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

Do Explicit Reasons Make Legal Intervention More Effective? An Experimental Study

Christoph Engel
Lilia Zhurakhovska

MAX PLANCK SOCIETY

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2013/16

Do Explicit Reasons Make Legal Intervention More Effective? An Experimental Study

Christoph Engel / Lilia Zhurakhovska

revised version January 2017

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

Do Explicit Reasons Make Legal Intervention More Effective? An Experimental Study

Christoph Engel
(Max Planck Institute for Research on Collective Goods, Bonn)

Lilia Zhurakhovska
(University of Duisburg-Essen)

Abstract

When judges or public authorities intervene in citizens' lives, they normally must give explicit reasons. Justification primarily serves the sense of justice. The law's subjects want to understand the intervention. But does justification also have a forward-looking effect? Are individuals more likely to change their behavior in the legally desired direction if the intervention is accompanied by explanation? And do authorities correctly anticipate the effect? To answer these questions under controlled conditions, we use a standard tool from experimental economics. We introduce central punishment to a public goods experiment. In the Baseline, authorities are requested to justify punishment decisions, but the reasons are kept confidential. In the Private treatment, only the addressee learns the justification. In the Public treatment, reasons are made public. Whenever reasons are communicated, there is less monetary punishment. Experimental authorities partly substitute words for action. Yet this is only effective, in the sense of mitigating the dilemma, if reasons are made public.

JEL: C91, D03, D62, D63, H41, K14, K40

Keywords: justification requirement, governance effect, public good, experiment

I. Introduction

The jury does not explain its verdict. The policeman may just shoot down the aggressor. The Supreme Court may deny certiorari without giving reasons (for more examples from the legal system, see Schauer, 1995). Sometimes, intervention is the only act of communication between the law and its subjects. Yet normally, procedural rules oblige courts and administrative authorities to justify their interventions. Their decision is accompanied by explicit reasons.

The justification requirement serves multiple purposes. It becomes easier for the addressee to accept the decision if she learns why the court or the authority had to intervene. Explaining the motives is an act of procedural fairness. It becomes easier for superior authorities to check whether the subordinate authority abides by the law. The legal order at large may improve in the light of the experiences from the concrete case. Anticipating the justification requirement, authorities make better decisions (Engel, 2007). In this paper, we focus on yet another potentially beneficial effect of explicit justification: legal intervention may become more effective. Hard interventions may partly be substituted by just making it clear which behavior is expected, and that other behavior is not acceptable. The addressee and other members of the community may in the future behave more in line with the law's expectations. Yet others may be more willing to trust people to follow the law, and resist the temptation to preempt being the sucker by breaking the law themselves.

These effects are important for the law. But they are not specific to legal intervention. Any authority must decide whether to give reasons. The effects just mentioned might be even more pronounced if the authority explicitly applies the law. But if they are relevant, the effects should already be found in a neutral setting. Removing the legal context is desirable for identification. One may randomly assign participants to settings that exclusively differ by the specification of the justification requirement. If one finds differences between these treatments, one may be sure that the justification requirement has indeed been causal.

This is why we study our research question in a lab experiment that adapts a standard design from experimental economics, a linear public good. Participants face a dilemma: individually they are best off keeping the endowment they receive from the experimenter for themselves. Yet the entire group to which they are randomly assigned is better off if all of them contribute their entire endowments to a public project. Usually in this setting, many participants initially make substantial contributions to the 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 et al., 2008).

In the interest of coming closer to the legal situation we want to understand, we slightly modify the standard design. Rather than giving group members the possibility to punish each other, we randomly select one participant to be an authority for a group of four active players (for a

comparison of central and de-central punishment in public-good games, see Nosenzo and Sefton, 2014). The participant in the role of the authority receives a fixed income (analogous to the judge's salary) and 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. 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). We implement a stranger design, i.e., group composition differs every period. This is analogous to a court in which not every trial is between the same judge and the same defendant.[1]

In all treatments, authorities are requested to justify their choices. Yet, in the *Baseline*, the reasons 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 at themselves and at all other group members. It is public knowledge who will be able to read the explanations.[2]

We have subtle, but interesting results. In the *Private* and *Public* treatments, there is significantly less punishment than in the *Baseline*. This suggests that experimental 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 goes 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. For the legal debate, we thus have a qualified message: a justification requirement does indeed serve the forward-looking purpose of making the law more effective at governing the lives of its subjects. The legal order may economize on outright sanctions and react less harshly to rule violations. Yet this beneficial effect requires that official reactions to rule violations become publicly known, including the reasons given to justify interventions.

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.

1 Additional technical reasons for this design choice are discussed in the design section of the paper.

2 Communication from subordinates towards an impartial decision maker has been studied, e.g., in Kleine et al. (2016).

II. Related Experimental Literature

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 (Seidenfeld, 2001, Tetlock, 1983).

To the best of our knowledge the effect of a justification requirement on punishment by an impartial authority on compliant behavior between subordinates has not previously been studied, neither theoretically nor experimentally. In the following we explain in which ways our design, beyond addressing our legal research question, also contributes to the experimental literature.

In treatments *Private* and *Public*, justification is a form of one-way communication from the authority to the active members. Communication in public-goods experiments among active players has generally been shown to increase cooperation (see the meta-analysis by Sally, 1995; the survey by Crawford, 1998; from the rich literature see, e.g., Bochet et al., 2006). 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 punished, and the forward looking function of promulgating an authority's punishment policy.

If all players hold (sufficiently pronounced) Fehr and Schmidt (1999) 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 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 et al. (2009), non-binding announcements by active members 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 implements a consistent policy, 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 her chosen norm. We might therefore see a positive effect.[3]

If active players learn the reasons, the authority may use justification to threaten free-riders in future periods. Masclet et al. (2013) have found that threats preceding decentralized punishment increase cooperation. Unlike our paper, they have not analyzed any substitution effects between justifications and punishment. Furthermore, justifications in their paper were mainly meant as announcements for future periods, while in our paper justifications are directly connected to chosen punishment levels in the current period.

Most importantly, we entrust punishment to a fifth player who does not benefit from the contributions to the public good. Engel and Zhurakhovska (2016) run an experiment with a similar design. Yet, in Engel and Zhurakhovska (2016) the authority is neither able nor requested to justify her decisions. The large majority of authorities aim at disciplining free-riders in the groups they happen to be assigned to. In that paper, we study whether and why authorities are willing to discipline free-riders, even if this is costly for them and yields no pecuniary benefit. By contrast in the current paper, we want to investigate whether explicit justification induces recipients to increase contributions to a public good, and if so, why. It should be noted that the experimental design in Engel and Zhurakhovska (2016) and in the present paper differ in several dimensions. In particular, the other paper uses a partner-matching.

