Engel, Christoph; Zhurakhovska, Lilia

Working Paper

Words substitute fists: Justifying punishment in a public good experiment

Preprints of the Max Planck Institute for Research on Collective Goods, No. 2013/16

Provided in Cooperation with:
Max Planck Institute for Research on Collective Goods

Suggested Citation: Engel, Christoph; Zhurakhovska, Lilia (2013) : Words substitute fists: Justifying punishment in a public good experiment, Preprints of the Max Planck Institute for Research on Collective Goods, No. 2013/16, Max Planck Institute for Research on Collective Goods, Bonn

This Version is available at:
https://hdl.handle.net/10419/84995

**Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2013/16**

Words Substitute Fists – Justifying Punishment in a Public Good Experiment

Christoph Engel
Lilia Zhurakhovska

MAX PLANCK SOCIETY

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2013/16

Words Substitute Fists – Justifying Punishment in a Public Good Experiment

Christoph Engel / Lilia Zhurakhovska

August 2013

Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, D-53113 Bonn
http://www.coll.mpg.de

Words Substitute Fists – Justifying Punishment in a Public Good Experiment[**]

Christoph Engel[*], Lilia Zhurakhovska[#]

Abstract

Punishees regularly ask for justification. But is justification also effective? To answer this question under controlled conditions, we have conducted a public goods experiment with central punishment. The authority is neutral – she does not benefit from contributions to the public good. Punishment is costly. Along with the punishment decisions the authority writes justifications for her decisions. In the *Baseline*, authorities are requested to justify punishment decisions, but the reasons are kept confidential. In the *Private* treatment, the addressee is only informed about the justification of the authority's decision affecting herself, not affecting others. In the *Public* treatment, all reasons are made public. Whenever reasons are communicated, there is less monetary punishment. Authorities partly substitute words for action. Contributions decay in later periods if the justification is only communicated to the addressee. In the remaining two treatments, contributions stabilize at a high level.

JEL: C91, D03, D62, D63, H41, K14

Keywords: justification, authority, central intervention, public good, experiment

[**] Helpful comments by Angela Dorrough and Paul Schempp and by the audience at the Workshop on Behavioral and Experimental Economics, at the European Economics Science Association and at the American Law and Economics Conference on an earlier version are gratefully acknowledged.

[*] Max Planck Institute for Research on Collective Goods, Bonn

[#] Max Planck Institute for Research on Collective Goods, Bonn, University of Cologne, University of Erlangen-Nuremberg

1. Introduction

Parents punish their children. Teachers punish their pupils. Masters punish their servants. Officers punish their soldiers. Guards punish their prisoners. Abbots punish their monks. Judges punish their defendants. In social life, authority frequently means authority to inflict punishment on a subordinate.

Sometimes, punishment is the only act of communication between the authority and the subordinate. The mother just slaps the child that has broken his toy. The teacher just sends the pupil out of the room that has disturbed class. The abbot just excludes the monk from high table that has missed morning prayer. Punishment without reasons is even frequent at the heart of the judicial system. Juries do not explain why they find the defendant guilty (for more examples from the legal system see Schauer 1995:634).

Yet often the subordinate comes back and asks: but why? Frequently authorities anticipate the question, and directly add justifying reasons to the sanction. The mother tells her boy: we have entrusted your toys to you. Be more heedful in the future. For this time, we will buy you a new one. But if you break it again, there will not be a new toy then.

The subordinate is not the only possible addressee of reasons. The authority may herself have a supervisor who asks her to justify the intervention. The headmaster finds the pupil walking idle in the corridor and calls upon the teacher to justify her decision. The prison warden wants the guard to explain why he used corporal punishment. The convict appeals his case. Another addressee of explicit justification is fellow subordinates. The mother punishes her elder boy and tells the younger one: be aware, this is what will happen if you do not look after your toys. Jeremy Bentham has built his entire utilitarian theory of criminal law on this point (Bentham 1830). Finally, explicit reasons may help those who have installed the authority to assess whether she should remain in office, or they may help the general public to form an opinion, and maybe call for political intervention. A case in point is criminal judges standing for re-election.

In the last decade, experimental economics has been very interested in punishment. The main field of application is linear public good games. If the experimenter does not provide any institutional framework, initially many participants make substantial contributions to the public project. Yet over time contributions decay. The trend reverses if participants are given the opportunity to punish each other, despite the fact that, in the typical implementation, punishment is costly (see only Fehr and Gächter 2000; Herrmann, Thöni et al. 2008).

We use this framework to test the effects of a justification requirement. In the interest of coming closer to the real world applications that motivate our research, we randomly select one participant to be an authority for a group of four active players. The participant in the role of the authority receives a fixed income (think of the judge's salary) and therefore does not benefit monetarily from the provision of the public good; in that sense we make the authority impartial. Yet to make her choices credible she has to pay for punishment points out of a small

additional endowment. Each punishment point she does not use increases her income by a small amount. That way we incentivize choices, despite the fact that the authority receives a fixed wage (think of additional effort or hassle, the more so the more severe the sanction).

In all treatments, authorities are requested to justify their choices. Yet in the *Baseline*, the reasons they give go to the experimenter only. In the *Private* treatment, each active player only learns the reasons for the decision affecting herself. We finally implement a *Public* treatment. In this treatment, all active players see the reasons directed to themselves and to all other group members.

We have subtle, but interesting results. In the *Private* and *Public* treatments, there is significantly less punishment than in the *Baseline*. If reasons are communicated, authorities partly substitute words for action. Contributions increase over time in the *Baseline* and in the *Public* treatment, while they do not in the *Private* treatment. Hence if justification is to the entire group, less monetary punishment is equally effective. In that setting, words also substitute action in terms of disciplining active players. Our data suggest however that there is a mismatch between the expectations of authorities and active players if reasons are only communicated to the addressee. While active players become even more sensitive to the severity of punishment, authorities reduce punishment, arguably because they expect reasons to serve as a partial substitute. By contrast if reasons are made public, active players become considerably more sensitive to the amount contributed by the remaining active players. Punishment combined with reasons stabilizes contributions on this indirect path.

The remainder of the paper is organized as follows: Section 2 relates the paper to the literature. Section 3 presents the design of the experiment. Section 4 contains the model and derives predictions. Section 5 reports results. Section 6 concludes.

2. Related Literature

To the best of our knowledge the effect of a justification requirement on punishment and contribution behavior in a public good has not previously been studied, neither theoretically nor experimentally.

In treatments *Private* and *Public*, justification is a form of one-way communication from the authority to the active members. Communication among active players has generally been shown to increase cooperation (see the meta-analysis by Sally 1995; the survey by Crawford 1998; the meta-analysis by Balliet 2010). Our design differs from this literature in that the only player allowed to communicate is the authority. Communication can therefore not serve as a vehicle for creating trust among the active players. It may merely serve the backward looking function of explaining why a player has been harmed, and the forward looking function of explaining an authority's punishment policy.

Duffy and Feltovich (2002) tested a prisoner's dilemma where active players either had a chance to send a pre-play cheap talk message, or where they could observe each other's choices in the previous period. Both had roughly the same, positive effect. We implement a stranger design. Therefore through feedback from earlier periods participants only learn about the population, not about the individual interaction partners in the next period. If communication by an authority is equally effective, we should expect a positive effect.

If all players hold Fehr-Schmidt preferences, the behavioral game has the character of a coordination game with multiple equilibria. It has been shown that, in coordination games, pre-play communication facilitates coordination on the Pareto-dominant equilibrium (Blume and Ortmann 2007). Communication by the exogenous authority might serve a similar function.

If reasons are communicated, the authority may use them to express disapproval. Masclet, Noussair et al. (2003) have shown that disapproval increases contributions, even if it is not backed up by monetary sanctions. They did not study the interaction of monetary and non-monetary sanctions, which is what we implement.

In treatments *Private* and *Public*, the authority may use the reasons she gives to announce a punishment policy. In Berlemann, Dittrich et al. (2009), non-binding announcements had practically no effect. There was a slight effect if, afterwards, it could be checked whether (active) participants behaved as announced. Yet in our experiment, active players cannot check whether the authority kept her word, given active players and authorities are re-matched every period.

Croson and Marks (2001), in a step level public good, introduced a recommendation by the experimenter how much to contribute. This only had a significant effect on contributions if participants benefitted heterogeneously from the provision of the public good. In our design, active players are homogeneous. Yet if the authority uses justifications to fix an expected contribution level, this is not a recommendation by the experimenter, but by another participant. Moreover the authority has power to enforce the norm. We might therefore see a positive effect.[1]

If active players learn the reasons, the authority may use justification to threaten freeriders in future periods. Masclet, Noussair et al. (2010) have found that threats preceding decentralized punishment increase cooperation.

We entrust punishment to a fifth player. In a companion paper, that only uses the data from an additional treatment without justifications, we show that the large majority of authorities is neither selfish nor spiteful. They also do not exploit punishment to equalize earnings with ac-

1 In our experiment we inform participants in the instructions about average contributions in a similar experiment; see instructions in the Appendix. That information could also be regarded as a subtle form of recommendation by the experimenter. However we neither expected ex ante nor found ex post that this information had a remarkable effect on the behavior of our participants. The only purpose of that information was to provide participants with one potential plausible contribution norm.

tive players. Instead they are motivated to manage the groups they happen to be assigned to (Engel and Zhurakhovska 2012). This strengthens a finding from Engel and Irlenbusch (2010), where an additional player had been given authority to discipline the group. Yet this player benefited from the success of the group, so that successfully managing the group was in the best pecuniary interest of the authority.

