

# The evolution of sanctioning institutions: an experimental approach to the social contract

Boyu Zhang <sup>a,b</sup>, Cong Li <sup>c</sup>, Hannelore De Silva <sup>d</sup>, Peter Bednarik <sup>e</sup> and Karl Sigmund <sup>b,f,\*</sup>

<sup>a</sup> School of Mathematical Sciences, Beijing Normal University, 100875 Beijing, China.

<sup>b</sup> Faculty of Mathematics, University of Vienna, 1090 Vienna, Austria.

<sup>c</sup> Key Laboratory of Animal Ecology and Conservation Biology, Centre for Computational Biology and Evolution, Institute of Zoology, Chinese Academy of Sciences, Beijing 100101, P. R. China.

<sup>d</sup> Department of Finance, Accounting and Statistics, Vienna University for Economics and Business, 1190 Wien, Austria.

<sup>e</sup> Courant Research Center Evolution of Social Behavior, University of Göttingen, D-37073 Göttingen, Germany.

<sup>f</sup> International Institute for Applied Systems Analysis (IIASA), A-2361 Laxenburg, Austria.

\*Author for correspondence: Telephone: +43 (0)1427750612

Fax: +43(0)142779506

E-mail: [karl.sigmund@univie.ac.at](mailto:karl.sigmund@univie.ac.at)

## Abstract

*A vast amount of empirical and theoretical research on public good games indicates that the threat of punishment can curb free-riding in human groups engaged in joint enterprises. Since punishment is often costly, however, this raises an issue of second-order free-riding: indeed, the sanctioning system itself is a common good which can be exploited. Most investigations, so far, considered peer punishment: players could impose fines on those who exploited them, at a cost to themselves. Only a minority considered so-called pool punishment. In this scenario, players contribute to a punishment pool before engaging in the joint enterprise, and without knowing who the free-riders will be. Theoretical investigations (Sigmund et al. 2010) have shown that peer punishment is more efficient, but pool punishment more stable. Social learning, i.e., the preferential imitation of successful strategies, should lead to pool punishment if sanctions are also imposed on second-order free-riders, but to peer punishment if they are not. Here we describe an economic experiment (the Mutual Aid game) which tests this prediction. We find that pool punishment only emerges if second-order free riders are punished, but that peer punishment is more stable than expected. Basically, our experiment shows that social learning can lead to a spontaneously emerging social contract, based on a sanctioning institution to overcome the free rider problem.*

## Keywords

*Public goods, experiments, collective action, punishment, institution, social learning.*

## JEL codes

*C73, C91, D03*

# 1. Introduction

The role of punishment in boosting cooperation is a well-studied topic in experimental economics. However, most investigations deal with so-called peer-punishment (see, e.g., Fehr and Gächter 2000, 2002; Fehr and Rockenbach 2003; Fowler 2005; Henrich et al. 2006; Casari 2007; Sigmund 2007; Carpenter 2007; Egas and Riedl 2008; Gächter et al. 2008; Chaudhuri 2011). Typically, the players in a public good game are allowed to impose fines on others, at a cost to themselves. The threat of punishment can lead to considerable increases in the level of cooperation in the collective action. Many players are willing (and frequently even eager) to shoulder the costs of imposing fines on cheaters. Some subjects, however, abuse sanctioning opportunities by engaging in antisocial punishment which harms cooperators (Cinyabuguma et al. 2006; Denant-Boemont et al. 2007; Dreber et al. 2008; Hermann et al. 2008; Nikiforakis 2008).

In most aspects of everyday life, the task of punishing exploiters has eventually been taken over by institutions (Ostrom 2005; Guala 2012). In developed societies, peer-punishment is not only less common (Balafoutas and Nikiforakis 2012), but often explicitly forbidden. Under conditions of anarchy (e.g., Hobbes's 'state of nature'), individuals have to take punishment into their own hands, but in all better-regulated communities, individuals engage in some form of social contract by delegating punishment to institutions. How can we envisage this important step in social development?

Evidently, this question can be approached from many different angles. We use an economic experiment to test how individuals who want to coerce their group to cooperate decide between inflicting pro-social punishment directly or using the intermediary of an institution. We do not offer them opportunities for antisocial punishment, or the bribery of institutions, but rather concentrate on the idealized versions. The foremost problem, in such an experiment, is how to implement the sanctioning institution (Tyler and Degoey 1995; Casari and Luini 2009; Kosfeld et al. 2009; Andreoni and Gee 2012). Which is the essential feature distinguishing institutional from peer-punishment? Some argue that it is the delegation of punishment. However, individuals who want to exert personal revenge can recur to 'hiring a gun', and this would still count as peer-punishment (Van Vugt et al. 2009). A more pronounced difference is that sanctioning institutions are established in advance, and thus entail running costs even in the case that no one commits a punishable offense. A county would have to pay its sheriff even if nobody commits a crime. We tried to model this as 'pool-punishment' (for experimental papers, see Yamagishi 1986; Guillen et al. 2006; Kamei

et al. 2011; Markussen et al. 2011; Traulsen et al. 2012; for theory, see Sigmund et al. 2010, 2011). Players who want to use such a sanctioning tool have to pay a fee, even before the joint enterprise takes place, or at least before they are informed of its outcome, and thus before they know whether there will be any exploiters to punish. Pool punishers can be viewed as paying a tax towards a police. We note that instead of pool- or peer-punishment, some authors use the terms ‘formal’ and ‘informal’ sanctions (Kamei et al. 2011; Markussen et al. 2011).