In a sender receiver game, Xiao and Tan (2014) compare three settings: a punishment authority receives a flat fee; the authority has a straightforward monetary incentive to punish senders who have not 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 punish for selfish reasons. 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. Most importantly, in our setting the authority has no monetary incentives for punishment and justifications.

3 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.

III. Design

III.1. The Game

We conduct a linear public good experiment with costly punishment by an additional participant who does not benefit from the provision of the public good. The additional participant provides reasons justifying her punishment for each of her decisions. All active players (players who can contribute to the public good) learn their own punishment and the punishment of all other group members as well as the contributions of all other group members. In addition, (depending on the treatment) active players do not or do receive the justification for their own punishment and the punishment of the other active group members. Specifically, the main experiment comes in three steps.[4]

Step 1:

N active players may contribute to the public good. In line with most of the experimental literature on public goods, (first stage) payoff of an active player i is defined by

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

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, and N is group size.

Step 2a:

An additional player A (the authority) learns about the contributions of all four active players in her group. The authority can assign punishment points to any of the active players. Therefore, the second stage payoff of an active player is given by

$$\pi_i^2 = e - c_i + \mu \sum_{n=1}^{N} c_n - \tau p \qquad (2)$$

where τ measures the negative effect of receiving one punishment point p on an active player's first stage income.

The authority has the payoff

$$\pi_A = x + y - \sum_{n=1}^{N} m p_n \qquad (3)$$

4 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.

where x is a fixed wage, y is an additional endowment she can use for punishing active players, $\sum_{n=1}^{N} p_n$is the total number of points which she assigns to any active player n multiplied by a marginal cost per point of *m*.

Step 2b:

Simultaneously to assigning the punishment points to the active players, the authority gives reasons justifying her punishment for each of her decisions. Note that the authority is also asked to explain her choice if she does not mete out punishment to an active player.

Step 3a:

Each active player is informed about the contributions made and of the number of punishment points received by each member of the group (including own contributions and punishment).

Step 3b – treatment variation:

Simultaneously, depending on treatment, each active player learns the reasons formulated by the authority for punishing herself and, in one treatment, other group members.

III.2. Treatments and procedure

At the beginning of the experiment, each subject is randomly assigned one of the two roles A (authority) or P (active player). Subjects are then matched in groups consisting of one player A in the role of an authority and N players P in the role of active players. In the experiment in all treatments, we set N=4, e=20, μ=0.4, τ=3, x=400, y=80, $m=\frac{1}{4}$. [5] In the *Public* treatment, the reasons the authority has given for all members of the group are communicated to all of them; in the *Private* treatment active players only learn the reasons concerning themselves; in the *Baseline*, reasons exclusively go to the experimenter, but active players know that the authority must formulate them. All rules and parameters are common knowledge from the beginning of the experiment on.

5 Technically, in the experiment, we use two different currencies. The income of active players is expressed in Taler, the pecuniary effect of punishment for authorities is expressed in Points. The above equation translates both into Taler. A Taler is worth 4 Euro-Cents. A Point is worth 1 Euro-Cent. The punishment ratio for the authority translated into Euro-Cents is 1:12, which makes punishment substantially cheaper than in most other public good experiments with punishment. 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.

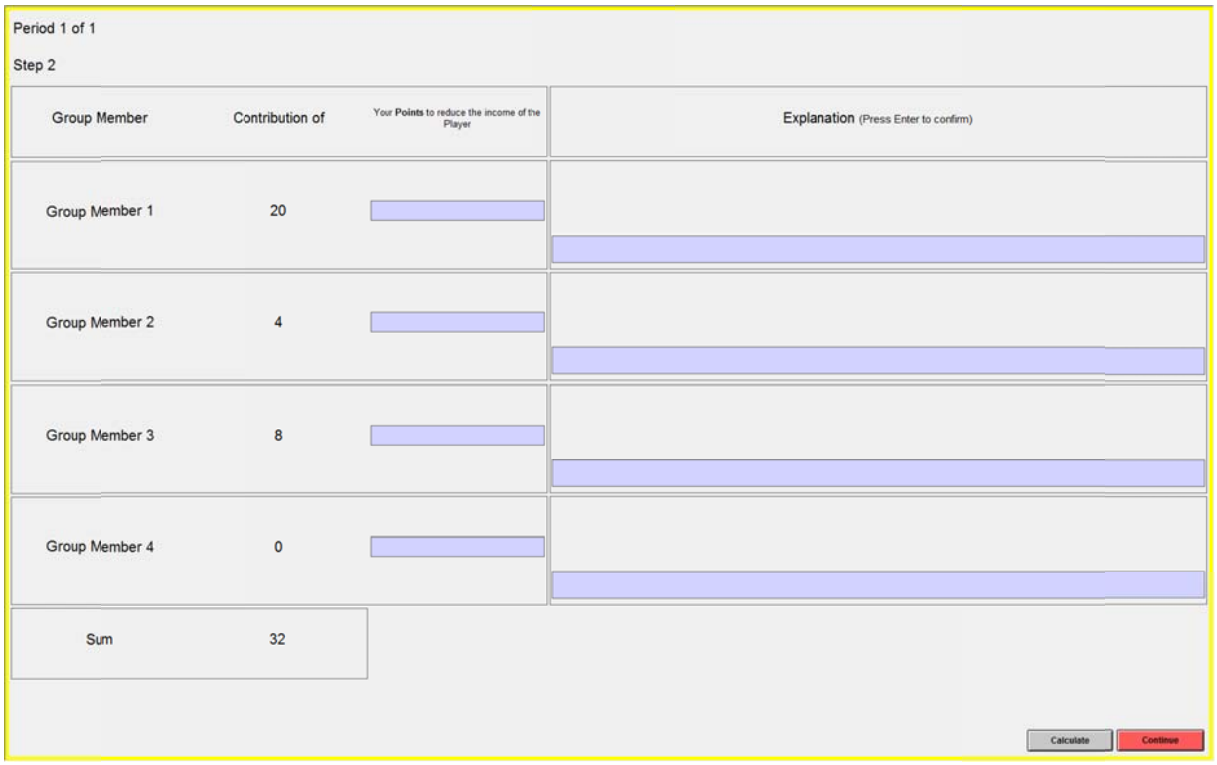


Figure 1

Decision screen of authority in Step 2

The authority must insert a number between 0 and 20 in each box in the column "*Your Points to reduce the income of the Player*". The total of her punishment points inflicted on one active group member cannot exceed 20. She may use the "*Calculate*" button to calculate the sum of points she would assign if pressing the Enter key. She is requested to type up to 500 characters in the boxes in the column "*Explanation*" to justify each of her punishment decisions. She cannot leave the stage before entering all punishment points and confirming each of her justifications by pressing the Enter key. The id number of each group member is re-shuffled each period as the group composition changes each period, as well.

