focus on two alternative ways to classify the punishers as cooperative or uncooperative. In the main part of the paper, we focused predominately on the punishers' initial contributions to the public good. The results with this classification can be interpreted causally for non-punishers as punishers have not observed the behavior of non-punishers yet. However, this measure does not fully represent the experience during the game of non-punishers. Therefore, we have also used the average of the first two, five and all rounds to classify punishers as cooperative and uncooperative, and have seen that the results are even stronger (see Table F.3). This, however, came at the cost of possible endogeneity. In this section, we introduce two further classification to distinguish between cooperative and uncooperative punishers. Note though that these two measures might be prone to endogeneity problems as well. Thus, the reader should keep this problem in mind and consider the results in this section with caution.

The first of the alternative classifications is how often a punisher has used his punishment power. To classify the experience of non-punishers to have a cooperative and uncooperative punishers, we focus on how often they have been punished in the game. Table F.6 reports the estimations of this alternative classification. We can see that punishers who punish more often also consider punishment more appropriate in both social and personal norms estimations. However, the amount of punishment used does not correlate with punishers' attitudes towards undercontribution: punishers who punish more do not differ in their social and personal norms in Questions 20 and 10 from punishers who contribute less.

For non-punishers, we see that the punishment received in the game is highly predictive of their social norm perception. Specifically, we see that non-punishers who were punished *more* consider undercontribution and punishment (*full contributing question*and *punishment question*) as *more* socially appropriate. However, the personal norm does not seem to change as a function of the own experience. These results support the insights presented in the main part of the paper: the more abusive the experience of non-punishers the more socially appropriate they consider this abusive behavior, while keeping the personal norm robust.

The second alternative classification relates to how often a punisher undercontributes in period t relative to the contribution of non-punishers in period $t-1$. Table F.7 reports on the estimations of this alternative classification. The results are very similar to the main part of the paper. Punishers who undercontribute consider this undercontribution and punishment as more socially as well as personally appropriate. Non-punishers who experienced a punisher who undercontributed consider punishment and undercontribution as socially more acceptable (for the *full contributing question* the results have the same sign but are not significant). But again, the personal norms of non-punishers do not change. These results again support the insights presented in the main part of the paper.

	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q
	Punishers			Non-punishers			Punishers			Non-punishers		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Constant	3.42***	4.39***	2.79***	3.14***	4.35***	2.76***	3.58***	4.07***	2.72***	3.28***	4.37***	2.98***
	(0.14)	(0.12)	(0.16)	(0.08)	(0.05)	(0.08)	(0.33)	(0.27)	(0.36)	(0.14)	(0.10)	(0.13)
NumberTimesPunUsed	−0.01	−0.01	0.05**				0.01	0.04	0.09***			
	(0.02)	(0.02)	(0.02)				(0.03)	(0.03)	(0.03)			
NumberTimesPunReceived				0.05***	0.01	0.03**				0.03	0.01	0.01
				(0.01)	(0.01)	(0.02)				(0.02)	(0.01)	(0.02)
Original experiment (Social Norms)	✓	✓	✓	✓	✓	✓	×	×	×	×	×	×
Group specific effects	×	×	×	✓	✓	✓	×	×	×	✓	✓	✓
Log Likelihood	-51.64	-40.41	-58.21	-160.72	-122.74	-191.19	-53.92	-46.48	-57.65	-169.46	-134.34	-158.12
Observations	53	53	53	159	159	159	41	41	41	123	123	123

Notes: *p<0.10;**p<0.05;***p<0.01;

Table F.6: Estimation of the average normative valence as a function of punishment.

The table depicts the average normative valence as a function of the punishers' punishment. The first 6 models depict the estimations of the original study were social norms where elicited, while the last 6 models depict the estimations of the follow-up study where personal norms were elicited. FC-Q, MC-Q, and Pun-Q denote the average normative valence in the *full contributing question*, *medium contributing question* and the *punishment question*, respectively. *NumberTimesPunUsed* denotes how often a punisher has used the punishment strategy. *NumberTimesPunReceived* denotes how often a non-punisher has received punishment. Heterogeneity on the group level is accounted for by group-specific random-intercept effects.

Panel A: Original experiment (Social Norms)

	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q
	Punishers			Non-punishers			Punishers			Non-punishers		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Constant	3.20***	4.23***	2.91***	3.22***	4.39***	2.77***	3.32***	4.29***	3.00***	3.26***	4.39***	2.81***
	(0.11)	(0.09)	(0.13)	(0.08)	(0.05)	(0.08)	(0.09)	(0.08)	(0.11)	(0.07)	(0.05)	(0.07)
$\#\{c_p^t < c_{-p}^{t-1}\}$	0.05***	0.03*	0.04*	0.02	−0.004	0.03*						
	(0.02)	(0.02)	(0.02)	(0.01)	(0.01)	(0.01)						
$\#\{c_p^t < c_{-p}^{t-1}\} \geq 10$							0.43*	0.21	0.45	0.27	−0.08	0.38**
							(0.24)	(0.20)	(0.29)	(0.18)	(0.12)	(0.17)
Group specific effects	×	×	×	✓	✓	✓	×	×	×	✓	✓	✓
Log Likelihood	-48.1	-39.26	-58.59	-165.71	-123.17	-191.63	-50.13	-40.17	-59.32	-163.13	-120.45	-188.51
Observations	53	53	53	159	159	159	53	53	53	159	159	159

Panel B: Follow-up experiment (Personal Norms)

	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q
	Punishers			Non-punishers			Punishers			Non-punishers		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Constant	3.24***	4.29***	2.92***	3.46***	4.45***	2.94***	3.72***	4.46***	3.46***	3.42***	4.42***	2.96***
	(0.21)	(0.19)	(0.24)	(0.14)	(0.10)	(0.12)	(0.16)	(0.13)	(0.18)	(0.09)	(0.07)	(0.08)
$\#\{c_p^t < c_{-p}^{t-1}\}$	0.10***	0.04	0.14***	−0.003	−0.01	0.02						
	(0.03)	(0.03)	(0.04)	(0.02)	(0.02)	(0.02)						
$\#\{c_p^t < c_{-p}^{t-1}\} \geq 10$							−0.12	0.14	0.94**	0.17	0.02	0.37*
							(0.41)	(0.35)	(0.46)	(0.24)	(0.18)	(0.21)
Group specific effects	×	×	×	✓	✓	✓	×	×	×	✓	✓	✓
Log Likelihood	-50.1	-47.02	-55.11	-170.31	-134.31	-157.71	-53.96	-47.79	-59.12	-167.7	-132	-154.15
Observations	41	41	41	123	123	123	41	41	41	123	123	123

