Universität Hamburg
DER FORSCHUNG | DER LEHRE | DER BILDUNG

FAKULTÄT
FÜR WIRTSCHAFTS- UND
SOZIALWISSENSCHAFTEN

# The joint benefits of observed and unobserved social sanctions

Andreas Glöckner
Sebastian Kube
Andreas Nicklisch

WiSo-HH Working Paper Series
Working Paper No. 13
March 2014

WiSo-HH Working Paper Series
Working Paper No. 13
March 2014

# The joint benefits of observed and unobserved social sanctions

Andreas Glöckner, Universität Göttingen, Germany
Sebastian Kube, Universität Bonn, Germany
Andreas Nicklisch, Universität Hamburg, Germany

ISSN 2196-8128

Font used: „TheSans UHH“ / LucasFonts

**Kontakt:**

WiSo-Forschungslabor
Von-Melle-Park 5
20146 Hamburg
E-Mail: experiments@wiso.uni-hamburg.de
Web: http://www.wiso.uni-hamburg.de/forschung/forschungslabor/home/

# The joint benefits of observed and unobserved social sanctions[*]

Andreas Glöckner[1,2], Sebastian Kube[1,3] & Andreas Nicklisch[1,4,‡]

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

[2] Department of Psychology, University of Göttingen,

[3] Department of Economics, University of Bonn

[4] School of Business, Economics and Social Science, University of Hamburg

## Abstract

Cooperation problems are at the heart of many environmental problems. Prominent solutions frequently rely on monitoring and punishment by central authorities. In recent years, the focus has shifted to decentralized approaches with mutual monitoring and social sanctions to foster cooperation. In this paper, we empirically test for the role of a specific form of social punishment, namely sanctions that are unobservable at first and only applied with a delay. We observe that in particular the combination of such unobservable sanctions with immediately observable sanctions strongly enhances cooperation within groups. Strikingly, this improvement is not caused by an extensive use of both forms of punishment. Our data suggest that the mere thread of unobservable sanctions increases the effectiveness of observable punishment.

Keywords: Unobserved Social Sanctions, Cooperation, Public Goods, Sanctioning Effectiveness

JEL: C92, H41, Q01, Q23

[*] 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 voluntary contribution mechanism.

[‡] Corresponding author: University of Hamburg, Von-Melle-Park 5, 20146 Hamburg, Germany, Phone: +49 (0)40 4283 88976, E-mail: andreas.nicklisch@wiso.uni-hamburg.de

## 1. Introduction

Cooperation problems are at the heart of many environmental problems. For instance, the tragedy of the commons (like overfishing, deforestation, wasting of water, pollution, littering) would not arise if people would cooperate and restrict their use of the resources in the first place. Consequently, most solution approaches to overcome environmental problems – although being manifold in their specific design – essentially aim at restricting selfish behavior one way or another. In order not to jeopardize their effectiveness, however, also these mechanisms themselves require a certain willingness to cooperate and to accept individual costs. This applies not only to centralized mechanisms (e.g., corruption of logging inspectors looking after forest concessions, see Amacher et al., 2012), but it holds particularly true for decentralized approaches with mutual monitoring and social sanctions (e.g., see Cox and Ross 2011 who explore collective-action problems in managing community irrigation systems; or Cason and Gangadharan 2013 studying mechanisms to reduce ambient pollution levels). In this paper, we focus on the latter and test for the role of a specific form of social punishment in enhancing cooperation, namely sanctions that are unobservable at first and only applied with a delay – as well as the combination of such unobservable sanctions with immediately observable sanctions.

A large number of experimental studies has analyzed the government of public goods via decentralized approaches which make use of mutual monitoring and social sanctions (see Ledyard, 1995, Zelmer, 2003, or Balliet et al., 2011 for overviews). A general finding is that social sanctions can be effective in fostering cooperation. Taking costly actions to punish someone who is observed exploiting common resources or freeriding on others' efforts to provide public goods leads to higher cooperation rates than in groups without punishment opportunities. This is found in controlled laboratory experiments where persons face a social dilemma and interact anonymously over a finite number of periods (e.g., Yamagishi, 1986, Ostrom et al., 1992, Fehr & Gächter, 2000, 2002), as well as in the field where anonymity is not necessarily granted and reputation might matter (e.g., common irrigation works in the Philippines or southern Spain, or forest management in Switzerland and Japan, see Ostrom, 1990). A major difference in the field, however, is the availability of different forms of social sanctions. In particular, not all instances of punishment in the field are immediately observable (as it is usually the case in the lab, see Chaudhuri 2011 for a survey), but some are unobservable at first and only applied with a delay.[1]

In this paper, we hypothesize that it is the joint availability of these two different forms of punishment that makes social sanctions such an effective instrument for fostering cooperation: i) punishment with immediate feedback and ii) punishment with delayed feedback. In the former case, players are instantly informed about received punishment, that is, subjects in the lab are informed at the end of each period about others' sanctions when playing a public goods game repeatedly over a finite number of periods. In the latter case, sanctioned persons are informed about the received punishment only after the very last period of interaction. The main motive for the use of immediate punishment is to train defectors, while unobserved punishment with delayed sanctions is mostly retributive since it eliminates strategic punishment motives (cp. Fudenberg and Pathak, 2010 and Vyrastekova et al. 2008, who compare both forms of punishment in isolation). Conceivably, it might be beneficial to have both forms of punishment available at the

[1] Examples for immediately observed sanctions include (physical or verbal) threats for non-cooperative users of a common pool resource and their property. Delayed sanctions include social exclusion, rumor spreading among neighbors, not passing on crucial (market) information to non-cooperators, or whistleblowing to central authorities, to name only a few examples.

same time – but only if they are complements rather than substitutes to one another, which ex ante is an open question.