As mentioned above, in Step 2b, the authority is requested to justify her punishment decisions. To do so, she is asked to type her reasons into four chat boxes, each box corresponding to one active player. Each box holds a maximum of 500 characters. This is made explicit in the instructions.[6] Figure 1 displays her decision screen.

After the end of the first period, there is a surprise restart, i.e., the game is repeated for another 10 periods. Participants receive additional instructions, which inform them that from now on they will be re-matched in every of the 10 periods, but that roles are kept constant throughout the experiment. We have matching groups of size 10, composed of eight active players and two authorities. Following the procedure that is standard in the experimental literature (see, e.g., Charness, 2000, Montero 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

6 In the instructions it is stated that authorities are not allowed to communicate any personal information, so as to preserve anonymity.

meant to guarantee independent observations, without inducing participants to second guess group composition.

Our main reason for implementing a stranger design is external validity. Judges are unlikely to meet the same defendant again. Another practical concern is that previous literature has shown contributions to be rather high in repeated public goods games with partner-matching and punishment (see, e.g., Fehr and Gächter, 2000). Had we used partner-matching, a ceiling effect might have made our manipulation meaningless. Note that a stranger design puts the socially beneficial effects of justification to a stronger test. Punished players know that they will not be in the same group in the next period. Therefore, they have less reason to feel guilty if the authority explains that they have misbehaved, and they have less of a chance to predict the effect of future authorities on other active players' future choices. This same authority and this same active player also only meet again with positive probability. Consequently, justification is less able to reduce the uncertainty about the next authority's punishment policy compared to a similar setting in the partner-design. Justification only reduces the uncertainty at the population level, not at the individual level. If we nonetheless find justification effects, we know that they are very robust.

We further increase external validity by providing subjects with a suggestion for a norm. We use a procedure that is meant to give participants as little reason as possible to interpret this norm as an expectation of the experimenter, rather than an expectation prevalent in the community of experimental participants: At the end of the instructions we inform subjects about average contributions in a similar previous experiment. A graph shows that average contributions were around c_i=14 in all periods.[7] Subjects are not told that they have to make use of this norm. It is just stated that this graph is "For [their] information".

In each session all instructions were read out aloud by the experimenter before the experiment started, to achieve common knowledge about the procedure. The experiment only started after all participants had completed a quiz about the rules and procedures, to insure that all participants had understood the instructions. Interaction was completely anonymous. The experiment was conducted in the Cologne Laboratory for Economic Research. The experiment is programmed in zTree (Fischbacher, 2007). Participants were invited using the software ORSEE (Greiner, 2015). 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 have 12 independent observations (matching groups of 10) in the *Baseline*, and 11 in each of the two treatments.[8]

7 The graph is reproduced in the Appendix, as part of the instructions.

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

IV. Hypotheses

In this experiment, we investigate in which ways differently specified justification requirements affect contribution choices. Obviously, punishment choices of the authorities and contribution choices of active players are related. In a similar experiment Engel and Zhurakhovska (2016) demonstrate that authorities do not act selfishly or anti-socially. Authorities typically punish free-riders, i.e. participants who contribute less than the average active player in their current group. We build on this result. We have no reason to believe that the justification requirement will destroy authorities' willingness to discipline the groups to whom they have been assigned. This expectation is in line with the literature on third party sanctioning (Almenberg et al., 2011, Carpenter et al., 2004, Charness et al., 2008, Fehr and Fischbacher, 2004, Leibbrandt and López-Pérez, 2012, Zhurakhovska, 2014). Based on this expectation, in this section we predict how active players react to the fact that authorities are obliged to justify punishment decisions, and how authorities respond to these reactions.

We distinguish a direct and an indirect effect. The direct effect results from the fact that an active player knows the authority is obliged to provide a written justification and that, depending on treatment, this justification is kept confidential, communicated to her in private, or communicated to the entire group of four. The indirect effect results from the fact that, irrespective of treatment, each active player learns how much the remaining members of her group have contributed to the joint project. We predict in which ways the reaction to this information differs, depending on the information an active player receives about the justification of the authority's decisions.

The direct effect may have two distinct sources: the addressee may dislike the negative characterization of her behavior inherent in the authority's statement; and the justification may affect the effect of monetary punishment. The monetary payoff is given by (2). If the first effect is relevant, utility is given by (4)

$$u_i = e - c_i + \mu \sum_{n=1}^{N} c_n - \tau p - b - \rho \tau p b \qquad (4)$$

where b is the disutility resulting from blame, and $\rho \in [-1,1]$ defines whether monetary and verbal punishment are substitutes ($\rho < 0$) or complements ($\rho > 0$). If this is the only or the dominant effect of justification, there should be no effect if reasons are not communicated to the addressee; since she never learns the words, she has no chance to be negatively affected by them. We predict

$$0 = b_{baseline} < b_{private} < b_{public}$$

In the *Private* treatment, the participant is exposed to the unpleasant experience of a reprimand. In the *Public* treatment, this disutility is heightened by the fact that all other group members read how this participant is scolded. If authorities see justification as verbal punishment, they dispose of a potential substitute for pecuniary punishment. Since pecuniary pun-

ishment is costly for them, in anticipation they reduce the amount of pecuniary punishment. As τ (the negative effect of one punishment point on the recipient's payoff) is fixed, the authority's decision variable is p (the number of punishment points). We predict

$$p_{baseline} > p_{private} > p_{public}$$

In the second interpretation of the justification requirement, the authority tries to use justification as a technology for making monetary punishment more effective. This works if active players dislike arbitrariness, and justification assures them that punishment is unlikely to be erratic.[9] On this assumption, utility is given by (5)

$$u_i = e - c_i + \mu \sum_{n=1}^{N} c_n - \tau p + \eta\sigma(\tau p) \qquad (5)$$

where σ stands for perceived variance, and η captures how important predictability is for this active participant. If the participant in question learns the reasons given for the decision, this should reduce perceived arbitrariness. This effect should be strongest if she knows the reasons given with respect to everyone. Consequently, we expect

$$\sigma_{baseline} > \sigma_{private} > \sigma_{public}$$

Now pecuniary punishment remains costly for authorities, which is why we predict that authorities anticipate the effect and adjust pecuniary punishment. We thus also predict

$$p_{baseline} > p_{private} > p_{public}$$

if the effect of the justification requirement is through reducing perceived inconsistency.