Notes: *p<0.10;**p<0.05;***p<0.01;

Table F.7: Estimation of the average normative valence as a function of how often the punisher uncontributed in the public good game

The table depicts the average normative valence as a function of of how often the punisher uncontributed in the public good game. The top panel depicts the estimations of the original study were social norms where elicited, while the bottom panel depicts the estimations of the follow-up study where personal norms were elicited. FC-Q, MC-Q, and Pun-Q denote the average normative valence in the *full contributing question*, *medium contributing question* and the *punishment question*, respectively. $\#\{c_p^t < c_{-p}^{t-1}\}$ denotes the number of times the punisher contribute in period t less than the average non-punisher contributed in $t-1$. $\#\{c_p^t < c_{-p}^{t-1}\} \geq 10$ denotes a dummy with value one if the punisher at least 10 times contribute in period t less than the average non-punisher contributed in $t-1$. Heterogeneity on the group level is accounted for by group-specific random-intercept effects.

F.6 Estimation of slopes

To make use of the detailed data on normative valences for all five hypothetical contributions of the punisher we can focus on the slope of the norm function formed by normative valences. However, as we can see in Figures E.3 and E.4, the norm functions are non-linear. Thus, to account for the non-linearity of the slopes we use a generalized additive model (GAM) to estimate the effect of cooperative vs. uncooperative punishers on normative valences. Table F.8 reports on the estimations of the GAM-model. We can see that all our results reported in the main part of the paper can be replicated using slopes. We find that cooperative punishers consider undercontribution and punishment less appropriate than uncooperative punishers. More importantly, we again find that non-punishers assigned an uncooperative punisher consider it *more* appropriate for the punisher to undercontribute and to punish than non-punishers assigned a cooperative punisher do. Further, we see that the personal norms do not differ between non-punishers assigned a cooperative or an uncooperative punisher but for punishers we observe a similar pattern as in the social norm elicitation.

Panel A: Original experiment (Social Norms)						
	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q
		Punishers			Non-punishers	
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	3.61**** (0.16)	4.38**** (0.13)	3.34**** (0.19)	3.49**** (0.10)	4.37**** (0.07)	3.05**** (0.11)
Cooperative Punisher	-0.33* (0.19)	-0.07 (0.16)	-0.39* (0.22)	-0.27** (0.12)	0.02 (0.09)	-0.25* (0.14)
Subject specific effects	✓	✓	✓	✓	✓	✓
Group specific effects	×	×	×	✓	✓	✓
Observations	265	265	265	795	795	795
Log Likelihood	-364.34	-416.31	-435.34	-1109.21	-1139.15	-1196.23
Panel B: Follow-up experiment (Personal Norms)						
	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q
		Punishers			Non-punishers	
	(1)	(2)	(3)	(4)	(5)	(6)
Constant	3.73**** (0.27)	4.50**** (0.23)	4.13**** (0.30)	3.38**** (0.16)	4.38**** (0.12)	2.94**** (0.14)
Cooperative Punisher	-0.04 (0.32)	-0.02 (0.27)	-0.76** (0.36)	0.10 (0.19)	0.05 (0.14)	0.10 (0.17)
Subject specific effects	✓	✓	✓	✓	✓	✓
Group specific effects	×	×	×	✓	✓	✓
Observations	205	205	205	615	615	615
Log Likelihood	-359.46	-360.69	-323.48	-972.4	-1029.24	-1064
Notes:						*p<0.10;**p<0.05;***p<0.01;

Table F.8: GAM regression on the slope of the normative valence.

The table depicts the estimation results of a generalized additive model (GAM) on the the slope of the normative valence. Thin plate regression splines are used to account for the non-linear function of the slope of the normative valences. The top panel depicts the estimations of the original study were social norms where elicited, while the bottom panel depicts the estimations of the follow-up study where personal norms were elicited. FC-Q, MC-Q, and Pun-Q denote the average normative valence in the *full contributing question*, *medium contributing question* and the *punishment question*, respectively. *Cooperative Punisher* denotes a dummy with value one if the punisher contributed above the median in the first round of the public good game. Heterogeneity on the group level is accounted for by group-specific random-intercept effects. Heterogeneity on the subject level is accounted for by subject-specific random-intercept effects.

F.7 Analysis of Free-Riding

As we know from the literature on public good games, free-riding (i.e., a contribution of zero) is sometimes considered differently than small contributions and undercontribution (Fischbacher et al., 2001). Therefore, one might ask whether there is a difference in the normative valences if we look at these two types of contributions separately. As we can see from Figure E.3, there is rather high agreement on free-riding being socially unacceptable, while full contribution is considered highly acceptable. The figure also indicates that most of the results are driven by the less clear situations (i.e., a contribution between 5 and 15). Nevertheless, we show the average normative valences for free-riding in both experiments for punishers and non-punishers in cooperative punisher-groups and uncooperative punisher-groups in Figure F.2. Figure F.3 shows the average normative valences for scenarios where the punisher contributed to the public good (i.e. contribution is bigger than zero). Further, Table F.9 reports the regressions of normative valences by using the punishers initial PGG contribution as a continuous measure.

We see very consistent agreement that free-riding is socially unacceptable (on a scale form 1 to 7). Further, we see that there is more disagreement in the personal norm than in the social norm, as the variances are substantially larger in the follow-up experiment compared to the original experiment. Further, we see that the only significant difference between cooperative punisher-groups and uncooperative punisher-groups is the perception of non-punishers in the personal norm statement. Non-punishers assigned a punisher who initially contributed more than the median punisher consider free-riding *more* socially acceptable than non-punishers assigned a punisher who initially contributed less than the median punisher.

More importantly, we see that all the results described in the main part of the paper are also found here if we focus on scenarios with the punisher who contributes less than the non-punishers but is not free-riding (i.e. is contributing between 5 and 20 points).