In a sender receiver game, Xiao and Tan (forthcoming) compare three settings: a punishment authority receives a flat fee; the authority has a straightforward monetary incentive to punish senders who have communicated the truth; this incentive is upheld, but authorities are obliged to justify their decision in a message that is communicated to the remaining two participants at the end of the experiment. With this obligation, authorities are less likely to abuse their power. Senders are less likely to lie. We test a different game. We make it impossible for authorities to be selfish. In our experiment, interest is not in taming corruption, but in improving the effectiveness of punishment. To that end we manipulate to whom reasons are communicated. We also derive hypotheses from a formal model.

The willingness of third parties with no monetary interest to punish others has also been studied in different games. Fehr and Fischbacher (2004) find that third parties are willing to punish dictators who give little, and players who defect in a prisoner's dilemma. They explain their findings with "strong reciprocity": third parties are willing to punish norm violations, even if they are not personally a victim or can expect potential monetary future benefits from punishment (Putnam 2001; Carpenter, Matthews et al. 2004). Questionnaire data suggest that the impulse to punish results from hurt emotions. Charness, Cobo-Reyes et al. (2008) play a trust game where a third player may either sanction the trustee for having sent back little, or may reward the trustor for having sent a lot. They find that many third parties are willing to use either option. They do not specify the third parties' motive. Leibbrandt and López-Pérez (2009) have an active player choose between two different allocations of a fixed pie between herself and a passive second player in 10 different games. A third party learns the choices and is allowed to reduce the payoff of either of the two parties, at a cost to herself. A substantial fraction of third parties use that power. The authors exploit the fact that they have multiple punishment choices per individual to classify the distribution norm each punisher adheres to. Almenberg, Dreber et al. (2010) add a third player to several variants of the dictator game. The third player may either punish or reward players of the dictator game, at a cost to herself. The majority of third parties use one of the options; a substantial minority even uses both. The authors explain this result with indirect reciprocity.

The authority may also be interpreted as being assigned the role of leading their current group. Different specifications of leadership have generally been shown to improve cooperation by the non-leading participants (Clark and Sefton 2001; Güth, Levati et al. 2007; Levati, Sutter et al. 2007; Gächter, Nosenzo et al. 2010; Glöckner, Irlenbusch et al. 2011). Yet recently Rivas and Sutter (2011) have found that leadership only increases contributions if the leader has volunteered, while it has no beneficial effect if the role is imposed – as in our experiment. Nikiforakis, Normann et al. (2010) do not find an increase in contributions if the effec-

tiveness of punishment is asymmetric. Yet, those with higher punishment power punish more. Likewise, in O'Gorman, Henrich et al. (2009), if only a single group member has punishment power, this one member stabilizes cooperation equally well. Rockenbach and Milinski (2006) show that the possibility to build up a reputation can be an effective substitute for direct monetary punishment.

In the legal literature, the obligation to justify decisions has been studied from a normative perspective (McCormac 1994; Schauer 1995). This literature expects explicit reasons to clarify the meaning of authoritative intervention, to authoritatively construct reality, to increase compliance, to enable control, to remove biases in addressees, to dissolve conflict (Engel 2007) and to make authorities more accountable (Tetlock 1983; Seidenfeld 2001).

3. Design

All rules of the experiment are common knowledge and the interaction is completely anonymous. The main experiment has three steps.[2]

Step 1:

We conduct a linear public good experiment with the standard payoff (π_i) function

$\pi_i = e - c_i + \mu \sum_{n=1}^{N} c_n$	(1)

where e is the endowment, c_i is the contribution of this player to the public good, $0 < \mu < 1 < N\mu$ is the marginal per capita rate, n is generic for any player, player i included, and N is group size. In the experiment $e = 20\,Taler,\ \mu = \frac{4}{10},\ N = 4$. Four active players may contribute to the public good in step 1.

Step 2:

We randomly assign a fifth player to each group. This player earns a fixed amount of 1 €. She receives 20 tokens that she may use for punishing any of the active players. Each punishment token assigned destroys three Taler of the active player's period income. Any punishment token the authority does not use is credited with .01 €. Given the exchange rate of 1 Taler = .04 €, we thus implement a fine to fee ratio of 1:12.[3] The fifth player learns about the contributions of all four active players in her group. After assigning the punishment the authority

2 We have two post-experimental tests, for social value orientation (Liebrand and McClintock 1988), and for relative risk aversion (Holt and Laury 2002), which we, however, do not use for the analysis since they do not turn out informative.

3 That ratio makes punishment substantially cheaper than in most other related experiments. Yet in our experiment, unlike in most earlier experiments, the authority does not benefit from contributions at all. Therefore any cost demonstrates intrinsic willingness for punishment and makes it meaningful.

types reasons justifying punishment into a chat box. The box holds a maximum of 500 characters. This is made explicit in the instructions.[4]

Step 3:

Active group members are informed about the contributions of all other active members of their current group, and of punishment tokens assigned to each of them, if any.[5] They also learn their income from this part of the experiment. Furthermore, (depending on treatment) active players learn the reasons formulated by the authority. In the *Baseline* the explanations given by the authority are not communicated to active players. In the *Private* treatment, each active player is only informed about the reasons given by the authority regarding herself, not the remaining active group members. Finally, in the *Public* treatment, all group members learn all reasons given for the decisions of the authority regarding any group member.

After the end of the first period, there is a surprise restart of another 10 periods of the same game. Participants learn that they will be re-matched every period, but that roles are kept constant throughout the experiment. Our main reason for implementing a stranger design is external validity. In the legal application that has triggered our research judges are unlikely to meet the same defendant again. So a matching group of 10 players comprises of two groups with one authority and four active players each. Every period, groups are randomly recomposed. Following the procedure that is standard in the experimental literature (see e.g. Charness 2000; Montero, Sefton et al. 2008), we only tell participants that they will be re-matched every period, not that matching groups have limited size. This procedure is meant to guarantee independent observations, without inducing participants to second guess group composition.

The experiment was conducted in the Cologne Laboratory for Economic Research in 2012. The experiment is programmed in zTree (Fischbacher 2007). Participants were invited using the software ORSEE (Greiner 2004). 340 student participants of various majors had mean age 24.31. 51.54 % were female. Participants on average earned 15.81 € (20.86 $ at the time of the experiment), 15.50 € for active players, and 17.04 € for authorities. We had 12 independent observations (matching groups of 10) in the *Baseline*, and 11 each in the two treatments.[6]

4. Hypotheses

Obviously, the punishment choices of the authorities and the contribution choices of the active players are related. To capture this, we present a model in which we derive reaction functions of the authorities to a certain level of contributions and vice versa. We begin with standard behavioral assumptions, which we relax step by step, to derive in which ways we expect

4 The only restriction was that authorities were not allowed to communicate any personal information, so as to preserve anonymity. See instructions in the Appendix for the exact wording.

5 In fact, they are informed about the amount of Talers subtracted from their income.

6 In *Private* and *Public*, we could not fill one matching group since invited participants did not show up.

our treatment manipulations to matter. The purpose of this section is to derive hypotheses about the effect of a differently specified justification requirement from a formal model. We want to be able to precisely define the three channels on which we expect the justification requirement to affect behavior. Yet our research question is understanding the behavioral effects of a justification requirement, not testing a general behavioral model.

For active players holding standard preferences, the introduction of the punishment option adds an additional term to the payoff function and changes (1) to

$\pi_i = e - c_i + \mu \sum_{n=1}^{N} c_n - \sigma(\hat{c} - c_i)$	(2)

where $\hat{c}$ is the contribution level the authority wants to implement, and σ is the severity of punishment. We assume punishment to be proportional to the deviation from the chosen norm. The authority has

$\pi_a = w - k \sum_{n=1}^{N} \sigma \max(\hat{c} - c_n, 0)$	(3)

where w is the authority's fixed wage, k is the cost per punishment token, and severity σ is the authority's decision variable. The authority's payoff strictly decreases in punishment, which is why it is her dominant strategy to choose $\sigma = 0$. In anticipation, active players' reaction function is determined by the first derivative $-1 + \mu < 0$. We have a corner solution. In the unique equilibrium of the stage game, active players keep their endowments. Through unraveling, this is also the unique equilibrium of the repeated game. Since authorities do not punish, it does not matter whether justifications for the punishment decisions are communicated to the active players. We therefore do not expect treatment differences.

For three behavioral reasons, we might have a more optimistic prediction: (1) active players might be conditional cooperators; (2) authorities might want to manage groups; (3) active players might be guilt averse, and authorities might exploit the justification statement to accentuate guilt.[7] Table 1 collects authorities' reaction functions, and first derivatives for active players, for all combinations of these behavioral effects.

The characteristic contribution patterns in public good games are usually explained with the fact that the majority, but not all, participants can be classified as conditional cooperators (Fischbacher, Gächter et al. 2001; Fischbacher and Gächter 2010). One prominent explanation for conditional cooperation is social preferences (for summaries see Fehr and Schmidt 2006; Cooper and Kagel 2013). In our context, outcome based social preferences, like inequity-

7 We explain below in which ways our notion of guilt aversion differs from Battigalli and Dufwenberg (2007).

aversion (Fehr and Schmidt 1999; Bolton and Ockenfels 2000), and intention based social preferences, like reciprocity (Charness and Rabin 2002; Falk, Fehr et al. 2008), point into the same direction. Therefore, we only present expectations based on the most prominent model, the one by Fehr and Schmidt (1999). We start initially with the assumption that all active players hold identical preferences and that this is common knowledge. Then active players' utility is given by

$u_i = \pi_i - \frac{1}{N-1}\alpha \sum_{j=1}^{N-1} \max(\pi_j - \pi_i, 0) - \frac{1}{N-1}\beta \sum_{j=1}^{N-1} \max(\pi_i - \pi_j, 0)$	(4)

Given the authority is selfish, active players do not expect any punishment. There is room for a cooperative equilibrium where all contribute a target amount $\hat{c}$ if $\beta > 1 - \mu$.