In our experiment, we investigated 18 small groups, or ‘toy-communities’, of 12 to 14 players. Each such group played 50 rounds of a Mutual Aid game, isolated from the other groups. The Mutual Aid Game is a variant of a Public Goods game, where players do not obtain any return from their own contribution, and hence are faced with an even more pronounced social dilemma. Within each group, players could decide, before each round, whether to join a Mutual Aid game without punishment (NoPun), with peer-punishment (Peer), with pool-punishment (Pool) or not to participate in the game at all (No). These games were played separately, i.e., the outcome of one game did not affect the outcomes of the other games in the group. Players were anonymous, and prevented from communicating. All that players learned, after each round, was how many opted, in their group, for each alternative, and which payoff they obtained. They then could choose whether to opt for (NoPun), (Peer), (Pool) or (No) in the next round. We thus observed, in each toy community, whether social learning led to institutional punishment or not.

It is clear that if pro-social punishment works as desired, i.e., if it leads to all-out cooperation, then peer- punishment is more efficient than pool-punishment, since it entails no running costs. However, theoretical considerations (Sigmund et al. 2010, see relevant theory in section 2) imply that pool-punishment is more stable, provided that it is also directed at those participants in the game who do not contribute to the punishment pool. Indeed, if cooperation is achieved, i.e., if no one needs to be punished, then a peer-punisher cannot be distinguished from a non-punisher. This means that second-order free-riders (defined as those who contribute to the Mutual Aid, but not to the sanctions) cannot be spotted, and thus cannot be punished. By contrast, those who do not contribute to the punishment pool are just as visible as those who do not contribute to the Mutual Aid, and can be punished just as well. A system implementing this is highly immune against exploitation, but requires payment of some sort of a tax to maintain the punishment pool. We wanted to compare the attractiveness of first-order and second-order pool punishment against the background of the same peer punishment and no punishment alternatives.

In our experiment, a clear majority chose peer punishment in the first round. Most players switched to pool punishment in later rounds, but (as predicted by theory) only if punishment was also imposed on second-order free-riders. The experiment involved 238 first-year students from universities in Vienna. Interactions were anonymous. Players were randomly allocated to 18 groups of 12 to 14 players each, for the duration of 50 rounds. We implemented 2 treatments with 9 groups each: in the ‘second-order treatment’, players were offered a pool punishment game which sanctioned second-order free riders, and in the ‘first-order treatment’ a pool punishment game which did not. The former treatment led to the emergence of pool punishment in six out of the nine groups, the latter in none. Peer punishment slowly declined over rounds in both treatments. Roughly speaking, it was not displaced by pool punishment, but eroded gradually. Contributions to the Mutual Aid game were vastly more frequent in the treatment with second-order pool punishment.

In a nutshell, players were allowed to ‘vote with their feet’ (the expression seems to be due to Tiebout 1956), and most of them decided in favor of a sanctioning institution, but only if this institution coerced participants to contribute not merely to the Mutual Aid, but also to its own upkeep. Under this additional commitment, the institution was adopted by the group, in a kind of ‘social contract’ which was achieved without explicit communication or deliberation, and was based on social learning from the own experience and that of others.

In section 2, we describe the theoretical background, and in section 3, the experiment. In section 4, we display the results, and in section 5, we offer a discussion and conclusions. The instructions for the players and the detailed results of every group are contained in the supplementary information of the Online Resource.

## 2. Theoretical background: a choice of games

In this section, we introduce the types of games used in the experiment and briefly sketch some of the relevant theory from Sigmund et al. (2010).

First of all, let us consider the Mutual Aid game (MA game) of type (NoPun) (no punishment). There are  $m$  players in the group. They can decide whether or not to contribute an amount  $c > 0$ , knowing that this will be multiplied by  $r > 1$  and divided among all *other* players in the group. If  $m_C$  is the number of those players who contribute, and  $m_D$  the number of those who don’t (with  $m_C + m_D = m$ ), then the payoff for a contributor is

$$P^C = rc \frac{m_C - 1}{m - 1} - c$$

and that for a defector

$$P^D = rc \frac{m_C}{m - 1}.$$

Clearly, we always have  $P^D > P^C$  (the difference is independent of  $m_C$ ) and thus the dominant strategy is to refuse to contribute. If all players contribute, their payoff is  $(r-1)c$ , which is independent of group size  $m$ . In our experiment,  $c=1$  monetary unit (MU),  $r=3$  and  $m \geq 2$  is variable. This game was first described in Wilson (1975), see also Sugden (1986), Yamagishi (1986), Fletcher and Zwick (2004), Sigmund (2010). It is very similar to the usual public good game (PG game), see e.g. Fehr and Gächter (2000). Whereas in the latter, an amount  $r/m$  of a player's contribution returns to the player, in the MA game players do not obtain any return from their own contribution. In Sigmund et al. 2010, 2011, the Mutual Aid game has been called the 'others only' variant of the Public Goods game. In our setup, the number of participants  $m$  can fluctuate. If  $m$  is smaller than  $r$ , the PG game is no longer a social dilemma: to contribute is the dominant strategy. Most PG games do explore the case when  $m < r$ , but we wanted to make it more difficult for cooperation to emerge: for the MA game, defection is always the dominant strategy. Note that if everyone contributes, the payoff is the same, namely  $(r-1)c$ , in both the MA and the PG game.

