**Preprints of the
Max Planck Institute for
Research on Collective Goods
Bonn 2011/30**

The Joint Benefits of Observed and Unobserved Punishment: Comment to Unobserved Punishment Supports Cooperation

Andreas Glöckner
Sebastian Kube
Andreas Nicklisch

Preprints of the
Max Planck Institute
for Research on Collective Goods

Bonn 2011/30

The Joint Benefits of Observed and Unobserved Punishment: Comment to Unobserved Punishment Supports Cooperation

Andreas Glöckner / Sebastian Kube / Andreas Nicklisch

November 2011

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

The Joint Benefits of Observed and Unobserved Punishment: Comment to Unobserved Punishment Supports Cooperation[¶]

Andreas Glöckner[*] / Sebastian Kube[†] / Andreas Nicklisch[‡]

November 18, 2011

Abstract

Laboratory experiments by Fudenberg and Pathak (2010), and Vyrastekova, Funaki and Takeuch (2008) show that punishment is able to sustain cooperation in groups even when it is observed only in the end of the interaction sequence. Our results demonstrate that the real power of unobserved punishment is unleashed when combined with observable punishment. Providing both unobserved and observed punishment strongly enhances cooperation within groups – strikingly, even with less intense sanctioning. This surprising result underlines the importance of the co-existence of observed and unobserved sanctioning mechanisms in social dilemmas.

Keywords: Unobserved Punishment, Public Goods, Sanctioning Effectiveness

JEL: C92, H40, H41

[¶] We are grateful for comments and suggestions made by Marie-Claire Villeval, Christian Thöni, Bernd Irlenbusch, and Simon Gächter. We also thank Benedikt Herrmann, Christian Thöni and Simon Gächter for providing us their data for the regular VCM.

[*] Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53113 Bonn, Germany. Email: gloeckner@coll.mpg.de.

[†] Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53113 Bonn, Germany. Department of Economics, University of Bonn, Adenauerallee 24-42, 53113 Bonn, Germany. Email: kube@coll.mpg.de.

[‡] Corresponding author. Max Planck Institute for Research on Collective Goods, Kurt-Schumacher-Str. 10, 53113 Bonn, Germany. School of Business, Economics and Social Science, University of Hamburg, von-Melle-Park 5, 20146 Hamburg, Germany. Email: nicklisch@coll.mpg.de.

1. Introduction

A large number of studies show that social sanctions are a successful mechanism to attain cooperation in social dilemmas. Generally, there are higher cooperation rates in groups whose members observe others’ behavior and have the opportunity to punish non-cooperation than in groups without punishment opportunity. This is even true if persons interact anonymously for a finite number of times and if punishment is costly for the punisher (e.g., Fehr, Gächter, 2000, 2002, Ostrom, 1990, Gürerk et al., 2006, Herrmann et al., 2008). In two recent articles, Drew Fudenberg and Parag Pathak (2010, hereafter FP) and Jana Vyrastekova, Yukihiko Funaki, and Ai Takeuch (2008, hereafter VFT) provide experimental evidence that subjects are willing to engage in costly punishment even if sanctioned persons are informed about the punishment only after the very last interaction (hereafter, we talk about *unobserved punishment*). Furthermore, they find that cooperation rates under *unobserved* punishment does not differ significantly from cooperation rates under *observed* punishment (i.e., when sanctioned persons are informed immediately that they are punished). As such, their results provide important insights into the motivation to punish non-cooperators. Given that players do still punish even if any “repeated play” explanations for punishment are ruled out by design, the results suggest a “pure preference”-explanation of costly punishment (cp. Quervain et al., 2004). That is, players seem to punish because they receive positive utility from carrying out punishment per se, rather than punishing strategically to increase one’s own monetary payoff in future periods.

The results of our experiment indicate that the real power of unobserved punishment is unleashed when combined with observed punishment. If subjects can use unobserved and observed punishment mechanisms at the same time, cooperation rates are enormously enhanced – and, moreover, with less intense sanctioning. This striking finding corresponds nicely to the coexistence of unobserved and observed punishment in real world social interactions: e.g., neighbors spread rumors (which later ruin the defector's reputation) while they immediately litter rubbish in the defector’s backyard; colleagues start workplace bullying and do not pass on crucial information as a reaction to peers' free-riding; etc. Moreover, our study provides a starting point for exploring the interaction between unobserved punishment and other mechanisms that potentially enhance cooperation in social dilemmas.

2. Experimental Design

To shed light on the interaction between observed and unobserved punishment, we conducted a series of laboratory experiments. Participants are anonymously matched in groups of four and play ten consecutive periods of the voluntary contribution game with punishment options.[1] Each

[1] Participants know that the experiment terminates after ten periods; the composition of the group remains constant throughout the entire 10 periods of the experiment, however, to prevent subjects from identifying each other across periods, they receive a random identification number between 1 and 4 at the beginning of each period.

period consists of two stages. In stage one, players are endowed with 20 tokens each and simultaneously decide how many of the tokens to transfer to a group account. The sum of contributions to the group account are multiplied by 0.4 and provided to each group member. Hence, each player profits equally from the group account, independent of his or her contribution. The total sum of income within the group is maximized if all group members contribute fully, but given any combination of other players' contributions, each player could increase his individual payoff by withholding his own contribution; so that we can interpret the contribution as the player's cooperation rate. In stage two, after having observed the contribution decisions of all group members, each player can assign up to ten sanctioning points to any other group member. Each point costs one token for the player who assigns it, while reducing the earnings of the player who receives the point by three tokens. Finally, all players are informed about their own income in this period, respectively about their total earnings after the last period. Assigning punishment points constitutes a second-order public good: all group members would jointly benefit from disciplining non-cooperators, but given that punishment is costly, each player has an incentive to free-ride on others' sanctions. (Not only) economists have frequently pointed out that selfish players should not be expected to punish in the finitely repeated version of this game. Consequently, players anticipating this should also be reluctant to contribute to the initial public good in stage one for the same reasons. The subgame perfect Nash equilibrium under the assumption of self-centered, money-maximizing preferences is thus i) no sanctions on the second stage, and ii) only minimum contributions in the first stage.

