**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 March 2018

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[†]& Lilia Zhurakhovska[‡]

Sunday 25th March, 2018

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

[*]We are grateful to Paul Schempp and two anonymous referees for helpful remarks on an earlier version.

[†]Max-Planck-Institute for Research on Collective Goods, Bonn; University of Bonn, Faculty of Law; Erasmus University Rotterdam, Law School; corresponding author: Kurt-Schumacher-Straße 10, D 53113 Bonn, engel@coll.mpg.de

[‡]University of Duisburg-Essen, Mercator School of Management

1 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 intervention may become more powerful, or it may be supplemented with merely verbal intervention. Specifically we investigate whether justification affects addressees on one of three channels: they become more sensitive to normative expectations an authority tries to impose, as explicit reasons make intervention more predictable; they dislike being blamed; they are more willing to trust that others will refrain from exploiting them. If justification makes hard punishment more effective, authorities might respond by reacting less harshly to norm violations.

In judicial practice, justification requirements come in different specifications. In the (rather rare) situations in which courts have power to decide without giving reasons to the addressee, for internal purposes a justification may still be expected.[1] Frequently, reasons are only communicated to the addressee of intervention. For instance, in many legal orders, the written reasons of rulings made by courts of first instance are not made publicly available. In other court cases, however, procedure is completely transparent. The complete ruling is made available on the court's website. We wonder whether these differences matter for the effectiveness of justification.

Justification effects are important for the law. But they are not specific to legal intervention. Any authority must decide whether to give reasons. A typical authority has the intention to govern the community for which she is responsible. She cares about the effectiveness of her intervention. She often also cares about ways to achieve her normative goal in a more heedful way, or at lower cost for that matter (for experimental evidence, see Engel and Zhurakhovska, 2017).

Generalising beyond the legal domain is helpful for designing a test. Removing the legal context improves identification. One may randomly assign participants to settings that exclusively differ by the specification of the justification requirement. If we find statisti-

[1] A case in point is the German Constitutional Court. According to § 93 I 3 BVerfGG, the court may reject a constitutional complaint without giving reasons. This decision is routinely taken by a chamber of three justices, § 93b,1 BVerfGG. In preparation, the justice referee prepares a written report. If her two colleagues agree with the proposed rejection, this report is merely taken on file.

cally significant differences between these treatments, we can conclude that the justification requirement has indeed caused the differences. Under the controlled conditions of a lab experiment, we also have the chance to discriminate between the three potential channels.

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 (Chaudhuri, 2011; Ledyard, 1995; Zelmer, 2003). The trend reverses if participants are given the opportunity to punish each other, despite the fact that, in the typical implementation, punishment is costly (Fehr and Gächter, 2000; Herrmann et al., 2008).

In the interest of coming closer to the legal situation we wish 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 in every period. This is analogous to a court in which not every trial is between the same judge and the same defendant.[2]

In all treatments, authorities are requested to justify their choices (including the decision not to punish a participant). 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.

We have subtle, but interesting results. If justifications are kept confidential, and if they are made available to everybody, contributions to the public good increase in early periods and then stabilize at a rather high level. This stabilizing effect is absent if reasons are only communicated to its addressee. In the *Private* and *Public* treatments, authorities continuously reduce punishment over time, while they increase punishment in later periods in the *Baseline*. In all treatments, the profit of active participants increases over time. But this beneficial effect is only about half as pronounced in the *Private* treatment. In this treatment, authorities discriminate less intensely between low and high contributions, while active players are even more sensitive to experienced severity of punishment. This suggests a mismatch between the expectations of authorities and of active players. By contrast, in the

[2]Additional technical reasons for this design choice are discussed in the design section of the paper.

Public treatment, the same governance effect as in the *Baseline* is reached with less central intervention.

For the legal debate, we thus have a qualified message: if one extrapolates from the experiment to the field, 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 develops a theoretical framework and derives predictions. Section 5 reports results. Section 6 concludes.

2 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 construct reality authoritatively, to increase compliance, to enable control, to remove biases in addressees, to dissolve conflict (Engel, 2007) and to make authorities more accountable (Tetlock, 1983; Lerner and Tetlock, 1999; Seidenfeld, 2002).

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 players. 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. This also distinguishes our paper from a recent experiment that, in one treatment, asked group members to tick one of three possible reasons for demanding a certain level of contributions when deciding whether to punish other group members (Andrighetto et al., 2016).

If all players are (to a sufficient degree) averse to inequity (Fehr and Schmidt, 1999), 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 players had practically no effect. There was a slight effect if afterwards it could be checked whether active players behaved as announced. Yet in our experiment, active players cannot check whether the authority implements a consistent policy, given that active players and authorities are re-matched every period.

Croson and Marks (2001), in a public good game, 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, in the sense of all having the same endowment and earning prospect. Moreover if the authority uses justifications to fix an expected contribution level, this is not a recommendation by the experimenter, but by another participant. Finally, the authority has power to enforce her chosen norm. Due to these differences, we might see a positive effect.

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.

If justifications are communicated to all group members, over time active group members more frequently observe punishment. This makes the possibility of being sanctioned more salient. Xiao and Houser (2011) have shown that, in a public good with automatic, but nondeterrent punishment, this richer information base increases contributions. In the *Public* treatment, a similar effect might obtain.

Most importantly, we entrust punishment to a fifth player who does not benefit from the contributions to the public good. Engel and Zhurakhovska (2017) run an experiment with a similar design. Yet, in that experiment the authority is neither able nor requested to justify her decisions.[3] 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.

[3] In that experiment, we also use a partner design, and do not inform active players about results from an earlier experiment. The two experiments should therefore not be read as different treatments of one and the same design.

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.[4]

3 Design

In section 4, we formally define the channels on which a justification requirement might affect the choices of active players and authorities. In section 5 we use differences between treatments, and differences in the development of dependent variables over time, to discriminate between these channels. But the purpose of this experiment is not testing some novel definition of utility, or the equilibrium in a Bayesian game. For these purposes, the design of the experiment would be too rich. We go to the lab with a different intention. We want to learn whether a justification requirement increases the effectiveness of punishment resulting from a monetary disincentive. It is this research question that guides the design of the experiment. We want to exploit the controlled conditions of a lab experiment to isolate a potentially socially beneficial effect of a justification requirement. We do so with the aim of making an empirical contribution to the normative debate about the desirability of justification. In this section, we explain in which ways we have translated this normative question into an experimental design.

3.1 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 choice whether to punish an active player, and if so, how severely. 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, and depending on the treatment, active players do not or do receive the justification for their own punishment and the punishment of the other active players. Specifically, the main experiment comes in three steps.

[4] Xiao (2017) is more remote. She is interested in the effect of a justification requirement on the willingness of participants to overcome the temptation to make a selfish lie. We start from the finding of our earlier experiment that experimental authorities with no monetary interest in the public project used their punishment technology in a reliable way to discipline the groups to which they have been randomly assigned, Engel and Zhurakhovska (2017).

Step 1 Active players i are randomly allocated to groups of size N to play a linear public good game. Their action space is constrained to making contributions to the public good. There is an authority a with no monetary interest in the public good, but power to punish active players. Each member i has the same endowment e_i and is free to invest any amount $c_i \leq e_i$ in a public project. μ is the marginal per-capita rate.

Step 2 Authority a has an endowment e_a which she may use to inflict punishment points p_i of defined severity τ on anyone of the active players. Each punishment point costs her m tokens. She may keep any point of endowment that she does not use for punishment. Hence her (period-)profit π_a[5] is given by

$$\pi_a = e_a - \sum_{i=1}^{N} m p_i \tag{1}$$

and the profit of active players π_i is given by

$$\pi_i = e_i - c_i + \mu \sum_{i=1}^{N} c_i - \tau p_i \tag{2}$$

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

Step 3 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). Simultaneously, depending on treatment, each active player learns the reasons formulated by the authority for punishing herself and, in one treatment, other group members.

3.2 Treatments, Parameters, and Procedure

At the beginning of the experiment, each subject is randomly assigned one of the two roles a (authority) or i (active player). Subjects are then matched in groups consisting of one player a in the role of an authority and N players i in the role of active players.