If active participants do not hold social preferences, the likely contributions of the remaining group members are not relevant information. A public good is dominance solvable. Participants do not contribute to the public good as long as $\tau p + b < 1 - \mu$. This changes with social preferences. For simplicity, we explain the point with inequity aversion à la Fehr and Schmidt (1999). But alternative definitions of social preferences would lead to the same prediction. Absent punishment, Fehr/Schmidt utility is given by (6)

$$u_i = e - c_i + \mu \sum_{n=1}^{N} c_n - \frac{1}{N-1} \sum_{j=1}^{N-1} \beta \max\{E(c_j) - c_i, 0\} \qquad (6)$$

Such participants contribute fully provided $E(c_j) = e, \forall j$ and $\beta > 1 - \mu$. Participants holding social preferences care about past contributions of other group members as a signal for their expected current contributions. If the direct effect (chiefly) results from verbal punishment (b), we predict a main effect of past contributions of others. But we have no reason to expect

9 Since we implement a stranger design, active players cannot directly verify whether a given authority sticks to her policy though. They must trust that the justification requirement disciplines them.

that this effect differs between treatments. We thus do not expect significant interaction effects.

This is different if the direct effect (chiefly) results from perceiving punishment as more predictable (σ). If the authority believes active participants to be the more sensitive to punishment the more punishment is predictable, the authority will aim at making a consistent impression. Yet its ability to manage impressions is a function of treatment. In all treatments and in each period, active participants learn how much each other member of her current group has contributed and how the authority has reacted to this choice. They can use this knowledge to form an expectation about the behavior of the current members of their newly composed group. If, in the *Baseline*, only the experimenter learns the reasons, the individual authority has no additional means to actively manage impressions. She can only rely on the generalized trust created by the fact that active participants know the justification requirement to be in place. By contrast in the *Private* treatment, active participants receive an additional private signal. But this signal is not very informative. The more the participant is well-behaved herself, the less she is likely to be punished. She does not learn how the authority thinks about antisocial behavior. By contrast if all justifications are made *Public*, all active participants have access to the complete punishment policy of the authority, and the reasons she gives for using it. The more the authority is consistent, the more an active participant must count on being punished if she violates the norm the authority aims to enforce. If this is anticipated by other active participants, they stand a better chance to predict the contributions of the remaining group members to the public project. Consequently, depending on treatment, past contributions of the other group members are a differently informative signal for their current contributions. Writing ψ for sensitivity of current choices to choices of the remaining group members in the previous period, we predict

$$\psi_{baseline} < \psi_{private} < \psi_{public}$$

We derive the following hypotheses:

H_1: Monetary punishment is highest in the *Baseline,* lower in the *Private* treatment, and lowest in the *Public* treatment.

H_2: Participants increase their contributions most intensely in reaction to having been punished in the previous period in treatment *Public*, lower in treatment *Private* and lowest in the *Baseline*.

H_3: Participants adjust their contributions most intensely to the contributions of the remaining group members in the previous period in treatment *Public*, less so in treatment *Private*, and least so in the *Baseline*.

V. Results

V.1. Treatment effects

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. Neither in non-parametric two-sided Mann-Whitney tests,[10] nor parametrically do we find any significant effects.[11] Since anticipation has no discernible effect, in the following we pool the data from the first and the second phases of the experiment.

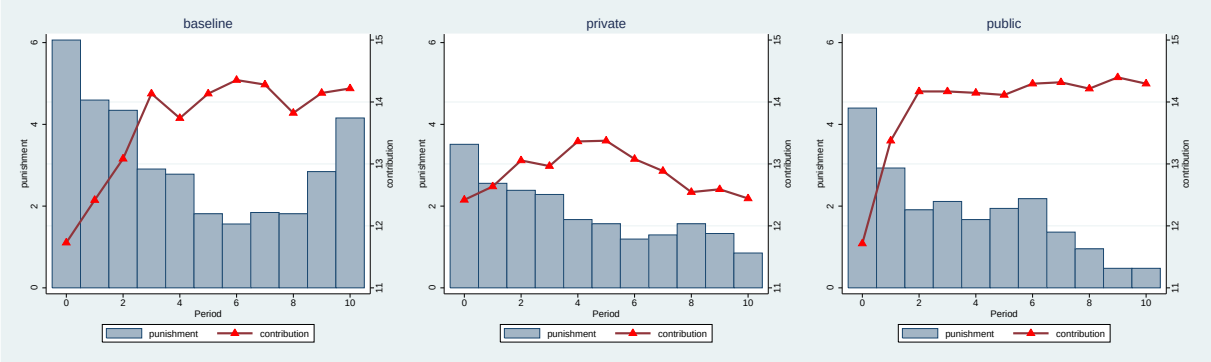


Figure 2

Treatment Effects

On the right vertical axis the mean contributions (in Taler) of the active players per period and treatment are displayed by a line. On the left vertical axis bars indicate mean punishment (in Taler) of the authorities per period and treatment. The horizontal axis shows the periods: the one-shot game is reported as period 0, the repeated game is reported as periods 1 – 10. In the graph, there is one panel per treatment.

Figure 2 reports treatment effects.[12] In all treatments, authorities use the punishment option.[13] Active players make substantial contributions to the public project.[14] The design of the experiment empowers authorities to perfectly deter free-riding. In each period each authority disposes of a maximum of 20 points for punishment. A complete free-rider is deterred if $p > .2$,

10 In the one-shot game, 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 are thus not independent. In these tests, we therefore work with mean punishment per authority as the dependent variable.

11 For parametric estimation, we use Tobit, capturing that many active participants are not punished at all.

12 In this section we present the data in Taler, which was the currency used in the experiment for contributions. We also translate punishment points into Taler, to make it easier to interpret the results.

13 Statistical tests are complicated by the fact that the design of the experiment excludes negative punishment, i.e., rewards. Therefore, technically Hypothesis 1 calls for a test at the limit of the support. We react by reporting the highest positive amount of punishment at which a one sample ranksum 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 ($N = 11$, $p = .010$).