To shed light on this question, we use the controlled environment of the laboratory. Keeping the environment constant, only the availability of unobserved and/or observed punishment opportunities is varied between treatments. This allows us to test for the use of the different types of social sanctions and their causal effects on cooperation behavior in a social dilemma situation (a typical public-good game). Our results reveal strong complementarities between the effects of unobserved and observed punishment. If individuals can use both mechanisms at the same time, cooperation rates are enormously enhanced – while, strikingly, overall there is even less intense sanctioning. A likely reason for the latter is found in the data on players' consecutive behavior after being punished. There is a large increase in the disciplining effect of observed punishment when it is accompanied by (the fear of) unobserved punishment. This implies that in order to increase cooperation rates, subjects need to spend less of their own private resources on immediately observed punishment. More precisely, to increase the contribution of non-cooperators in the treatment where only observed punishment is available by the same amount as in the treatment where both forms of punishment are jointly available, one has to spend three times as many observed punishment points (and destroy substantially more payoffs since punishment is costly).

The increased effectiveness of social sanctions is particularly interesting since a major drawback is that social sanctions as implemented in the standard public-good paradigm in the lab frequently come at severe costs, because significant amount of subjects' private resources are spoiled for the sake of sanctioning. The negative effect of punishment is usually so severe that the average payoff of players falls below the level achieved in the same game without punishment (at least in the short run, see Gächter et al, 2008). This is supported by our data from the treatments where only one type of punishment is available. However, if observed and unobserved punishment mechanisms are jointly available, we find that punishment is highly efficient and players' overall payoffs increase. This might help to explain the seeming discrepancy between field and lab evidence on social sanctions, because field evidence suggests that cooperation can also be sustained without harsh punishment and that instances of harsh real world punishment are the rare exception rather than a usual practice (see Ostrom 1990, Agrawal & Ostrom, 2001). In view of our findings, we would argue that it is the availability of different forms of social sanctions in the field that make decentralized approaches such an effective instrument for fostering cooperation.

We proceed as follows. We begin by describing our experimental design, which tests for the effectiveness of decentralized approaches with mutual monitoring and social sanctions with three different forms of punishment in a voluntary contribution mechanism. We then present our experimental results, focusing on differences in contribution behavior, punishment behavior, and sanctioning effectiveness between treatments. We conclude with a discussion and potential avenues for future research.

## 2. The game

Our experimental tool is the standard voluntary contribution mechanism (VCM) with and without punishment. This design has been widely tested (see Zelmer, 2003, for an overview) and represents a framework that incorporates many important features that are at the heart of most environmental problems. It allows to investigate cooperation and punishment behavior, and to compare the efficiency of different punishment institutions in a clear and concise manner.

In our VCM game, four players interact repeatedly over ten periods. Each period consists of two stages. In stage one each player receives an endowment of 20 ECU (experimental currency units; in the instruction for the participants we refer to this units as “Taler”). Players choose simultaneously how many ECU to contribute to a public good, $g_i$, $g_i \in \{0, 1, 2,\ldots, 20\}$. Each ECU contributed to the public good yields a benefit of 1.6 ECU to the entire group that is equally distributed among the players in the group. Therefore, the marginal per capita return from player’s own contributions to the public good is 0.4.

In stage two, players are informed about individual contributions in this period. Then they may or may not punish any of the three other players. For this purpose, each player may assign punishment points to a particular player. Each punishment point assigned leads to a deduction of three ECUs from the punished player's account, but also reduces the punisher's income by one ECU. In sum, each player can spend up to 10 ECU on (total) punishment in each period. There are two types of punishment points distributed by player $i$ to player $k$, observed punishment $p_{ik}$ and unobserved punishment $s_{ik}$. Points $p_{ik}$ lead to immediate feedback on received payoff deduction after each period, whereas points $s_{ik}$ lead to delayed feedback only after the final period of the experiment. Formally, player $i$’s payoff equals

$$\pi_i = 20 - g_i + 0.4\sum_k g_k - \sum_{k\neq i} p_{ik} - \sum_{k\neq i} s_{ik} - 3\sum_{k\neq i} p_{ki} - 3\sum_{k\neq i} s_{ki}. \qquad (1)$$

After each period, players learn their own payoff from stage one, their payoff reduction due to distributed punishment points and received punishment points with immediate feedback. Players then proceed to the next period; payoffs accrue over ten periods. All parameters and payoff functions are common knowledge.

Assigning punishment points (irrespectively whether they lead to immediate or delayed feedback) 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. Therefore, selfish players under common knowledge of rationality should not expect to be punished 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. Hence, the subgame perfect Nash equilibrium under the assumption of self-centered, money-maximizing preferences is thus i) no social sanctions on stage two, and ii) no contributions on stage one.

We implement three treatment conditions: Our first treatment implements a regular sanctioning mechanism with *observed* punishment only (treatment *O* in the following). That is, punishment points are commonly restricted such that, for each $i$ and each $k$, $p_{ik} \leq 10$ and $s_{ik} \equiv 0$ in every period (cp., Herrmann et al., 2008). The second treatment implements a sanctioning mechanism with *unobserved* punishment only (treatment *U*). That is, punishment points are commonly restricted such that, for each $i$ and each $k$, $s_{ik} \leq 10$ and $p_{ik} \equiv 0$ in every period, so that only after the final period of the experiment, subjects get to learn the accrued points and corresponding sanctions are deducted (cp., Vyrastekova et al., 2008). Finally, the third treatment features *both* mechanisms at the same time (treatment *O+U*): punishment points are commonly restricted such that, for each $i$ and each $k$, $s_{ik} + p_{ik} \leq 10$ in every period. Thus, players can choose in each period of the *O+U* treatment how many observed and how many unobserved punishment points they want to carry out. We would like to stress that in all treatments, a player can at most assign 10 sanctioning points in each period to each player, and this is common knowledge (i.e., there is no threat of additional punishment points in the *O+U* treatment compared to the other treatment conditions). The only difference is the feedback channel(s) by which players are informed about the punishment.