In all treatments of the experiment, we set $N = 4, e_i = 20, \mu = .4, \tau = 3, e_a = 80, m = .25$. The authority earns a fixed wage w, which is the equivalent of 400 units of experimental currency.[6]

[5] For her participation in the experiment, the authority additionally receives a fixed wage w that is unrelated to her choices.

[6] 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 Eurocent. A Point is worth 1 Eurocent. The punishment ratio for the authority translated into Eurocent 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

As mentioned, the authority is requested to justify her punishment decisions.[7] 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.[8]

At the beginning of the experiment, participants are informed that the experiment will consist of several phases, but that they will only learn the exact content of each phase immediately before it starts.[9] The first phase of the experiment is a one-shot version of the experiment described above. In this phase, we can test whether active players anticipate the effects of a justification requirement.

After the end of the first phase, participants receive additional instructions for the second phase of the experiment. Now they learn that the experiment conducted in the first phase will be repeated for another 10 periods. Further, they are informed that from now on they will be re-matched in every of the 10 periods (stranger design), 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 meant to guarantee independent observations, without inducing participants to second-guess group composition.

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. This same authority and this same active player only meet again with positive probability. Consequently, justification is less able to reduce the uncertainty about the next authority's punishment policy. Justification only reduces the uncertainty at the population level. Punished players know that they will not be in the same group in the next period. Therefore, they have less reason to be concerned about their social image 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. If we nonetheless find justification effects, we can be certain that they are very robust.

---

experiments, the authority does not benefit from contributions at all. Therefore, any cost demonstrates intrinsic willingness for punishment and makes it meaningful.

[7] We have also run a treatment with no justification requirement. Results look similar to the *Baseline*, Figure A4, which is why in this paper we focus on the differential effect of different specifications of a justification requirement.

[8] See Figure A1.

[9] Phases are called "Parts" in the instructions. We have two post-experimental tests. "Part Three of the Experiment" is a test for social value orientation (Liebrand and McClintock, 1988), and "Part Four of the Experiment" is a test for relative risk aversion (Holt and Laury, 2002). We did not expect treatment differences in these measures. We included the measures as a safeguard. In case subjects' heterogeneity in these preferences would have been high, we might have been forced to control for these standard personality traits in order to find treatment differences in our main variables of interest. However, we were able to find treatment differences without using these additional controls.

Public good games are normatively ambiguous. The efficient choice is for all group members to contribute their entire endowments. But if one member does, and another does not, the former is exploited. By the design of the game, exploitation is likely if the norm is full contributions. A single deviation by one group member suffices. This suggests a less demanding norm. But ex ante, it is not clear which lower level would be appropriate. To reduce the resulting normative ambiguity, 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. 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 correctly answered control questions about the rules and procedures, to ensure 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 a mean age of 24.31. 51.54% were female. Participants on average earned 15.81 € (US$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.[10]

4 Theoretical Framework

In a public good, marginal per-capita rate $\mu < 1 < N\mu$ constitutes the dilemma: if there is no punishment ($p_i = 0$) and $\mu < 1$, each active player maximizes profit by not contributing to the project. Yet the group is best off if all members contribute $c_i = e_i$, as $N\mu > 1$.

If the authority maximizes profit, she does not mete out punishment, irrespective of the active players' choices c_i. The authority keeps her endowment e_a. Yet, in an earlier paper we showed that, in a closely related setting, experimental authorities aim at disciplining the groups to which they have been randomly assigned (Engel and Zhurakhovska, 2017). This implies that an authority experiences disutility if an active player's choice c_i differs from the norm $\tilde{c}$ that the authority deems appropriate. If the authority expects punishment to increase the probability that $c_i \geq \tilde{c}$, she must trade off disutility from norm deviations in the group she wants to control against a reduction of her endowment. Then her decision problem is

[10] As described above, each matching group consisted of 10 participants (2 authorities and 8 active players) who were split into two sub-groups (each with 1 authority and 4 active players) and re-matched in each period. In most sessions, 3 matching groups of 10 participated simultaneously. Hence most sessions consisted of 30 participants. In the *Private* and *Public* treatments, there was one matching group we could not fill, since invited participants did not show up.

given by (3):

$$u_a = e_a - \sum_{i=1}^{N} (mp_i + f(\dot{c}_i(p_i), \tilde{c})) \tag{3}$$

The second term in the sum operator captures disutility from doing a bad job. For generality we use a generic loss function f.[11]. When a decides how many punishment points to inflict on group member i, she observes the actual contribution c_i. But actual punishment does not have a forward-looking effect. To achieve the normative goal $\tilde{c}$, the critical parameter is the punishment group member i expects if she contributes c_i. We denote a's belief about i's reaction to (expected) punishment by $\dot{c}_i$.[12] The authority finds the optimal number of punishment points p_i^* by solving the first-order condition of (3) for p_i:[13]

$$-\frac{\partial f}{\partial \dot{c}_i}\frac{\partial \dot{c}_i}{\partial p_i} = m \tag{4}$$

If an active player is selfish, her utility is given by (2). Were she to know the authority's punishment policy with certainty, she would contribute $c_i = \tilde{c}$ if $p_i \geq 1 - \mu$, and $c_i = 0$ otherwise.[14] Now the design of the experiment does not make it possible for the authority to commit to a punishment policy. When deciding how much to contribute to the public good, i must work with an expectation about the authority's conditional choice $\dot{p}_i|c_i$, and about the norm the authority wants to impose $\dot{\tilde{c}}$. In the one-shot game, the only proxy is the information about choices in the previous experiment. In the repeated game, there is an additional proxy: the punishment $p_{i,t-1}|c_{i,t-1}$ that this participant i has received for her own contribution in the previous period $t-1$, as well as the punishment $p_{-i,t-1}|c_{-i,t-1}$ that other members $-i$ of her then group have received for their choices. We thus posit $\dot{p}_{i,t}(P_{t-1})$: The belief about punishment in the current period is a function of the punishment P_{t-1} that all members of previous period's group have received for their choices.

Predictability This opens up a first channel on which treatments might matter. In the *Baseline*, active players only observe the authority's actions. In the *Private* treatment, the authority also has a chance to explain her intentions. Writing $j_{i,t-1}$ for the explicit justification given for the intervention $p_{i,t-1}$, we thus posit $\dot{p}_{i,t}(P_{t-1}, j_{i,t-1})$. In the *Public* treatment, the authority has even more scope, as active players also learn the authority's intentions with respect to choices other than their own.[15] This makes it easier for them to predict what would happen to them, were they to change their contributions in the next

[11] One straightforward option would be quadratic loss: $f(\cdot) = (c_i - \tilde{c})^2$.

[12] Throughout the paper, we denote first-order beliefs with one dot, and second order beliefs with two dots.

[13] In the experiment, given experiences from earlier experiments with a comparable setting, e_a is so large that the budget constraint does not bind. We therefore abstain from defining the authority's problem as optimization under constraints.

[14] Since (2) is linear in c_i, the optimal choice is a corner solution.

[15] Note that active players learn punishment *choices* with respect to all group members in all treatments. If we find an effect of treatment *Public*, it can therefore not merely result from the fact that active players have seen more punishment choices; Xiao and Houser (2011) show that this increases contributions. The effect

period. The former effect can also be characterised as individual learning, and the latter effect as vicarious learning (cf. Bandura, 1977). More formally, we posit $\dot{p}_{i,t}(P_{t-1}, J_{t-1})$, where J are all justifications given to all members of the respective group. Treatments matter if the additional information from making intentions explicit makes predictions $\dot{p}_i$ more precise. As a shorthand we use $\sigma \in \mathbb{R}^+$ for the noise ratio in the expectations of active players about $p_i|c_i$ and $\tilde{c}$, and predict

$$\sigma_{base} > \sigma_{priv} > \sigma_{pub}$$

If selfish active players anticipate $p_i > 1 - \mu$, they adjust $c_i = \dot{\tilde{c}}$. Consequently, they have less reason to adjust choices in reaction to the experiences from the previous period. We therefore expect reactions to past punishment to be most pronounced in the *Baseline*, less pronounced in the *Private*, and least pronounced in the *Public* treatment.