Our first treatment implements a regular sanctioning mechanism with *observed* punishment, that is, at the end of each period each subject receives immediate feedback on the amount of sanctions that he received (treatment *O* in the following) (cp., Herrmann et al., 2008). The second treatment replicates FT and VFP, implementing a sanctioning mechanism with *unobserved* punishment. Received punishment points are not immediately revealed, but are accumulated over all periods. Only after the final period of the experiment, subjects get to learn the accrued points and corresponding sanctions are deducted (treatment *U*). Finally, the third treatment is novel as it features *both* mechanisms at the same time (treatment *O+U*): players can choose in each period how many observable and how many unobservable sanctioning points they want to carry out.

In total, we ran 9 sessions with 23 groups (92 subjects), resulting in 8 independent group observations each in the *O* and *O+U* treatments, and 7 independent group observations in the *U* treatment.[2] The computerized experiments were conducted in Bonn in 2008.[3] For comparison, we also include the data of a regular voluntary contribution mechanism experiment without any punishment (treatment *VCM*).[4]

[2] Supporting materials and methods are reported in the Appendix.

[3] We used zTree (Fischbacher, 2007) for the experiments, and ORSEE (Greiner, 2004) for the recruitment.

[4] Data for the *VCM* was provided by Herrmann et al. (2008), who ran the *VCM* in the same laboratory using exactly the same set of parameters.

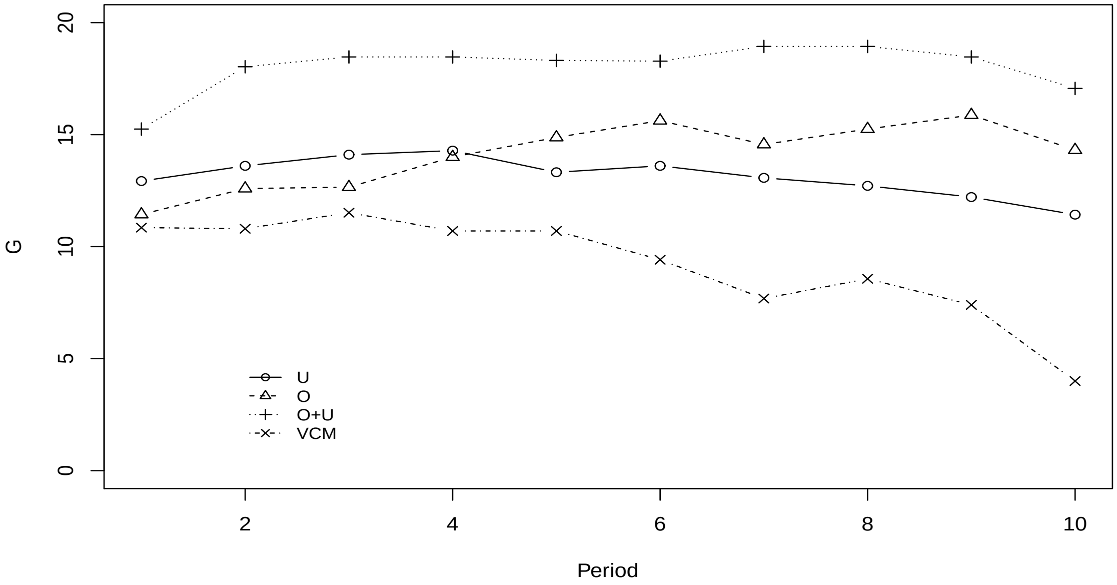


Fig 1. Average contributions over periods and treatment conditions

3. Results

Figure 1 illustrates the average contributions over time for the treatment conditions. While contributions in the absence of punishment exhibit the usual decline over time, the three sanctioning mechanisms foster cooperation. Average distributions in all three punishment conditions are higher than in *VCM*.[5] Contributions in *O* rise over time and are maintained almost over the entire course of the experiment. A similar (though less distinct) effect is observed in *U*. There is no significant difference between average contribution levels in treatments *O* and *U*.[6] Even if we focus on the first five or on the last five periods, there are no significant differences.[7] Strikingly, comparisons between *O+U* and *O* as well as between *O+U* and *U* reveal significant differences, economical as well as statistical. Contributions in the *O+U* treatment are higher than in the other treatments throughout the entire experiment.[8] Interestingly, contribution levels are already higher from the outset, which suggests that subjects (correctly) anticipate that the combination of both mechanisms is an extremely effective disciplining device.

[5] We use exact two-sided rank-sum tests, here and in the following with group averages over all ten periods as independent observations; comparing *VCM* to *O+U, O,* and *U*: $p<0.001$, $p=0.03$, and $p=0.07$.

[6] Exact rank-sum test, two-sided; $p=0.69$.

[7] Exact rank-sum tests, two-sided; $p=0.87$, and $p=0.23$.

[8] Exact rank-sum tests, two-sided; overall $p=0.005$ and $p=0.015$, first period $p=0.006$ and $p=0.087$.

We observe no significant differences between the number of sanctioning points distributed in *O* (on average 0.43 points) and *U* (on average 0.5).[9] In *O+U*, where both types of points are available, subjects assign on average 0.13 observable points and 0.09 unobservable points. Observable and unobservable points seem to be substitutes for one another, rather than a complementary means, because only 16% of all punishment decisions in *O+U* use both type of points simultaneously.

Interestingly, our results suggest that although the cooperation is strikingly higher in *O+U*, this increased cooperation is associated with less intense sanctioning. Specifically, Probit regressions indicate a significantly lower probability that players assign either observed or unobserved sanctioning points (irrespective of the number of points). Furthermore, Tobit regression results show that the number of both types of points assigned decreases significantly in *O+U* compared to the other treatment conditions, even when controlling for the differences in contributions between the treatments.[10] This finding is also reflected in the sanctioning effectiveness of observed sanctioning points. Sanctioning effectiveness is defined as the change in players' contribution (between the period where they were punished and the subsequent period) per observed sanctioning point. Average sanctioning effectiveness in conditions *O* and *O+U* are shown in Figure 2. We find an average sanctioning effectiveness of 0.67 in the *O* condition, i.e., the sanctioned player increases his or her average contribution by 0.67 tokens in the subsequent period. In contrast, we find a significantly higher average sanctioning effectiveness of 2.12 in the *O+U* condition.[11] The effect of punishment on contributions is more than tripled when observed sanctions are accompanied by (the fear of) unobserved sanctions, making punishment highly productive in the *O+U* condition compared to the *U* condition.