Next assume (mild) preference uncertainty. While three active players hold Fehr/Schmidt preferences with identical, known parameters, one player with probability q also holds such preferences, while this player is selfish with counter-probability $1 - q$. This constitutes a Bayesian game. The uncertainty has two drawbacks: if a player defects who is inequity-averse herself, she only expects to feel the full effect of inequity-aversion with probability q. With counter-probability $1 - q$ not only the beneficial effect of aversion against advantageous inequity is reduced to 2/3. The player also feels the counterproductive effect of aversion against exploitation. The more pronounced the uncertainty, the more difficult it becomes to sustain the cooperative equilibrium. Of course, if this player deems it possible that more than one of the remaining players holds standard preferences, demands on her own aversion against exploiting others, i.e. on β, become even stronger.

In the companion paper we analyze the behavior of authorities and show that the large majority of our authorities aims at managing the groups to which they are randomly assigned, despite the fact that this is costly (Engel and Zhurakhovska 2012).[8] This resonates with earlier findings on third party punishment (Carpenter, Matthews et al. 2004; Fehr and Fischbacher 2004; Charness, Cobo-Reyes et al. 2008; Leibbrandt and López-Pérez 2009; Almenberg, Dreber et al. 2010). First, assume that all authorities have this desire and that this is common knowledge, which we capture by

$u_a = \pi_a + \sum_{n=1}^{N} (4\mu - 1)(\hat{c} - c_n)$	(5)

8 Note that the subjects face a stranger matching. Therefore, the term "managing the group" should not be over-interpreted. In our context it means that an authority may strive at making the active players, who are punished by her, cooperate on a certain level in future periods even though group composition changes and neither the authority nor the active players are informed that they might meet again at some point. "Managing" should therefore be understood as "influencing players in a sustainable way".

Forcing any active player tempted to defect by investing σ gives the authority the social benefit from increasing all four active players' income by μ, minus the social cost of 1. The authority punishes provided $4\mu - 1 > k$, i.e. provided benefit outweighs cost. Note that the authority's problem does not depend on the target level $\hat{c}$. The authority is free to enforce any contribution level she deems fit. If the authority enforces $\hat{c}$, active players' first order condition changes to $-1 + \mu + \sigma$. This is positive provided $\sigma > 1 - \mu$. Hence in equilibrium, the authority imposes this sanction. All active players contribute $\hat{c}$.

Next reintroduce inequity-aversion, and assume all active players hold Fehr/Schmidt preferences. Since the authority wants to manage groups, active players' first derivative changes to $-1 + \mu + \beta + \sigma$. In anticipation, the authority reduces severity to $\sigma > 1 - \mu - \beta$. The more pronounced inequity-aversion, the smaller the need for costly punishment.[9]

Now, reintroduce uncertainty about active players' preferences. Active players' first order condition and the corresponding severity level react to the degree of uncertainty $1 - q$, as in Table 1. The more pronounced the uncertainty, the more the authority must compensate it with higher severity.

In the next step, we introduce heterogeneity of authorities. With common prior p, the authority wants to manage her group, i.e. holds preferences as in (5). With counter-probability $1 - p$, the authority maximizes payoff, as in (3). This too constitutes a Bayesian game. Active players' profit function changes to

$\pi_i = e - c_i + \mu \sum_{n=1}^{N} c_n - p\sigma(\hat{c} - c_i)$	(6)

Those authorities, who want to manage their groups, react to this uncertainty by raising severity to $\sigma > \frac{1-\mu}{p}$. Knowing this, active players still all contribute $\hat{c}$.

At this point, we for the first time see treatment effects. The critical issue is *perceived* uncertainty. In the *Baseline*, active players might reason: "if authorities have to explain their choices, they are induced to develop a deliberate punishment policy, even if I do not learn this policy". In *Private* and *Public*, active players additionally receive an individual signal about the punishment policies prevalent in this population of authorities. This helps them update their beliefs about the certainty and the severity of punishment. Therefore we should have $p_{base} < p_{priv} \approx p_{pub}$. The larger p, the less authorities (who want to manage their groups) must react to low contributions by increasing severity. This is the first (direct) channel on which we expect the justification requirement to matter. Following the same logic as before, we can add

9 We refrain from explicitly modeling interactions between motivating forces. Technically, they are easy to introduce by a multiplicative term in (4). If the correlation coefficient is positive, both motivating forces are substitutes. If it is negative, they are complements. In the interest of keeping the model simple, we assume the correlation coefficient to be 0, i.e. motivating forces to be additively separable.

active players' inequity-aversion and uncertainty about the type of one active player (see Table 1).

As is well established, experimental participants are sensitive to framing manipulations (Kahneman and Tversky 2000). One powerful type of frame labels the opportunity structure such that it triggers normative expectations (Elliott, Hayward et al. 1998), like the labeling of a public good as a "community" rather than a "Wall Street" game (Ross and Ward 1996; Liberman, Samuels et al. 2004). One explanation for the effect is guilt aversion. If the authority may make its disapproval of a participant's choice explicit, she may stress her normative expectations, and thereby increase the effect of guilt. In the framework of public goods, guilt aversion has been formalized as disutility from falling below a normative expectation (Dufwenberg, Gächter et al. 2011), which translates into an additional term in (2):

$u_i = e - c_i + \mu \sum_{n=1}^{N} c_n - \sigma(\hat{c} - c_i) - \gamma(\hat{c} - c_i)$	(7)

In the *Baseline*, the authority has no chance to label freeriding, while she does in the remaining two treatments. Yet in the *Private* treatment, an additional element of shaming is missing. Shaming should even increase guilt. We therefore expect $\gamma_{base} = 0 < \gamma_{priv} < \gamma_{pub}$. This constitutes the second (direct) channel on which we expect the justification requirement to matter. Since monetary punishment is costly, if available, the authority replaces it by guilt. Severity is reduced (see Table 1).

Finally, we expect the justification requirement to matter on an indirect channel: The more this player believes the authority to credibly deter freeriding (the larger p) and the more she believes the authority to trigger guilt (the larger γ), the more she will also expect other players to contribute (the larger q). Since we have explained how our treatments affect the former two parameters, these direct effects should translate into an additional indirect effect, such that $q_{base} < q_{priv} < q_{pub}$. If correctly anticipated by authorities, the indirect effect reduces σ accordingly.

In line with the prevalent approach in the literature (see e.g. Fehr and Schmidt 1999), we assume that not only payoff functions, but also utility functions are linear in the decision variables. We therefore predict corner solutions. If all actors hold standard preferences; if authorities hold preferences as in (5), but if their desire to manage groups is not pronounced enough to outweigh the cost; if the uncertainty about authorities' types is too pronounced; if active players are not sufficiently averse to advantageous inequity; if the uncertainty about active players' type is too pronounced, we expect

H_1: Authorities do not punish. Active players keep their endowments.

If at least one of the aforementioned conditions is not fulfilled, we expect

H_2: Active players contribute the amount that the authority tries to implement.

We only expect treatment differences in punishment, not in contributions; in equilibrium, authorities that want to manage their groups compensate by higher severity for any deficiency in active players' social preferences, for type uncertainty and for the impossibility to make guilt salient. For the reasons explained we expect

H_3: Punishment is most severe in the *Baseline*, less so in *Private* and least severe in *Public*.

Table 1 summarizes the first derivatives of active players' utility functions (whenever the term is positive, they contribute the amount the authority tries to implement) and predicted punishment choices of authorities.

authority	active player	first derivative active player	σ
standard	standard	$-1+\mu$	0
	certain FS	$-1+\mu+\beta$	0
	uncertain FS	$-1+\mu+q\beta+(1-q)\left(\frac{2}{3}\beta-\frac{1}{3}\alpha\right)$	0
certain manage	standard	$-1+\mu+\sigma$	$>1-\mu$
	certain FS	$-1+\mu+\beta+\sigma$	$>1-\mu-\beta$
	uncertain FS	$-1+\mu+q\beta+(1-q)\left(\frac{2}{3}\beta-\frac{1}{3}\alpha\right)+\sigma$	$>1-\mu-q\beta-(1-q)\left(\frac{2}{3}\beta-\frac{1}{3}\alpha\right)$
uncertain manage	standard	$-1+\mu+p\sigma$	$>\frac{1-\mu}{p}$
	certain FS	$-1+\mu+\beta+p\sigma$	$>\frac{1-\mu-\beta}{p}$
	uncertain FS	$-1+\mu+q\beta+(1-q)\left(\frac{2}{3}\beta-\frac{1}{3}\alpha\right)+p\sigma$	$>\frac{1-\mu-q\beta-1-\mu-q\beta-(1-q)\left(\frac{2}{3}\beta-\frac{1}{3}\alpha\right)}{p}$
uncertain manage + guilt	standard	$-1+\mu+p\sigma+\gamma$	$>\frac{1-\mu-\gamma}{p}$
	certain FS	$-1+\mu+\beta+p\sigma+\gamma$	$>\frac{1-\mu-\beta-\gamma}{p}$
	uncertain FS	$-1+\mu+q\beta+(1-q)\left(\frac{2}{3}\beta-\frac{1}{3}\alpha\right)+p\sigma+\gamma$	$>\frac{1-\mu-q\beta-1-\mu-q\beta-(1-q)\left(\frac{2}{3}\beta-\frac{1}{3}\alpha\right)-\gamma}{p}$

Table 1
Model Predictions
assuming all determinants of choices to be additively separable
otherwise add interaction terms
first and second column: assumptions about individuals' preferences
legend: standard: standard preferences; manage: authority wants to manage group; guilt: authority may make guilt salient; FS: active player holds Fehr/Schmidt preferences; certain FS: it is common knowledge that all active players hold FS preferences; uncertain FS: the probability that active players hold FS preferences, with known parameters, is common knowledge
shaded area: choices are sensitive to treatment manipulations

5. Results

The first, one-shot phase of the experiment was meant to test whether active players anticipate the effects of a justification requirement. This is not the case. In non-parametric Mann-Whitney tests, we do not find any significant effects.[10] Parametrically we find a weakly sig-

10 In the first period, individual choices of active players are still independent. But one authority simultaneously decides about punishing four members of her first group. In this dimension, punishment decisions

nificant difference between the *Baseline* and *Private* in terms of punishment (p = .069).[11] Punishment is less severe in the *Private* treatment. This translates into significantly higher profit. Since anticipation at most has a very small effect, in the following we pool the data from the first and the second phases of the experiment.