According to (3), the authority has no positive utility from group members even contributing more than $\tilde{c}$. Yet the authority wants to avoid that $c_i < \tilde{c}$. Consequently, the less the authority is sure about $c_i|\dot{p}_i$, the more she must overshoot by punishing more severely. Conversely, the more the authority is confident that active players will correctly anticipate p_i and $\tilde{c}$, the more she will reduce punishment, to save money. We therefore expect

$$p_{base} > p_{priv} > p_{pub}$$

In the one-shot game, active players cannot build on prior information to generate beliefs. But they know whether the authority will have to justify her choices, and to whom these justifications will be communicated. Within the confines of the model, for this to matter we would have to posit $\dot{p}_i(\dot{j}_i)$ and $\dot{p}_i(\dot{J})$: not only observed justification improves the formation of beliefs about punishment; active players also anticipate the disciplining effect of the justification requirement on authorities' choices. If authorities were to anticipate this (anticipation) effect on the choices of active players, they would also reduce punishment in the one-shot game, the more so the more pronounced the justification requirement.

From these considerations we derive

Hypothesis 1 Predictability:

a) **individual learning:** *Active participants contribute more if they have been punished more severely in previous periods. The effect is more pronounced in the* **Baseline** *than in the* **Private** *and* **Public** *treatments.*

b) **vicarious learning:** *Active participants contribute more if the group to which they had been assigned has been punished more severely in the past. The effect is more pronounced* **Baseline** *in the* **Private** *treatment than in the* **Public** *treatment.*

Active Participants' Utility The effects leading to Hypothesis 1 do not require active players to be motivated by anything but profit. Yet, even in the absence of any intervention, in public good experiments many participants make substantial contributions, in particular at the beginning of repeated interaction (Ledyard, 1995; Chaudhuri, 2011). To rationalize such findings, one must shift from profit to utility. In that spirit we now develop hypotheses

must result from the fact that additionally learning justifications makes information about punishment more informative.

that assume utility as in (5)

$$u_i = g(c_i, c_{-i}) - p_i - \mathbb{1}b_i \tag{5}$$

where $g(.)$ is generic for profit or utility as a function of i's own contribution and the contributions of her group members $-i$, and $b_i \in \mathbb{R}^+$ stands for image concerns. The indicator variable $\mathbb{1}$ points to the fact that image concerns are possible, but need not be present.

Image Concerns Identity (Akerlof and Kranton, 2000; Bénabou and Tirole, 2011) in the form of self-image (Ariely et al., 2009; Cappelen et al., 2017) and social-image concerns (Andreoni and Bernheim, 2009)(Lacetera and Macis, 2010) constitute the second channel on which treatments might matter. In the *Baseline*, the authority has no technology for increasing either concern. By contrast, in the *Private* treatment, she may use justification to induce bad conscience, and thereby to heighten self-image concerns. Hence formally we posit $b_i(\dot{j}_i)$: disutility from image concerns is a function of expected justification j_i given to this participant. In the *Public* treatment, the authority may additionally also make social-image concerns more salient, by signalling out low contributions as a violation of social norms, or as the unfair exploitation of other group members, and by making this assessment known to all members of the current group. More formally we posit $b_i(\dot{J})$: disutility from image concerns is a function of the expected justification J given to all members of the respective group. We therefore expect

$$0 = b_{base} < b_{priv} < b_{pub}$$

If active players are sensitive to image concerns, the experience of having been blamed in the previous period should alert them to the risk of being blamed again in the current period. Hence we expect $c_{i,t}(b_{i,t-1})$. The more powerful the technology for heightening image concerns, the more the authority can substitute words for (costly) action. This gives us

Hypothesis 2 Verbal Punishment: *Active participants contribute more if their choice has been criticised in the previous period. The effect is more pronounced in the* **Public** *than in the* **Private** *treatment.*

Conditional Cooperation If the only non-standard element in (5) is image concerns, $g(.) = e_i - c_i + \mu \sum_{i=1}^{N} c_i$. Such active players trade off profit against self- or social-image. Now a rich experimental literature shows that the majority of a typical experimental population consists of conditional cooperators. They make high contributions c_i themselves, if they know or expect high contributions c_{-i} from other group members as well (Fischbacher et al., 2001; Fischbacher and Gächter, 2010). For the purposes of our project, it does not matter whether $g(\cdot)$ is inequity aversion (Fehr and Schmidt, 1999), a concern for reciprocity (Rabin, 1993), or guilt from violating a norm (Dufwenberg et al., 2011), as long as the weight of the concern for the well-being of other group members is conditional on their own behavior.

The design of the experiment forces active players to define their own contribution c_i before they observe the contributions c_{-i} made by other group members. Active participants must

therefore decide based on their beliefs $\dot{c}_{-i}$ about the choices of others. It has been shown that experimental participants hide selfish behavior behind uncertainty (Dana et al., 2007). Relying on these findings, we posit $\dot{c}_{-i}(\dot{\xi})$: the more group member i expects the choices of other group members to be unpredictable (the higher $\dot{\xi} \in \mathbb{R}^{+}$), the more sceptical her expectations about their choices $\dot{c}_{-i}$. If i expects $-i$ to cooperate only conditionally, her second-order belief $\ddot{c}_i$ also becomes relevant. It depends on the predictability of her own choice. Hence we also posit $\ddot{c}_i(\ddot{\xi})$: second-order beliefs are sensitive to the degree by which other group members $-i$ perceive member i to be predictable.

In the one-shot game, the only signal active players have about choices of other active players are the results from the previous experiment. In the repeated game, active players can additionally use the contributions the remaining members of the group of the last period have made as a proxy for the choices of others they expect in the current period, $\dot{c}_{-i,t}(c_{-i,t-1})$. This signal is more credible the more predictable the punishment policy of the current authority.

These considerations open up a third, indirect channel for treatments to matter. In Hypothesis 1, we have explained in which ways treatments may affect the predictability of authorities' choices. The more punishment is predictable, the more i may trust that $-i$ will comply with the norm $\dot{\tilde{c}}$ she expects the authority to impose. This may help i create more positive (first-order) beliefs about $\dot{c}_{-i}$. Likewise, i has more reason to believe that $-i$ will trust her own loyalty. More formally we posit $\dot{\xi}(\sigma)$ and $\ddot{\xi}(\sigma)$ and expect

$$\dot{\xi}_{base} > \dot{\xi}_{priv} > \dot{\xi}_{pub}$$
and
$$\ddot{\xi}_{base} > \ddot{\xi}_{priv} > \ddot{\xi}_{pub}$$

If the social preferences of group members are strong enough, there is no need for punishment in the first place. Yet it has also been shown that weak (non-deterrent) punishment may supplement insufficiently strong, but positive social preferences (Engel, 2014). Either way, the authority can reduce punishment the more she expects social preferences to be at play. This gives us

Hypothesis 3 Conditional Cooperation: *Active participants contribute more if contributions have been high in the groups to which they have been assigned in previous periods. The effect is least pronounced in the* **Baseline***, more pronounced in the* **Private** *treatment, and most pronounced in the* **Public** *treatment.*

On each of the three channels, treatments induce authorities to reduce (hard) punishment, the more so the more intensely justifications are communicated. Hence we posit

Hypothesis 4 Severity of Punishment: *Authorities punish more severely in the* **Baseline** *than in the* **Private** *treatment than in the* **Public** *treatment.*

5 Results

5.1 Predictability

In Hypothesis 1 a) we had predicted that active players would contribute more the more severely they have been punished themselves in the previous period. In Hypothesis 1 b) we had predicted that they would contribute more the more severely the group has been punished to which they have been attached in the previous period. Yet we had expected that the former effect would be less pronounced as soon as participants learn the justification the authority had given for punishing them individually. We had expected that the latter effect would be less pronounced as soon as participants learn the justifications the authority had given for punishing all group members. We only have partial support for these hypotheses.