Panel A: Zero contribution by the punisher

	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q
	Punishers			Non-punishers			Punishers			Non-punishers		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Constant	1.09***	1.37***	1.38***	1.02***	1.12***	1.03***	2.04***	2.19***	1.78***	0.97***	1.33***	1.47***
	(0.07)	(0.16)	(0.23)	(0.04)	(0.07)	(0.03)	(0.50)	(0.41)	(0.35)	(0.16)	(0.19)	(0.20)
Cont.Pun$_{t\in\{1\}}$	−0.001	−0.01	−0.02	0.001	0.01	−0.0002	−0.04	−0.04	−0.03	0.03**	0.02	−0.01
	(0.01)	(0.01)	(0.02)	(0.003)	(0.01)	(0.002)	(0.04)	(0.03)	(0.03)	(0.01)	(0.02)	(0.02)
Original experiment (Social Norms)	✓	✓	✓	✓	✓	✓	×	×	×	×	×	×
Group specific effects	×	×	×	✓	✓	✓	×	×	×	✓	✓	✓
Log Likelihood	-4.6	-47.48	-64.8	-26.39	-109.95	43.25	-77.03	-68.9	-61.61	-167.47	-185.32	-190.93
Observations	53	53	53	159	159	159	41	41	41	123	123	123

Panel B: Non-zero contribution by the punisher

	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q	FC-Q	MC-Q	Pun-Q
	Punishers			Non-punishers			Punishers			Non-punishers		
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)	(9)	(10)	(11)	(12)
Constant	4.29***	5.21***	3.67***	4.04***	5.22***	3.60***	4.48***	5.43***	4.57***	3.73***	5.06***	3.30***
	(0.21)	(0.17)	(0.26)	(0.16)	(0.10)	(0.15)	(0.30)	(0.28)	(0.36)	(0.19)	(0.15)	(0.19)
Cont.Pun$_{t\in\{1\}}$	−0.03*	−0.01	−0.01	−0.01	−0.004	−0.02**	−0.02	−0.02	−0.04	0.02	0.01	0.01
	(0.01)	(0.01)	(0.02)	(0.01)	(0.01)	(0.01)	(0.02)	(0.02)	(0.03)	(0.01)	(0.01)	(0.01)
Original experiment (Social Norms)	✓	✓	✓	✓	✓	✓	×	×	×	×	×	×
Group specific effects	×	×	×	✓	✓	✓	×	×	×	✓	✓	✓
Log Likelihood	-60.62	-49.67	-71.44	-199.58	-151.41	-226.01	-56.1	-53.03	-63.37	-183.62	-153.64	-182.91
Observations	53	53	53	159	159	159	41	41	41	123	123	123

Notes: *p<0.10;**p<0.05;***p<0.01;

Table F.9: Estimation of normative valence for a zero contribution and a non-zero contribution by the punishers.

The table depicts the normative valence for a zero contribution by the punishers and a non-zero contribution by the punishers as a function of the punishers' initial public good contributions. The top panel depicts the normative valence for a zero contribution by the punishers, while the bottom panel depicts the normative valence for a non-zero contribution by the punishers. The first 6 models depict the estimations of the original study were social norms where elicited, while the last 6 models depict the estimations of the follow-up study where personal norms were elicited. FC-Q, MC-Q, and Pun-Q denote the average normative valence in the *full contributing question*, *medium contributing question* and the *punishment question*, respectively. *Cont.Pun*$_{t\in\{1\}}$ denotes the punishers contribution in the first round. Heterogeneity on the group level is accounted for by group-specific random-intercept effects.