Altogether, 92 subjects, mostly students from the University of Bonn majoring in various fields, participated in the experiment (10 percent were non-students) in October and November 2009. Mean age was 24.5 years (standard deviation 5.5 years), 62 percent were females. Each subject participated only once in the experiment and none of them had participated in a public good experiment before. In total, we ran 9 sessions with 23 groups, resulting in 8 independent group observations each in the *O* and *O+U* treatments, and 7 independent group observations in the *U* treatment.[2] For comparison, we also include data of a regular voluntary contribution mechanism experiment without any punishment (treatment *VCM*).[3]

A session lasted for about 60 minutes. 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 (see Appendix). Afterwards, participants could pose clarifying questions to the experiment supervisor in private. After questions were answered individually, participants had to answer a set of control questions to ensure that everybody had understood the game.[4] Control questions were corrected individually, and wrong answers were explained privately. Then, participants were randomly and anonymously matched in groups of four players by the computer. Participants knew that the

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

[3] Representative 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

[4] Questions are almost identical to the control questions of Herrmann et al. (2008). In one of the original questions, the sum of punishment points exceeds ten which was not possible in our study. We adjusted numbers in such a way that they sum up to a value below ten (i.e., nine).

experiment terminates after ten periods; the composition of the group remained constant throughout the entire 10 periods of the experiment (partner design), however, to prevent participants from identifying each other across periods, they received a random identification number between 1 and 4 at the beginning of each period. At the end of the experiment, ECU earned were accrued over all periods and converted at an exchange rate of 3 Euro per 100 ECU. Participants were paid out individually to ensure their anonymity. They earned on average 13.86 Euro[5] (standard deviation 1.45 Euro), including a show-up fee of 5 Euro.[6]

## 3. Results

CONTRIBUTIONS: Figure 1 illustrates the development of average contributions over the entire course of the experiment for the different 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*. To test differences for statistical significance, we take a very conservative approach and 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 the treatments *O+U*, *O*, and *U* revealed significant differences for the first two comparisons ($p$<0.001, $p$=0.03) and a marginally significant difference between *VCM* and *U* ($p$=0.07). 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

[5] Corresponds to $20.40 (in November 2009)

[6] 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 ECU at the beginning of the experiment. However, no player accrued an overall negative payoff at the end of the experiment.

treatments *O* and *U* (*p*=0.69). Even if we focus on the first five or on the last five periods, there are no significant differences (*p*=0.87, and *p*=0.23). 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 (overall *p*=0.005 and *p*=0.015, first period *p*=0.006 and *p*=0.087). 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.

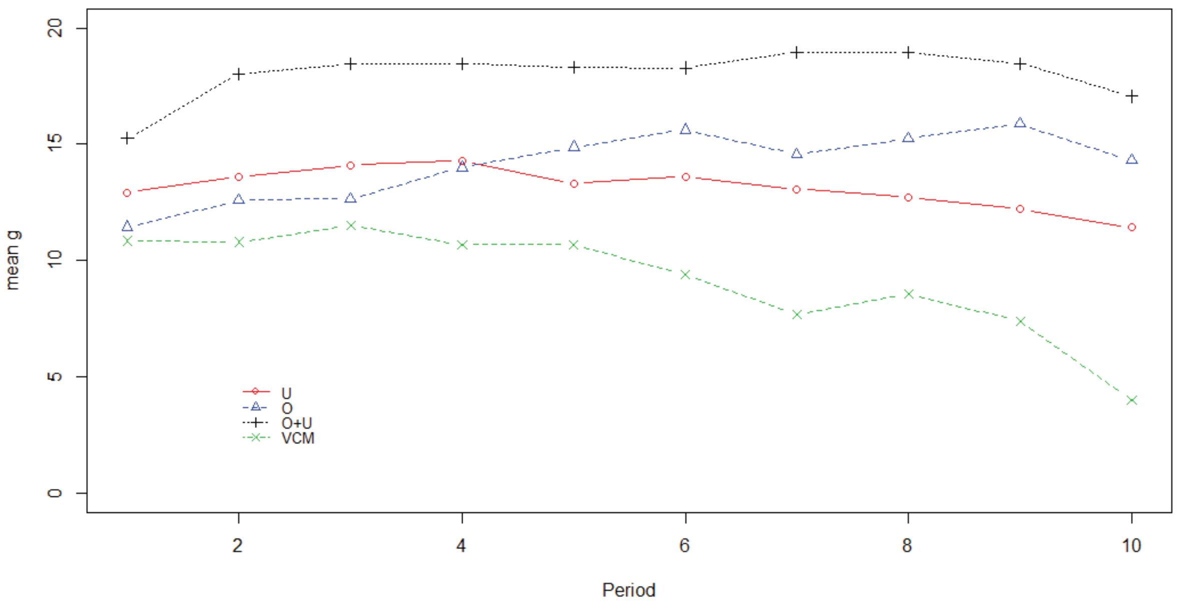


**Figure 1.** Average contributions (*g*) over periods and treatment conditions unobserved punishment only (*U*), observed punishment only (*O*), combination of observed and unobserved punishment (*O+U*), and the standard voluntary contribution game without punishment (*VCM*).

PUNISHMENT BEHAVIOR: Concerning the difference between the two treatments with a single punishment channel, we observe no significant differences between the number of sanctioning points distributed in *O* (on average 1.29 points) and *U* (on average 1.51 points) (*p*=0.61). In *O+U*, where both types of points are available, subjects assign on average 0.40

observable points and 0.28 unobservable points, in sum 0.68. However, the difference between the *O*+*U* treatment and the *O* treatment (the *U* treatment) fails to be significant (*p*=0.12, and *p*=0.24, respectively).

For a more specific look at punishment, let us define pro-social (anti-social) punishment as the distribution of punishment points by a player whose contribution is larger (smaller) than the contribution of the punished player (compare Herrmann et al., 2008). Figure 2 shows the average number of punishment points distributed depending on the difference between the contribution of the punished and the punishing player. The four most left classes of differences fall into the category of pro-social punishment, while the two most right classes are instance of anti-social punishment. Obviously, the severity of pro-social punishment changes with the differences in contributions across all treatment conditions. Overall, less pro-social punishment is distributed in *O*+*U*, while there seems to be little difference between punishment in *O* and *U* (a more detailed analysis follows below).

With respect to anti-social punishment, we find occasional instances in *O* and in *U* confirming earlier research (e.g., Herrmann et al., 2008). Yet, there is less anti-social punishment in *O*+*U* than in *O* and *U*, whereas we do not find a prominent channel for it. That is, anti-social punishment (if there is any) and pro-social punishment in *O*+*U* are executed both with immediate and latent feedback punishment.