Finally, Figure 3 shows the development of efficiency – defined as players' average monetary payoff – over time. On average, players earn 28.1 tokens (out of a maximum of 32) in *O+U*, 21.8 in *U*, 23.3 in *O*, and 25.5 in *VCM*. Thus, average efficiency is highest in the *O+U* condition,[12] while, compared to treatment *VCM*, both sanctioning mechanisms in isolation do not lead to better efficiency rates.[13] Still, the mere fact that both sanctioning mechanisms are jointly available tremendously increases the efficiency of group cooperation within only a few periods – and furthermore does so without the substantial short-run efficiency losses due to punishment. In our view, this ultimately underlines the benefits of a combination of observed and unobserved punishment in social dilemmas.

[9] Exact rank-sum test, two-sided; p=0.61.

[10] See the supplementary material in the Appendix for regression details.

[11] Exact rank-sum test, two-sided; p=0.04.

[12] Exact rank-sum tests, two-sided, comparing *O+U* to *O*, *U* and *VCM*: p=0.02, p=0.01, and p=0.03.

[13] Exact rank-sum tests, two-sided, comparing *VCM* to *O* and *U*: p=0.27 and p=0.12; we might expect, however, the sanctioning mechanisms to enhance efficiency if the number of periods were sufficiently large, see for this Gächter et al., 2008.

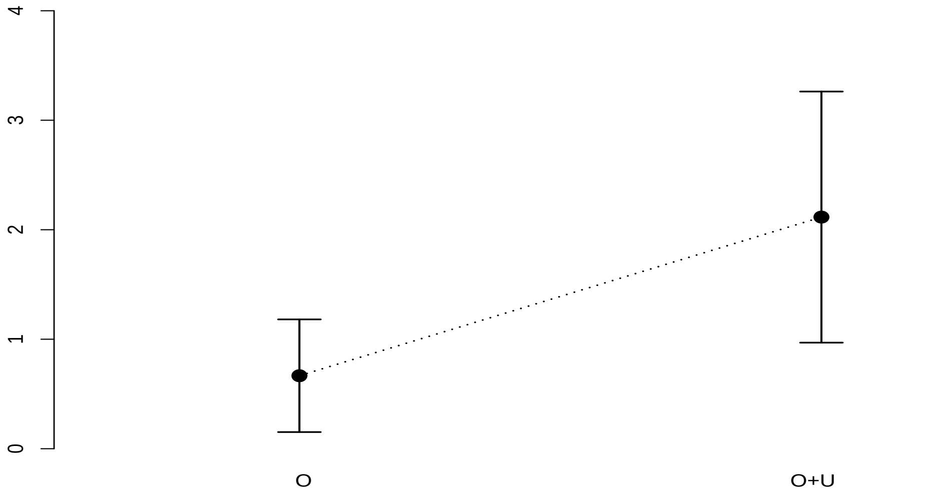


Fig. 2. Average sanctioning effectiveness of immediate sanctioning points
Note: *Bars indicate the 95% confidence intervals*

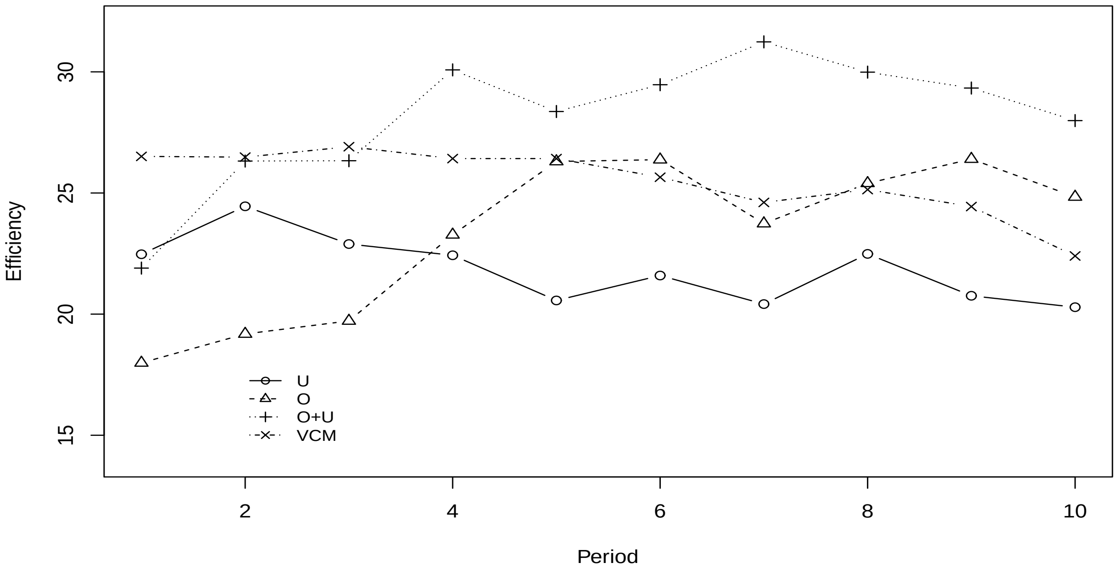


Fig. 3. Average efficiency over periods and treatment conditions

4. Discussion

The laboratory experiments by Fudenberg and Pathak as well as Vyrastekova, Funaki and Takeuchi show that the availability of unobserved punishment leads to stable cooperation in social dilemmas, however, like in the case of observable punishment, at the cost of efficiency losses in the short run. Our findings demonstrate that the co-existence of both potential threats leads to higher cooperation and yields tremendous efficiency gains. The mere existence of unobserved punishment more than triples the sanctioning effectiveness of observed punishment, while overall sanctioning expenditures are significantly reduced.

We interpret our results such that observed sanctions serve as warning signs of harsh unobserved consequences. One might hypothesize that this mechanism relies on some kind of deeply rooted human experience to avoid provoking latent payback (e.g., a later passionate burst of anger). In this spirit, our experiment demonstrates the beneficial effects of the co-existence of observed and unobserved sanctioning mechanisms in social dilemmas. Due to its appealing simplicity and practicability, one should consider unobserved punishment as an important complement that efficiently stabilizes cooperation in societies.