Result 1 *In the* **Baseline** *and in the* **Private** *treatment, active players contribute more the more severely the group has been punished in the previous period.*

Support We operationalize punishment severity as

$$\frac{p_i}{20 - c_i}$$

,i.e., as the number of punishment points divided by the difference between the social optimum and the actual contribution.[16] We turn to regression analysis for testing our hypotheses. We first discuss regressions that isolate each channel. The final regression (Model 6 of Table 1) will test all channels together, using the respective other channels as control variables. As Model 1 of Table 1 shows, without further control variables the severity of punishment experienced by an active player herself in the previous period does not explain her contribution choice. Experienced severity does not seem to determine choices. Interactions between the measure for individual severity and treatments are insignificant.[17]

By contrast, active players strongly and positively react to the severity of punishment meted out to all members of their group in the previous period (Model 2). We operationalize severity at the group level by the mean of the measure at the individual level. Yet this effect is only present in the *Baseline* and the *Private* treatment: the interaction with treatment *Public* completely neutralises the main effect of experienced group severity. If active players do not only see punishment choices, but also learn punishment intentions, they no longer react to experienced punishment. This result is in line with justifications making future punishment more predictable.

[16] If $c_i = 20$ and $p_i = 0$, we set the severity index to 0. We drop observations from the analysis where $c_i = 20, p_i > 0$. In these 16 (of 2720) cases (10 in the *Baseline*, 4 in the *Private* and 2 in the *Public* treatments), we have no logical way of expressing severity. What we observe in these rare cases cannot follow from (3).

[17] We again supplement all Tobit regressions with linear mirror models and report differences in footnotes. These additional regressions are available from the authors upon request. With contributions, the assumptions underlying a Tobit model are tenable. Participants who keep their entire endowment might even have wanted to take money from other participants; participants who contribute their entire endowment to the public project might even have wanted to contribute some of their income from earlier periods, or their show-up free, if the design of the experiment had not made this impossible.

If we control for experienced individual severity, the effect of experienced group severity becomes even stronger (Model 3). Now the effect of experienced individual severity even turns significantly negative. What active players care about is not what they have experienced themselves, but what has on average happened with low contributions. Actually, if a participant has been punished severely herself, while others have been punished more leniently for similarly low contributions, she becomes less sensitive to punishment. Active participants care about punishment policies being equitable, and they use information from the previous round as a proxy for this.

5.2 Image Concerns

In Hypothesis 2 a), we had predicted that an active player would contribute more if the authority of the previous period had used the justification requirement to criticise her behavior. We expected the effect to be more pronounced in the *Public* than in the *Private* treatment. We do not have support for this hypothesis.

To quantify the ways in which authorities have used the justification requirement, we had two independent coders rate the statements. We first had them classify statements as made with the intention to govern the group, as opposed to selfish concerns.[18] To test Hypothesis 2 a), we use this rating. We generate a dummy variable that is 1 if a participant has received any punishment points and the authority has given a non-selfish justification.[19] As Model 4 of Table 1 shows, this variable does not explain choices. We even find a negative interaction with the *Public* treatment.[20] This suggests that verbal punishment is not an important channel on which a justification requirement helps govern a group.

5.3 Conditional Cooperation

In Hypothesis 3 a) we had predicted that active players would react to the cooperativeness of others in the previous period. We expected this effect to be most pronounced in the *Public*, less pronounced in the *Private* treatment, and least pronounced in the *Baseline*. We have partial support for this hypothesis.

Result 2 *Active participants contribute more the more other members of their group have contributed in the previous period. The effect is more pronounced in the* **Public** *treatment.*

Support In Model 5 of Table 1, we find a strong significant positive effect of the contributions of other group members in the previous period on contribution choices in the current

[18] For the coding scheme, please see the Appendix.
The fact that authorities are less likely to be selfish in the *Private* treatment stands in contrast to Xiao (2017). She finds that a justification requirement induces participants to be more in line with prevalent normative expectations. The *Public* treatment creates a much bigger audience than the *Private* treatment.

[19] For the development of this variable over time please see Figure A3 in the Appendix.
In the *Baseline*, justifications are not communicated so that this explanatory variable would not be meaningful. This is why Model 4 of Table 1 is confined to the *Private* and *Public* treatments.

[20] We do, however, note that the interaction effect is only weakly significant ($p = .091$) in the linear mirror model.

period. The model predicts that active players contribute approximately one more token of their endowment to the public project if the remaining members in the previous period have, on average, contributed two more tokens. This result is all the more remarkable as we have implemented a stranger design. Experiences from the previous period are therefore only informative at the population level, not at the group level. This socially beneficial effect is more than 50% stronger in the *Public* treatment (significant positive interaction effect), i.e., if justifications are communicated to all group members.[21]

Model 6 analyzes all effects in conjunction with each other. When controlling for all other effects, and in particular for the signal of other participants' cooperativeness, the effect of experience about the severity of punishment at the group level (group severity at $t-1$) is even stronger than in Models 2 and 3. It is extremely strong in the *Private* treatment (the interaction effect almost doubles the main effect). This shows that, in conjunction with information about cooperativeness, information about punishment policies has a very strong effect. Participants learn how strongly cooperation pays, and they learn what may happen if they disregard prevalent normative expectations. As with earlier specifications, past individual punishment has a negative effect. If a participant has been punished severely herself, while others have been punished more leniently for similarly low contributions, this deters cooperation. Participants care about punishment policies being equitable. In this specification, verbal punishment even has a negative effect.[22] Finally, also with all controls information about past cooperativeness still has a strong, significant positive effect. When information about all justifications is made available, this effect is even stronger (positive and significant interaction with the *Public* treatment.[23]).

[21] Over all periods, contributions are not significantly different from 14, which was the approximate level of contributions in the previous experiment. (The additional regression showing this is available from the authors upon request). Informing participants about this history of play may therefore have had an effect. Yet authorities have only very rarely expressed the intention to enforce this norm. Our coders have only found 14 of 1021 instances in the *Baseline* where at least one of them saw a link to history of play. In the *Private* treatment, this held for 31 of 835 instances, and in the *Public* treatment for 60 of 943 instances.

[22] In the interest of being able to analyze data from all treatments, for this specification we also include data from the *Baseline*. In this treatment, we set variable "l.scorn" to 0, as justification is not communicated. We note that the negative effect of this variable is not significant in the linear mirror model.

[23] In the linear mirror model, however, this interaction effect is insignificant. But if we calculate average marginal effects of Model 6 (which are also properly identified in a non-linear model), we find that contributions are significantly higher in the *Public* treatment if the other group members in the previous period had contributed 18 or more tokens on average. We also find a weakly significant effect of treatment *Public* if the average contribution of the remaining group members in the previous period was 16 or 17 tokens.

	Model 1	Model 2	Model 3	Model 4	Model 5	Model 6
Private	-.880	-.713	-.755		.097	-.164
	(2.406)	(2.375)	(2.385)		(1.670)	(1.557)
Public	1.249	1.550	1.545	2.264	-3.438*	-3.086
	(2.406)	(2.374)	(2.384)	(2.646)	(1.723)	(1.621)
Individual severity t-1	.426		-1.402*			-3.893***
	(.613)		(.704)			(.787)
Private * l.indsev	1.130		.621			-.862
	(1.291)		(1.460)			(2.282)
Public * l.indsev	-1.141		.693			.941
	(.955)		(1.023)			(1.315)
Group severity t-1		4.309***	5.356***			7.338***
		(.901)	(1.047)			(1.203)
Private * l.grpsev		.979	.358			7.007**
		(1.593)	(1.821)			(2.591)
Public * l.grpsev		-4.438***	-5.391***			.653
		(1.015)	(1.151)			(2.217)
Blamed t-1				-.131		-1.594**
				(.386)		(.518)
Public * l.blame				-1.428*		-.820
				(.574)		(.795)
Mean contribution -i,t-1					.541***	.595***
					(.045)	(.045)
Private * l.avothcontr					-.041	-.023
					(.066)	(.067)
Public * l.avothcontr					.323***	.316***
					(.071)	(.072)
cons	14.361***	13.983***	14.049***	13.584***	7.032***	5.884***
	(1.665)	(1.643)	(1.650)	(1.872)	(1.157)	(1.076)
N	2704	2704	2704	1760	2704	2704
left censored at 0	187	187	187	124	187	187
right censored at 20	612	612	612	435	612	612
reference category	*Baseline*	*Baseline*	*Baseline*	*Private*	*Baseline*	*Baseline*

Table 1: Determinants of Contributions

dv: contribution
Model 4: data from *Private* and *Public* treatments only
Tobit, left censoring at 0, right censoring at 20
standard errors (for choices nested in 136 individuals nested in 34 matching groups) in parenthesis
*** p $<$ .001, ** p $<$ .01, * p $<$ .05, $^{+}$ p $<$ .1

5.4 Severity of Punishment

Hypothesis 4 predicts that authorities punish more severly in the *Baseline* than in the *Private* treatment than in the *Public* treatment. We only have partial support for this prediction.