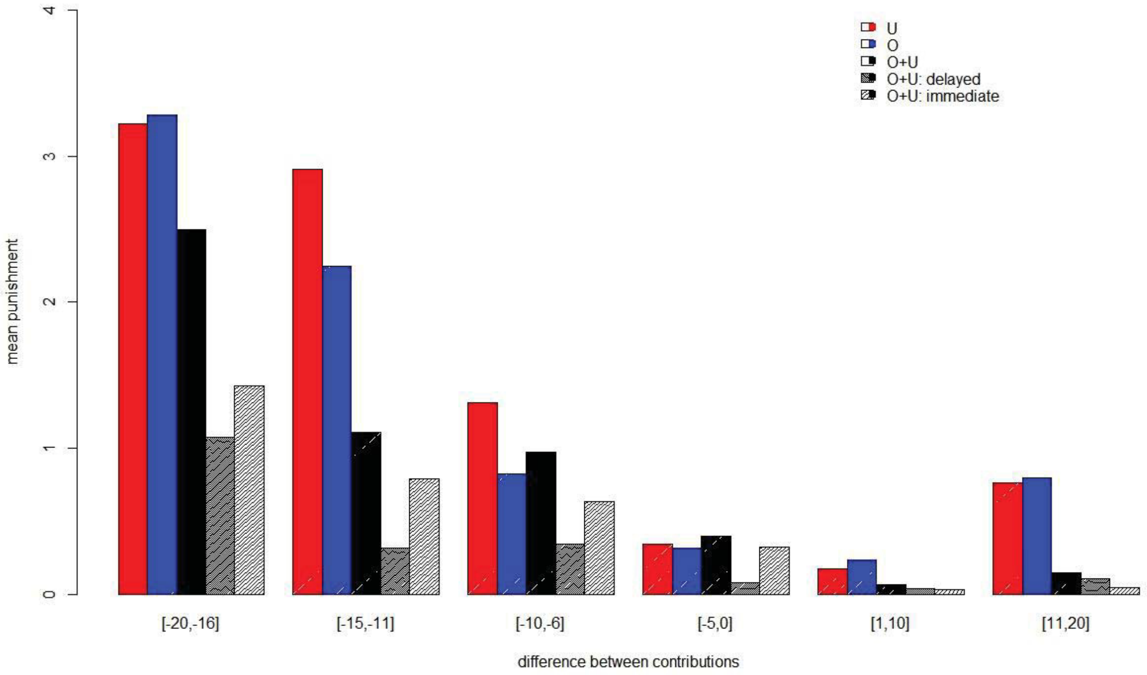


**Figure 2.** Average amount of punishment divided by differences in contributions between punished and punishing person; negative (positive) numbers indicate that the punished person contributed less (more) than the punishing person. Treatment conditions are unobserved punishment only (*U*), observed punishment only (*O*), combination of observed and unobserved punishment (*O+U*), the punishment points with delayed feedback in *O+U*, and the punishment points with immediate feedback in *O+U*.

To underpin our observation that punishment in *O+U* is less intense, we use several regression analyses. First, we would like to focus on punishment instances, but not the severity of punishment. For this purpose, let us define the two dummy variables $\mathrm{I}_{p>0}$ and $\mathrm{I}_{s>0}$ which equal one if player *i* assigns points with immediate (*p*), resp. delayed (*s*) feedback to player *k*, and zero otherwise. $\mathrm{I}_{p>0}$ and $\mathrm{I}_{s>0}$ are the dependent variables in two distinct estimations. Further, as independent variables, we introduce the contribution $g_\mathrm{k}$ of the person punished, the contributions of the remaining two group members $G_\mathrm{jl}=g_\mathrm{j}+g_\mathrm{l}$, $j,l \neq i,k$, and the absolute difference between contributions $d_\mathrm{ik}^+ = |\max(g_\mathrm{i} - g_\mathrm{k}, 0)|$ and $d_\mathrm{ik}^- = |\min(g_\mathrm{i} - g_\mathrm{k}, 0)|$. We also add a dummy variable $\mathrm{I}_{\mathrm{O+U}}$ indicating the *O+U* condition, and interaction terms $d_\mathrm{ik}^+\ \mathrm{I}_{\mathrm{O+U}}$ and $d_\mathrm{ik}^-\ \mathrm{I}_{\mathrm{O+U}}$. Therefore, $g_\mathrm{k}$

indicates the effect of the contribution of player *k* on the probability of being punished, while $G_{\mathrm{jl}}$ 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 ($d_{\mathrm{ik}}^{+}$) and negative differences ($d_{\mathrm{ik}}^{-}$). Significant positive coefficients for $d_{\mathrm{ik}}^{+}$ suggest pro-social punishment, whereas significant positive coefficients for $d_{\mathrm{ik}}^{-}$ suggest anti-social punishment. Finally, $\mathrm{I}_{\mathrm{O+U}}$, $d_{\mathrm{ik}}^{-}\ \mathrm{I}_{\mathrm{O+U}}$, and $d_{\mathrm{ik}}^{+}\ \mathrm{I}_{\mathrm{O+U}}$ show differences between the *O+U* and the *U* (for the dependent variable $\mathrm{I}_{s>0}$) and between the *O+U* and the *O* condition (for the dependent variable $\mathrm{I}_{p>0}$), respectively.[7] Table 1 reports the estimation results for the mean marginal effects of the independent variables in a probit regression.[8]

We find evidence for pro-social and anti-social punishment both with immediate and delayed feedback 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 $d_{\mathrm{ij}}^{-}$ and $d_{\mathrm{ij}}^{+}$. However, the punished player's absolute contribution level influences the probability that punishment occurs with both types of points. 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 immediate and delayed feedback points as soon as we control for the contribution situation, that is, for the contribution differences across treatments.

[7] 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.

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

**Table 1:** Mean marginal effects of the Probit estimation.

| dependent<br>independent | $I_{p>0}$ | $I_{s>0}$ |
|---|---|---|
| $g_k$ | –0.007* | –0.010** |
| | (0.004) | (0.005) |
| $d_{ik}^{+}$ | 0.020*** | 0.017** |
| | (0.007) | (0.007) |
| $d_{ik}^{-}$ | 0.010*** | 0.012** |
| | (0.003) | (0.005) |
| $G_{jl}$ | 0.005** | 0.007*** |
| | (0.002) | (0.002) |
| $I_{O+U}$ | –0.121** | –0.120** |
| | (0.047) | (0.056) |
| $d_{ik}^{+}$ $I_{O+U}$ | –0.019 | –0.023 |
| | (0.070) | (0.121) |
| $d_{ik}^{-}$ $I_{O+U}$ | –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. *$p$<0.1; **$p$<0.05; ***$p$<0.01. Marginal effects are evaluated at the means. The constant terms of the models are –1.579*** (0.394) and –1.584** (0.682). The Wald test indicates the significance of the estimation's improvement against the null model.