References

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

Fehr, E. & S. Gächter (2002), Altruistic punishment in humans, Nature 415, 137–140.

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

Fudenberg, D. & P. Pathak (2010), Unobserved Punishment Supports Cooperation, Journal of Public Economics, 94, 78-86.

Gächter, S., E. Renner, & M. Sefton (2008), The long-run benefits of punishment, Science 322, 1510.

Greiner, B. (2004), An online recruitment system for economic experiments, in: Kremer, K. & Macho, V. (eds.), Forschung und wissenschaftliches Rechnen 2003, Bericht der Gesellschaft für wissenschaftlichen Dateverarbeitung Göttingen 63, 79–93.

Herrmann, B., C. Thöni & S. Gächter (2008), Antisocial punishment across societies, Science 319, 1362–1367.

Gürerk, Ö., B. Irlenbusch & B. Rockenbach (2006), The competitive advantage of sanctioning institutions. Science, 312, 108–111.

Ostrom, E. (1990), Governing the commons, Cambridge University Press, Cambridge.

de Quervain, D., U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck and E. Fehr (2004), The Neural Basis of Altruistic Punishment. Science, 305, 1254-1258.

Vyrastekova, J., Y. Funaki & A. Takeuchi (2008), Strategic versus Non-Strategic Motivations of Sanctioning, Tilburg University Discussion Paper.

Appendix

Appendix A provides a detailed description of the experiment and the instructions. We report a detailed analysis of sanctions in Appendix B.

Appendix A: Experimental method

At the beginning of each session, participants had to draw lots, in order to assign each of them to a cubicle, where we asked them to take their seats immediately. Once all subjects were seated, instructions were distributed and read out aloud. Afterwards, participants could pose clarifying questions to the experiment supervisor in private. Participants then had to answer a set of control questions to ensure that everybody had understood the game.[14] Control questions were corrected individually, and wrong answers were explained privately. Participants were randomly and anonymously matched in groups of four players each. The composition of the group remained constant throughout the entire 10 periods of the experiment. To prevent subjects from identifying each other across periods, each received a random identification number between 1 and 4 at the beginning of each period.

Altogether, 92 subjects, mostly students from the University of Bonn majoring in various fields, participated in the experiment (10 percent were non-students). Mean age was 24.5 years (standard deviation 5.5 years), 62 percent were females. Each subject participated only once in the experiment, that is, there were different subjects in each treatment. None of the subjects had participated in a public-good experiment before. A session lasted for about 60 minutes. Tokens earned were accrued over all periods and converted at an exchange rate of 3 Euro per 100 tokens. Participants were paid out individually to ensure their anonymity. They earned on average 13.86 Euro (standard deviation 1.45 Euro), including a show-up fee of 5 Euro.[15]

English translation of the German instructions for the *O+U* condition[16]

General explanations for participants

You are taking part in an economic science experiment. If you read the following explanations closely, you can earn a rather significant sum of money, depending on the decisions you make. It is therefore very important that you pay attention to the following points.

[14] Questions are almost identical to the control questions of Herrmann et al. (2008).

[15] 13.86 Euro corresponds to $17.60 (as of November 2008). Notice that since actual period payoffs could be negative due to costs for deduction points or the punishment of deduction points (which rarely occurred), all players received an additional endowment of 50 tokens at the beginning of the experiment. No player accrued an overall negative payoff at the end of the experiment.

[16] Instructions for the *O*, resp. for the *U* condition, were identical except for the omitted parts referring to immediate, resp. mediate punishment points. Screens differed accordingly.

The instructions you have received from us are intended solely for your private information. During the experiment, you will not be allowed to communicate with anyone. Should you have any questions, please direct them directly to us. Not abiding by this rule will lead to exclusion from the experiment and from any payments.

In this experiment, we calculate in Taler, rather than in Euro. Your entire income will therefore initially be calculated in Taler. The total sum of Taler will later be exchanged into Euro as follows:

1 Taler = 3 Euro cent

The accumulated amount will be paid to you in cash at the end of the experiment.

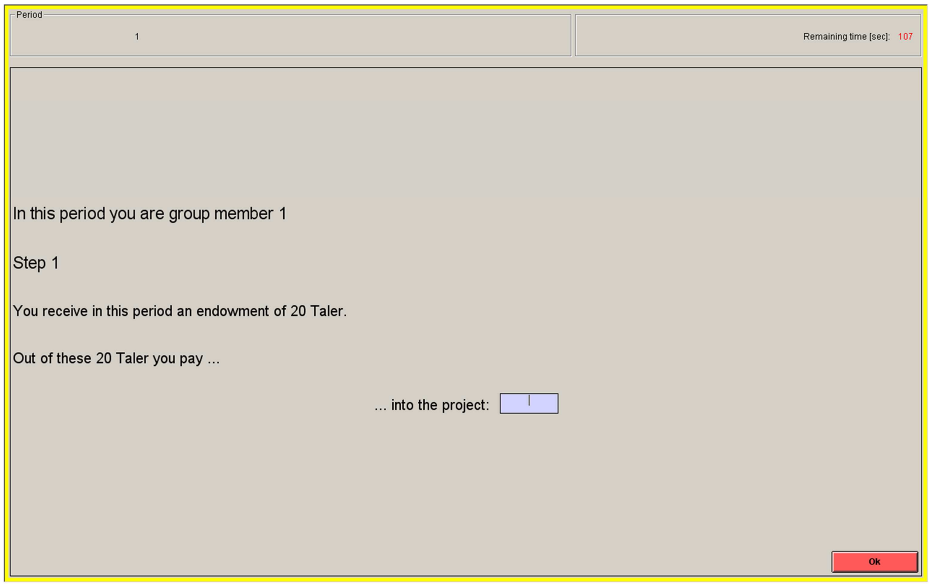
The experiment is divided into separate periods. It consists of a total of 10 periods. Participants are randomly assigned into groups of four. Each group, thus, has three further members, apart from you. During these 10 periods, the constellation of your group of four will remain unaltered. For 10 periods you will therefore be in the same group. Please note that the identification number assigned to you and the other members of the group changes randomly in each period. Group members can therefore not be identified as the periods progress. Each participant will receive from us 50 Taler, with which possible losses can be counterbalanced. The following pages outline the exact procedure of the experiment.