Figure 1 reports treatment effects. As expected, we reject $\mathbf{H_1}$. In all treatments, authorities do use the punishment option.[12] Active players make substantial contributions to the public project.[13]

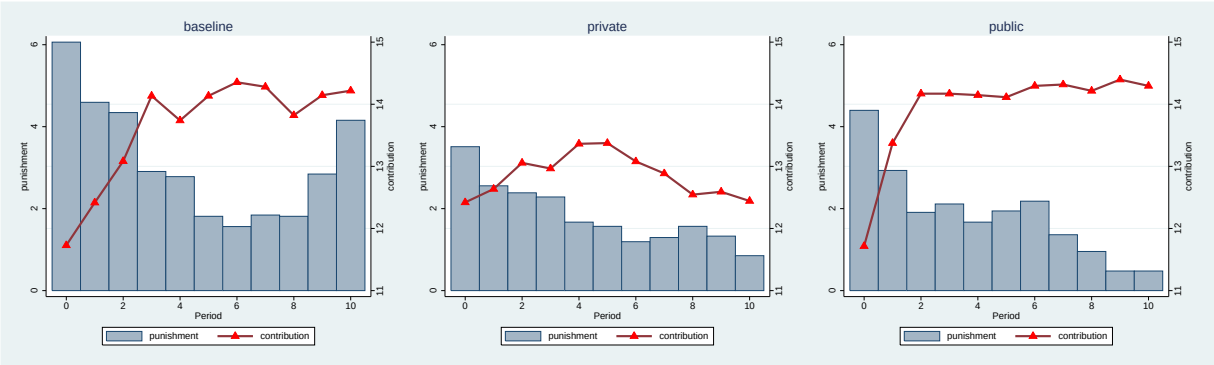


Figure 1
Treatment Effects
right vertical axis: mean contributions (in Taler) of the active players per period and treatment
left vertical axis: mean punishment (in Taler) of the authorities per period and treatment
horizontal axis: one-shot game reported as period 0, repeated game reported as periods 1 – 10
one panel per treatment

The design of the experiment empowers authorities to perfectly deter freeriding. In each period each authority disposes of a maximum of 20 tokens for punishment. Each punishment token destroys 3 Taler of the punishee's period profit. A complete freerider is deterred if she is

---

are thus not independent. In these tests, we therefore work with mean punishment per authority, in the first period, as the dependent variable. Results are the same if instead we work with total punishment tokens meted out by any one authority.

11 In this paper we do not analyze in detail the determinants of authorities' punishment policies, and the contents of the reasons they give. Readers interested in these results are referred to our companion paper (Engel and Zhurakhovska 2012). The most prominent explanation for punishment is the fact that an active player has contributed less than the mean of his group. Some authorities also try to impose an idiosyncratic standard, typically 10 Taler. Yet, others stress that the punishee has acted unfairly.

12 Statistical tests are complicated by the fact that the design of the experiment excludes negative punishment, i.e. rewards. Therefore technically **H1** calls for a test at the limit of the support. We react by reporting the highest positive amount of punishment at which a signed-rank test still rejects at conventional levels. All tests are over means at the highest level of dependence, i.e. matching groups. The test still rejects the hypothesis that mean punishment is 1.5 Taler per active player in the *Baseline* (N = 12, p = .031), 1 Taler per active player in treatment *Private* (N = 11, p = .004), and.5 Taler per active player in treatment *Public* (N = 11, p = .010).

13 Using the same procedure as in the previous footnote we find that signed-rank tests still reject the hypothesis that mean contributions are 8 Taler in the *Baseline* (N = 12, p = .023), 7 Taler in treatment *Private* (N = 11, p = .021), and 8 Taler in treatment *Public* (N = 11, p = .016). In all treatments, mean contributions are significantly above those limits.

punished by more than 4*3 = 12 Taler. Yet effectively in all treatments punishment is frequently non-deterrent, even if there is punishment at all.[14]

We now turn to hypothesis **H_3** that expects our treatments to matter for punishment. Descriptively there is indeed less punishment in treatments *Private* and *Public*, i.e. when active players learn reasons (see Figure 1). Moreover, in these treatments punishment decays over time, whereas it goes up again in the *Baseline*.

Non-parametrically, we find a weakly significant difference between the *Baseline* and *Public* (Mann Whitney over means per matching group, N = 23, p = .0524). This difference is significant at conventional levels for the last four periods (N = 23, p = .0354), as well as any smaller number of the final periods. For the last two periods, we also find a significant difference between the *Baseline* and treatment *Private* (N = 23, p = .0303).

Parametrically, we see strong effects[15] (see models 1 and 2 of Table 3). In these models, we not only control for the fact that, in all treatments, punishment is most pronounced in the beginning; this is captured by the time trend and its interaction with treatment. We also take into account that in the *Baseline,* punishment is U-shaped; it goes up again in the end. This is not the case in *Private* and *Public*. The different patterns of the time trends we capture by the square of the time trend, and interactions with treatments. In the mixed effects model we see a strong negative main effect of both treatments where reasons are revealed to punishees. In these treatments, punishment decays less rapidly over time (the positive interaction effects neutralize most of the negative main effect of period), and it hardly goes up in the end (the negative interaction effects neutralize most of the positive main effect of period squared). When there is room for it (since reasons are communicated), authorities pecuniary partly substitute words for pecuniary punishment, as expected.

This interpretation is further supported by models 3 and 4 in Table 3. In these models, we control for the respective active player's contributions, i.e. we estimate authorities' empirical reaction functions. The substitution effect is directly visible in the positive interaction between contribution and treatment: in both treatments, the level of punishment is less sensitive to differences in contributions.

---

14 We of course face the same technical challenge as in the previous two footnotes, and tackle it the same way. The maximum mean contribution that is still rejected at conventional levels is 13 Taler in the *Baseline* (N = 12, p = .028), 14 Taler in treatment *Private* (N = 11, p = .006), and 14 Taler in treatment *Public* (N = 11, p = .013).

15 For parametric estimation, we have challenging data. Every period each authority has power to punish four active group members. The authority stays the same over time, and she remains assigned to the same matching group (with different active players per group in each period, though). Punishment data is therefore from periods nested in authorities and these in turn are nested in matching groups. This data generating process is captured by a mixed effects model. Yet most active players most of the time do not get punished at all. Therefore the data is also left censored. This can be captured by a random effects Tobit model where the authority is the cross-section, and punishment choices directed to individual members of the current group of active participants constitute the "time" dimension. Since there is no generally acknowledged mixed effects Tobit estimator, in Table 2 we report both specifications. Note that results look similar if, instead, we estimate models with matching group fixed effects; of course then the treatment main effects are not identified, but interactions with treatment are.

	model 1	model 2	model 3	model 4
dv: size of deduction through punishment	mixed effects	random effects Tobit	mixed effects	random effects Tobit
Private	-4.106*** (1.128)	-7.130 (4.814)	-6.511*** (1.699)	-10.338** (3.022)
Public	-3.719** (1.128)	-9.052^{+} (4.891)	-4.457* (1.739)	-5.852^{+} (3.214)
period	-1.733*** (.247)	-5.297*** (.872)		
*Private**period	1.216** (.357)	2.935* (1.276)		
*Public**period	1.210** (.357)	3.949** (1.347)		
period^2	.124*** (.020)	.385*** (.071)		
*Private**period^2	-.098** (.029)	-.271* (.106)		
*Public**period^2	-.105*** (.029)	-.433*** (.115)		
contribution			-.952*** (.032)	-2.067*** (.123)
*Private**contribution			.345*** (.046)	.406* (.182)
*Public**contribution			.246*** (.052)	.094 (.193)
cons	7.868*** (.780)	.914 (3.336)	16.138*** (1.181)	18.293*** (2.072)
N	2992	2992	2992	2992
left censored		2300		2300
Wald chi2	105.06	118.10	1525.39	470.63
p model	<.0001	<.0001	<.0001	<.0001

Table 2
Treatment Effects on Punishment
mixed effects model: punishment points received nested in period nested in authority nested in matching group
random effects Tobit: lower censoring at 0
standard errors in parenthesis
reference category: *Baseline*
*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

This leads to

> *Result 1: If the reasons for punishing are communicated to punishees, punishers partly substitute them for monetary sanctions.*

Descriptively, from Figure 1 we see that there is not a pronounced difference across treatments regarding the level of contributions, as predicted by our theory. Yet contributions are not stable in the *Private* treatment, while the time trend remains positive in the *Baseline* and in the *Public* treatment. The visual impression is corroborated by statistical analysis.[16] If we compare mean contributions of the active players per matching group, in non-parametric tests we do not find any significant treatment differences. By contrast, in a parametric test of all treatments, we have a significant interaction between treatment *Public* and the time trend. In a

16 Again alternative models with matching group fixed effects look very similar.

Wald test, the difference between the interaction of treatment *Private* with period, and the interaction of treatment *Public* with period, is also significant (p = .0006).

dv: contribution	mixed effects	random effects Tobit
Private	.547 (1.707)	.897 (1.084)
Public	.551 (1.707)	.645 (1.082)
period	.204*** (.035)	.235*** (.048)
*Private**period	-.223*** (.051)	-.271*** (.070)
*Public**period	-.045 (.051)	.036 (.071)
cons	12.415*** (1.181)	12.970*** (.747)
N	2992	2992
left censored		206
right censored		677
Wald chi2	53.31	54.24
p model	<.0001	<.0001

Table 3
Treatment Effect on Contributions
mixed effects: choices nested in individuals nested in matching groups
random effects Tobit: lower censoring at 0, upper censoring at 20
standard errors in parenthesis
reference category: *Baseline*
*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

This gives us

> *Result 2: If authorities are obliged to justify punishment decisions, in a linear public good contributions stabilize over time if these reasons are kept confidential or if they are made public; there is no stabilizing effect if reasons are only communicated to the punishee in private.*

Our theory was not confined to the effects of a justification requirement on punishment and contributions. From theory we also derived expectations about the underlying forces. We test these expectations in the regressions of Table 4.[17] Two forces independently and significantly explain choices: experienced severity of punishment[18], and experienced cooperativeness of the remaining active players[19].