Now let us consider the MA game of type (Peer) (peer punishment). After deciding whether to contribute or not to the Mutual Aid in an MA game, players can punish defectors. We assume that only players who contribute can punish. If they do, they have to punish all defectors in the group. As the focus of our study is the decision between different types of pro-social punishment, this assumption makes the issue as clear-cut as possible.

Let us suppose that  $m_{Pe}$  is the number of players who contribute and punish those who do not contribute,  $m_C$  is the number of players who contribute, but do not punish, and  $m_D$  is the number of those who neither contribute nor punish, i.e., the defectors or free-riders. Thus  $m_{Pe} + m_C + m_D = m$ . Every peer punisher has to punish every free rider. Let  $\beta$  be the size of the fine that each non-contributor has to pay *per punisher*, and  $\gamma$  the fee each punisher has to pay for each non-contributor he or she punishes. The cost of punishment is uncertain as it depends on the number  $m_D$  of free-riders, but the fine-to-fee ratio is fixed to  $\beta: \gamma$ . Then we obtain as payoff values

$$\begin{aligned}
P^C &= rc \frac{m_C + m_{Pe} - 1}{m - 1} - c \\
P^{Pe} &= rc \frac{m_C + m_{Pe} - 1}{m - 1} - c - \gamma m_D \\
P^D &= rc \frac{m_C + m_{Pe}}{m - 1} - \beta m_{Pe}
\end{aligned}$$

There is no dominant strategy. The group optimum is obtained whenever  $m_D = 0$ . In this case, every player obtains  $(r-1)c$ , just as in the MA game of type (NoPun). Clearly, we have  $P^C \geq P^{Pe}$  (with equality if and only if  $m_D = 0$ ). The state when no one contributes is a strict Nash equilibrium. Other (non-strict) equilibria exist for  $m_D = 0$  and  $m_{Pe} \geq (c+\beta)/\beta$ . In our experiment,  $\beta = 1$  MU and  $\gamma = 0.5$  MU so that states with two or more peer punishers, but no defector are also Nash equilibria.

Finally, let us consider the MA game of type (Pool) (pool punishment). Just as in the (Peer) game, we assume that only contributors can punish. At the same time that they decide whether to contribute to the Mutual Aid or not, they also decide whether to contribute to the punishment pool. There are  $m_C$  players who contribute to the Mutual Aid, but not to the punishment pool,  $m_{Po}$  players who contribute to both pools, and  $m_D$  players who contribute to neither pool (with  $m_{Po} + m_C + m_D = m$ ). Pool punishers have to contribute an amount  $c$  to the Mutual Aid and an amount  $F$  to the punishment pool. The fine-to-fee ratio depends on the number of free-riders.

The (Pool) game is played in two variants. In the first-order variant, everyone who does not contribute to the Mutual Aid is fined by an amount  $Bm_{Po}$ , whereas in the second-order variant, everyone who does not contribute to the punishment pool (i.e., who does not contribute to both Mutual Aid and the sanctioning) is fined by that amount. The payoff values are

$$P^{Po} = rc \frac{m_C + m_{Po} - 1}{m - 1} - c - F$$

and in the first-order variant (Pool1)

$$\begin{aligned}
P^C &= rc \frac{m_C + m_{Po} - 1}{m - 1} - c \\
P^D &= rc \frac{m_C + m_{Po}}{m - 1} - Bm_{Po}
\end{aligned}$$

resp. in the second-order variant (Pool2)

$$P^C = rc \frac{m_C + m_{P_o} - 1}{m - 1} - c - Bm_{P_o}$$

$$P^D = rc \frac{m_C + m_{P_o}}{m - 1} - Bm_{P_o}$$

In our experiment, we used  $B=1$  MU and  $F=0.5$  MU. In the first-order variant, we have  $P^C > P^{P_o}$  so that  $m_D = m$  is the only Nash equilibrium. In the second-order variant,  $m_{P_o} = m$  is another equilibrium (as long as  $c + F \leq B(m-1)$ ), which for our parameter values means that there are at least three punishers). This equilibrium is not efficient (i.e., pareto-optimal), since  $m_C = m$  provides a higher per capita payoff and actually is the group optimum.

In each of these games, the all-defector state is a strict Nash equilibrium. In the context of evolutionary games, it represents a social trap. How, then, could cooperation emerge? In Sigmund et al. 2010, it was shown that if the game is optional, i.e., if players have also the possibility of abstaining from it, then cooperation based on peer- or pool punishment can emerge. Indeed, the participation in a Mutual Aid game (or a public good game) can be viewed as a risky enterprise, which only succeeds if enough co-players cooperate.

Non-participants can be viewed as risk-averse players who, rather than engage in such an uncertain interaction, prefer to engage in some other activity whose payoff  $\sigma$  does not depend on what the others are doing. We assume that this payoff is somewhere between the payoff obtained if no one contributes to the MA game, and that obtained if everyone contributes to the common pool (and, in the Pool game, to the punishment pool). This means  $0 < \sigma < (r-1)c$  and, in the Pool game,  $0 < \sigma < (r-1)c - F$ . Whenever defectors thrive and cooperation breaks down, the option of abstaining from the game yields a higher payoff and non-participants take over. This in turn gives players willing to invest in the Mutual Aid game the chance to re-establish cooperation. The possibility of abstaining from the game therefore offers an escape from the social trap. In the theoretical model of Sigmund et al. 2010, free riders took over if participation was not optional, but compulsory.