Information on the exact procedure of the experiment

Step 1

At the beginning of each period, each participant is allotted 20 Taler, which we shall henceforth refer to as his endowment. Each player than has to decide how to use his endowment. You have to decide how many of the 20 Taler you wish to pay into a project and how many you wish to keep for yourself. The consequences of your decision are explained in greater detail below.

At the beginning you will see the following contribution screen:



In this period you are group member 1

Step 1

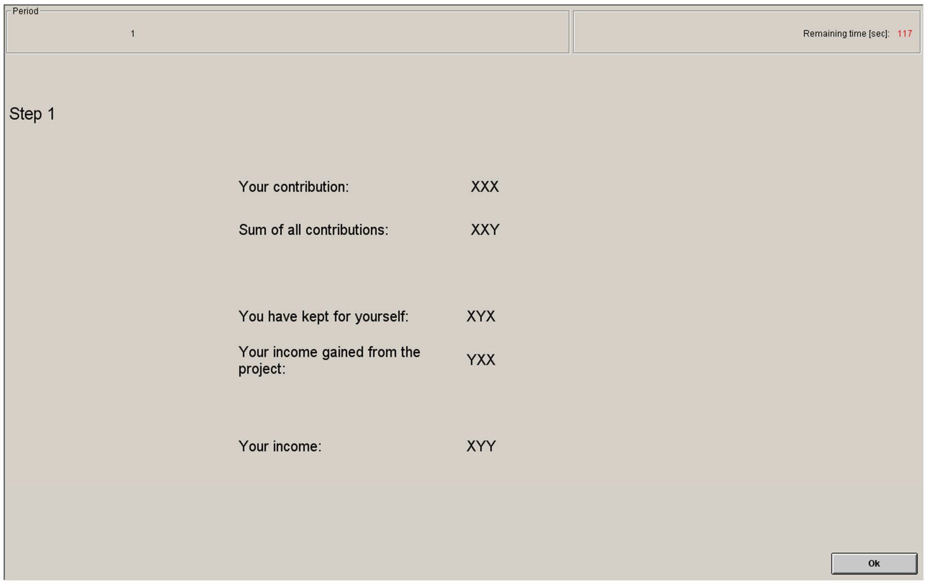
You receive in this period an endowment of 20 Taler.

Out of these 20 Taler you pay ...

... into the project:

In the left upper corner of the screen you will find the period number. In the right upper corner you will find the remaining time for your decision in seconds.

Your endowment is 20 Taler in each period. You make a decision on your project contribution by typing any integer number between 0 and 20 into the appropriate field on your screen. This field can be accessed using the mouse. As soon as you have determined your contribution, you have also decided on how many Taler to keep for yourself, i.e., 20 – your contribution. Once you have typed in your contribution, please click on Continue, again using the mouse. Once you have done this, your decision for this period is irreversible.



Step 1

Your contribution: XXX

Sum of all contributions: XXY

You have kept for yourself: XYX

Your income gained from the project: YXX

Your income: XYY

Once all members of the group have made their decisions, you will be told how high the total sum of contributions from all group members (including your own) to the project is. In addition, you are informed about your own contribution and the number of Taler kept by you; you are also told how many Taler you have earned in total during Step 1.

Your income therefore consists of two parts, namely:

(1) the Taler you have kept for yourself

(2) the "income gained from the project". Your income from the project is calculated as follows:

Income from the project

= .4 × total sum of all contributions to the project

Your income in Taler in each period thus equals

(20 – Your contribution to the project) +.4× (total sum of contributions to the project)

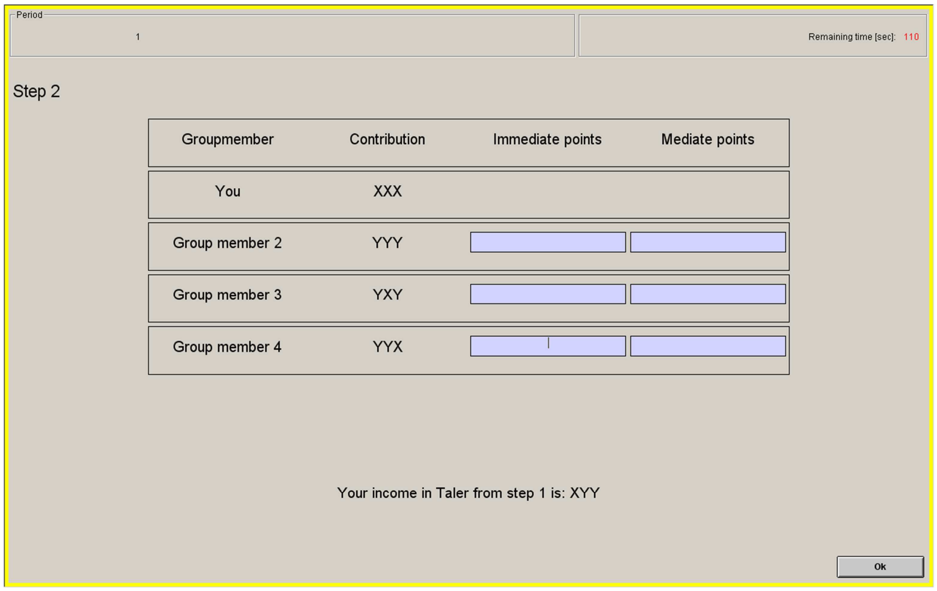
The total income at the end of Step 1, in Taler, is calculated according to the same formula for each member of the group. If, for example, the sum of the contributions from all group members adds up to 60 Taler, you and all other members each receive a project income of .4× 60 = 24 Taler. If the group members have contributed a total of 9 Taler to the project, you and all other members each receive an income of .4× 9 = 3.6 Taler from the project.