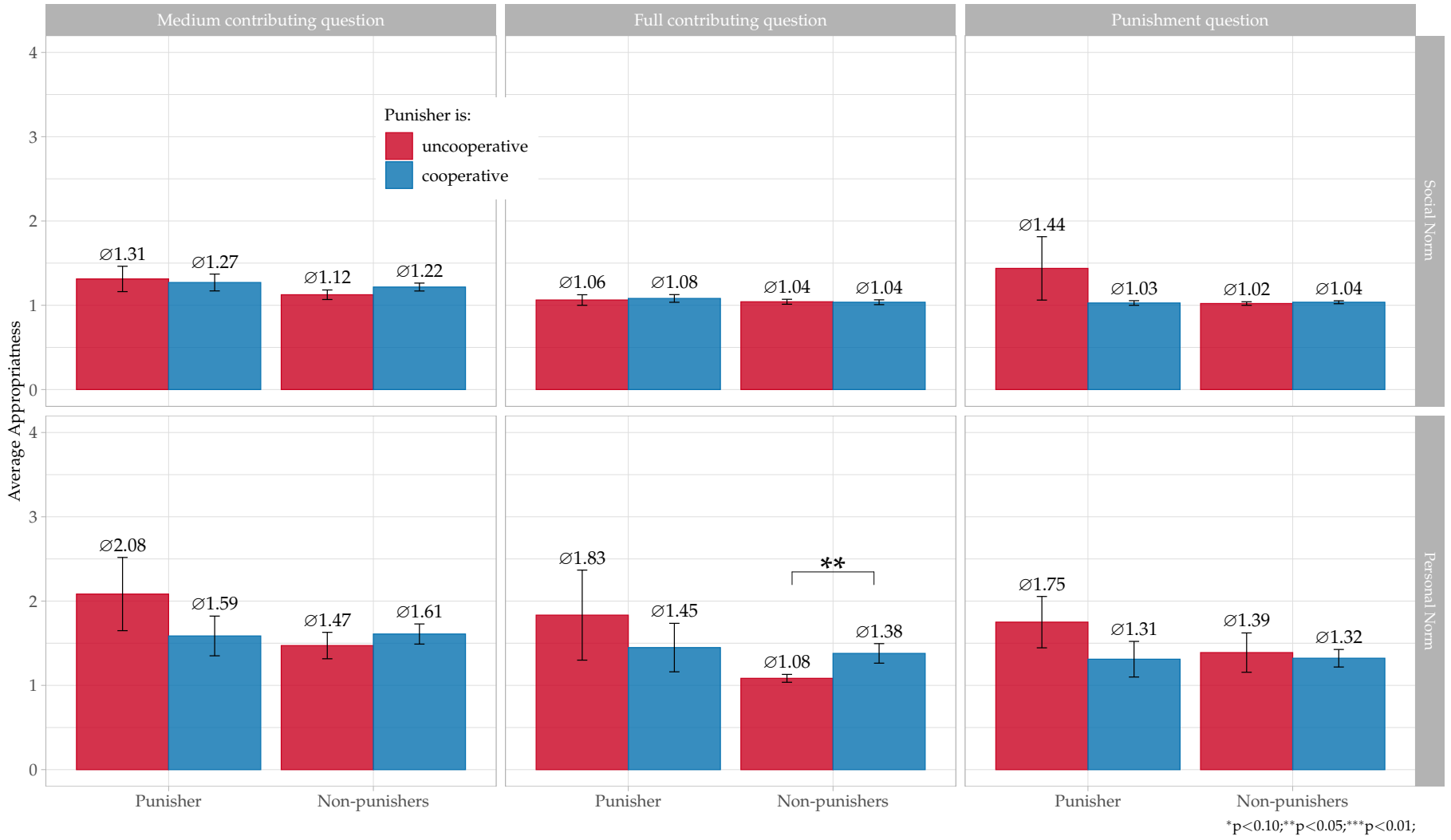


Figure F.2: Average normative valences for full freerding behavior (i.e. punisher contributes zero).

The figure depicts the normative valences of full free-riding (i.e. punisher contributes zero) reported by participants. The top panel depicts the social norms while the bottom panel depicts the personal norms. Left panels denote the normative valences for the *medium contributing question*, the mid-panels denote the normative valences for the *full contributing question*, while the right panels denote the normative valences for the *punishment question*. Blue bars present the averages normative valences in the cooperative punisher-groups (i.e. the punisher contributed above the median in the first round of the PGG), while red bars present the average normative valences in uncooperative punisher-groups. Error bars denote standard errors.

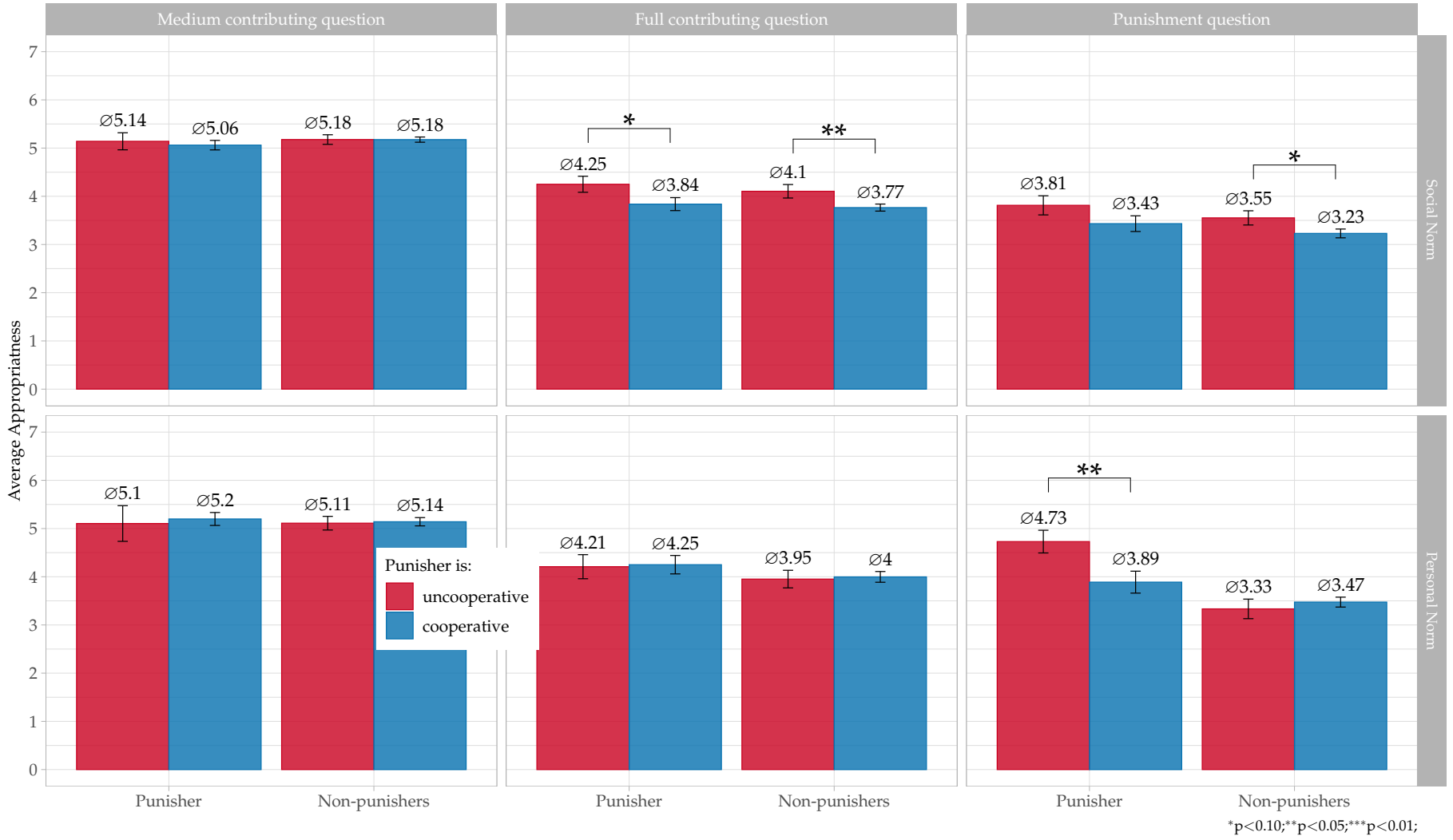


Figure F.3: Average normative valences for not full freerding behavior (i.e. punisher contributes at least 5).

The figure depicts the normative valences of not-full free-riding (i.e. punisher contributes at least 5 to the PGG) reported by participants. The top panel depicts the social norms while the bottom panel depicts the personal norms. Left panels denote the normative valences for the *medium contributing question*, the mid-panels denote the normative valences for the *full contributing question*, while the right panels denote the normative valences for the *punishment question*. Blue bars present the averages normative valences in the cooperative punisher-groups (i.e. the punisher contributed above the median in the first round of the PGG), while red bars present the average normative valences in uncooperative punisher-groups. Error bars denote standard errors.