To test for effects of guilt, i.e. for a possible first direct effect of justifying punishment decisions, treatments are recoded the following way: "communication" is a dummy that is 1 whenever the authority had to communicate her reasons to the punishee, i.e. in treatments *Pri-*

17 Again results look similar if we add matching group fixed effects; of course the main effects of the first three regressors are not identified in these models.

18 Our measure for severity is generated the following way: in auxiliary regressions, for each individual and period we regress received punishment on contributions, for all periods until the previous. The coefficient of this regressor is our measure for severity. For the ease of interpretation, we multiply the resulting coefficient in the auxiliary regressions by -1, so that a higher coefficient of regressor „experienced severity" in the final regression implies that participants are more sensitive to the severity of punishment.

19 We operationalize experienced cooperativeness by the average contribution of the remaining group members, in the previous period.

vate and *Public*. "Transparency" is a dummy that is 1 if all reasons given by the current authority are made publicly available, i.e. in treatment *Public*. Per se, exposing punishees to higher levels of guilt is not instrumental. In the mixed effects model all effects are insignificant. In the Tobit model, the effect of transparency even turns out significantly negative. Guilt is not driving the results.

The coefficient for experienced severity informs us about the second direct effect of a justification requirement.[20] As Table 4 shows, monetary punishment is most effective if reasons are communicated, but not made public (treatment *Private*). This result is supported by the regressions. The interaction between "communication" and "experienced severity" shows that severity is significantly more effective in treatment *Private*, compared with the *Baseline*.

dv: contribution	mixed effects	random effects Tobit
communication	-.029 (1.211)	-.958 (1.135)
transparency	-1.124 (1.261)	-3.029* (1.229)
experienced severity	.254*** (.064)	.319*** (.083)
communication*exp_sev	.354* (.163)	.584** (.655)
transparency*exp_sev	-.330^{+} (.179)	-.286 (.256)
experienced cooperativeness	.467*** (.034)	.603*** (.044)
communication*exp_coop	-.051 (.050)	.028 (.064)
transparency*exp_coop	.150** (.053)	.318*** (.071)
period	-.012 (.020)	-.005 (.028)
cons	7.319*** (.836)	5.954*** (.785)
N	2720	2720
Wald chi2	576.05	674.76
p model	<.0001	<.0001

Table 4
Driving Forces
mixed effects: choices nested in individuals nested in matching groups
random effects Tobit: lower censoring at 0, upper censoring at 20
communication: treatment not *Baseline*, transparency: treatment = *Public*
experienced cooperativeness (exp_coop): mean contribution of other group members in previous period
experienced severity (exp_sev): coefficient of local regression of received punishment on contribution, for this participant, from period 1 until previous period
standard errors in parenthesis
reference category: *Baseline* (no communication, no transparency)
*** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1

20 A numeric example may help interpret the result. Assume that a participant had contributed nothing in the first period, and received 4 punishment tokens. She had contributed 5 Taler in the second period, and had received 3 punishment tokens. In the third period she had contributed 10 Taler and had received 2 punishment tokens. The local regression equation then becomes 4 - .2*contribution. Let's assume this participant contributes 11 Taler in period 4, and there are no other explanatory factors. The regression of contribution would then have to find very strong sensitivity to past severity of punishment. Period 4 choices of this one participant would be perfectly predicted if the coefficient for past severity was -5, and if the regression read 10 – 5 * (-.2 [severity coefficient from the local regression]) = 11.

From the strong and highly significant coefficient of experienced cooperativeness we learn that this is an important driver of cooperation even if justifications for punishment choices are not communicated. Per se, communicating reasons to the addressee does not make participants more sensitive to experienced cooperativeness (the interactions with "communication" are insignificant. This is different if reasons are made publicly known; there is a significant and strong positive interaction between "transparency" and experienced cooperativeness.

This leads to

> *Results 3: If, in a linear public good, authorities are obliged to justify punishment decisions, this affects contributions on a direct and on an indirect channel. On the direct channel, participants become more sensitive to the severity of punishment, whenever reasons are communicated to them. On the indirect channel, participants become more sensitive to experienced cooperativeness of the remaining group members, if reasons are made publicly known.*

We may now also explain why communicating justifications individually is less successful than communicating them publicly. In both treatments, authorities punish less, presumably because they see reproach as a partial substitute for monetary punishment. If reasons are made public, this strategy works, while it does not if reasons remain private. In that case active players expect others even more to be disciplined financially. This gives us

> *Result 4: The reasons given for punishment work as a partial substitute for monetary sanctions only if they are made public.*

6. Conclusion

In social interaction, punishers are usually expected to justify their interventions. By contrast, the standard protocol exposes experimental punishees to sanctions without reasons. In this paper, we test in which ways punishment choices and contributions change if authorities are obliged to formulate explicit reasons for punishing active players in a linear public good. In our *Baseline*, authorities are requested to justify punishment decisions, but the reasons are kept confidential. In the first treatment, the addressee is informed about the justification of the authority's decision affecting her, but each active player only learns the reasons regarding herself. In the second treatment, all reasons are made public. Whenever reasons are communicated, there is less monetary punishment. Authorities partly substitute words for actions. However contributions decay in later periods if the justification is only communicated to the addressee. In our *Public* treatment contributions are stabilized at a high level by a combination of low monetary punishment and justification, while in the *Baseline* without communication a high level of punishment is needed to achieve the same stable level of contributions.

In all treatments experienced cooperativeness and experienced severity significantly explain contribution choices. However these experiences have a differently strong effect, depending

on how the justification requirement is specified. Seeing the remaining participants make substantial contributions is the most important factor. This factor carries most weight if reasons are made public. If reasons are only communicated to the addressee, punishment authorities punish significantly less, but active players are even more sensitive to the severity of punishment. This suggests that there is a mismatch between the expectations of authorities (assuming reproach to be a partial substitute for monetary harm) and the expectations of active players (waiting for freeriders to be severely punished). This mismatch dissolves if reasons are made publicly available.

One should be cautious when extrapolating from the lab to the field. Lab experiments are tools for identifying effects and explaining them. In the interest of achieving this, they deliberately abstract from a host of contextual factors that are very likely to matter in the field. All of our motivating examples have features that are likely to affect the effectiveness of justification and were not present in our experiment. Specifically, in the experiment interaction was anonymous, whereas in all examples the authority and the potential recipient of punishment are identified. Moreover, in the experiment authorities and active group members were rematched every period, whereas in many examples, the relationship is stable over time. Notably, this is, however, different in the legal example. In the experiment, the role of an authority was randomly assigned, whereas in all examples authorities hold a position that has been given to them by some higher authority (which is nature in the case of parents). By using a stranger matching, we not only come closer to the characteristic situation in courts. We also put our theory to a harder test. As is well known, cooperation is easier to achieve in experiments with partner matching and punishment is more effective. In our design, we deliberately exclude any reputation and reciprocity-effects and thereby isolate the effect of communicating reasons. In all examples, authorities have superior competence. In the experiment, if communication is permitted it is strictly unilateral. In all examples, the potential recipient of punishment may at least explicitly ask for a justification. In many examples she even has some right to be heard.

It will be interesting, in future work, to test some of these moderating factors. Nonetheless, even based on this first experimental investigation of a justification requirement in a public good game, tentative normative conclusions can be drawn. It seems that giving reasons is not necessarily a good idea. If these reasons are not made public, the authority may overly focus on educating the addressee, whereas bystanders become skeptical that others who are tempted to misbehave are effectively disciplined. By contrast, if the authority is transparent about the reasons, words may indeed partly substitute acts, to everybody's benefit.

References

ALMENBERG, JOHAN, ANNA DREBER, COREN L. APICELLA and DAVID RAND (2010). Third Party Reward and Punishment. Group Size, Efficiency and Public Goods http://papers.ssrn.com/sol3/papers.cfm?abstract_id=1715305.

BALLIET, DANIEL (2010). "Communication and Cooperation in Social Dilemmas: A Meta-analytic Review." Journal of Conflict Resolution **54**(1): 39-57.

BATTIGALLI, PIERPAOLO and MARTIN DUFWENBERG (2007). "Guilt in Games." American Economic Review **97**(2): 170-176.

BENTHAM, JEREMY (1830). The Rationale of Punishment. London,, R. Heward.

BERLEMANN, MICHAEL, MARCUS DITTRICH and GUNTHER MARKWARDT (2009). "The Value of Non-binding Announcements in Public Goods Experiments. Some Theory and Experimental Evidence." Journal of Socio-Economics **38**(3): 421-428.

BLUME, ANDREAS and ANDREAS ORTMANN (2007). "The Effects of Costless Pre-Play Communication. Experimental Evidence from Games with Pareto-Ranked Equilibria." Journal of Economic Theory **132**: 274-290.