14 Using the same procedure as in the previous footnote we find that ranksum tests still reject the hypothesis that mean contributions are 11 Taler in the Baseline ($N = 12$, $p = .049$), 9 Taler in treatment Private ($N = 11$, $p = .021$), and 10 Taler in treatment Public ($N = 11$, $p = .016$). In all treatments, mean contributions are significantly above those limits.

i.e. if she expects to lose at least .2 Taler for every Taler that her contribution falls short of the norm she expects the authority to enforce. Hence the authority needs at most 16 punishment points to enforce full contribution. Yet effectively in all treatments punishment is frequently non-deterrent, even if there is punishment.[15]

We now turn to hypothesis 1 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 2). 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 test over means per matching group, N = 23, p = .0564). 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, in model 1 of Table 1 we mirror the visual impression from Figure 2.[16] In this model, 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 (*period*) 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*. We capture the different patterns of the time trends by the square of the time trend ($period^2$), and interactions with treatments. In treatments *Private* and *Public*, 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 in treatment *Private*, and the interaction effect reverses the main effect in treatment *Public*). When there is room for it (since reasons are communicated), authorities partly substitute words for pecuniary punishment, as expected.

15 In the Baseline, 27.71% of those participants who contributed less than 20 Taler to the public projects were punished such that their second stage payoff was below the payoff they would have had, had they fully contributed. In the Private treatment, this held for 21.39% of all participants in this situation. In the Public treatment, it held for 22.13%.

16 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 are therefore from choices nested in periods nested in authorities 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 are also left censored. Thus, we report mixed effects Tobit models. Right censoring (maximum punishment) plays no role in this data, which is why we confine the specification to left censoring.

Table 1

Treatment Effect on Punishment

	Model 1	Model 2
Private	-4.472	-10.616
	(4.243)	(6.562)
Public	-5.527	-3.518
	(4.279)	(6.668)
Period	-4.536***	
	(.743)	
*Private**Period	2.401*	
	(1.084)	
*Public**Period	3.073**	
	(1.139)	
$Period^2$	.385***	
	(.072)	
*Private**$Period^2$	-.272*	
	(.107)	
*Public**$Period^2$	-.432***	
	(.116)	
Contribution		-2.666***
		(.099)
*Private**Contribution		.777***
		(.132)
*Public**Contribution		.189
		(.155)
Constant	-3.997	22.664***
	(2.968)	(4.594)
N	2992	2992
Left Censored	2300	2300

Mixed effects Tobit models, allowing punishment choices to be left censored. Standard errors for punishment choices regarding individual active participants, nested in period, nested in authority, nested in matching group. Choices from one-shot and repeated game pooled. Standard errors in parenthesis. Reference category *Baseline*. *** p < .001, ** p < .01, * p < .05.

This interpretation is further supported by model 2 in Table 1. In this model, we control for the respective active players' contributions, i.e., we estimate authorities' empirical reaction functions. The substitution effect is supported by the positive interaction between contribution and treatment (*Private*contribution*; *Public*contribution*). The interaction effect is, however, only significant for treatment *Private.* Hence, support for a substitution effect is even stronger if reasons are not made public. Yet note that authorities also have to write reasons in the *Baseline*. The decline of punishment over time in treatments *Private* and *Public* can therefore not be attributed to a reticence to explicate their choices. The only difference between the *Baseline* and the two treatments is the addressee of these explications. Based on the results from Table 1, we therefore feel confident to conclude

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

Table 2 *Treatment Effect on Contributions*	
Private	.618
	(2.288)
Public	.713
	(2.287)
Period	.233***
	(.048)
*Private**Period	-.270***
	(.070)
*Public**Period	.039
	(.070)
Constant	13.139***
	(1.582)
N	2992
Left Censored	206
Right Censored	677

Mixed effects Tobit models, allowing contribution choices to be left censored at 0, and right censored at 20. Standard errors for contribution choices nested in individual active members, nested in matching group. Choices from one-shot and repeated game pooled. Standard errors in parenthesis. Reference category *Baseline*. *** p < .001, ** p < .01, * p < .05.

Descriptively, from Figure 2 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. If we compare mean contributions of the active players per matching group, in two-sided non-parametric tests we do not find any significant treatment differences. By contrast, in a parametric test (mixed effects Tobit regression) of all treatments, we have a significant negative interaction between treatment *Private* and the time trend (Table 2). This has the following interpretation: From the significant positive main effect of *Period* it follows that contributions increase over time in the *Baseline*. Since the interaction between treatment *Public* and *Period* is insignificant, the same also holds for treatment *Public*. By contrast, the significant negative interaction between treatment *Private* and *Period* completely neutralizes the positive main effect of period. In treatment *Private*, contributions do not increase over time. This gives us partial support for our hypothesis 2:

> **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 punished player in private.

V.2. Driving Forces

In our hypothesis section we discuss potential driving forces. We test these motives in the regressions of Table 3. In both models we find a straightforward, strong effect of experienced cooperativeness of the remaining active players (the ψ factor from our theory section).[17] In both models we also find a strong and highly significant positive interaction between experienced cooperativeness and treatment *Public* ("Transparency"). Yet the interaction with the element that treatments *Private* and *Public* have in common, i.e., with the fact that reasons are communicated to the addressee of punishment, is close to zero and far from significant. This qualifies our theoretical expectation and leads to

$$\psi_{base} \approx \psi_{priv} < \psi_{pub} \ :$$

if the reasons authorities give for their punishment choices are made public, participants become much more sensitive to good experiences they make in their group.

Communicating a justification puts blame on the recipient of punishment. If these reasons are made public, an element of shaming is added. In our theory, we had expected that the increasing power of non-monetary disutility makes it possible for authorities to reduce (costly) monetary punishment, more in treatment *Private* than in the *Baseline*, and yet more in treatment *Public*. In model 1, we do not find clear effects of the variable that captures the experienced severity of punishment.[18] Yet the picture clears if, in model 2, we interact this variable with the variable that measures experienced cooperativeness. This suggests that severity and cooperativeness have to be considered as (partial) substitutes: the more a participant has experienced that other active participants are willing to cooperate, the less harsh sanctions are needed to induce her to make substantial contributions to the joint project. Yet this substitution relationship is not present if the authority's reasons are kept confidential: the two-way interaction between experienced severity and experienced cooperativeness (which measures this relationship for the *Baseline*) is close to zero and insignificant.

In fact, model 2 not only demonstrates that communicating reasons creates a relationship between experienced cooperativeness and experienced severity; this relationship critically depends on the addressees of this communication. If justifications are only communicated to those who are sanctioned themselves (*Private*), severity becomes absolutely critical (significant positive interaction between experienced severity and communication). This effect is the more important the less the group is cooperative (significant negative three-way interaction between experienced severity, experienced cooperativeness, and communication).

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

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.