G Instructions

G.1 Public Goods Game Instructions

General information
You are about to participate in a decision-making experiment. If you follow the instructions carefully, you can earn a considerable amount of money depending on your decisions and the decisions of the other participants. Your earnings will be paid to you in cash at the end of the experiment.

This set of instructions is for your private use only. During the experiment, you are not allowed to communicate with anybody. In case of questions, please raise your hand. Then we will come to your seat and answer your questions. Any violation of this rule excludes you immediately from the experiment and all payments. The funds for conducting this experiment have been provided by Max Planck Institute for Research on Collective Goods.

Throughout the experiment, you will make decisions about amounts of tokens. At the end of the experiment, all tokens you have will be converted into Euros at the exchange rate 0.20 Euro per token and paid you in cash in addition to the show-up fee of 5 Euros.

During the experiment, all your decisions will be treated confidentially. This means that none of the other participants will be able to associate your decisions with your personal identity.

PART I

Part I of the experiment will consist of 15 decision-making periods. At the beginning of the experiment, you will be matched with 3 other people in this room. Therefore, there are 4 people, including yourself, participating in your group. You will be matched with the same people during the entire Part I of the experiment. For the purpose of the experiment, you and the other group members will be randomly assigned labels A, B, C, and D, which will identify you and the others throughout Part I of the experiment. None of the participants knows your personal identity in the group.

First Stage of a Period
Before each period, you and each other person in your group, will be given the endowment of 20 tokens. At the first stage of each period, you will be asked to allocate your endowment between a private account and a group account. The other members of your group will be asked to do the same. The tokens that you place in the private account have a return of 1. This means that at the end of the first stage of each period your private account will contain exactly the amount of tokens you put into the private account at the beginning of the period. Nobody except yourself benefits from your private account. The tokens that you place to the group account are added to the tokens that the other three members of your group have placed to the group account. The tokens in the group account have a return of 2. Every member of the group benefits equally from the group account. Specifically, the total amount of tokens placed to the group account by all group members is multiplied by 2 and then is equally divided among the four group members. Hence, your share of the group account is

$$2 * (\text{sum of tokens in the group account})/4$$

Thus, at the end of the first stage of each period, the number of tokens that you have is equal to the number of tokens you place in your private account plus your share of the group account.

$$\text{Payoff} = 20 - \text{tokens you put into the group account} + 2 * (\text{sum of tokens in the group account})/4$$

Here are three examples to make this clear:

1. Suppose you place 0 tokens to the group account and 20 tokens in the private account, and the other members of your group place a total of 45 tokens to the group account. The sum of tokens in the group account is 45. Your share of the group account would be 2 * 45 / 4 = 22.5 tokens. Each other member of the group would also receive a share of the group account equal to 22.5 tokens. The amount of tokens that you have at the end of the first stage is, thus, equal to 20 + 22.5 = 42.5 tokens. Each other member of your group receives on average 27.5 tokens.

2. Suppose you place 15 tokens to the group account and 5 tokens in the private account, and the other members of your group place a total of 45 tokens to the group account. The sum of tokens in the group account is 60. Your share of the group account would be 2 * 60 / 4 = 30 tokens. Each other member of the group would also receive a share of the group account equal to 30 tokens. The amount of tokens that you have at the end of the first stage is, thus, equal to 5 + 30 = 35 tokens. Each other member of your group receives on average 35 tokens.
3. Suppose you place 15 tokens to the group account and 5 tokens in the private account, and the other members of your group place a total of 10 tokens to the group account. The sum of tokens in the group's account is 25. Your share of the group account would be 2 * 25 / 4 = 12.5 tokens. Each other member of the group would also receive a share of the group account equal to 12.5 tokens. The amount of tokens that you have at the end of the first stage is, thus, equal to 5 + 12.5 = 17.5 tokens. Each other member of your group receives on average 29.1 tokens.

Second Stage of a Period
In the second stage of each period, only the member of your group who was labeled D is active. The group members who received labels A, B, and C do not make any decisions in the second stage of each period.

If your label in the group is D, you will be asked to react to the decisions made by group members A, B, and C during the first stage of each period. At this point, you will already know the decisions taken by each group member at the first stage and the number of tokens they have after the first stage. You will decide whether you want to subtract tokens from any other group member or not. The group members that you decide to subtract tokens from will lose the amount of tokens you choose. The decisions you make at this stage will not change the amount of tokens that you have after the first stage.

You may subtract different amounts of tokens from different group members. The total amount of tokens that you choose to subtract from the group members A, B, and C may not exceed 30 tokens. Any group member can only lose maximum the amount of tokens he or she has. For example, if at the end of the first stage group members A, B, and C have 10, 15, and 20 tokens, respectively, and you choose to subtract 15, 10, and 0 tokens from them, then group members A, B, and C will be left with 0, 5, and 20 tokens.

Information about the Choices and Tokens in the End of a Period
At the end of each period, each member of the group will be informed about:

- His/her contribution to the group account;
- The amount of tokens contributed by all group members individually to the group account;
- His/her share of the group account (remember, it is the same for all group members);
- If you are member A, B, or C: how many tokens were subtracted from you by member D;
- If you are member A, B, or C: the number of tokens at the end of the period, which is equal to the number of tokens in the private account plus the share of tokens from the group account minus the number of tokens subtracted by D;
- If you are member D: the number of tokens at the end of the period, which is equal to the number of tokens in the private account plus the share of tokens from the group account.

Structure of Part I of the Experiment
The structure of the experiment in all 15 periods is identical. In the first stage of each period, each group member A, B, C, and D chooses how to split 20 tokens between private and group accounts. Then all group members receive the returns from both accounts. In the second stage of the period, group member D can subtract tokens from group members A, B, and C. At the end of the period, all members are informed about the decisions of others in the group, and the number of tokens they have.

Money Earned in Part I of the Experiment
In the end of the experiment, the computer will randomly choose one period for which you and other members of

your group will be paid. Your income at the end of Part I of the experiment is equal to the amount of tokens at the end of this randomly chosen period times the exchange rate of 0.20 Euro for 1 token.