BOLTON, GARY E. and AXEL OCKENFELS (2000). "ERC: A Theory of Equity, Reciprocity and Competition." American Economic Review **90**: 166-193.

CARPENTER, JEFFREY P., PETER HANNS MATTHEWS and OKOMBOLI ONG'ONG'A (2004). "Why Punish? Social Reciprocity and the Enforcement of Prosocial Norms." Journal of Evolutionary Economics **14**(4): 407-429.

CHARNESS, GARY (2000). "Self-Serving Cheap Talk. A Test of Aumann's Conjecture." Games and Economic Behavior **33**: 177-194.

CHARNESS, GARY, RAMÓN COBO-REYES and NATALIA JIMÉNEZ (2008). "An Investment Game with Third-party Intervention." Journal of Economic Behavior & Organization **68**(1): 18-28.

CHARNESS, GARY and MATTHEW RABIN (2002). "Understanding Social Preferences with Simple Tests." Quarterly Journal of Economics **117**: 817-869.

CLARK, KENNETH and MARTIN SEFTON (2001). "The Sequential Prisoner's Dilemma. Evidence on Reciprocation." Economic Journal **111**(468): 51-68.

COOPER, DAVID J. and JOHN H. KAGEL (2013). Other-Regarding Preferences. The Handbook of Experimental Economics II. John H. Kagel und Alvin E. Roth**:** ***.

CRAWFORD, VINCENT (1998). "A Survey of Experiments on Communication via Cheap Talk." Journal of Economic Theory **78**(2): 286-298.

CROSON, RACHEL T.A. and MELANIE MARKS (2001). "The Effect of Recommended Contributions in the Voluntary Provision of Public Goods." Economic Inquiry **39**: 238-249.

DUFFY, JOHN and NICK FELTOVICH (2002). "Do Actions Speak Louder Than Words? An Experimental Comparison of Observation and Cheap Talk." Games and Economic Behavior **39**(1): 1-27.

DUFWENBERG, MARTIN, SIMON GÄCHTER and HEIKE HENNIG-SCHMIDT (2011). "The Framing of Games and the Psychology of Play." Games and Economic Behavior **73**(2): 459-478.

ELLIOTT, CATHERINE S, DONALD M HAYWARD and SEBASTIAN CANON (1998). "Institutional Framing. Some Experimental Evidence." Journal of Economic Behavior and Organization **35**(4): 455-464.

ENGEL, CHRISTOPH (2007). The Psychological Case for Obliging Judges to Write Reasons. The Impact of Court Procedure on the Psychology of Judicial Decision Making. Christoph Engel und Fritz Strack. Baden-Baden, Nomos**:** 71-109.

ENGEL, CHRISTOPH and BERND IRLENBUSCH (2010). Turning the Lab into Jeremy Bentham's Panopticon. The Effect of Punishment on Offenders and Non-Offenders http://ssrn.com/abstract=1555589.

ENGEL, CHRISTOPH and LILIA ZHURAKHOVSKA (2012). You are in Charge. Experimentally Testing the Motivating Power of Holding a (Judicial) Office

FALK, ARMIN, ERNST FEHR and URS FISCHBACHER (2008). "Testing Theories of Fairness - Intentions Matter." Games and Economic Behavior **62**: 287-303.

FEHR, ERNST and URS FISCHBACHER (2004). "Third-Party Punishment and Social Norms." Evolution and Human Behavior **25**: 63-87.

FEHR, ERNST and SIMON GÄCHTER (2000). "Cooperation and Punishment in Public Goods Experiments." American Economic Review **90**: 980-994.

FEHR, ERNST and KLAUS M. SCHMIDT (1999). "A Theory of Fairness, Competition, and Cooperation." Quarterly Journal of Economics **114**: 817-868.

FEHR, ERNST and KLAUS M. SCHMIDT (2006). "The Economics of Fairness, Reciprocity and Altruism. Experimental Evidence and New Theories." Handbook on the Economics of Giving, Reciprocity and Altruism **1**: 615-691.

FISCHBACHER, URS (2007). "z-Tree. Zurich Toolbox for Ready-made Economic Experiments." Experimental Economics **10**: 171-178.

FISCHBACHER, URS and SIMON GÄCHTER (2010). "Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Good Experiments." American Economic Review **100**: 541-556.

FISCHBACHER, URS, SIMON GÄCHTER and ERNST FEHR (2001). "Are People Conditionally Cooperative? Evidence from a Public Goods Experiment." Economics Letters **71**: 397-404.

GÄCHTER, SIMON, DANIELE NOSENZO, ELKE RENNER and MARTIN SEFTON (2010). "Sequential vs. Simultaneous Contributions to Public Goods. Experimental Evidence." Journal of Public Economics **94**(7): 515-522.

GLÖCKNER, ANDREAS, BERND IRLENBUSCH, SEBASTIAN KUBE, ANDREAS NICKLISCH and HANS-THEO NORMANN (2011). "Leading With (Out) Sacrifice? A Public-Goods Experiment With A Privileged Player." Economic Inquiry **49**(2): 591-597.

GREINER, BEN (2004). An Online Recruiting System for Economic Experiments. Forschung und wissenschaftliches Rechnen 2003. Kurt Kremer und Volker Macho. Göttingen**:** 79-93.

GÜTH, WERNER, VITTORIA M. LEVATI, MATTHIAS SUTTER and ELINE VAN DER HEIJDEN (2007). "Leading by Example With and Without Exclusion Power in Voluntary Contribution Experiments." Journal of Public Economics **91**: 1023-1042.

HERRMANN, BENEDIKT, CHRISTIAN THÖNI and SIMON GÄCHTER (2008). "Antisocial Punishment Across Societies." Science **319**: 1362-1367.

HOLT, CHARLES A. and SUSAN K. LAURY (2002). "Risk Aversion and Incentive Effects." American Economic Review **92**: 1644-1655.

KAHNEMAN, DANIEL and AMOS TVERSKY (2000). Choices, Values, and Frames. Choices, Values, and Frames. Daniel Kahneman und Amos Tversky. Cambridge, Cambridge University Press**:** 1-16.

LEIBBRANDT, ANDREAS and RAÚL LÓPEZ-PÉREZ (2009). An Exploration of Third and Second Party Punishment in Ten Simple Games http://www.uam.es/personal_pdi/economicas/ralopez/LeibbrandtLopez_18June09.pdf.

LEVATI, VITTORIA M., MATTHIAS SUTTER and ELINE VAN DER HEIJDEN (2007). "Leading by Example in a Public Goods Experiment with Heterogeneity and Incomplete Information." Journal of Conflict Resolution **51**(5): 793-818.

LIBERMAN, VARDA, STEVEN M SAMUELS and LEE ROSS (2004). "The Name of the Game. Predictive Power of Reputations versus Situational Labels in Determining Prisoner's Dilemma Game Moves." Personality and Social Psychology Bulletin **30**(9): 1175-1185.

LIEBRAND, WIM B. and CHARLESG. MCCLINTOCK (1988). "The Ring Measure of Social Values. A Computerized Procedure for Assessing Individual Differences in Information Processing and Social Value Orientation." European Journal of Personality **2**: 217-230.

MASCLET, DAVID, CHARLES NOUSSAIR, STEVEN TUCKER and MARIE-CLAIRE VILLEVAL (2003). "Monetary and Non-Monetary Punishment in the Voluntary Contributions Mechanism." American Economic Review **93**: 366-380.

MASCLET, DAVID, CHARLES NOUSSAIR and MARIE-CLAIRE VILLEVAL (2010). Threat and Punishment in Public Good Experiments http://halshs.archives-ouvertes.fr/docs/00/52/26/55/PDF/1019.pdf.

MCCORMAC, JOHN W. (1994). "Reason Comes Before Decision." Ohio State Law Journal **55**: 161-166.

MONTERO, MARIA, MARTIN SEFTON and PING ZHANG (2008). "Enlargement and the Balance of Power. An Experimental Study." Social Choice & Welfare **30**: 69-87.

NIKIFORAKIS, NIKOS S., HANS-THEO NORMANN and BRIAN WALLACE (2010). "Asymmetric Enforcement of Cooperation in a Social Dilemma." Southern Economic Journal **76**(3): 638-659.

O'GORMAN, RICK O., JOSEPH HENRICH and MARK VAN VUGT (2009). "Constraining Free Riding in Public Goods Games. Designated Solitary Punishers can Sustain Human Cooperation." Proceedings of the Royal Society B: Biological Sciences **276**(1655): 323-329.

PUTNAM, ROBERT D. (2001). Bowling Alone. The Xollapse and Revival of American Community, Simon and Schuster.

RIVAS, M. FERNANDA and MATTHIAS SUTTER (2011). "The Benefits of Voluntary Leadership in Experimental Public Goods Games." Economics Letters **112**: 176-178.

ROCKENBACH, BETTINA and MANFRED MILINSKI (2006). "The Efficient Interaction of Indirect Reciprocity and Costly Punishment." Nature **444**(7120): 718-723.

ROSS, LEE and ANDREW WARD (1996). Naive Realism in Everyday Life. Implications for Social Conflict and Misunderstanding. Values and Knowledge. Edward S. Reed und Elliott Turiel. Maywah, Erlbaum**:** 103-135.

SALLY, DAVID (1995). "Conversation and Cooperation in Social Dilemmas. A Meta-analysis of Experiments from 1958 to 1992." Rationality and Society **7**(1): 58-92.

SCHAUER, FREDERICK (1995). "Giving Reasons." Stanford Law Review **47**: 633-659.

SEIDENFELD, MARK (2001). "The Psychology of Accountability and Political Review of Agency Rules." Duke Law Journal **51**: 1059-1095.