For each Taler you keep for yourself, you earn an income of 1 Taler. If, on the other hand, you contribute one Taler from your endowment to your group's project instead, the sum of the contributions to the project increases by one Taler and your income from the project increases by .4× 1 = .4 Taler. However, the income of each individual group member also increases by .4 Taler, so that the group's total income increases by .4× 4 = 1.6 Taler. The other group members thereby also profit from your contributions to the project. In turn, you profit from other members' contributions to the project. For each Taler contributed to the project by another group member, you earn .4× 1 = .4 Taler.

Step 2

In Step 2, you can decrease, or leave as it is, the income of each individual group member by giving points. You have the opportunity to assign two different types of points, immediate and mediate points. The income reduction through immediate points takes place at the end of each period. The income reduction through mediate points takes place only at the end of the experiment. This means that mediate points you have received throughout the experiments will be accumulated and deducted from your total income at the end of the experiment. All other group members are allowed to decrease your income, too, if they so wish. You will see this when considering the input screen of the second step.

You will be shown on the screen, along the number of periods and the remaining time, how many Taler each individual group member has contributed to the project. Your contribution will be shown in the row "You", while the contributions of the other three group members will be shown in randomly changing rows over periods.



Period

1

Remaining time [sec]: 110

Step 2

Groupmember	Contribution	Immediate points	Mediate points
You	XXX		
Group member 2	YYY		
Group member 3	YXY		
Group member 4	YYX		

Your income in Taler from step 1 is: XYY

You now have to decide for every group member about the combination of two types of points you wish to assign to them. It is compulsory to enter a number at this stage. If you do not wish to alter a certain group member's income, please insert 0. If you want to assign points you have to choose a number greater than 0. You can operate within the fields by using the tab key or the mouse.

When assigning points, you incur costs in Taler which depend on the number of points you assign to the individual players. The sum of immediate and mediate points per group member and period need not to exceed 10. The more points you assign to an individual player, the higher your costs are. Your total costs in Taler are calculated as the sum of the costs of points that you assigned to all other group members. The following formula shows the connection between the points distributed to an individual group member and the costs of such distribution:

Costs for assigned points = sum of immediate and mediate points (in Taler)

Each assigned point costs you 1 Taler. For example, if you have assigned 2 points to one member, your costs are 2 Taler; if, in addition, you assign 9 points to another group member, your costs are 9 Taler; if you assign the final group member 0 points, you have no costs. Your total costs are therefore 11 Taler (2+11+0). As long as you have not yet clicked on Continue, you may still change your decision.

If you assign 0 points to a certain group member, you do not alter this group member's income. If you assign 1 point (choosing 1) to a group member, you decrease this particular group member's income from Step 1 by 3 Taler. If you assign 2 points to a group member (choosing 2), you decrease his income by 6 Taler etc. Each point allocated by you to a particular group member reduces the group member's income from step 1 by 3 Taler.

By how much a group member's income from Step 1 is reduced overall depends on the total number of points received. If, for instance, one member receives a total of 3 points from all other members, the income in Step 1 is reduced by 9 Taler. If a member receives a total of 4 points, the income in step 1 is reduced by 12 Taler.

A person who receives immediate points will be informed about the income reduction immediately at the end of each period, but without knowing who assigned these points to him. The reduction of income by mediate points will be revealed not after each period, but only after the final period of the experiment. This means that all received mediate points are accumulated over periods and are deducted from the total income after the experiment, without detailed information on the period and the group member who has assigned these points. For your total income at the end of step 2, it follows that:

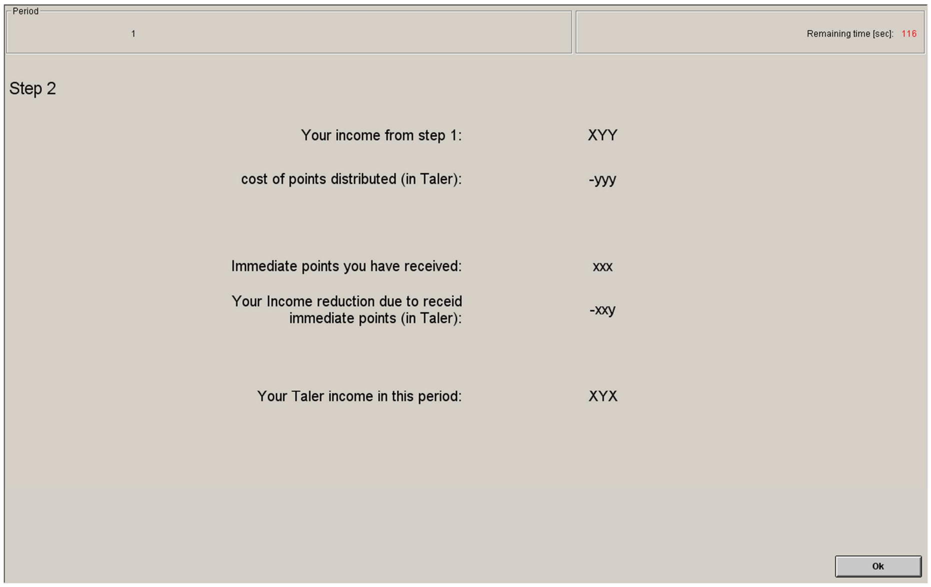
Total income at the end of step 2 = Period income

= Income after step 1

– 3× (sum of received immediate points)
– cost of points assigned by you

Please note that your total income at the end of step 2 can become negative if your costs for assigned points exceed your income after step 1 minus the reduction of your income due to received immediate points.

Once all members of the group have made their decisions, you will be informed about your period income in the following screen:



Step 2

Your income from step 1: XYY

cost of points distributed (in Taler): -yyy

Immediate points you have received: xxx

Your Income reduction due to receid immediate points (in Taler): -xxy

Your Taler income in this period: XYX

Your total income at the end of the experiment equals the sum of all period incomes minus the sum of mediate points:

Total income (in Taler)

= Total sum of period incomes (1)

– 3× (sum of received mediate points) (2)

(If the deduction (2) is larger than the sum of period incomes (1), your income is 0 Taler.)

Do you have any further questions?

Appendix B: Analysis of punishment

We focused on the intensity and effectiveness of observed and unobserved sanctions in our main text. To supplement the reported findings, this section describes subjects' sanctioning behavior in more detail.