The same picture emerges if we investigate the number of points rather than the decision to punish or not; that is, when analyzing the severity of punishment. To see this, let us consider $p$ and $s$ (the number of immediate, resp. delayed feedback points) 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 (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 $\mathrm{I}_{O+U}$, $d_{ik}^{-}$ $\mathrm{I}_{O+U}$, and $d_{ik}^{+}$ $\mathrm{I}_{O+U}$, indicate differences between the *O+U* and the *U* (*O*) condition. Table 2 reports the estimation results for the mean marginal effects of the independent variables in a robust least square regression.[9]

**Table 2:** Mean marginal effects of the Tobit regressions.

| dependent<br>independent | *p* | *s* |
|---|---|---|
| $g_k$ | –0.119* | –0.178** |
| | (0.073) | (0.079) |
| $d_{ik}^{+}$ | 0.294*** | 0.365** |
| | (0.089) | (0.181) |
| $d_{ik}^{-}$ | 0.173*** | 0.230*** |
| | (0.041) | (0.083) |
| $G_{jl}$ | 0.094*** | 0.108*** |
| | (0.030) | (0.033) |
| $\mathrm{I}_{O+U}$ | –2.290*** | –2.163** |
| | (0.775) | (0.980) |
| $d_{ik}^{+}$ $\mathrm{I}_{O+U}$ | –0.043 | –0.236* |
| | (0.046) | (0.127) |
| $d_{ik}^{-}$ $\mathrm{I}_{O+U}$ | –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. *$p$<0.1; **$p$<0.05; ***$p$<0.01. The constant terms of the models are –4.581*** (1.164) and –4.547** (2.177). The *F*-test indicates the significance of the joint coefficients.

Results again indicate important treatment differences with respect to the number of immediate and delayed feedback points assigned. Significant negative marginal effects for $\mathrm{I}_{O+U}$

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

show that players assign less punishment points. Moreover, the weakly significant marginal effect of the interaction $d_{ik}^{+}I_{O+U}$ indicates less pro-social punishment for delayed feedback points. Interestingly, there is no evidence that anti-social punishment is mainly done using delayed feedback points if both sanctioning mechanisms are available (i.e., in *O+U*): the marginal effect of $d_{ik}^{-}I_{O+U}$ is neither significantly negative in regression model for immediate feedback points nor significantly positive in the regression model for delayed feedback points. Concerning the other independent variables, qualitatively similar results as in the Probit regression are found.

Let us summarize our findings so far: the punishment channels in *O+U* work in a complementary way, while total punishment expenditures are substantially lower in this treatment condition. We would like to stress that complementarity in our experiment does not mean that players use both type of points simultaneously, although some players actually do this (in 16% of all punishment decisions in *O+U* both types of points are distributed at once). Rather, in most cases the immediate feedback points function as a kind of warning for non-cooperators that – given no correction in behavior – unobservable sanctions might be used such that consecutive delayed punishment might be up-coming. This idea is reflected by the "temporal order" in the use of both punishment channels: Table 3 reports the correlation between received immediate feedback and delayed feedback points in *O+U* across two subsequent periods (i.e., in *t-1* and *t*). We observe that the correlation coefficients are generally smaller in the last row, that is, in those cells where we look at delayed sanctions taking place first. Moreover, we find the strongest (significantly positive) correlation between the two channels for immediate feedback points received in period *t-1* and delayed feedback points in period *t.* This suggests that many subjects indeed first use the immediately observable sanctions and only later switch to unobserved, delayed punishment if cooperation behavior does not change. In the next paragraph, we will see that behavior frequently does change after receiving social sanctions when testing for the relation between received punishment and consecutive contributions of the punished person.

**Table 3:** Correlation between different punishment points in *O+U*.

| | *immediate feedback points in t* | *delayed feedback points in t* |
|---|---|---|
| *immediate feedback points in t* | | 0.30*** |
| *immediate feedback points in t-1* | 0.32*** | 0.31*** |
| *delayed feedback points in t-1* | 0.15** | 0.19*** |

*Note*. Pearson's product-moment correlation test; *p<0.1; **p<0.05; ***p<0.01.

SANCTIONING EFFECTIVENESS: Earlier on, we showed that although the contributions are strikingly high in *O+U*, this increased cooperation is associated with less intense but more effective sanctioning. In other words, non-cooperators are more responsive to received (immediate feedback) punishment in *O+U* than in *O*. To formalize this, let us define sanctioning effectiveness as the change in players' contribution (between the period where they are punished and the subsequent period) per observed sanctioning point. Average sanctioning effectiveness in conditions *O* and *O+U* are shown in Figure 3. We find an average sanctioning effectiveness of 0.67 in the *O* condition, that is, the sanctioned player increases his average contribution by 0.67 tokens in the subsequent period after receiving one point of punishment. In contrast, we find a significantly higher average sanctioning effectiveness of 2.12 in the *O+U* condition ($p$=0.04). 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 *O* condition. In other words, investing one immediate feedback point to train a non-cooperator redeems immediately in terms of contributions by the punished player the *O+U* treatment, while this takes more than one period in *O*. As a consequence, less punishment and fewer periods are necessary to discipline non-cooperators in the former treatment.