For the competition between peer-punishment and pool-punishment, Sigmund et al (2010, 2011) show, using arguments from evolutionary game theory, that in the first-order variant, peer-punishers prevail most of the time, but sometimes second-order free-riders invade. In this case, defectors and then non-participants take over before peer-punishment is reestablished. In contrast, in the second-order version, pool-punishers eventually establish a very stable regime, although it is less efficient.

### 3. The experiment

The 18 groups of 12 to 14 players (our ‘toy-communities’) were the independent sample points of our experiment. Players in different groups were not allowed to communicate with each other and interacted only within their group. The players were not told that the number of rounds was fixed beforehand at 50, so as to prevent end-round effects. In each round, players were given 3 MU and asked to choose one of three variants of the Mutual Aid (MA) games: (NoPun) MA without punishment; (Peer) MA with peer punishment; (Pool) MA with pool punishment. The players could also decide (No) not to participate in any of these games. Such non-participants received an additional 0.5 MU. The idea, here, was that when not participating in a joint enterprise, an individual can engage in some useful activity which does not depend on the decisions of others.

Once players had chosen one of these games, they played one round of the game they had chosen with those group-members who had chosen the same game. Players who opted for one of the games (NoPun), (Peer) or (Pool), but found no co-players to join them, were treated as non-participants (No), and received an additional 0.5 MU, independently of what the others did. Once the round was over, the players learned how many (in their group) had played (NoPun), (Peer), (Pool) or (No), how many in each game had chosen which strategy, and which payoff they had obtained. They could use this information to decide for which game to opt in the next round. Players did not learn about who did what, so there was no possibility to establish a reputation. Players knew that they would be paid immediately after the experiment, at a rate of 10 cents (euro) per MU, without having to give away their identity (as players) to their co-players or to the experimenters. The guaranteed minimal payoff was 10 euro.

Players participating in a MA game of type (NoPun) could decide whether or not to contribute 1 MU to the common pool, knowing that their contribution would be multiplied by 3 and divided equally among all *other* players in their game, irrespective of whether these co-players had contributed or not. Thus contributors did not benefit from their own contribution. If all cooperate, everyone gains 2 MU.

Players choosing to participate in an MA game of type (Peer) would first play an MA game as described above, and then, in a second stage of the same round, be shown the number of non-contributors (i.e., defectors) in their game. Contributors could then decide whether or not to punish these free-riders. The fine-to-fee ratio is fixed to 2:1 in (Peer), as used in Carpenter (2007) and Nikiforakis and Normann (2008). Each punisher would have to pay a



fee of 0.5 MU per defector, and each defector would have to pay a fine of 1 MU per punisher. Again, if all cooperate, everyone gains 2 MU.

Players participating in a MA game of type (Pool) had to choose between three options: (i) not to contribute anything, (ii) to contribute to the Mutual Aid (i.e., to pay 1 MU so that 3 MU would be shared among all other members who had chosen (Pool)), or (iii) to contribute to *both* the Mutual Aid and the punishment pool. This last alternative requires the players to pay 1 MU to the Mutual Aid and an additional 0.5 MU into the punishment pool. Thus if all cooperate, everyone gains 1.5 MU. This MA game was played in two variants, denoted as ‘first-order variant’ resp. ‘second-order variant’. In the first-order variant, players knew that everyone who had not contributed to the Mutual Aid would be fined 1MU per punisher. In the second-order variant, players knew that everyone who had not contributed to *both* Mutual Aid and punishment pool would be fined 1MU per punisher. Hence, in the second-order variant of game (Pool), the institution punishes all of these free-riders (i.e., those who opted for (i) or (ii)) irrespective of whether they contributed or not to the Mutual Aid game), while in the first-order variant (Pool1), second-order free-riders (those who opted for (ii)) were not punished. The fine to fee ratio can greatly vary, in this game, depending on the number of defectors and pool punishers. In groups 1-9 (with altogether 120 subjects), the game of type (Pool) was offered in the first-order variant, and in groups 10-18 (with 118 subjects) in the second-order variant.

We note that this is a complex game, without obvious money-maximizing strategies for the individuals choosing (Peer) and (Pool), since payoff depends on how many decide for the different options. In order to provide the players with an appreciation of the issues involved, they were given, at the start of the session, 25 practice rounds (see Online Resource). They knew that these rounds would not count towards their score and that groups would be reshuffled before the experiment started. More precisely, players were first given, via computer screen, a brief introduction into game (NoPun), then played five rounds of the game (NoPun). The same then happened with games (Peer) and (Pool). Finally, they all played 10 rounds with the option, in each round, to choose between the three games (NoPun), (Peer) (Pool), or (No), which meant to abstain from participation (exactly as later in the actual experiment). Thus players could familiarize themselves with their options, in the practice rounds, but were precluded from sharing their experiences through communication. Immediately after the practice rounds, the ‘toy communities’ were re-assembled randomly.

After each round, players were shown the payoffs for all strategies used in their group, and had 15 seconds to decide which game (NoPun), (Peer), (Pool) or (No) to join next. The tightness of the schedule and the complexity of the task provided a strong motivation to be guided by the size of the payoffs, i.e., to engage in social learning. We also did use loaded language in the instructions, for instance by calling punishment ‘punishment’. Since our main aim was to compare different pro-social sanctioning mechanisms, we felt justified in acknowledging the underlying, common intention to uphold norms of collaboration. In the same spirit, asocial punishment or revenge were not offered as options to our players. Moreover, to reduce the complexity, we avoided the issue of increasing or decreasing group returns (i.e., we assumed that the Mutual Aid was proportional to the number of contributors). Furthermore, we considered only pure strategies. Players could either make a full contribution or none, and either punish all exploiters or none. This need not correspond to real-life situations, but rather aims at eliciting clear responses from the candidates facing choices between pro-social punishment mechanisms. Admittedly, these experimental strictures limit the generalizability of the results.