Result 3 *In the Private treatment, authorities react less severely to low contributions.*

Support Figure 1 shows that, in the *Private* treatment, the slope of the authorities' reaction functions is considerably flatter, both in the one-shot game (upper panels) and in all data (lower panel). This in particular results from a lower intercept. If active players contribute nothing or little, the reaction is less harsh, compared with either the *Baseline* or the *Public* treatment.

The visual impression is supported by statistical analysis (Table 2). The interaction between the *Private* treatment and contribution is positive and significant, both for the one-shot game, and even more pronouncedly if we consider choices from all periods. The positive coefficient implies that, in this treatment, authorities reduce punishment substantially less if an active player contributes one more unit. By contrast, the interaction between the *Public* treatment is far from significant, and even has opposite sign in the one-shot game. Hence, in the *Public* treatment, authorities react as sensitively to differences in contributions as they do in the *Baseline*.[24]

[24] The significant positive interaction effect implies that punishment is less intensely reduced if the participant contributes one more unit of her endowment to the public project.
One may wonder whether Tobit is the appropriate functional form. It implies that authorities would have wanted to reward active players for particularly high contributions, had the design of the experiment not made this impossible. This may be true for some authorities who do not punish a participant. But for others, the fact that they do not punish will simply follow from the intention to maximize their own profit. A more conservative approach that allows for this possibility is a double hurdle model. It yields qualitatively similar results. The additional regression is available from the authors upon request.
Strictly speaking, the coefficient and the standard errors of the interaction effect cannot be interpreted as the Tobit model is non-linear (Ai and Norton, 2003). Yet we also replicate the effect in a linear mirror model (that ignores left censoring). This additional regression is also available from the authors upon request.

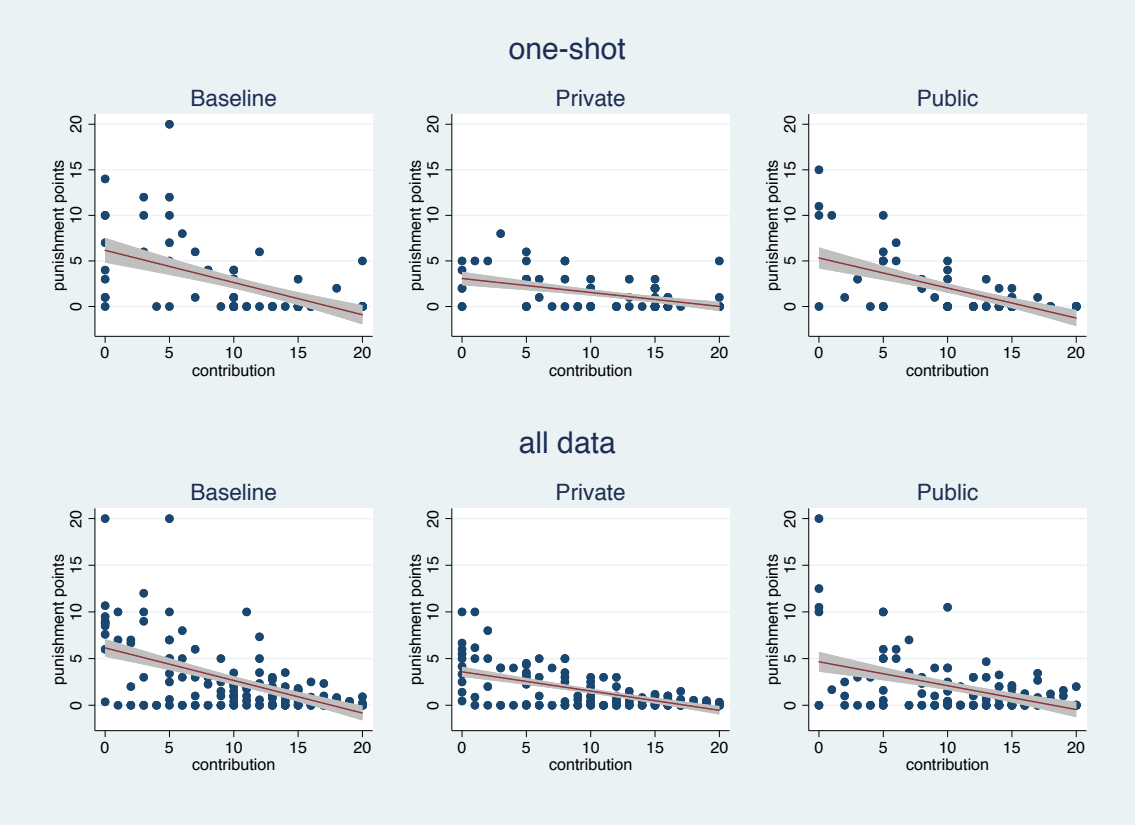


Figure 1: Severity of Punishment

Dots are observations, line is regression line, with 95% confidence interval

	one shot	all data
Private	-4.200	-3.539
	(2.107)	(2.187)
Public	-.459	-1.173
	(2.351)	(2.223)
Contribution	-.777***	-.889***
	(.138)	(.033)
Private * contribution	.362*	.259***
	(.162)	(.044)
Public * contribution	-.044	.063
	(.199)	(.052)
cons	7.464***	7.555***
	(1.645)	(1.531)
N	272	2992
left censored	174	2300

Table 2: Severity of Punishment

dv: amount of punishment points given to an active player
Tobit, left censoring at 0
standard errors (Model 1: for choices nested in 68 authorities, Model 2: for choices nested in 11 periods nested in 68 authorities nested in 34 matching groups) in parenthesis
*** $p < .001$, ** $p < .01$, * $p < .05$, $^{+}$ $p < .1$

5.5 Effectiveness

If they know that justification is communicated individually to its addressee, authorities discriminate less intensely between high and low contributions (Result 3), while active players are more sensitive to experienced severity (Result 1). This suggests a mismatch between the expectations of authorities and of active players. By contrast, if justifications are communicated to all group members, active players react more intensely to experienced cooperativeness (Result 2). This suggests that transparency is effective on this indirect channel.

Figure 2 indeed shows that punishment patterns are similar in the *Private* and *Public* treatments: punishment is relatively high at the beginning, but becomes less and less severe over time. This is different in the *Baseline*, where punishment goes up again by the end of the interaction. However, contribution paths resemble each other in the *Baseline* and in the *Public* treatment. They start at a relatively low level, but stabilize at a high level after some initial periods. This is different in the *Private* treatment. Contributions are lower overall, and they do not stabilize. In the *Baseline* and in the *Public* treatment, for a long time profit closely matches contributions. Yet, in the end, profit decreases in the *Baseline*, while it increases in the *Public* treatment. This difference follows from the fact that, in the *Public* treatment, active players achieve similarly high contributions with much less punishment. While contributions are substantially lower in the *Private* treatment, profit looks more similar to the remaining treatments.