This is the end of the instructions for Part I. If you have any questions, please raise your hand and an experimenter will come by to answer them.

G.2 Norm Elicitation Instructions for the PGG subjects

PART II

Description of the Task (Screen 1)
On the following screens, you will read the descriptions of a series of hypothetical situations that could have taken place in Part I of the experiment. These descriptions correspond to situations in which a person, acting in the role of member D (who will be called Individual D), makes decisions about the amounts of tokens to be placed to the group account and decisions to subtract tokens from members A, B, and C. For each situation, you will be given a description of the decision faced by Individual D. This description will include several possible choices available to this Individual.

After you have read the description of the decision, you will be asked to evaluate the different possible actions available and to decide, for each of the actions, whether taking that action would be "socially appropriate" and "consistent with moral or proper social behavior" or "socially inappropriate" and "inconsistent with moral or proper social behavior." By socially appropriate, we mean behavior that most people agree is the "correct" or "ethical" thing to do. Another way to think about what we mean is that if Individual D were to select a socially inappropriate choice, then someone else might be angry at Individual D for doing so.

In each of your responses, we would like you to answer as truthfully as possible, based on your opinion of what constitutes socially appropriate or socially inappropriate behavior.

To give you an idea of how the experiment will proceed, we will go through an example and show you how you will indicate your responses. On the next screen you will see an example of a situation. Click OK when you are ready to go on.

Example Situation (Screen 2)
Bob is at a café. While there, Bob notices that someone has left a wallet at one of the tables. Bob must decide what to do. He has four possible choices: take the wallet, ask others nearby if the wallet belongs to them, leave the wallet where it is, or give the wallet to the bartender. Bob can choose only one of these four options. The table on the right of the screen presents a list of the possible actions available to Bob. For each of the actions, please indicate on the scale from 1 to 7 how socially appropriate you believe choosing that option is. To indicate your response, please click on the corresponding cell. Please make sure you make an assessment for each possible choice in each row of the table.

Screen 3
In what follows, you will be asked to assess the appropriateness of the actions in three situations that could have arisen in Part I of the experiment. For each action in each situation please indicate the extent to which you believe taking that action would be "socially appropriate" and "consistent with moral or proper social behavior" or "socially inappropriate" and "inconsistent with moral or proper social behavior." By socially appropriate we mean behavior that most people agree is the "correct" or "ethical" thing to do.

Payment
For each situation that follows, you will read its description. You will then indicate your appropriateness rating by placing a check mark in the corresponding cell.

At the end of Part II of the experiment, in order to determine your payment, we will randomly select one of the situations. For this situation, we will also randomly select one of the possible choices that Individual D could make. Thus, we will select both a scenario and one possible choice at random. This means that when you make your choices you should make each of them as if it is the one for which you will be paid.

Your payment in this part of the experiment will depend on whether your response to the choice thus selected is the same as the response made by the most people with the same role as you in Part I of the experiment (who are in this room). In particular, if in Part I of the experiment you were member A, B, or C, then your response to a selected choice will be compared to the responses of all people in this room who were members A, B, and C in Part I. If you were member D, then your response to a selected choice will be compared to the responses of all people in this room who were members D. If you give the same response as that most frequently given by other members with the same role, then you will receive € 8. This amount will be paid to you, in cash, at the conclusion of the experiment.

For instance, there are overall $N/4$ participants who were members D in the previous part of the experiment and $3N/4$ participants who were members A, B, or C (including you). Suppose we were to select the example situation from the last screen and the possible choice "Leave the wallet where it is," and your response had been 3, "somewhat socially inappropriate." Then, if you are member D, you would receive € 8 if this was the response selected by most of other $N/4 - 1$ members D in today's session. If you were member A, B, or C, you would receive € 8 if this was the response selected by most of other $3N/4 - 1$ members A, B, and C in today's session. If your response is not the same as that of the majority of others with the same role as you, you will receive nothing in this part of the experiment.

Please click OK when you are ready to go on. If you have any questions, please raise your hand and wait for the experimenter to come.

Screen 4
Imagine that members A, B, C have each placed 10 tokens (out of 20) to the group account in the previous period. Look at the table on the right-hand side of the screen and consider five possible amounts that Individual D could place to the group account (presented in rows). Please indicate on the scale from 1 to 7 how socially appropriate you believe choosing each of these amounts to be, given the amounts that others contributed to the group account in the previous period.

Remember: when we select a scenario and an action for payment, you will only receive € 8 if your response is the same as the most frequent response made by other ⟨NUMBER⟩ members ⟨ROLE⟩ in this room.

Screen 5
Imagine that members A, B, C have each placed 20 tokens (out of 20) to the group account in the previous period. Look at the table on the right-hand side of the screen and consider five possible amounts that Individual D could place to the group account (presented in rows). Please indicate on the scale from 1 to 7 how socially appropriate you believe choosing each of these amounts to be, given the amounts that others contributed to the group account in the previous period.

Remember: when we select a scenario and an action for payment, you will only receive € 8 if your response is the same as the most frequent response made by other ⟨NUMBER⟩ members ⟨ROLE⟩ in this room.

Screen 6
Imagine that members A, B, C, and D have made their choices in the first stage of a period. Namely, members A, B, and C placed 10 tokens each to the group account and individual D placed the amount of tokens equal to one of the five options listed on the right part of the screen. For each of the amounts that individual D could have placed to the group account, please indicate how socially appropriate you believe subtracting tokens from individuals A, B, and C is, given the amount that members A, B, C, and D contributed to the group account.

Remember: when we select a scenario and an action for payment, you will only receive € 8 if your response is the same as the most frequent response made by other ⟨NUMBER⟩ members ⟨ROLE⟩ in this room.

PART III

Description of the Task (Screen 1)
In this final part of the experiment we ask you to evaluate the social appropriateness of actions in the same three situations as before. The only difference is that now you will be paid if your evaluation is the same as the evaluation of the majority of two groups of participants who have already made their evaluation decisions. The first group is the participants who had other role than you (members ⟨OTHER ROLE⟩ in this room) who have just made their evaluations in Part II. The second group is a separate group of other participants who took part in the experiment before and who evaluated the same situations as in the previous part but without actually making real choices as in Part I. In particular, these other participants were given the same instructions of Part I as you did and then evaluated social appropriateness in exactly same way that you just did, with the only difference that for the payment they were matched with everyone in their respective sessions.