## 4. Results: social learning of social control

In the actual experiment, we observed strong changes in behavior in most of the 18 groups, especially during the initial phase. 12 of the groups eventually settled down, in the sense that the same game was chosen by the majority for each of the last 10 rounds. Six of these groups settled down for pool punishment. All six belonged to the second-order treatment. In three groups playing the second-order treatment, and three groups playing the first-order treatment, players settled for peer punishment. The null hypothesis that pool punishment is equally likely in both treatments can be rejected with a significance of  $p < 0.05$  ( $n_1=9$ ,  $n_2=9$ , two-sided binomial sample test). Based on the theoretical model, we had indeed expected pool punishment to emerge in the second-order treatment only.

The average frequency of pool punishment increased during the first rounds, in the second-order treatment, and overtook the frequency of peer punishment. In fact, the initial frequencies of (NoPun), (Peer), (Pool) and (No), in the first-order treatment, corresponded closely to the initial frequencies in the second-order treatment, but then the frequencies evolved very differently (see Figure 1). Frequencies of peer punishers decreased in both treatments, but only slowly. Frequencies of pool punishment decreased in the first-order

treatment, but increased in the second-order treatment. (See Online Resources for the regression equations)

More precisely, in the first round of the second-order treatment, 55 per cent of players choose the peer punishment game and 36 per cent the pool punishment game. The initial frequencies in the first-order version were 56 per cent and 31 per cent, respectively. However, in the first-order treatment, both frequencies declined to reach 48 per cent and 19 per cent, respectively, by round 50. By contrast, the evolution in the second-order treatment reversed frequencies, so that after 50 rounds, 63 per cent of players opted for the pool punishment game but only 33 per cent for the peer punishment game (Figure 1b). This reversal took place in the first 20 rounds. The regression equation is  $y=0.326+0.0146x$  (where  $y$  represents the frequency of pool-punishment and  $x$  the round), with coefficient of determination  $R^2=0.9167$  (which measures how well the regression line represents the data) and  $P\text{-value}<0.0001$ . Obviously, players approached both first- and second-order treatments with similar expectations, but then underwent a very different learning experience.

If we average over all 50 rounds, we find a significant preference for peer punishment in the first-order treatment, and a less significant preference for pool punishment in the second-order treatment (Figure 3a). The latter treatment leads to a very pronounced cooperative behavior. Indeed, the frequency of contributions was significantly higher in the second-order treatment than in the first-order treatment (88.2 per cent vs. 48.9 per cent, Mann-Whitney U-test,  $n_1=9$ ,  $n_2=9$ ,  $p=0.0373$ ), and it hardly changed over the 50 rounds (Figure 2b). We can see (Figure 3c and Online Resource) that average payoff values differ by little, but that peer punishment clearly yields the highest payoff in the first-order treatment, whereas it shares front rank with pool punishment, in the second-order treatment.

In the first-order treatment, peer punishment was preferred by a wide margin: game (Peer) was chosen in 55.6 per cent of all decisions, game (Pool) in 20.2 per cent, game (NoPun) in 11.7 per cent and non-participation (No) in 12.6 per cent (Figure 3a). A majority (62 per cent) of decisions in the peer punishment game (Peer) were to contribute to the Mutual Aid, but not to punish, only 12 percent were for punishment. (In both treatments, the average number of peer punishers in the peer games is decreasing with the number of defectors. See Table S4 in Online Resource). The payoff of these first-order defectors was higher than that of the punishers (4.636 MU vs. 4.1 MU, Mann-Whitney U-test,  $n_1=9$ ,  $n_2=9$ ,  $p=0.077$ ). It is obvious that within any round, this has to hold, if some players defect; we see here that it also holds on average. The non-contributors in the peer punishment game earned

marginally more than the non-participants (3.61 MU vs 3.5 MU, the difference is not significant). All in all, 48.9 per cent of all decisions were in favor of contributing to the Mutual Aid, rather than defecting (35.6 per cent) or abstaining from the game (15.5 per cent). But the time evolution over 50 rounds tells a more pessimistic story (Figure 2a). Three-fourth of players cooperated in the first round but half of them gave up in later rounds. The regression equation is  $y=0.669-0.0065x$  (where  $y$  represents the frequency of cooperation and  $x$  the round), with coefficient of determination  $R^2=0.8812$  and P-value  $<0.0001$ . Moreover, in the first-order pool punishment games, neither contributions to the Mutual Aid nor to the sanctioning took off. In particular, only a tiny fraction of the decisions (54 out of 1149) favored investing into the punishment pool.

In the second-order treatment, the preferences for the games change drastically (the hypothesis that the preferences for the four games are the same can be rejected, using a Chi-square test, P-value $<0.0001$ ). The game (Pool), was chosen in 54.1 per cent of all decisions, and almost always (namely, in 3155 of 3174 cases) was combined with a decision to actually contribute to the punishment pool. The peer punishment game (Peer) was chosen in 41 per cent of the decisions. Interestingly, players who chose the peer punishment game rarely decided to actually punish (only 9 per cent did), and the average payoff for those who actually engaged in peer punishment, 3.78 MU, was significantly less than that of second-order free-riders (4.77 MU, Mann-Whitney U-test,  $n_1=9$ ,  $n_2=9$ ,  $p=0.0106$ ). But this minority of peer-punishers sufficed to keep free-riding in the Mutual Aid game down to 16 per cent. Few decisions (4.5 per cent) were in favor of the alternative (NoPun), i.e., joining a Mutual Aid game without punishment, and only 0.4 per cent were in (No). The average payoff for the peer punishment game (Peer) was insignificantly larger than for the pool-punishment game (Pool) (4.49 MU vs. 4.46 MU), but those who actually peer-punished had a significantly lower payoff than those who actually contributed to the punishment pool (3.78 MU vs. 4.49 MU, Mann-Whitney U-test,  $n_1=9$ ,  $n_2=9$ ,  $p=0.004$ ).