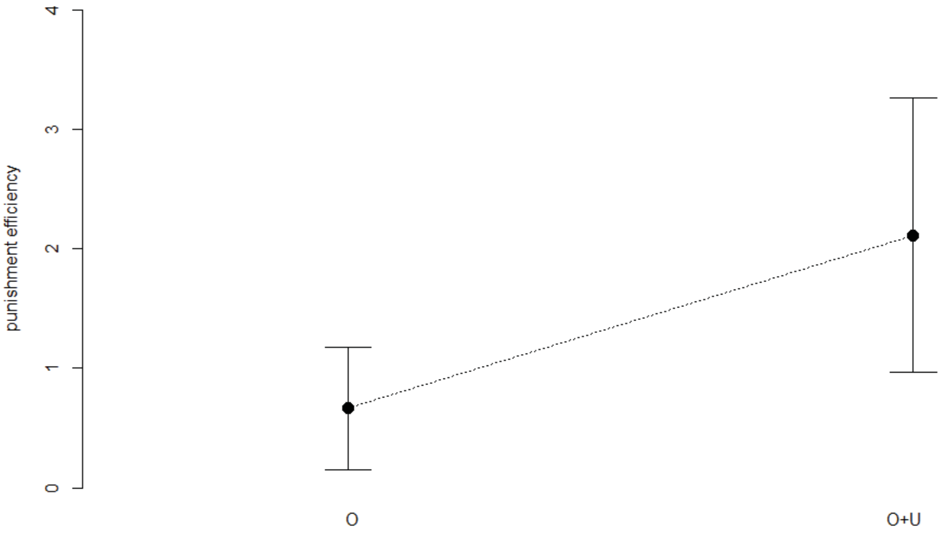


**Figure 3.** Average sanctioning effectiveness of immediate feedback points (y-axis) indicating the average effect of one point of punishment on contributions of the participant receiving punishment in the subsequent round in the *O* and the *O+U* treatments, respectively. Bars indicate the 95% confidence intervals.

PAYOFF EFFICIENCY: Given our previous findings, there is no surprise that we find superior efficiency in the *O+U* treatment. Figure 4 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 (comparing *O+U* to *O*, *U* and *VCM*: *p*=0.02, *p*=0.01, and *p*=0.03), while, compared to treatment *VCM*, both sanctioning mechanisms in isolation do not lead to better efficiency rates (comparing *VCM* to *O and U*: *p*=0.27 and *p*=0.12).[10] 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.

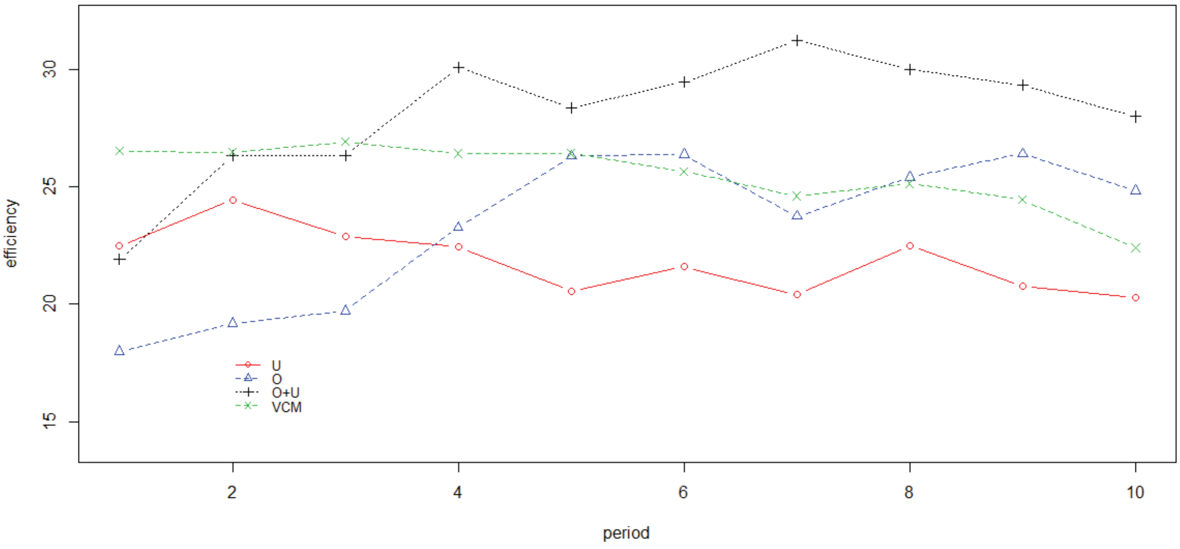


**Figure 4.** Average group efficiency over periods and treatment conditions *U, O, O+U*, and the standard voluntary contribution game without punishment (*VCM*).

## 5. Discussion

Today, a growing number of public goods and natural resources are successfully governed via decentralized approaches. These decentralized approaches heavily rely on mutual monitoring and involve sanctioning of non-cooperative behavior. Anecdotic evidence, however, suggests that actual sanctions are an exception rather than a usual practice (Ostrom, 1990). This observation

[10] 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).

contrasts usual findings of laboratory experiments on decentralized approaches, which show that there are substantial acts of punishment. Moreover, punishment in the lab frequently results in lower payoff efficiencies; that is, having expenses for punishment eating up all efficiency gains generated by the training of non-cooperative group members. To resolve the seeming discrepancy between lab and field evidence, we test for the importance of having different punishment channels available at the same time: in particular the combination of social sanctions with immediate feedback and unobservable sanctions with delayed feedback.

Indeed, our findings demonstrate that the co-existence of both potential threats leads to higher cooperation and yields strong payoff efficiency gains. The mere existence of unobserved sanctions with delayed feedback more than triples the sanctioning effectiveness of immediately observed punishment, while at the same time overall sanctioning expenditures are significantly reduced. Adding the Damocles sword of unobserved punishment enhances cooperation and group efficiency without increasing punishment per se. Due to its appealing simplicity and practicability, the combination of observed and unobserved punishment is likely to serve as an important element for decentralized approaches. The multiplicity of punishment channels seem to stabilize cooperation efficiently and provide another crucial argument in favor of decentralized mechanisms that build on mutual monitoring and social sanctions.

While not the main focus of our study, our findings are also of interest for the ever-growing research on the "dark side" of social punishment, namely on anti-social punishment (e.g., Cinyabuguma et al., 2006, Denant-Boemont, 2007, Nicklisch & Wolff, 2011, Kamei & Putterman, 2013). While punishment of cooperative players is observed in many studies and is particularly pronounced in specific cultures (Herrmann et al., 2008), there is still an ongoing debate about the channels underlying this phenomenon. Some have argued that it is mere blind revenge (e.g., Ostrom et al., 1992), but it could also be driven by a taste for increasing payoff differences (e.g., Fehr and Schmidt 1999), or be meant strategically to prevent pro-social punishers from sanctioning free-riders in subsequent periods (e.g., Nikiforakis, 2008). Depending on the channel, subjects in our treatment with both punishment mechanisms being jointly available should either tend to use immediate or delayed sanctions for their anti-social punishment. Yet, we do not observe that one channel is preferred over the other for anti-social punishment, indicating that the driving forces behind anti-social punishment are heterogeneous. Furthermore, the possibility to use two channels of punishment does not increase the total amount of anti-social punishment.