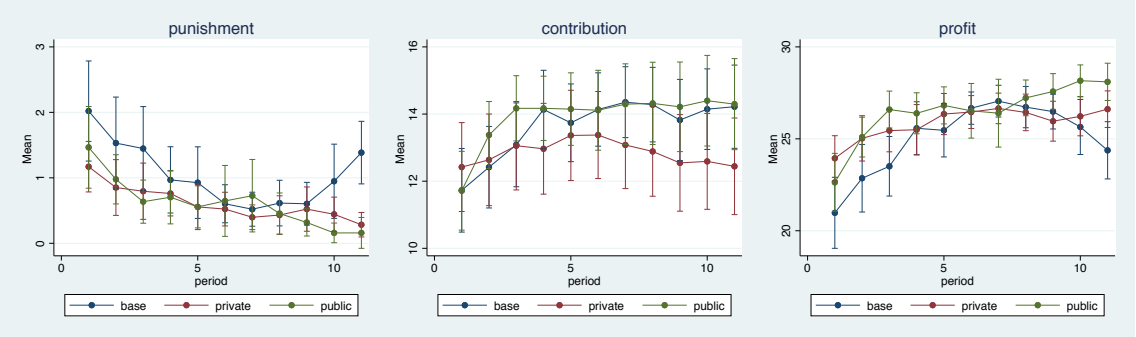


Figure 2: Effectiveness

Error bars from 95% confidence intervals

The visual impression is supported by statistical analysis. In all treatments, the severity of punishment decreases over time (main effect of period in Model 1 of Table 3). But the decay is much more pronounced in the *Public* treatment (interaction between *Public* and period).[25] In the *Baseline*, contributions increase over time (main effect of period in Model 2). This also holds for the *Public* treatment (insignificant interaction between *Public* and period). In the *Private* treatment, by contrast, the interaction effect completely neutralizes the positive main effect of period: contributions do not increase over time. Consequently, punishment without communicating justifications to addressees is effective in increasing contributions. The same holds if justifications are fully transparent. But punishment does not have the socially beneficial effect of increasing contributions if justifications are only communicated to their individual addressees. From a normative perspective, the effect of treatment on the profit active players is even more relevant. We find the same pattern as with contributions. In the *Baseline* (main effect of period in Model 3) and in the *Public* treatment (insignificant interaction between *Public* and period), profit increases over time. Society is on a socially beneficial path. This effect is considerably weaker if justifications are only communicated individually. But even in the *Private* treatment, profit increases over time (Wald test, period + interaction, $p = .0004$). Hence, if an institutional designer cares about welfare, making justifications fully transparent puts society on the most beneficial path. Intervention is as effective as in the *Baseline*. But the beneficial effect is reached at a lower cost.

[25] The interaction effect is insignificant in the linear mirror model. Yet all effects replicate, both with Tobit and in the linear specification, if we add period^2. These additional regressions are available from the authors upon request.

	Model 1 punishment	Model 2 contribution	Model 3 profit
Private	-.133	.888	1.884^{+}
	(1.392)	(2.300)	(1.132)
Public	.474	.675	1.316
	(1.404)	(2.299)	(1.132)
Period	-.247***	.233***	.372***
	(.070)	(.048)	(.054)
Private * period	-.108	-.270***	-.173*
	(.105)	(.070)	(.077)
Public * period	-.394***	.039	.026
	(.113)	(.070)	(.077)
cons	-2.876**	12.906***	22.795***
	(.980)	(1.590)	(.783)
N	2992	2992	2992
left censored at 0	2300	206	
right censored at 20		677	

Table 3: Effectiveness

dv: Model 1: amount of punishment points given to an active player
Model 2: contributions; Model 3: profit of active players
Model 1 and 2: Tobit, Model 3: linear
Model 1: left censoring at 0
Model 2: left censoring at 0, right censoring at 20
standard errors in parenthesis
Model 1: SE for choices nested in 11 periods nested in 68 authorities nested in 34 matching groups
Model 2 and 3: SE for choices nested in 136 individuals nested in 34 matching groups
*** $p < .001$, ** $p < .01$, * $p < .05$, $^{+}$ $p < .1$

6 Conclusions

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 players' reactions. Authorities seem to believe erroneously that private communication suffices for the substitution effect.

Our experiment has not been designed to test new behavioral theory. But we have based our hypotheses on (formal) theory. The theory predicts in which ways alternative specifications of a justification requirement might matter for the choices made by authorities and by active players. We have partial support for these theoretical expectations. We had expected that the experience of punishment would induce individual and vicarious learning. We do find an effect of vicarious learning: if the authority of the previous period has reacted more intensely to low contributions, active players contribute more in the subsequent period. We do not find a similar effect for the experience of having been punished oneself. This suggests that active players care about information regarding the punishment policy of an authority, not so much about the effect of this policy on themselves. We had further expected that the learning effect would be less pronounced the better the information about justifications: experience becomes less important for correctly predicting punishment policies. This is indeed what we find, but only for the *Public* treatment. Against expectations, we do not find a disciplining effect of verbal punishment. We finally had predicted that justification is effective on an indirect channel: learning justifications could help conditional cooperators make better predictions about the choices of other group members, and could help them generate better second-order predictions about the expectations other group members hold about their own choices. This effect should be the more pronounced the better the information about justifications. We find support for this claim, but the treatment effect is again confined to the *Public* treatment. We finally had expected that authorities would anticipate the effect of justification on contributions, and reduce the severity of monetary punishment accordingly. This expectation is only borne out in the *Private* treatment.

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 causal effects and explaining them. In the interest of achieving identification, 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 personally known. 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. 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 stranger matching, we come closer to the characteristic situation in courts; the typical judge does not know the defendant beforehand, and is unlikely to meet her again. 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.

References

Chunrong Ai and Edward C. Norton. Interaction terms in logit and probit models. *Economics Letters*, 80:123–129, 2003.

George A. Akerlof and Rachel E. Kranton. Economics and identity. *Quarterly Journal of Economics*, 115:715–753, 2000.

James Andreoni and B. Douglas Bernheim. Social image and the 50-50 norm: A theoretical and experimental analysis of audience effects. *Econometrica*, 77:1607–1636, 2009.

Giulia Andrighetto, Jordi Brandts, Rosaria Conte, Jordi Sabater-Mir, Hector Solaz, Áron Székely, and Daniel Villatoro. Counter-punishment, communication, and cooperation among partners. *Frontiers in Behavioral Neuroscience*, 10, 2016.

Dan Ariely, Anat Bracha, and Stephan Meier. Doing good or doing well? Image motivation and monetary incentives in behaving prosocially. *American Economic Review*, 99:544–555, 2009. ISSN 0002-8282.

Albert Bandura. *Social Learning Theory*. Prenctice Hall, Englewood Cliffs, 1977.

Jeremy Bentham. *The Rationale of Punishment*. R. Heward, London,, 1830. By Jeremy Bentham. 22 cm.

Michael Berlemann, Marcus Dittrich, and Gunther Markwardt. The value of non-binding announcements in public goods experiments: Some theory and experimental evidence. *Journal of Socio-Economics*, 38(3):421–428, 2009. ISSN 1053-5357.

Andreas Blume and Andreas Ortmann. The effects of costless pre-play communication: Experimental evidence from games with pareto-ranked equilibria. *Journal of Economic Theory*, 132:274–290, 2007.

Oliver Bochet, Talbot Page, and Louis Putterman. Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization*, 60: 11–26, 2006.

Roland Bénabou and Jean Tirole. Identity, morals, and taboos: Beliefs as assets. *Quarterly Journal of Economics*, 126(2):805–855, 2011. ISSN 0033-5533.

Alexander W Cappelen, Trond Halvorsen, Erik Ø Sørensen, and Bertil Tungodden. Face-saving or fair-minded: What motivates moral behavior? *Journal of the European Economic Association*, 15(3):540–557, 2017.

Gary Charness. Self-serving cheap talk: A test of Aumann's conjecture. *Games and Economic Behavior*, 33:177–194, 2000.

Ananish Chaudhuri. Sustaining cooperation in laboratory public goods experiments: A selective survey of the literature. *Experimental Economics*, 14(1):47–83, 2011. ISSN 1386-4157.

Vincent Crawford. A survey of experiments on communication via cheap talk. *Journal of Economic Theory*, 78(2):286–298, 1998. Journal of Economic Theory.

Rachel T.A. Croson and Melanie Marks. The effect of recommended contributions in the voluntary provision of public goods. *Economic Inquiry*, 39:238–249, 2001.

Jason Dana, Roberto A Weber, and Jason Xi Kuang. Exploiting moral wiggle room: Experiments demonstrating an illusory preference for fairness. *Economic Theory*, 33(1):67–80, 2007. ISSN 0938-2259.

Martin Dufwenberg, Simon Gächter, and Heike Hennig-Schmidt. The framing of games and the psychology of play. *Games and Economic Behavior*, 73:459–478, 2011. ISSN 0899-8256.

Christoph Engel. *The Psychological Case for Obliging Judges to Write Reasons*, pages 71–109. Nomos, Baden-Baden, 2007.