The average payoff for those choosing a given game is almost the same for both first order and second order treatments, with one exception: the payoff for choosing pool punishment has substantially increased in the second-order treatment, because almost all players contributed to the Mutual Aid in the second-order treatment, but less than a third did so in the first-order version (Figure 3c).

It remains to show that players, in this experiment, were essentially guided by social learning, defined here as preferential copying of the strategies with the highest payoff

observed among other players in the same group. There were 2850 decisions to switch to another strategy. Of these, 2012 decisions (70.6 percent) were in favor of a strategy having currently a strictly better payoff in the group. There were 7446 decisions not to switch, but to repeat the former move. In 76.1 percent of the cases, the payoff was optimal. Conversely, when the payoff was optimal, 85.7 percent (i.e., 5666 out of 6615) of the decisions were not to switch.

The clearest form of social learning occurs when a player who switches adopts the strategy with currently highest payoff. This occurs indeed for 1700 of the 2850 switches. The frequency of players who adopt a strategy which is less than optimal decays exponentially in  $d$ , the difference between the current optimum and the current payoff of the strategy that the switching player will adopt in the next round. (Nonlinear regression  $\nu = 0.4058 \times 0.2888^d$ , correlation coefficient  $R=0.8439$  and  $P\text{-value}<0.001$ , see Figure S3 in Online Resource). Needless to say, in the next round both the optimal payoff and the payoff of the newly adopted strategy may be different.

If we define a player as social learner when at least 90 percent of that player's decisions could be explained either as (a) switching to a strategy with currently higher payoff, or as (b) sticking with a strategy of currently highest payoff, then 78.6 percent of the players were social learners. We stress that the players had only 15 seconds between rounds, which hardly offered them time for strategic calculations.

## 5. Discussion and Conclusion

Coercion plays an essential role in overcoming social dilemmas. The corresponding line of reasoning goes back at least as far as Hobbes' 'Leviathan' from 1651, and the practical implementation can be traced throughout history. The selfish motivations endangering collective actions have to be suppressed by positive and negative incentives (Olson 1965; Boyd and Richerson, 1992; Andreoni et al., 2003; Rockenbach and Milinski, 2006). In particular, the threat of punishment curbs the temptation to free-ride, i.e., to exploit the contributions of others without offering an adequate return.

Institutions can be viewed as tools for providing incentives (Ostrom, 2005). It has been shown that even in small-scale societies far removed from 'Leviathan'-like states, grass-root institutions can deal, often efficiently, with the tasks of monitoring joint efforts and sanctioning defectors (Ostrom, 1990; Boehm, 2000; Henrich, 2006; Baldassarri and

Grossman, 2011). We wanted to test how players could opt for such a rudimentary institution, modeled as pool-punishment.

Our experiment is close in spirit and design to an experiment by Gürer et al. 2006. In that experiment, players were given the choice between a Public Good game with and one without peer punishment. The majority started with a clear preference for the treatment without punishment, but switched after a few rounds to the peer-punishment treatment, apparently guided by payoff considerations. Essentially, we kept the three-staged structure (choice of treatment, decision to contribute, decision to punish), but added pool punishment and non-participation as additional choices. (In contrast to the paper by Gürer et al. 2006, we did not allow for rewarding; a related endogenous choice between peer punishing and rewarding has been investigated by Sutter et al. (2010), who found that reward was often chosen.)

The option of pool punishment adds an important element, as it essentially provides the opportunity for a tacit social contract establishing a sanctioning institution. To our knowledge, this is the first experiment demonstrating that such a social contract can emerge through social learning based on comparing the (frequency dependent) payoff values of diverse options. ‘Social contract’ means that players can submit to a sanctioning authority, captured here as a ‘punishment pool’. By contrast, peer punishment corresponds to self-justice, and belongs to an anarchic ‘state of nature’. Both philosophers and experimental game theorists have shown that peer punishment can lead to an escalation of conflicts, i.e., to a ‘war of all against all’. For example, John Locke wrote in §126 of his ‘Two Treatises of Government’ from 1689 that in the state of nature, ‘... resistance [by defaulters] many times makes the punishment dangerous, and frequently destructive, to those who attempt it’. In a similar vein, experiments such as those by Denant-Boemont, Masclet, Noussair (2007), Nikiforakis (2008) or Nikiforakis and Engelmann (2011) show that peer punishment can invite counter-punishment and lead to costly feuds. We have deliberately excluded the possibility of counter-punishment, which threatens self-justice; we also have excluded the possibility of a corrupt authority, which threatens sanctioning institutions. The different punishment regimes offered in our experiment were, in this sense, idealized, pro-social versions.

The lesson for institution design is clear: pool punishment requires the sanctioning of second-order free-riders. The important role of second-order free-riding is well-known (Oliver 1980), and our experiment confirms it. In the second-order treatment, pool punishment effectively prohibits this possibility, whereas in the first-order treatment, it does not.