We find the completely opposite relationship if reasons are communicated to the entire group (*Public*). Here severity is *less* important than in the *Baseline* (and, *a fortiori,* in treatment *Private*): the highly significant negative interaction between *Transparency* and experienced severity not only neutralizes the two-way interaction with *Communication,* but is so strong that the effect reverses signs. Yet it follows from the significant three-way interaction between *Transparency*, experienced severity and experienced cooperativeness: the difference between *Private* and *Public* is the less important the more participants are cooperative anyway. Whether reasons are only communicated to the addressee of punishment, or to the whole community, is most important if the social context (the ad hoc group) is not very cooperative. Then with only privately communicated reasons, severity becomes key. By contrast if reasons are made publicly known, communicating reasons does indeed substitute for monetary punishment. We thus find

$$b_{priv} < b_{base} < b_{pub}.$$

The justifications given for punishment are most effective in disciplining behavior if these reasons are made public. The effect is less pronounced if these reasons are kept completely confidential. The effect is least strong if justifications are only communicated to the addressee of monetary punishment. This treatment difference is the more pronounced the less the group that the participant had been assigned to in the previous period has been cooperative. The striking difference between our findings and our theory concerns reasons that are privately communicated to the addressee: communicating them does not help, but hurt.

We finally had hypothesized an effect of communicating reasons on the perceived predictability of punishment policies (our theory variable σ). Since we control for the experience channel (experienced cooperativeness) and the severity channel (experienced severity), this effect would have to show up in the main effects of communication and transparency. In fact, we have coded treatments such that variable "communication" captures a setting that gives group members an intermediate chance to perceive predictability (they learn how their own choices are treated over time), while variable transparency captures the add on resulting from, additionally, communicating these reasons to the entire group (so that a skeptical individual has a much larger knowledge base to assess whether punishment authorities do indeed act consistently). As we have shown, it is important to take the relationship between experienced cooperativeness and experienced severity into account. In model 2, which does this, we do not find an effect of communication, and we only find a weakly significant effect of transparency. This suggests that perceived consistency is at most a secondary channel on which a justification requirement affects the provision of public goods. As it happens, the signs of both communication and transparency are negative: conditional on experienced cooperativeness and on experienced severity, the effect of perceived consistency is, if at all, negative. At any rate, perceived consistency is not the main channel through which a justification requirement increases contributions in a social dilemma. We note, however, that this (non)result may be due to the fact that we have implemented a stranger design. This design makes it difficult for ac-

tive participants to assess the consistency of authorities themselves. They would have to trust the disciplining effect of the obligation to formulate explicit reasons.

Table 3
Driving Forces

	Model 1	Model 2
Communication	-.871	-1.565
	(1.868)	(1.892)
Transparency	-4.206*	-3.306^{+}
	(1.983)	(1.992)
Experienced Severity	-.183^{+}	-.449
	(.109)	(.387)
Communication*Exp_Sev	.362^{+}	1.528**
	(.189)	(.527)
Transparency*Exp_Sev	.033	-2.797***
	(.256)	(.699)
Experienced Cooperativeness	.437***	.424***
	(.048)	(.052)
Communication*Exp_Coop	-.002	.048
	(.070)	(.074)
Transparency*Exp_Coop	.422***	.341***
	(.081)	(.083)
Exp Sev*Exp Coop		.019
		(.027)
Communication*Exp Sev *Exp Coop		-.088*
		(.036)
Transparency*Exp Sev*Exp Coop		.223***
		(.050)
Period	-.016	-.030
	(.029)	(.029)
Constant	8.773***	9.025***
	(1.303)	(1.331)
N	2448	2448
Left Censored	168	168
Right Censored	567	567

Dependent variable: contributions. Mixed effects Tobit models, allowing contribution choices to be left censored at 0, and right censored at 20. Standard errors for contribution choices nested in individual active members, nested in matching group. Choices from one-shot and repeated game pooled. Standard errors in parenthesis. Reference category *Baseline*. The *Communication* dummy equals 1 for all observations which are not in the *Baseline*, the *Transparency* dummy equals 1 for all observations of the *Public* treatment. Experienced cooperativeness (*Exp_Coop*) is the mean contribution of other group members in the previous period. Experienced severity (*Exp_Sev*) is the coefficient of a local regression of received punishment on contribution, for this participant, from period 1 until the previous period. *** p < .001, ** p < .01, * p < .05, $^{+}$ p < .1.

We may now also explain why communicating justifications privately 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 3: The reasons given for punishment work as a partial substitute for monetary sanctions only if they are made public.

V.3. Explicit Reasons

Compared with the *Baseline,* in treatment *Private*, authorities reduce punishment, but contributions do not stabilize. We have explained this result by a mismatch between authorities' expectations and active players' reactions. As further support for this explanation, we have had two independent raters rate the explicit reasons along the eight dimensions listed in Table 4.[19] Inter-rater reliability is good (mean Cohen's kappa is .766).[20] For descriptives and analysis, we work with the mean rating of the two raters, per reason given.

As Table 4 demonstrates, most of the time most authorities make statements that are not selfish, and even less focused on their personal profit. Both effects are most pronounced in treatment *Private.* The difference between treatment *Private* and the *Baseline* is significant for both dimensions. In the *Baseline,* the profit motive is also significantly more pronounced than in treatment *Private*. In this treatment, authorities clearly want to use the verbal channel for disciplining their assigned groups. But in this treatment, authorities are also significantly more likely to insist on efficiency, to impose an idiosyncratic behavioral rule, or to make other unspecified (non-selfish) statements than in both other treatments, and they are significantly less likely to justify punishment with the fact that others in the group have contributed more. This pattern fits the idea that authorities who only communicate with the individual addressee of punishment overestimate their ability to govern by words.

19 The instructions given to raters are reported in the Appendix.

20 Cohen's kappa starts from the probability that the two raters come to the same conclusion if both randomize. Since all individual codes are binary, and since we have two raters, this probability if .25 for each reason. Cohen's kappa is ((1 – mean rating) - .25) / (1 - .25). kappa > .6 is regarded to be substantial, kappa > .8 is regarded to be near complete.

TABLE 4

Explicit Reasons

	selfish	profit	efficiency	history	relative	idiosyncratic	ethical	unspecified
Baseline	27.95	24.24	25.22	0.93	19.00	23.95	20.76	10.68
Private	19.25	16.59	28.92	1.92	11.63	29.16	23.35	16.11
Public	27.67	20.52	24.44	3.50	17.87	25.45	17.07	7.37
Baseline vs. Private	.002	.002	.085		.001	.006		.002
Baseline vs. Public							.071	.053
Private vs. Public	.026		.035		.008	.044	.003	<.001

Lines 1-3: each statement has been classified by two independent raters in all eight dimensions. Mean statements over both raters and all reasons are reported, in percent of authorities in respective treatment using that reason.