Christoph Engel. Social preferences can make imperfect sanctions work: Evidence from a public good experiment. *Journal of Economic Behavior & Organization*, 108:343–353, 2014.

Christoph Engel and Lilia Zhurakhovska. You are in charge: Experimentally testing the motivating power of holding a judicial office. *Journal of Legal Studies*, 46:1–50, 2017.

Ernst Fehr and Simon Gächter. Cooperation and punishment in public goods experiments. *American Economic Review*, 90:980–994, 2000.

Ernst Fehr and Klaus M. Schmidt. A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114:817–868, 1999.

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

Urs Fischbacher and Simon Gächter. Social preferences, beliefs, and the dynamics of free riding in public good experiments. *American Economic Review*, 100:541–556, 2010.

Urs Fischbacher, Simon Gächter, and Ernst Fehr. Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters*, 71:397–404, 2001.

Ben Greiner. Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, 1:1–12, 2015. ISSN 2199-6776.

Benedikt Herrmann, Christian Thöni, and Simon Gächter. Antisocial punishment across societies. *Science*, 319:1362–1367, 2008.

Charles A. Holt and Susan K. Laury. Risk aversion and incentive effects. *American Economic Review*, 92:1644–1655, 2002.

Nicola Lacetera and Mario Macis. Social image concerns and prosocial behavior: Field evidence from a nonlinear incentive scheme. *Journal of Economic Behavior & Organization*, 76(2):225–237, 2010.

John O. Ledyard. *Public Goods: A Survey of Experimental Research*, pages 111–194. Princeton University Press, Princeton, NJ, 1995.

Jennifer S. Lerner and Philip E. Tetlock. Accounting for the effects of accountability. *Psychological Bulletin*, 125:255–275, 1999.

Wim B. Liebrand and Charles G. McClintock. The ring measure of social values: A computerized procedure for assessing individual differences in information processing and social

value orientation. *European Journal of Personality*, 2:217–230, 1988.

David Masclet, Charles Noussair, Steven Tucker, and Marie-Claire Villeval. Monetary and non-monetary punishment in the voluntary contributions mechanism. *American Economic Review*, 93:366–380, 2003.

David Masclet, Charles Noussair, and Marie-Claire Villeval. Threat and punishment in public good experiments. *Economic Inquiry*, 51:1421–1441, 2013.

John W. McCormac. Reason comes before decision. *Ohio State Law Journal*, 55:161–166, 1994.

Maria Montero, Martin Sefton, and Ping Zhang. Enlargement and the balance of power: An experimental study. *Social Choice & Welfare*, 30:69–87, 2008.

Daniele Nosenzo and Martin Sefton. *Promoting Cooperation: The Distribution of Reward and Punishment Power*, pages 87–114. Oxford University Press, Oxford, 2014.

Matthew Rabin. Incorporating fairness into game theory and economics. *American Economic Review*, 83:1281–1302, 1993.

David Sally. Conversation and cooperation in social dilemmas: A meta-analysis of experiments from 1958 to 1992. *Rationality and Society*, 7(1):58–92, 1995.

Frederick Schauer. Giving reasons. *Stanford Law Review*, 47:633–659, 1995.

Mark Seidenfeld. Cognitive loafing, social conformity, and judicial review of agency rulemaking. *Cornell Law Review*, 87:486–548, 2002.

Philip E. Tetlock. Accountability and complexity of thought. *Journal of Personality and Social Psychology*, 45:74–83, 1983.

Erte Xiao. Justification and conformity. *Journal of Economic Behavior & Organization*, 136: 15–28, 2017.

Erte Xiao and Daniel Houser. Punish in public. *Journal of Public Economics*, 95(7):1006–1017, 2011.

Erte Xiao and Fangfang Tan. Justification and legitimate punishment. *Journal of Institutional and Theoretical Economics*, 170:168–188, 2014.

Jennifer Zelmer. Linear public goods: A meta-analysis. *Experimental Economics*, 6:299–310, 2003.

Appendix

A1 Instructions

In order to highlight differences between the instructions in the respective treatments, we present all instructions in one file here. Highlights in different colors indicate the treatment differences. The differences in the instructions were not highlighted in the original instructions. Instructions that are not highlighted are seen by participants in all treatments, including the *Baseline*. Light blue parts of the instructions are only presented in the Baseline. Darker blue parts of the instructions are only presented to subjects in the *Private* treatment. The darkest blue highlights instructions that are specific to the *Public* treatment. Note that 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.

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. Player B's income 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.

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.

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.

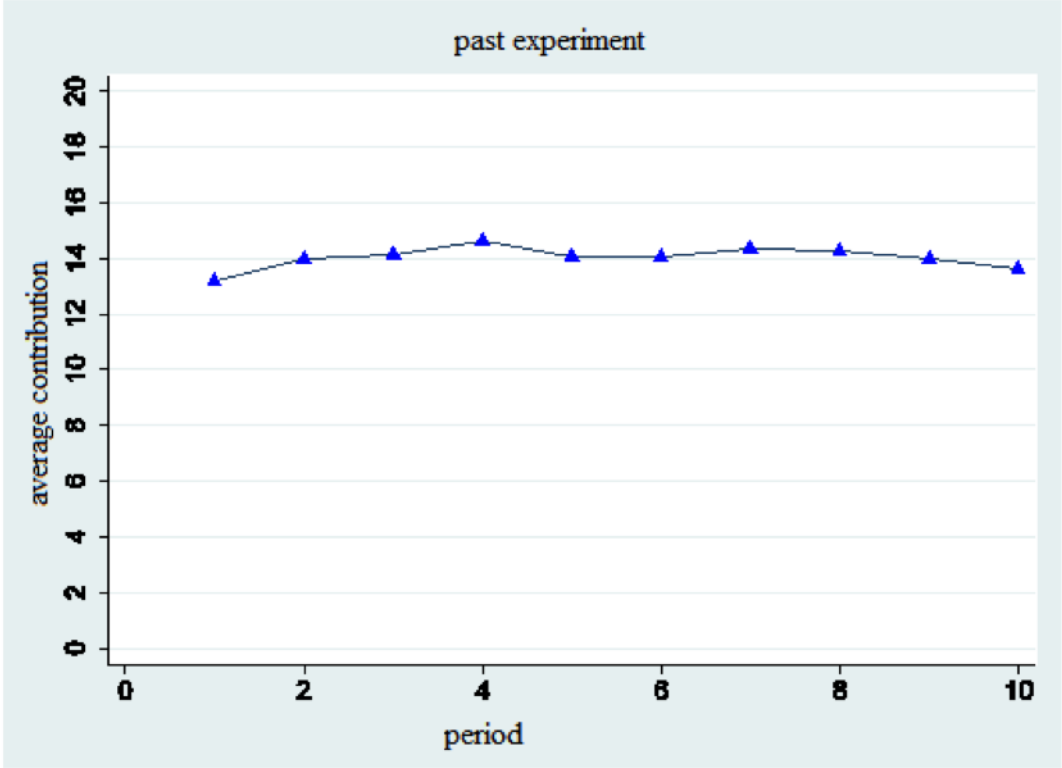
In addition, you will be told player B's reason for distributing whatever amount of points you got. This information goes only to you. The other players do not know this reason. They are only aware of the reasons they have been given for their own allocation of points.

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.

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.

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.

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 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 € with a chance of 10%, or 1.60 € with a chance of 90%	3.85 € with a chance of 10%, or 0.10 € 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.