TETLOCK, PHILIP E. (1983). "Accountability and Complexity of Thought." Journal of Personality and Social Psychology **45**: 74-83.

XIAO, ERTE and FANGFANG TAN (forthcoming). "Justification and Legitimate Punishment." Journal of Institutional and Theoretical Economics **170**: ***.

Appendix: Instructions

The instructions for the *Baseline* and the other treatments differ only in Step 2 of Part One and in Part Two of the Experiment. The rest is identical. Therefore we report first the full instructions of the *Baseline* and afterwards only Step 2 of Part One and in Part Two of the Experiment of the other treatments. To make it easier to see the changes across treatments, the parts of the instructions that differ across treatments are shaded her; they were not shaded in the original instructions.

a) Baseline

General Instructions

In the following experiment, you can earn a substantial amount of money, depending on your decisions. It is therefore very important that you read these instructions carefully.

During the experiment, any communication whatsoever is forbidden. If you have any questions, please ask us. Disobeying this rule will lead to exclusion from the experiment and from all payments.

You will in any case receive 4 € for taking part in this experiment. In the first two parts of the experiment, we do not speak of €, but instead of Taler. Your entire income from these two parts of the experiment is hence initially calculated in Taler. The total number of Taler you earn during the experiment is converted into € at the end and paid to you in cash, at the rate of

1 Taler = 4 Eurocent.

The experiment consists of four parts. We will start by explaining the first part. You will receive separate instructions for the other parts.

Part One of the Experiment

In the first part of the experiment, there are two roles: A and B. Four participants who have the role A form a group. One participant who has the role B is allocated to each group. The computer will randomly assign your role to you at the beginning of the experiment.
On the following pages, we will describe to you the exact procedure of this part of the experiment.

Information on the Exact Procedure of the Experiment

This part of the experiment has two steps. In the first step, role A participants make a decision on contributions to a project. In the second step, the role B participant can reduce the role A participants' income. At the start, each **role A** participant receives **20 Taler**, which we refer to in the following as the **endowment**. **Role B** participants receive 20 points at the start of step 2. We explain below how role B participants may use these points.

Step 1:

In Step 1, **only the four role A participants** in a group make a decision. Each role A member's decision influences the income of all other role A players in the group. The income of player B is not affected by this decision. As a role A participant, you have to decide how many of the 20 Taler you wish to invest in a **project** and how many you wish to keep for yourself.
If you are a **role A** player, **your income** consists of two parts:
(1) the Taler you have kept for yourself (**"income retained from endowment")**
(2) the **"income from the project"**. The income from the project is calculated as follows:

Your income from the project =
0.4 *times* the total sum of contributions to the project

Your **income** is therefore calculated as follows:
(20 Taler – your contribution to the project) + 0.4* (total sum of contributions to the project).
The income **from the project** of all role A group members is calculated according to the same formula, i.e., each role A group member receives the same income from the project. If, for example, the sum of the contributions from all role A group members is 60 Taler, then you and all other role A group members receive an income from the project of 0.4*60 = 24 Taler. If the role A group members have contributed a total of 9 Taler to the project, then you and all other role A group members receive an income from the project of 0.4*9 = 3.6 Taler.
For every Taler that you keep for yourself, you earn an income of 1 Taler. If instead you contribute a Taler from your endowment to your group's project, the

sum of the contributions to the project increases by 1 Taler and your income from the project increases by 0.4*1 = 0.4 Taler. However, this also means that the income of all other role A group members increases by 0.4 Taler, so that the total group income increases by 0.4*4 = 1.6 Taler. In other words, the other role A group members also profit from your own contributions to the project. In turn, you also benefit from the other group members' contributions to the project. For every Taler that another group member contributes to the project, you earn 0.4*1 = 0.4 Taler.

Please note that the role B participant cannot contribute to the project and does not earn any income from the project.

Step 2:

In Step 2, **only the role B participant** makes decisions. As role B participant, you may **reduce or maintain** the income of **every** participant in Step 2 by distributing **points**.

At the beginning of Step 2, the four role A participants and the role B participant are told how much each of the role A participants has contributed to the project.

As a role B player, you now have to decide, for **each** of the four role A participants, whether you wish to distribute points to them and, if so, how many points you wish to distribute to them. You are obliged to enter a figure. If you do not wish to change the income of a particular role A participant, please enter 0. Should you choose a number greater than zero, you reduce the income of that particular participant. **For each point that you allocate to a participant, the income of this participant is reduced by 3 Taler.**

The total Taler income of a role A participant from both steps is hence calculated using the following formula:

Income from Step 1 – 3* (sum of *points* received)

Please note that Taler income at the end of Step 2 can also be negative for role A participants. This can be the case if the income-subtraction from points received is larger than the income from Step 1. However, the role B participant can distribute a maximum of 20 points to all four role A members of the group. 20 points are the maximum limit. As a role B participant, you can also distribute fewer points. It is also possible not to distribute any points at all.

If you have role B, please state your reasons for your decision to distribute (or not to distribute) points, and why you distributed a particular number of points, if applicable. In doing this, please try to be factual. Please enter your statement in the corresponding space on your screen. You have 500 characters max. to do this. Please note that, in order to send your statement, you will have to press "Enter" once each time. As soon as you have done this, you will no longer be able to change what you have written.

The reasons you give will remain confidential. This means that only the experimenter knows them. Of course, the reasons will remain anonymous – the experimenter will therefore not know which of the participants gave what reason.

The income of the role B participant does not depend on the income of the other role A participants, nor on the income from the project. For taking part in the first part of the experiment, he or she receives a fixed payment of

1 €.

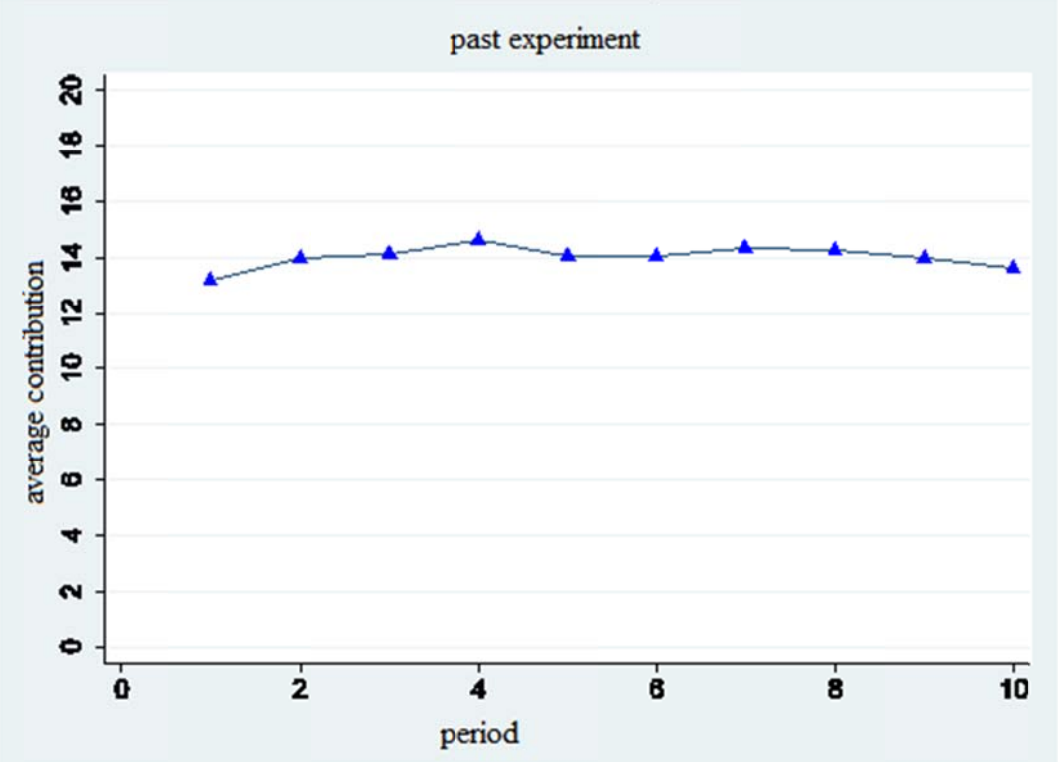
In addition, the role B participant receives the sum of 0.01 € for each point that he or she did not distribute. Once all participants have made their decisions, your screen will show your income for the period and your total income so far.

After this, the first part of the experiment ends. You will then be told what your payment is for this part of the experiment. Hence, you will also know how many points you and all other participants have been given by player B.

Experiences from an Earlier Experiment

For your information, we give you the following graph, which tells you the average contributions made in a very similar experiment that was conducted in this laboratory.

In this experiment, too, there were groups of 4 role A participants and one role B participant each. The role A participants' income was calculated in exactly the same way. The experiment had 10 equal periods. The role B participant also had 20 points at his disposal in each period. At the end of each period, the role A participants were told how much each of the other participants had contributed and how



Part Two of the Experiment

The second part of the experiment consists of 10 repetitions of the first part. **Throughout the entire second part, all participants keep the role they had in the first part of the experiment. The computer randomly re-matches the groups of four in every period. In each period, the computer randomly assigns a role B participant to each group.**

As a reminder:

In each period, each role A participant receives 20 Taler, which may be contributed to the project entirely, in part, or not at all. For each period, calculating the income from the project for the role A participants in a group happens in exactly the same way as it did in the first part of the experiment. In each period, each role B participant receives 20 points, which may be used to reduce the income of the players A in the group. For each point that a role A participant receives in a period, 3 Taler are subtracted. For each point that a role B participant does not use, he or she is given the sum of 0.01 €. In addition to the income from the points retained, each role B participant receives a flat fee of 10 € for participating in this second part of the experiment.

At the beginning of Step 2 of each period, the four role A participants and the role B participant are told how much each of the role A participants contributed to the project.