Compared with data from other societies, however, there are only few incidences of anti-social punishment overall. It might thus be interesting to test the decentralized approach to governing the commons via multiple channels of social sanctions in societies where the thread of anti-social punishment is more severe. Speaking of possible extensions, future studies might also look at situations where the mixture of social sanctions is composed of punishment mechanisms with different “technologies”. In our study, both forms of punishment reduced the other’s payoff by three tokens per token invested, but one might study environments where immediate sanctions are more expensive than delayed sanctions, or vice versa. This would allow us to test if demand for punishment is shifted toward the more cost effective punishment channel, and if so, how this affects the sanctioning effectiveness of immediate punishment. Along similar lines, it would also be interesting to see how the sanctioning effectiveness develops when subjects play the game repeatedly, that is, receive feedback on the delayed sanctions in between. Maybe in that case, the mere ambiguous threat of unobserved punishment has even stronger positive effects for society by increasing cooperation since the thread is sometimes resolved and players who did not acknowledge the threat in the first now do take care of it and start cooperating. We leave those important questions open to future studies and invite other researchers to follow us along this track of research.

## References

Agrawal, A., & Ostrom, E. (2001). Collective action, property rights, and decentralization in resource use in India and Nepal. *Politics & Society*, 29, 485-514.

Amacher, G., Ollikainen, M., & Koskela, E. (2012). Corruption and Forest Concessions. *Journal of Environmental Economics and Management*, 63, 92-104.

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin,* 137, 594-615.

Cason, T. & Gangadharan, L. (2013). Empowering neighbors versus imposing regulations: An experimental analysis of pollution reduction schemes. *Journal of Environmental Economics and Management*, 65, 469-484.

Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, 14, 47-83.

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Can second-order punishment deter perverse punishment? *Experimental Economics*, 9, 265-279.

Cox, M. & Ross, J. (2011). Robustness and vulnerability of community irrigation systems: The case of the Taos Valley Acequias. *Journal of Environmental Economics and Management*, 61, 254-266.

Denant-Boemont, L., Masclet, D., & Noussair, C. N. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory*, 33, 145-167.

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

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

Fehr, E. & Schmidt, K. (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics*, 114, 817-868.

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

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

Gächter, S., Renner, E., & Sefton, M. (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 Datenverarbeitung Göttingen, 63, 79–93.

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

Kamei, K., & Putterman, L. (2013). In broad daylight: Fuller information and higher-order punishment opportunities can promote cooperation. *Working paper.*

Ledyard, J. (1995). Public goods: A survey of experimental research. In: Kagel, J. & Roth, A. (Eds.), *Handbook of Experimental Economics*. Princeton University Press, 111-194.

Nicklisch, A., & Wolff, I. (2011). Cooperation norms in multiple-stage punishment. *Journal of Public Economic Theory*, 13, 791-827.

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

Ostrom, E. (1990). *Governing the commons: The evolution of institutions for collective action.* Cambridge University Press.

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: Self-governance is possible. *The American Political Science Review*, 404-417.

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

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology,* 51, 110-116.

Zelmer, J. (2003). Linear public goods experiments: A meta-analysis. *Experimental Economics*, 6, 299-310.

## Appendix: English translation of the German instructions for the *O+U* condition[11]

*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.

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

[11] 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.

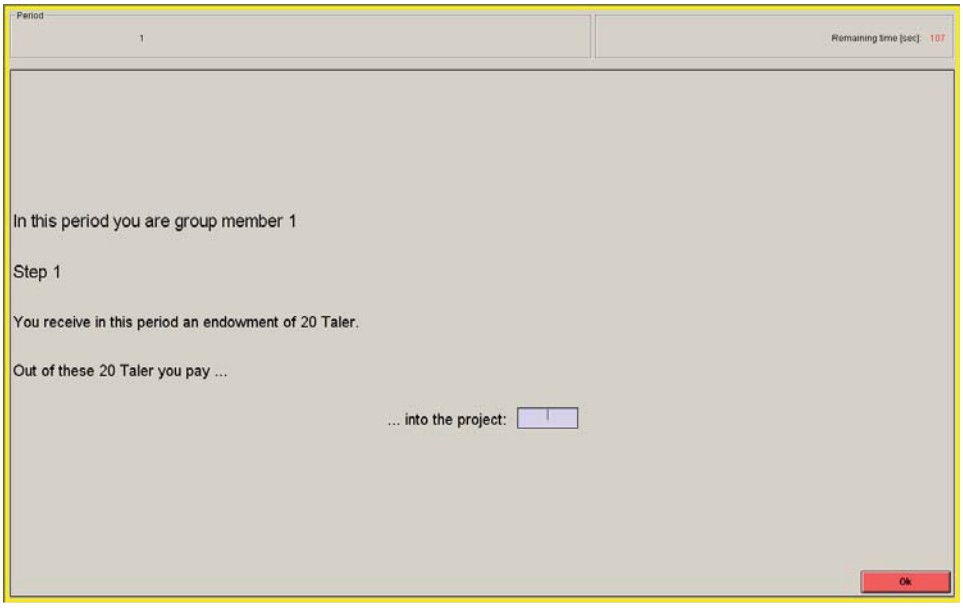
members can therefore not be identified as the periods progress. Each participant will receive from us 50 Taler, with which possible losses can be counterbalanced. The following pages outline the exact procedure of the experiment.

*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:



Period 1

Remaining time [sec]: 107

In this period you are group member 1

Step 1

You receive in this period an endowment of 20 Taler.