Lines 4 and 5: p-values of treatment coefficients in multivariate regression, explaining classification in all eight dimensions with treatment, standard errors for statements nested in authorities nested in matching groups.

Line 6: p-values of Wald tests, testing for the respective difference between both treatment coefficients being 0. The multivariate model is in order since each statement is rated in all dimensions, so that there are multiple dependent variables per statement.

VI. Conclusion

As a default, courts and administrative authorities must justify their decisions. These reasons routinely go to the addressee. Often, decisions are also made publicly known, in recent years quite frequently even online. The justification requirement serves many purposes. In this experimental paper we focus on one of them: we test whether the law becomes more effective in governing people's lives if the intervention comes with an explicit justification. We have a qualified result: if the justification exclusively comes to the attention of the addressee of the intervention, it is not only pointless, but even counterproductive. Punishment with a justification that remains confidential is more effective at disciplining free-riders in a dilemma situation. However, there is a socially beneficial effect of justification if these reasons become publicly known. Then, verbal intervention partly substitutes for monetary intervention. Authorities can act more moderately to achieve the same stabilizing effect. We show that the counterproductive effect of privately communicated reasons results from a mismatch between the authorities' expectations and the active participants' reactions. Authorities seem to believe erroneously that private communication suffices for the substitution effect.

One should be cautious when extrapolating from the lab to the field. Caution is even more in order if one relies on experiments with students for analyzing the effect of institutions in the courtroom or in administrative procedure. 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. Specifically, in the experiment, interaction was anonymous, whereas in the courtroom and in administrative procedure, the

authority and the potential recipient of punishment are identified. In the experiment, the role of an authority was randomly assigned, whereas judges are elected or appointed, as are administrators. Arguably, legal authorities have superior competence, while our authorities are randomly selected. We do not give the authority an explicit rule to enforce.[21] Instead, our design provides two reasonable behavioral norms: on the one hand, from the opportunity structure it is obvious that full contribution is the only efficient choice; on the other hand, participants receive information about contributions in a previous similar experiment, as a hint to behavior that is socially acceptable and therefore unlikely to attract punishment. By using a stranger matching, we not only come closer to the characteristic situation in courts. We also put our hypothesis 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, thereby isolating the effect of communicating reasons. In the experiment, if communication is permitted it is strictly unilateral. In the courtroom, the defendant may at least explicitly ask for a justification, and usually has the legal 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 focus excessively 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. With these data, Jeremy Bentham's quest for making punishment decisions public (Bentham, 1830) gains support. One sees that this is not only desirable because would-be perpetrators realize the threat with punishment. They also better understand what the authority is after, and they learn that the antisocial behavior of others does not go unchecked. A possible policy implication concerns the promulgation of justifying words. Our experiment suggests that justifications should not only be addressed to the perpetrator, but that they should be made publicly available.

21 We thus avoid a strong experimenter demand effect (Zizzo, 2010).

References

Almenberg, Johan; Anna Dreber; Coren L. Apicella and David Rand. 2011. "Third Party Reward and Punishment. Group Size, Efficiency and Public Goods," N. Palmetti and J. P. Russo, *Psychology of Punishment.* New York: Nova, 73-92.

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

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), 421-28.

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, 274-90.

Bochet, Oliver; Talbot Page and Louis Putterman. 2006. "Communication and Punishment in Voluntary Contribution Experiments." *Journal of Economic Behavior & Organization,* 60, 11-26.

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), 407-29.

Charness, Gary. 2000. "Self-Serving Cheap Talk. A Test of Aumann's Conjecture." *Games and Economic Behavior*, 33, 177-94.

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.

Crawford, Vincent. 1998. "A Survey of Experiments on Communication Via Cheap Talk." *Journal of Economic Theory*, 78(2), 286-98.

Croson, Rachel T.A. and Melanie Marks. 2001. "The Effect of Recommended Contributions in the Voluntary Provision of Public Goods." *Economic Inquiry,* 39, 238-49.

Engel, Christoph. 2007. "The Psychological Case for Obliging Judges to Write Reasons," C. Engel and F. Strack, *The Impact of Court Procedure on the Psychology of Judicial Decision Making.* Baden-Baden: Nomos, 71-109.

Engel, Christoph and Lilia Zhurakhovska. 2016. "You Are in Charge. Experimentally Testing the Motivating Power of Holding a (Judicial) Office," https://papers.ssrn.com/sol3/papers.cfm?abstract_id=2214525.

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, 980-94.

Fehr, Ernst and Klaus M. Schmidt. 1999. "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, 114, 817-68.

Fischbacher, Urs. 2007. "Z-Tree. Zurich Toolbox for Ready-Made Economic Experiments." *Experimental Economics*, 10, 171-78.

Greiner, Ben. 2015. "Subject Pool Recruitment Procedures. Organizing Experiments with Orsee." *Journal of the Economic Science Association*, 1, 1-12.

Herrmann, Benedikt; Christian Thöni and Simon Gächter. 2008. "Antisocial Punishment across Societies." *Science*, 319, 1362-67.

Holt, Charles A. and Susan K. Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review*, 92, 1644-55.

Kleine, Marco; Pascal Langenbach and Lilia Zhurakhovska. 2016. "Fairness and Persuasion. How Stakeholder Communication Affects Impartial Decision Making." *Economics Letters*, 141, 173-76.

Leibbrandt, Andreas and Raúl López-Pérez. 2012. "An Exploration of Third and Second Party Punishment in Ten Simple Games." *Journal of Economic Behavior & Organization*, 84, 753-66.

Liebrand, Wim B. and Charles G. 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, 217-30.

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, 366-80.

Masclet, David; Charles Noussair and Marie-Claire Villeval. 2013. "Threat and Punishment in Public Good Experiments." *Economic Inquiry*, 51, 1421-41.

McCormac, John W. 1994. "Reason Comes before Decision." *Ohio State Law Journal*, 55, 161-66.

Montero, Maria; Martin Sefton and Ping Zhang. 2008. "Enlargement and the Balance of Power. An Experimental Study." *Social Choice & Welfare*, 30, 69-87.

Nosenzo, Daniele and Martin Sefton. 2014. "Promoting Cooperation. The Distribution of Reward and Punishment Power," P. v. Lange, B. Rockenbach and T. Yamagishi, *Social Dilemmas. New Perspectives on Reward and Punishment.* Oxford: Oxford University Press, 87-114.

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, 633-59.