Please note that the groups are re-matched anew in each period.

After each period, you are told about your individual payoff. You are therefore also informed how many points you and the other participants have been assigned by the role B participant.

Part Three of the Experiment

We will now ask you to make some decisions. In order to do this, **you will be randomly paired with another participant.** In several distribution decisions, you will be able to allocate points to this other participant and to yourself by repeatedly **choosing between two distributions, 'A' and 'B'**. The points you allocate to yourself will be paid out to you at the end of the experiment at a rate of **500 points = 1 €**. At the same time, you are also randomly assigned to **another** participant in the experiment, who is, in turn, also able to allocate points to you by choosing between distributions. This participant is **not the same participant** as the one to whom you have been allocating points. The points allocated to you are also credited to your account. The **sum** of all points you have allocated to yourself and those allocated to you by the other participant are paid out to you at the end of the experiment at a rate of 500 points = 1 €.

Please note that the participants assigned to you in this part of the experiment are **not the members of your group** from the preceding part of the experiment. You will therefore be dealing with other participants.

The individual decision tasks will look like this:

Possibility A:		**Possibility B:**	
Your points	The points of the experiment participant allocated to you	Your points	The points of the experiment participant allocated to you
0	500	304	397

A		B

In this example: If you click 'A', you give yourself 0 points and 500 points to the participant allocated to you. If you click 'B', you give yourself 304 points and 397 points to the participant allocated to you.

Part Four of the Experiment

In this part of the experiment, you **do not form a pair** with another participant. Your decisions are therefore only significant to you and **only influence your own payoff**. The other participants' decisions only influence their own payoffs.

In this part of the experiment, you are requested to decide, **in 10 different cases (lotteries)** between **Option a and Option b**. Both options consist of **two possible payments** (one high and one low), which are paid with varying possibilities.

Options a and b are presented to you on your screen, as in the following example:

Lottery	Option a	Option b	Your decision
1	2.00 Euro with a chance of 10%, or 1.60 Euro with a chance of 90%	3.85 Euro with a chance of 10%, or 0.10 Euro with a chance of 90%	Option a ☐
			Option b ☐

The computer will ensure that these payments occur with exactly the possibilities that have been indicated.

For the above example, this means:

If option a is chosen, the winnings of 2 € have a 10 % chance of occurring, and the winnings of 1.60 € have a 90 % chance of occurring.

If option b is chosen, the winnings of 3.85 € have a 10 % chance of occurring, and the winnings of 0.10 € have a 90 % chance of occurring.

In the right-hand column, please indicate which option you would like to choose.

Please note that at the end of the experiment **only one** of the 10 cases becomes relevant for your payment. All cases are **equally possible**. The computer will randomly choose **one payment-relevant case**.
After this, the computer determines, for the payment-relevant case and with the possibilities indicated above, whether the higher (2 € or 3.85 €) or the lower winnings (1.60 € or 0.1 €) will be paid to you.

b) Private

Step 2:

In Step 2, **only the role B participant** makes decisions. As role B participant, you may **reduce or maintain** the income of **every** participant in Step 2 by distributing **points**.
At the beginning of Step 2, the four role A participants and the role B participant are told how much each of the role A participants has contributed to the project.
As a role B player, you now have to decide, for **each** of the four role A participants, whether you wish to distribute points to them and, if so, how many points you wish to distribute to them. You are obliged to enter a figure. If you do not wish to change the income of a particular role A participant, please enter 0. Should you choose a number greater than zero, you reduce the income of that particular participant. **For each point that you allocate to a participant, the income of this participant is reduced by 3 Taler.**
The total Taler income of a role A participant from both steps is hence calculated using the following formula:

Income from Step 1 – 3* (sum of *points* received)

Please note that Taler income at the end of Step 2 can also be negative for role A participants. This can be the case if the income-subtraction from points received is larger than the income from Step 1. However, the role B participant can distribute a maximum of 20 points to all four role A members of the group. 20 points are the maximum limit. As a role B participant, you can also distribute fewer points. It is also possible not to distribute any points at all.
If you have role B, please state your reasons for your decision to distribute (or not to distribute) points, and why you distributed a particular number of points, if applicable. In doing this, please try to be factual. Please enter your statement in the corresponding space on your screen. You have 500 characters max. to do this. Please note that, in order to send your statement, you will have to press "Enter" once each time. As soon as you have done this, you will no longer be able to change what you have written.
Each role A participant is informed of the reasons that you have given him/her for your decision. Of course, the reasons will remain anonymous – neither the experimenter nor the participants will therefore know which of the participants gave what reason.
The income of the role B participant does not depend on the income of the other role A participants, nor on the income from the project. For taking part in the first part of the experiment, he or she receives a fixed payment of

1 €.

In addition, the role B participant receives the sum of 0.01 € for each point that he or she did not distribute. Once all participants have made their decisions, your screen will show your income for the period and your total income so far.
After this, the first part of the experiment ends. You will then be told what your payment is for this part of the experiment. Hence, you will also know how many points you and all other participants have been given by player B.
In addition, you will be told player B's reason for distributing whatever amount of points you got. This information goes only to you. The other players do not know

this reason. They are only aware of the reasons they have been given for their own allocation of points.

Part Two of the Experiment

The second part of the experiment consists of 10 repetitions of the first part. **Throughout the entire second part, all participants keep the role they had in the first part of the experiment. The computer randomly re-matches the groups of four in every period. In each period, the computer randomly assigns a role B participant to each group.**

As a reminder:

In each period, each role A participant receives 20 Taler, which may be contributed to the project entirely, in part, or not at all. For each period, calculating the income from the project for the role A participants in a group happens in exactly the same way as it did in the first part of the experiment. In each period, each role B participant receives 20 points, which may be used to reduce the income of the players A in the group. For each point that a role A participant receives in a period, 3 Taler are subtracted. For each point that a role B participant does not use, he or she is given the sum of 0.01 €. In addition to the income from the points retained, each role B participant receives a flat fee of 10 € for participating in this second part of the experiment.

At the beginning of Step 2 of each period, the four role A participants and the role B participant are told how much each of the role A participants contributed to the project.

Please note that the groups are re-matched anew in each period.

After each period, you are told about your individual payoff. You are therefore also informed how many points you and the other participants have been assigned by the role B participant.

In addition, you will be told player B's reason for distributing whatever amount of points you got. This information goes only to you. The other players do not know this reason. They are only aware of the reasons they have been given for their own allocation of points.

c) Public

Step 2:

In Step 2, **only the role B participant** makes decisions. As role B participant, you may **reduce or maintain** the income of **every** participant in Step 2 by distributing **points**.

At the beginning of Step 2, the four role A participants and the role B participant are told how much each of the role A participants has contributed to the project.

As a role B player, you now have to decide, for **each** of the four role A participants, whether you wish to distribute points to them and, if so, how many points you wish to distribute to them. You are obliged to enter a figure. If you do not wish to change the income of a particular role A participant, please enter 0. Should you choose a number greater than zero, you reduce the income of that particular participant. **For each point that you allocate to a participant, the income of this participant is reduced by 3 Taler.**

The total Taler income of a role A participant from both steps is hence calculated using the following formula:

Income from Step 1 – 3* (sum of *points* received)

Please note that Taler income at the end of Step 2 can also be negative for role A participants. This can be the case if the income-subtraction from points received is larger than the income from Step 1. However, the role B participant can distribute a maximum of 20 points to all four role A members of the group. 20 points are the maximum limit. As a role B participant, you can also distribute fewer points. It is also possible not to distribute any points at all.

If you have role B, please state your reasons for your decision to distribute (or not to distribute) points, and why you distributed a particular number of points, if applicable. In doing this, please try to be factual. Please enter your statement in the corresponding space on your screen. You have 500 characters max. to do this. Please note that, in order to send your statement, you will have to press "Enter" once each time. As soon as you have done this, you will no longer be able to change what you have written.

All reasons are told to all role A participants in the group. Of course, the reasons shall remain anonymous – neither the experimenter nor the participants will therefore know which of the participants gave what reason.

The income of the role B participant does not depend on the income of the other role A participants, nor on the income from the project. For taking part in the first part of the experiment, he or she receives a fixed payment of

1 €.

In addition, the role B participant receives the sum of 0.01 € for each point that he or she did not distribute. Once all participants have made their decisions, your screen will show your income for the period and your total income so far.

After this, the first part of the experiment ends. You will then be told what your payment is for this part of the experiment. Hence, you will also know how many points you and all other participants have been given by player B.

In addition, you will be told player B's reasons for distributing whatever amount of points you and the other participants got. The other players also know these reasons.

Part Two of the Experiment

The second part of the experiment consists of 10 repetitions of the first part. **Throughout the entire second part, all participants keep the role they had in the first part of the experiment. The computer randomly re-matches the groups of four in every period. In each period, the computer randomly assigns a role B participant to each group.**

As a reminder:

In each period, each role A participant receives 20 Taler, which may be contributed to the project entirely, in part, or not at all. For each period, calculating the income from the project for the role A participants in a group happens in exactly the same way as it did in the first part of the experiment. In each period, each role B participant receives 20 points, which may be used to reduce the income of the players A in the group. For each point that a role A participant receives in a period, 3 Taler are subtracted. For each point that a role B participant does not use, he or she is given the sum of 0.01 €. In addition to the income from the points retained, each role B participant receives a flat fee of 10 € for participating in this second part of the experiment.

At the beginning of Step 2 of each period, the four role A participants and the role B participant are told how much each of the role A participants contributed to the project.

Please note that the groups are re-matched anew in each period.

After each period, you are told about your individual payoff. You are therefore also informed how many points you and the other participants have been assigned by the role B participant.

In addition, you will be told player B's reasons for distributing whatever amount of points you and the other participants got. The other players also know these reasons.