Payment (Screen 2)
As before, for your payment we will choose one random situation and one random action that you evaluate. This

means that when you make your choices you should make each of them as if it is the one for which you will be paid. Your payment in this part of the experiment will depend on whether your response to the selected choice is the same as the response made by the most people in a group who have already chosen. For example, if you are matched with members ⟨OTHER ROLE⟩, then your payment depends on how members ⟨OTHER ROLE⟩ chose in the previous part of the experiment. Remember, the members ⟨OTHER ROLE⟩ when choosing in Part II were paid if they chose the same answer as the majority of other members ⟨OTHER ROLE⟩. The same holds for the separate group of other participants. If you are matched with them, then your payment depends on how they chose in a separate experiment. Remember, these participants were paid if they chose the same answer as the majority of other participants in their session.

If you give the same response as that most frequently given by other members in one of the two groups, then you will receive € 8. This amount will be paid to you, in cash, at the conclusion of the experiment. Please click OK when you are ready to go on. If you have any questions, please raise your hand and wait for the experimenter to come.

Screen 4
Put yourself in the shoes of MEMBERS ⟨OTHER ROLE⟩ in this room who have just provided their evaluations of social appropriateness of the actions of Individual D in the following situation that you have also seen. Remember, that they were paid if they guessed as the majority in their own group of members ⟨OTHER ROLE⟩. Imagine that members A, B, C have each placed 10 tokens (out of 20) to the group account in the previous period. Look at the table on the right-hand side of the screen and consider five possible amounts that Individual D could place to the group account (presented in rows). Please indicate on the scale from 1 to 7 how socially appropriate you believe choosing each of these amounts to be, given the amounts that others contributed to the group account in the previous period.

Remember: when we select a scenario and an action for payment, you will only receive € 8 if your response is the same as the most frequent response made by MEMBERS ⟨OTHER ROLE⟩ in this room in the previous part of the experiment.

Screen 5
Put yourself in the shoes of MEMBERS ⟨OTHER ROLE⟩ in this room who have just provided their evaluations of social appropriateness of the actions of Individual D in the following situation that you have also seen. Remember, that they were paid if they guessed as the majority in their own group of members ⟨OTHER ROLE⟩.

Imagine that members A, B, C have each placed 20 tokens (out of 20) to the group account in the previous period. Look at the table on the right-hand side of the screen and consider five possible amounts that Individual D could place to the group account (presented in rows). Please indicate on the scale from 1 to 7 how socially appropriate you believe choosing each of these amounts to be, given the amounts that others contributed to the group account in the previous period.

Remember: when we select a scenario and an action for payment, you will only receive € 8 if your response is the same as the most frequent response made by MEMBERS ⟨OTHER ROLE⟩ in this room in the previous part of the experiment.

Screen 6
Put yourself in the shoes of MEMBERS ⟨OTHER ROLE⟩ in this room who have just provided their evaluations of social appropriateness of the actions of Individual D in the following situation that you have also seen. Remember, that they were paid if they guessed as the majority in their own group of members ⟨OTHER ROLE⟩.

Imagine that members A, B, C, and D made their choices in the first stage of a period. Namely, members A, B, and C placed 10 tokens each to the group account and individual D placed the amount of tokens equal to one of the five options listed on the right part of the screen. For each of the amounts that individual D could have placed to the group account, please indicate how socially appropriate you believe subtracting tokens from individuals A, B, and C is, given the amount that they contributed to the group account.

Remember: when we select a scenario and an action for payment, you will only receive € 8 if your response is the same as the most frequent response made by MEMBERS ⟨OTHER ROLE⟩ in this room in the previous part of the experiment.

Screen 7
Put yourself in the shoes of OTHER PARTICIPANTS who gave evaluations in the previous experiment who have provided their evaluations of social appropriateness of the actions of Individual D in the following situation that you have also seen. Remember, that they were paid if they guessed as the majority in their own group.

Imagine that members A, B, C have each placed 10 tokens (out of 20) to the group account in the previous period. Look at the table on the right-hand side of the screen and consider five possible amounts that Individual D could place to the group account (presented in rows). Please indicate on the scale from 1 to 7 how socially appropriate you believe choosing each of these amounts to be, given the amounts that others contributed to the group account in the previous period.

Remember: when we select a scenario and an action for payment, you will only receive € 8 if your response is the same as the most frequent response made by OTHER PARTICIPANTS in a separate the experiment.

Screen 8
Put yourself in the shoes of OTHER PARTICIPANTS who gave evaluations in the previous experiment who have provided their evaluations of social appropriateness of the actions of Individual D in the following situation that you have also seen. Remember, that they were paid if they guessed as the majority in their own group.

Imagine that members A, B, C have each placed 20 tokens (out of 20) to the group account in the previous period. Look at the table on the right-hand side of the screen and consider five possible amounts that Individual D could place to the group account (presented in rows). Please indicate on the scale from 1 to 7 how socially appropriate you believe choosing each of these amounts to be, given the amounts that others contributed to the group account in the previous period.

Remember: when we select a scenario and an action for payment, you will only receive € 8 if your response is the same as the most frequent response made by OTHER PARTICIPANTS in a separate the experiment.

Screen 9
Put yourself in the shoes of OTHER PARTICIPANTS who gave evaluations in the previous experiment who have provided their evaluations of social appropriateness of the actions of Individual D in the following situation that you have also seen. Remember, that they were paid if they guessed as the majority in their own group.

Imagine that members A, B, C, and D made their choices in the first stage of a period. Namely, members A, B, and C placed 10 tokens each to the group account and individual D placed the amount of tokens equal to one of the five options listed on the right part of the screen. For each of the amounts that individual D could have placed to the group account, please indicate how socially appropriate you believe subtracting tokens from individuals A, B, and C is, given the amount that they contributed to the group account.

Remember: when we select a scenario and an action for payment, you will only receive € 8 if your response is the same as the most frequent response made by OTHER PARTICIPANTS in a separate the experiment.