Apparently, pool-punishers notice that they are exploited, in the first-order treatment, and react against this breach in equity (Bolton and Ockenfels, 2006). Voting for the second-order treatment implies a higher commitment.

We now discuss several aspects of the experimental design which may limit the generalizability of our results.

We did allow for players to abstain from the game. Clearly, there exist joint enterprises or common resources from which one cannot abstain: the global climate is the best example. Such compulsory interactions do not belong to the class considered here, since we have allowed players to opt for non-participation. Nevertheless, it could well be that the main ‘efficiency vs. stability’ result still holds for compulsory games. It was for two reasons that we decided to consider only voluntary interactions in our experiment: first, because the theoretical results guiding our predictions were derived for this class of games only, and second because, in the course of the experiment, we sometimes (but rarely) encounter a player who is the only individual choosing a given Mutual Aid game of type (NoPun), (Peer) or (Pool). In this case, it is practical to assign them option (No), namely ‘non-participation’. This, incidentally, hardly affects the statistics.

The fact that we chose a Mutual Aid game, rather than the Public Good game, should not overly restrict the range of applications. We did so because the number of players was variable, and that the Public Good game stops being a social dilemma if the number of players is smaller than the multiplication factor  $r$ . In contrast, a Mutual Aid game always is a social dilemma: this makes the issue of altruism vs. free-riding more clear-cut. As a real-world example of a Mutual Aid game, we refer to the ‘sick clubs’ or ‘friendly societies’ run by working men in nineteenth-century England (see Sugden, 1986). If one of the members fell ill, the others could contribute to his aid. Such joint enterprises were informal fore-runners of state insurance schemes. If we assume that in a given period, everyone is equally likely to fall ill, and that the others could decide between contributing a fixed amount or not, we obtain exactly the structure of a Mutual Aid game.

In our experimental design, we did not allow for punishment of non-punishers in the peer punishment game. The reason is twofold. On the one hand, theoretical models predict that it has no effect on the outcome (Sigmund et al, 2010). On the other hand, economic experiments have confirmed this in similar situations (Cinyabuguma et al. 2006; Kiyonari and Barclay 2008; Traulsen et al. 2012). We do not expect second-order peer punishment to affect the outcome.

We have reduced all individual decisions to choices between two, three or four alternatives. It would be interesting to investigate scenarios where players have a larger range of strategies, for instance by allowing them to choose between ten levels of contribution to the Mutual Aid, or different degrees of punishment. Similarly, we have proposed only one, extremely rudimentary form of institution. It is easy to think of better designs, for instance by allowing part, at least, of the unused funds to return to the players who have contributed to the punishment pool. We refrained from doing this, because we did not want to make it too easy for institutional punishment to emerge. The fact that as many groups ended up with peer- as with pool-punishment suggests that we succeeded in this ‘calibration’. Moreover, our experiment is already complex enough, and we feared to make it cognitively too demanding by adding more choices. As it was, the practice rounds needed to familiarize the players with their options took almost one hour (as long as the subsequent experiment).

Our main objective was to compare two different versions of pool punishment (rather than pool with peer). We note that there exist at least three experiments (independently conceived and in part not yet published) comparing pool with peer punishment, or ‘informal’ with ‘formal’ sanctions (Kamei et al. 2011; Markussen et al. 2011; Traulsen et al. 2012). Kamei et al. 2011 and Markussen et al. 2011 adopt a continuous version of first-order pool punishment, and Traulsen et al. 2012 consider both first- and second-order pool punishment (same as our experiment). In Markussen et al. 2011, fixed groups of five players play for 24 rounds, and can vote, at specific instants, between two different regimes (corresponding, in our setup, to decisions between (NoPun) and (Peer), (Peer) and (Pool), or (NoPun) and (PoolC)). In Kamei et al. 2011, the choice is between (Peer) and (Pool) with various parameters for the sanctions. Informal sanctioning does remarkably well, both from the viewpoint of frequency and payoff. (The experiment by Ertan et al. 2009 and the theoretical model by Boyd et al. 2010 confirm that peer punishment works well when players have an opportunity for coordinating.) In contrast, formal sanctions (which did not include second-order punishment) fared poorly.

The experiment by Traulsen, Röhl and Milinski 2012 investigates three treatments. In (a), players are offered the possibility of peer punishment, first 10 rounds without, then 15 rounds with second-order punishment. In (b), players are offered the possibility of pool punishment, again first without, then with second-order punishment. In (c), players could use both types of punishment in each round, by investing into a punishment pool before the public good interaction, and exerting peer punishment after the public good interaction. Again, this was first played without, and then with second-order punishment. In (a), switching from first-



to second-order punishment had very little effect; in (b) and (c), the switch strongly boosted pool punishment. In our experiment, players were not allowed to use both types of punishment simultaneously. They had to decide between one or the other (or none), thereby going separate ways, and building communities with different rules.

In all these experiments (including ours) peer punishment did reasonably well. This may in part be due to the higher efficiency in the optimal situation when all contribute to the Mutual Aid, or the Public Good. It may also be due to the fact that possibilities for retaliation and feuding were excluded in the experimental designs.

Since we wanted to favor conditions for social learning, we provided the players with information on the frequencies and payoffs obtained by the various strategies in their group (See Online Resource). However, we refrained from giving them opportunities to build up individual profiles, for instance reputations, or significant differences in resources. Needless to say, this does not imply that reputations or differences in resource holding power are irrelevant for the evolution of institutions. Similarly, we did not consider other-regarding preferences (Fehr and Schmidt 1999) or contests between groups, although such struggles played doubtlessly an important role in human evolution (Choi and Bowles 2007).