We report in the paper that the strikingly high degree of cooperation in *O+U* is obtained by less intense sanctioning: In the absence of additional controls, a comparison of the average sums of points is insignificant.[17] However, an important – but often neglected – aspect in comparing the average punishment over all observations is the fact that contribution levels also differ between treatments. The question then is whether sanctions are less intense ceteris paribus, i.e., after taking the high cooperation levels in treatment *O+U* into account. Our analysis of sanctioning effectiveness already point into this direction. To stress it further, we compare the sanctioning intensity across treatments for a given kind of "conduct" or "norm violation". Following the previous literature on social sanctions (e.g., Herrmann et al., 2008), we define the conduct by the difference in contributions between the punisher (denoted as g_i) and the sanctioned person (g_j) – the idea being that the difference in contributions (i.e., g_j–g_i) measures the severity of the norm violation.

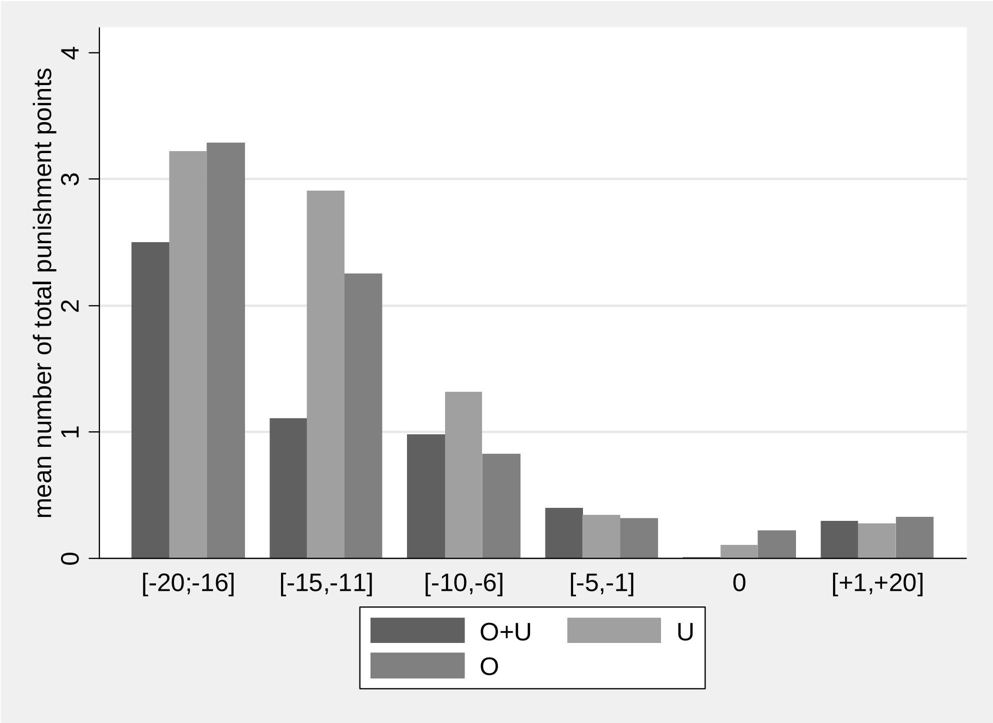


Fig. A1: Average amount of punishment (observable and unobservable)

[17] Exact rank-sum test, comparing *U* and the average sum in *O+U*, resp. comparing *O* and the average sum in *O+U* yields p=0.24 and p=0.12 (two-sided).

Figure A1 compares the punishment intensity for fixed intervals of norm violations between our three treatments. We observe that punishment in *O+U* is less intense for almost all categories of norm-violations, suggesting that, ceteris paribus, less intense sanctioning is required to obtain the cooperation in *O+U*. As we will see below in the regression analysis, this even holds true after controlling for the other group members' contribution levels.

Before turning to this issue, let us turn to the phenomenon of anti-social punishment as it is visible in the right-hand bars in Figure A1 above. Earlier research already demonstrated that punishment is sometimes used anti-socially, that is, actions which benefit the society are occasionally sanctioned (e.g., Herrmann et al., 2008). In our game, this happens if a player receives punishment points from someone who contributed less (or the same) amount than he or she did. We denote this as anti-social punishment, while we refer to pro-social punishment if points are assigned from a player who contributed more than the recipient. Interestingly, neither unobservable (treatment *U*) nor the combination of unobservable and observable sanctions (treatment *O+U*) seems to increase the amount of anti-social punishment in our groups (in the rightmost interval of Figure A1). This can also be seen in Table A1 below, where we report descriptive statistics of distributed sanctioning points (as well as sanctioning effectiveness) across treatment conditions (we report observable points as *p* points, while unobservable points as *s* points) both for pro-social and anti-social instances separately.

Furthermore, we conducted a regression analysis in which we differentiate between pro-social and anti-social sanctions. Our regression results stress that punishment is significantly less intense in *O+U*. Interestingly, we also find no evidence that anti-social punishment is predominantly executed by observable or unobservable punishment points if both sanctioning mechanisms are applicable (i.e., in treatment *O+U*).

Table A1: Descriptive statistics of sanctioning points

	U	*O*	*O+U*
p (observ. sanctions)		0.43	0.13
p (pro-social)		0.76	0.64
p (anti-social)		0.27	0.03
s (unobserv. sanctions)	0.50		0.09
s (pro-social)	0.90		0.34
s (anti-social)	0.22		0.04
sanctioning effectiveness	0.22[18]	0.67	2.12

For our regression analysis, let us define the two dummy variables $I_{p>0}$ and $I_{s>0}$ which equal one if player *i* assigns observable, resp. unobservable, punishment points to player *j*, and zero otherwise. $I_{p>0}$ and $I_{s>0}$ are the dependent variables in two distinct estimations. Further, as independent

[18] Notice that this number represents the "anticipated" sanctioning effectiveness of an unobserved point in *U*. Of course, this number is purely hypothetical as punished players do not receive feedback on unobservable points. For *O* and *O+U*, we report the sanctioning effectiveness of observable points.