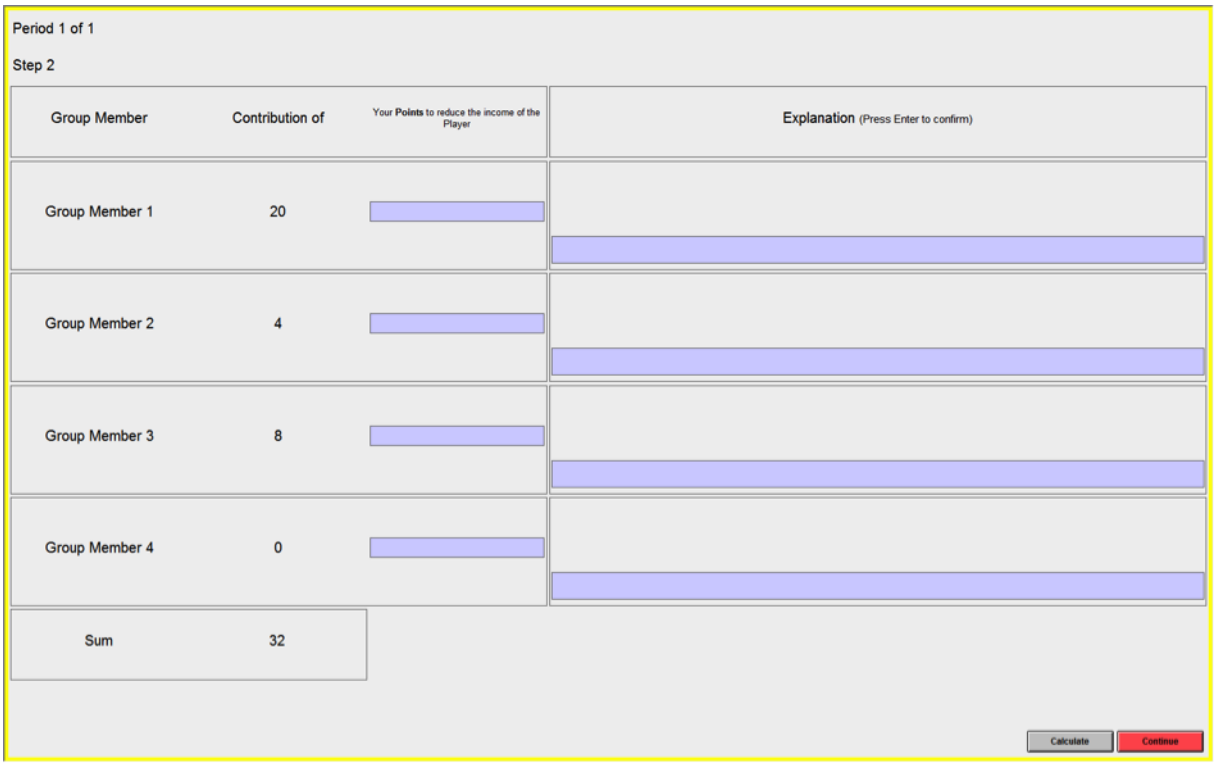


Figure A1: Decision Screen for Justifications

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 player cannot exceed 20. She may use the "Calculate" button to calculate the sum of points she would assign by 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.

A2 Supplementary Data Analysis

A2.1 Coding of Reasons

The following coding scheme has been given to the two independent raters.

"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:

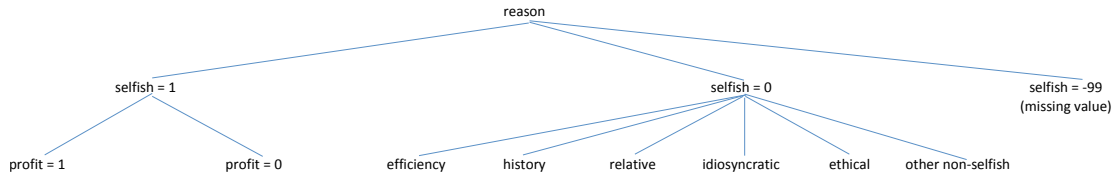


Figure A2: Coding Scheme

We have had two independent raters rate the explicit reasons along the eight dimensions listed in Figure A3. If a justification has been classified as "not selfish" it can contain aspects that fall under the subcategories "efficiency", "history", "relative", "idiosyncratic", "ethical", and/or "unspecified". The suboptions are not mutually exclusive. Inter-rater reliability is good (mean Cohen's kappa is .766). 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 is .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. Note that we had no ex-ante hypothesis on explicit reasons. Nonetheless, we are convinced that it is worth providing this information to the reader.

As Table A1 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 the *Private* treatment. The difference between the *Private* treatment and the *Baseline* is significant for both dimensions. In the *Baseline*, the profit motive is also significantly more pronounced than in treatment *Private*. In the *Public* treatment, authorities seem 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.

	selfish	profit	efficiency	history	relative	idio-syncratic	ethical	unspecified
Baseline	27.95	24.24	25.22	.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	$<.001$

Table A1: Explicit Reasons

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

The following are a few illustrations of the reasons the experimental authorities have given for their choices (translated into English):

- "Approximately the average. Fair enough."
- "Exceptionally high contribution"
- "You have not contributed enough to the project"
- "You have invested three quarter of your endowment. This I consider fair."
- "You have only invested half of your endowment. This is not enough".
- "For the most greedy person".
- "Others have given even more. They shall not have a disadvantage".
- "Still free riding, after so many rounds".

- "Good group member".
- "By meting out points, I reduce my own profit. I do not care how much others earn".

A2.2 Supplementary Data Analysis

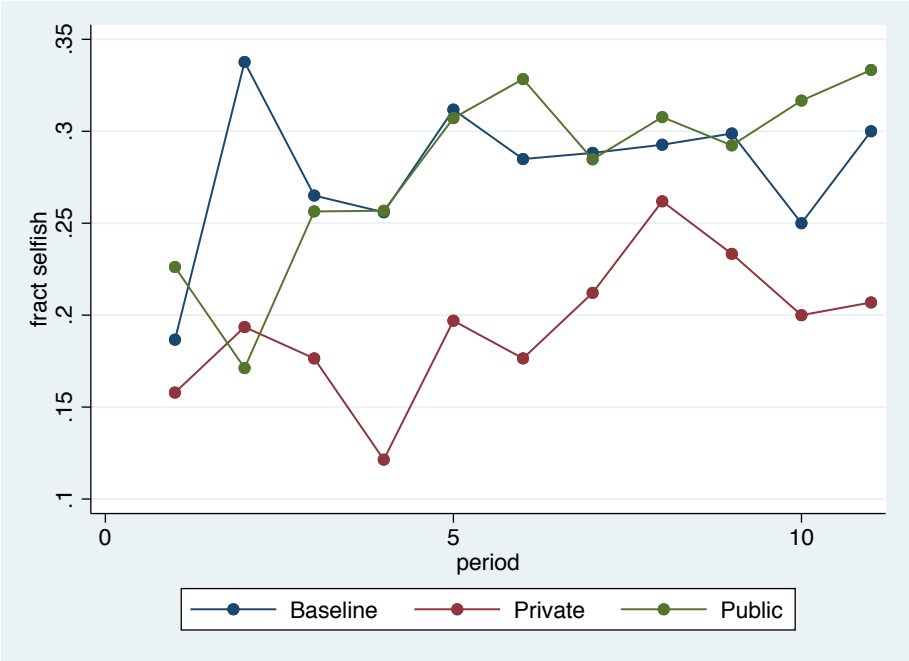


Figure A3: Selfish Reasons

Frequency of reasons that have been classified as selfish over time, per treatment.

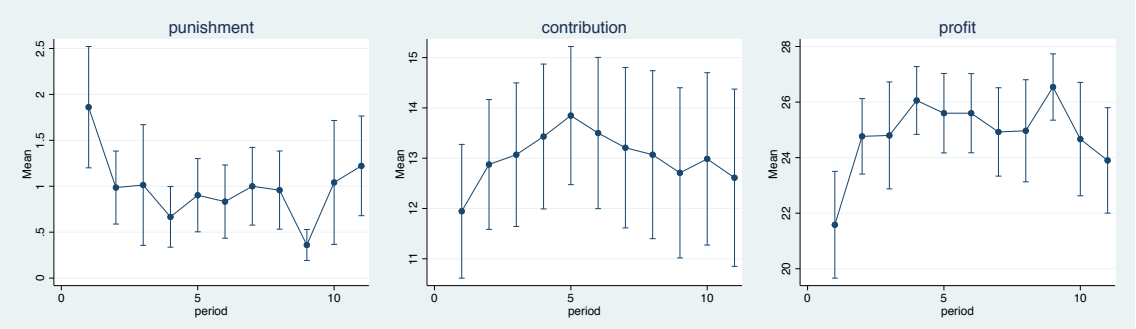


Figure A4: Treatment Without Justification Requirement

For comparison see Figure 2