Seidenfeld, Mark. 2001. "The Psychology of Accountability and Political Review of Agency Rules." *Duke Law Journal*, 51, 1059-95.

Tetlock, Philip E. 1983. "Accountability and the Perseverance of First Impressions." *Social Psychology Quarterly*, 46, 285-92.

Xiao, Erte and Fangfang Tan. 2014. "Justification and Legitimate Punishment." *Journal of Institutional and Theoretical Economics*, 170, 168-88.

Zhurakhovska, Lilia. 2014. "Strategic Trustworthiness Via Unstrategic Third-Party Reward – an Experiment," http://ssrn.com/abstract=2449286.

Zizzo, Daniel John. 2010. "Experimenter Demand Effects in Economic Experiments." *Experimental Economics*, 13(1), 75-98.

Appendix A. 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.1. 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 €.

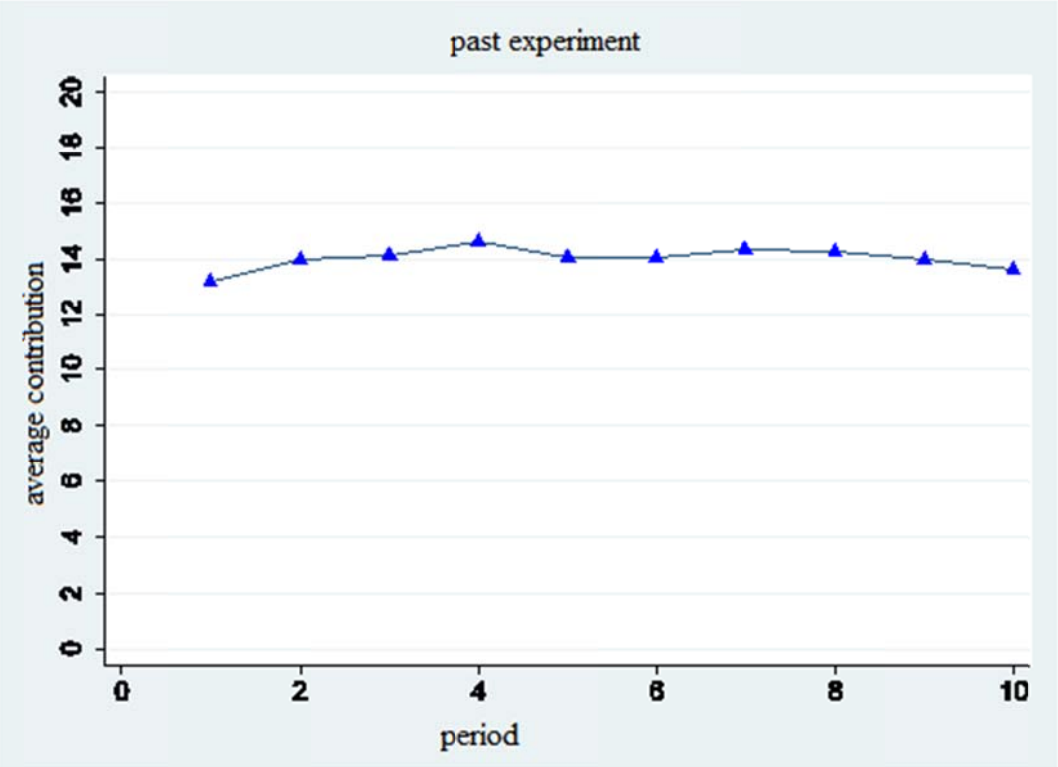
In addition, the role B participant receives the sum of 0.01 € for each point that he or she did not distribute. Once all participants have made their decisions, your screen will show your income for the period and your total income so far.

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 the role B participant had reacted to this.



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.

A.2. 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.

A.3. 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 state-

ment, 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.

Appendix B. Coding Scheme

Here, the coding scheme that had been given to the two independent raters is presented.

Coding Scheme

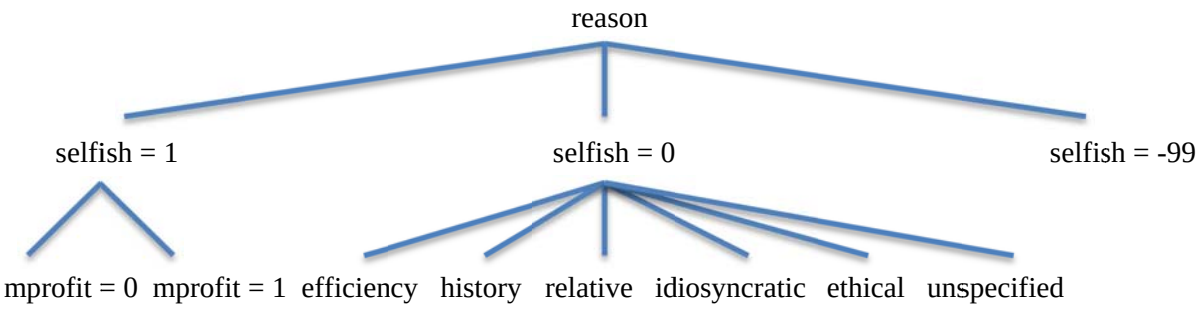
The coding scheme has two levels. At the first level, each reason is classified to be "selfish" (dummy = 1) or not. At the second level, selfish reasons, on the one hand, and non-selfish reasons, on the other hand, are classified.

If an authority does not give any reason for one of her choices, for this choice all classification variables should be put to missing (-99).

There are two options for selfish reasons: either the authority expresses that she cares about her own profit (profit = 1), or she voices any other reason that is not concerned with the good governance of the group to which they have been attached (profit = 0)

There are several suboptions for classifying reasons that, at the first level, have been classified as not selfish (selfish = 0), with "other non-selfish" being the fall back option. Hence, for each reason that is classified as "selfish = 0" on the first level, one of the suboptions should be 1. At this level, more than one dummy may be 1. Hence the suboptions are not mutually exclusive.

The following is a graphical representation of the classification scheme. It also introduces variable names. All variables are dummy variables, with -99 marking a missing value:



Does the authority justify her choice with:

mprofit:	concern for her own profit?
efficiency:	efficiency?
history:	choices of participants in the previous experiment?
relative:	choices of other participants in the current experiment?
idiosyncratic:	her own idiosyncratic standard?
ethical:	ethical / moral concerns?
unspecified:	non-selfish concerns that do not fall in any of the previous categories?

Please do not leave any variables blank.

The variables mprofit, efficiency, history, relative, idiosyncratic, ethical, and unspecified should all be either 0 or 1.

The variable selfish should be either 0, 1, or -99 (if the authority has not given a reason at all).