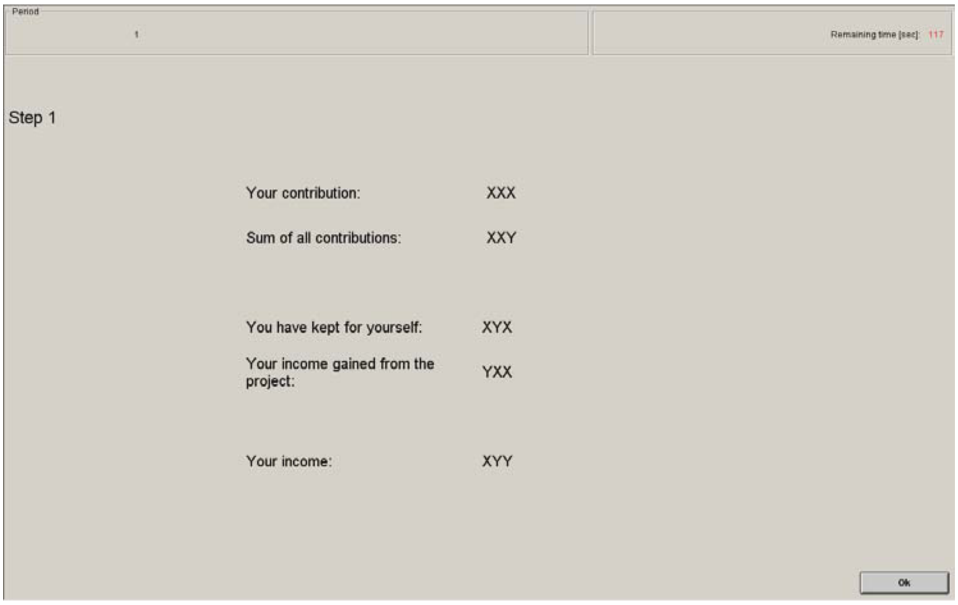
Out of these 20 Taler you pay ...

... 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.

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.



Period

1

Remaining time [sec]: 117

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 |

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)

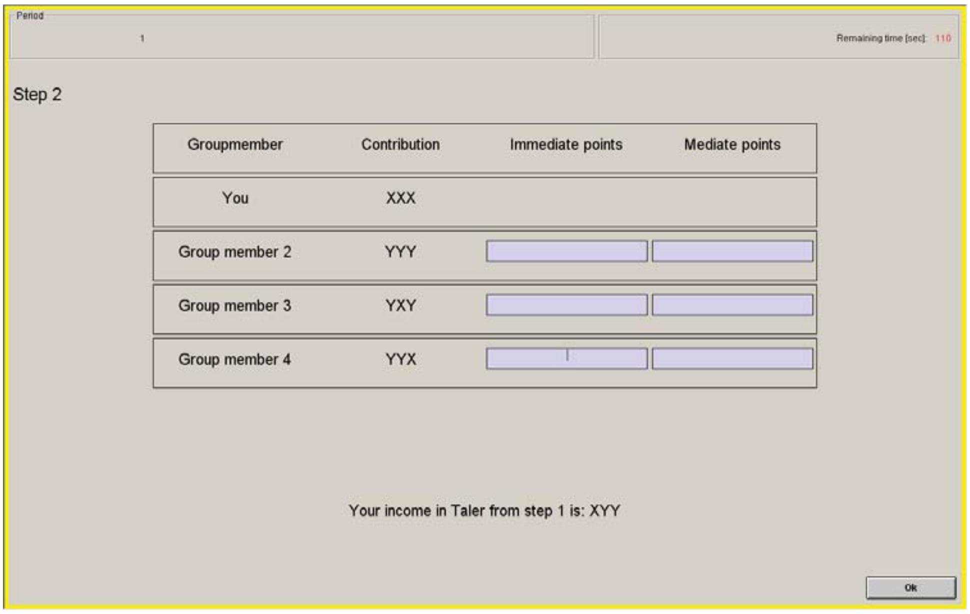
The total income at the end of Step 1, in Taler, is calculated according to the same formula for each member of the group. If, for example, the sum of the contributions from all group members adds up to 60 Taler, you and all other members each receive a project income of .4× 60 = 24 Taler. If the group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

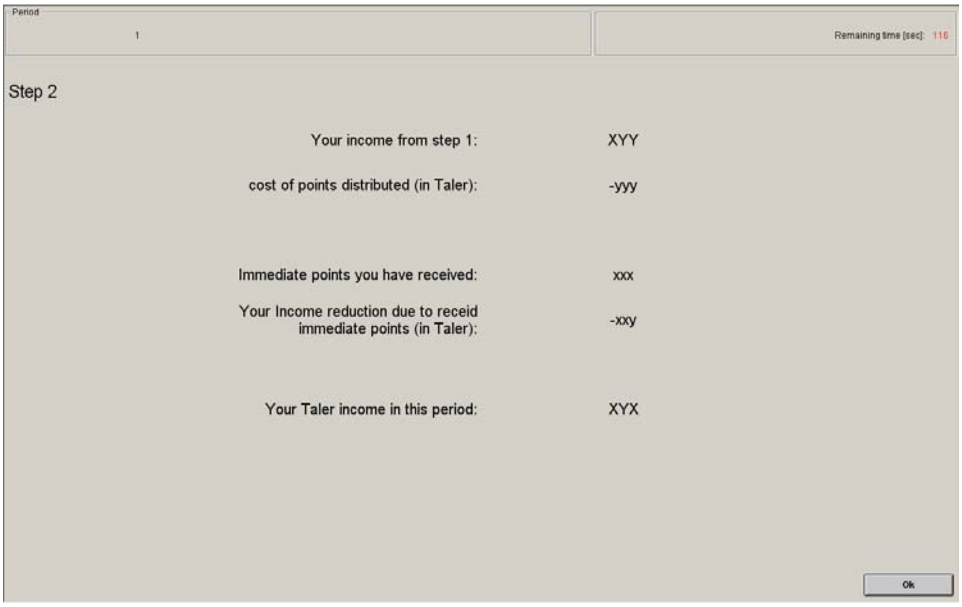
By how much a group member's income from Step 1 is reduced overall depends on the total number of points received. If, for instance, one member receives a total of 3 points from all other members, the income in Step 1 is reduced by 9 Taler. If a member receives a total of 4 points, the income in step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



Period

1

Remaining time [sec]: 116

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?