variables, we introduce the contribution g_j of the person punished, the contributions of the remaining two group members $G_{kl}=g_k + g_l$, $k,l \neq i,j$, and the absolute difference between contributions $d_{ij}^{+} = |\max(g_i - g_j, 0)|$ and $d_{ij}^{-} = |\min(g_i - g_j, 0)|$. We also add a dummy variable I_{O+U} indicating the *O+U* condition, and interaction terms d_{ij}^{+} I_{O+U} and d_{ij}^{-} I_{O+U}. Therefore, g_j indicates the effect of the contribution of player *j* on the probability of being punished, while G_{kl} shows the effect of the contributions of other group members, indicating whether being in a group of free riders or a group of full contributors affect *i's* punishment decision. The two difference measures allow us to estimate how the (absolute) difference between the punished player's and punishing player's contributions affects the decision to assign points. We differentiate between positive differences (d_{ij}^{+}) and negative differences (d_{ij}^{-}). Significant positive coefficients for d_{ij}^{+} suggest pro-social punishment, whereas significant positive coefficients for d_{ij}^{-} suggest anti-social punishment. Finally, I_{O+U}, d_{ij}^{-} I_{O+U}, and d_{ij}^{+} I_{O+U} show differences between the *O+U* and the *U* (for the dependent variable $I_{s>0}$) and between the *O+U* and the *O* condition (for the dependent variable $I_{p>0}$), respectively.[19] Table A2 reports the estimation results for the mean marginal effects of the independent variables in a probit regression.[20]

The econometric results underline what we observed above in Figure A1. We find evidence for pro-social and anti-social punishment both with observed and unobserved points. The difference between contributions influences the decision whether to punish or not to punish, as it is indicated by the significant positive marginal effects of d_{ij}^{-} and d_{ij}^{+}. However, the punished player’s absolute contribution level influences the probability that unobserved and observed punishment occurs. As one would expect, the probability decreases for higher contributions. Likewise, increasing the contributions of the other players significantly increases the probability that punishment occurs. Most importantly, the significant negative marginal effect of the treatment dummy shows that there is a significantly lower probability in *O+U* for unobserved and observed punishment as soon as we control for the contribution situation, i.e., for the contribution differences across treatments.

The same picture emerges if we look at the number of points rather than at the decision to punish or not. To see this, let us consider *p* and *s* as dependent variables in our second regression analysis. Notice that *p* and *s* are censored in the interval zero to ten, so that we apply a Tobit regression (note that the results are qualitatively the same if we use different regression models, e.g., OLS). We use two distinct estimations: one for *p* and one for *s*; as independent variables, we use the same variables as in the first two regressions. Again, the variables I_{O+U}, d_{ij}^{-} I_{O+U}, and d_{ij}^{+} I_{O+U}, indicate differences between the *O+U* and the *U* (*O*) condition. Table A3 reports the estimation results for the mean marginal effects of the independent variables in a robust least square regression.[21]

[19] Of course, the first estimation contains only observations from the *O+U* and the *U* conditions, while the second estimation contains only observations from the *O+U* and the *O* conditions.

[20] Standard errors are clustered for each group over the entire 10 periods.

[21] Again, standard errors are clustered for each group over the entire 10 periods.

Table A2: Mean marginal effects of the Probit estimation

dependent	$I_{p>0}$	$I_{s>0}$
independent		
g_j	–0.007*	–0.010**
	(0.004)	(0.005)
d_{ij}^{+}	0.020***	0.017**
	(0.007)	(0.007)
d_{ij}^{-}	0.010***	0.012**
	(0.003)	(0.005)
G_{kl}	0.005**	0.007***
	(0.002)	(0.002)
I_{O+U}	–0.121**	–0.120**
	(0.047)	(0.056)
$d_{ij}^{+} I_{O+U}$	–0.019	–0.023
	(0.070)	(0.121)
$d_{ij}^{-} I_{O+U}$	–0.012	–0.011
	(0.045)	(0.061)
number of observations	1920	1800
logLik	–545	–509
PseudoR²	0.28	0.29
Wald test (7)	619***	295***

Note: Standard errors are reported in parenthesis. *** indicates significance at a $p < 0.01$ level, ** at a $p < 0.05$ level and * at a $p < 0.1$ level. Marginal effects are evaluated at the means. The constant terms of the models are –1.579*** (0.394) and –1.584** (0.682). The number of observations is reported along the log likelihood (logLik) and the fitness of the estimation by means of the PseudoR². Finally, the Wald test indicates the significance of the estimation's improvement against the null model.

Results again indicate important treatment differences with respect to the number of observable and unobservable punishment points assigned. Both significant negative marginal effects for I_{I+L} show that players assign less observable and unobservable punishment points. Moreover, the weakly significant marginal effect of the interaction $d_{ij}^{+}I_{O+U}$ indicates less pro-social punishment for unobservable points. Interestingly, there is no evidence that anti-social punishment is mainly done using unobservable points if both sanctioning mechanisms are available (i.e., in *O+U*): the marginal effect of $d_{ij}^{-}I_{O+U}$ is neither significantly negative in regression model for observable points nor significantly positive in the regression model for unobservable points. Concerning the other independent variables, qualitatively similar results as in the probit regressions are found.

Table A3: Mean marginal effects of the Tobit regression

dependent independent	*p*	*s*
g_j	–0.119*	–0.178**
	(0.073)	(0.079)
d_{ij}^{+}	0.294***	0.365**
	(0.089)	(0.181)
d_{ij}^{-}	0.173***	0.230***
	(0.041)	(0.083)
G_{kl}	0.094***	0.108***
	(0.030)	(0.033)
I_{O+U}	–2.290***	–2.163**
	(0.775)	(0.980)
$d_{ij}^{+}\, I_{O+U}$	–0.043	–0.236*
	(0.046)	(0.127)
$d_{ij}^{-}\, I_{O+U}$	–0.129	–0.053
	(0.078)	(0.101)
number of observations	1920	1800
logLik	–981	–915
Pseudo R²	0.18	0.19
F test (7, number of observations)	26.46***	45.21***

Note: Standard errors are reported in parenthesis. *** indicates significance at a p < 0.01 level, ** at a p < 0.05 level and * at a p < 0.1 level. The constant terms of the models are –4.581*** (1.164) and –4.547** (2.177). The number of observations is reported along the log likelihood (logLik) and the fitness of the estimation by means of the PseudoR². Finally, the F-test indicates the significance of the joint coefficients.