G.3 Norm Elicitation Instructions in the follow-up study

PART II

Description of the Task (Screen 1)
On the following screens, you will read descriptions of a series of hypothetical situations that could have taken place in Part I of the experiment. These descriptions correspond to situations in which one person, acting in the role of member D (who will be called Individual D), makes decisions about the amounts of tokens to be placed in the group account and decisions to subtract tokens from members A, B, and C. For each situation, you will be given a description of the decision faced by Individual D. This description will include several possible choices available to this Individual.

After you read the description of the decision, you will be asked to evaluate the possible actions available to Individual D and to decide, for each of the actions, whether taking that action would be "socially appropriate" and "consistent with moral or proper social behavior" or "socially inappropriate" and "inconsistent with moral or proper social behavior." By socially appropriate, we mean behavior that most people agree is the "correct" or "ethical" thing to do. Another way to think about what we mean is that if Individual D were to select a socially inappropriate choice, then someone else might be angry at Individual D for doing so.

In each of your responses, we would like you to answer as truthfully as possible, based on your opinion of what constitutes socially appropriate or socially inappropriate behavior.

To give you an idea of how the experiment will proceed, we will go through an example and show you how you will indicate your responses. On the next screen you will see an example of a situation. Click OK when you are ready to go on.

Example Situation (Screen 2)
Bob is at a café. While there, Bob notices that someone has left a wallet at one of the tables. Bob must decide what to do. He has four possible choices: take the wallet, ask others nearby if the wallet belongs to them, leave the wallet where it is, or give the wallet to the bartender. Bob can choose only one of these four options. The table on the right of the screen presents a list of the possible actions available to Bob. For each of the actions, please indicate on the scale from 1 to 7 how socially appropriate you believe choosing that option is. To indicate your response, please click on the corresponding cell. Please make sure you make an assessment for each possible choice in each row of the table.

Screen 3
In what follows, you will be asked to assess the appropriateness of the actions in three situations that could have arisen in Part I of the experiment. For each action in each situation please indicate the extent to which you believe taking that action would be "socially appropriate" and "consistent with moral or proper social behavior" or "socially inappropriate" and "inconsistent with moral or proper social behavior." By socially appropriate we mean behavior that most people agree is the "correct" or "ethical" thing to do.

Payment
For each situation that follows, you will read its description. You will then indicate your appropriateness rating by placing a check mark in the corresponding cell.

At the end of the experiment, you will receive €2 for this part of the experiment.

Please click OK when you are ready to go on. If you have any questions, please raise your hand and wait for the experimenter to come.

Screen 4
Imagine that individuals A, B, C have each placed 10 tokens (out of 20) to the group account in the previous period. Look at the table on the right side of the screen and consider five possible amounts that Individual D could place in the group account (presented in rows). Please indicate on the scale from 1 to 7 how socially appropriate you believe

choosing each of these amounts is, given the amounts that others contributed to the group account in the previous period.

Screen 5
Imagine that individuals A, B, C have each placed 20 tokens (out of 20) to the group account in the previous period. Look at the table on the right side of the screen and consider five possible amounts that Individual D could place in the group account (presented in rows). Please indicate on the scale from 1 to 7 how socially appropriate you believe choosing each of these amounts is, given the amounts that others contributed to the group account in the previous period.

Screen 6
Imagine that members A, B, C, and D have made their choices in the first stage of a period. Namely, members A, B, and C placed 10 tokens each to the group account and individual D placed the amount of tokens equal to one of the five options listed on the right part of the screen. For each of the amounts that individual D could have placed to the group account, please indicate how socially appropriate you believe subtracting tokens from individuals A, B, and C is, given the amount that members A, B, C, and D contributed to the group account.

PART III

Description of the Task (Screen 1)
In this final part of the experiment, we ask you to make decisions that are potentially payoff-relevant for participants in role D (Individuals D) who took part in a previous experiment that has already been finished. All participant of this previous experiment have already been paid. Your task will be to decide the amount of money (between 0 Euro and 10 Euro) that additionally should be transferred to an Individual D.

Your task
In this task you will see three possible scenarios (the same as you have just seen) and you will be asked to determine the amount of money between 0 Euro and 10 Euro that should be transferred to an Individual D depending on his/her behavior in the previous experiment.

Specifically, you will see five descriptions of behaviors by Individuals D. For example, "D contributed 10 tokens to the Group account" means that some Individual D contributed 10 tokens to the group account in the previous experiment. For each of possible scenarios, please indicate how much money between 0 Euro and 10 Euro you think they should receive.

Payment of D
If one of your decisions is chosen to be implemented, it will affect a real Individual D. The scenario and situation which is the closest to the average behavior of this Individual D will be implemented. Thus, all your decision might influence the payment of another participants.

Please click OK when you are ready to go on.

Screen 2
Imagine that individuals A, B, C have each placed 10 tokens (out of 20) to the group account in a previous experiment. Look at the table on the right side of the screen and consider five possible amounts that Individual D could place in the group account (presented in rows). For each of the amounts that individual D could have placed to the group account, please indicate how much money D should receive as an additional payment (between 0 and 10€) given the amount that members A, B, C, and D contributed to the group account.

Remember: a participant from a previous experiment who had been assigned the role D might receive the additional payment in line with your decision if your decision is chosen to be payoff relevant.

Screen 3
Imagine that individuals A, B, C have each placed 20 tokens (out of 20) to the group account in a previous experiment. Look at the table on the right side of the screen and consider five possible amounts that Individual D could place in the group account (presented in rows). For each of the amounts that individual D could have placed to the group account, please indicate how much money D should receive as an additional payment (between 0 and 10€) given the amount that members A, B, C, and D contributed to the group account.

Remember: a participant from a previous experiment who had been assigned the role D might receive the additional payment in line with your decision if your decision is chosen to be payoff relevant.

Screen 4
Imagine that individuals A, B, C have each placed 10 tokens (out of 20) to the group account in a previous experiment. Look at the table on the right side of the screen and consider five possible amounts that Individual D could place in the group account (presented in rows). For each of the amounts that individual D could have placed to the group account, please indicate how much money D should receive as an additional payment (between 0 and 10€) given the amount that members A, B, C, and D contributed to the group account and that D reduced the payoff of A, B, or C.

Remember: a participant from a previous experiment who had been assigned the role D might receive the additional payment in line with your decision if your decision is chosen to be payoff relevant.