What are the roots of sanctioning institutions? Our players were given the choice between one type of peer and one type of pool punishment. Needless to say, such an approach cannot tell how such opportunities for sanctioning emerge. Cooperation has frequently arisen through biological evolution (Maynard Smith and Szathmary 1995), often via subtle mechanisms suppressing competition (Frank 1995), and there exist many examples of animals punishing each other (Clutton-Brock and Parker 1995). In particular, parents repress competition between their offspring, in many species, and it may be that this eventually led, in human populations, to institutionalized sanctioning. Offspring would simply have to remain with their parents (a costly option which provides some safety) rather than leave and defend their interests single-handedly. This fits with Jean Jacques Rousseau's claim, in 'Du Contrat Social' from 1762, that *'the family is the first model of a political society: its head is the image of the father'* (Book 1, Chapter 2). However, such a scenario was not addressed in our experiment. We presumed that the interactions are symmetric, and that all participants have equal resources. This better fits with Boehm 2000, who proposed another origin for sanctioning institutions. Its first instances, accordingly, were coalitions directed against alpha-males in the group, and the first social contract aimed at suppressing bullying behavior, in order to guarantee an egalitarian distribution of big-game meat.

It seems that institutions, once they have arisen, apply themselves to curb the vengeful and aggressive instincts fuelling peer-punishment. It would be interesting to explore this, both by modeling and by experiment. In our experimental setup, we have not allowed pool-punishers to sanction peer-punishers, or punished players to retaliate (Cinyabuguma et al. 2006; Nikiforakis 2008). We also excluded communication and deliberation, although theoretical models, field observations and experiments alike have stressed the importance of communication in sanctioning exploiters (Walker et al. 2000; Bochet et al. 2006; Ertan et al. 2009). If individuals can look for allies, or deliberate with their peers, stable systems of incentives can arise (Casari and Luini 2009; Ertan et al. 2009; Boyd et al. 2010). We aimed for a minimalistic scenario based on social learning, and showed that it can lead to the emergence of a rudimentary type of institutionalized coercion helping to overcome individuals' selfish preferences.

#### Acknowledgements

Our software was based on z-Tree, for which we thank Urs Fischbacher. We are grateful to Dirk Semmann, Jean Pierre Tyran, Matthias Sutter and Simon Gächter for comments, and to Stephan Aigner for help during the experiment. We acknowledge support from the Austrian Science Funds and the European Science Foundation through TECT I-104 G11 and Grant #RFP-12-21 from Foundational Questions in Evolutionary Biology Fund. We also thank for the support of the Chinese Scholarship Council (CSC).

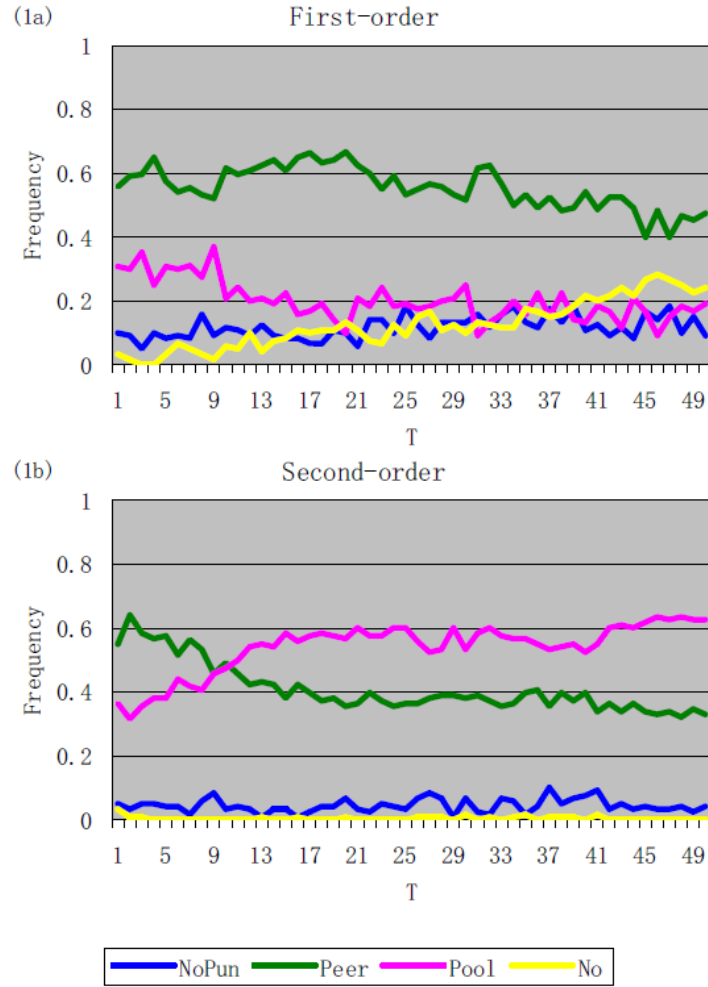
#### References

- Andreoni, J. and Gee, L. L. (2012). Gun for hire: Delegated enforcement and peer punishment in public goods provision. *Journal of Public Economics*, 96, 1036-1046.
- Andreoni, J., Harbaugh, W. and Vesterlund, L. (2003). The carrot or the stick: rewards, punishments, and cooperation. *American Economic Review*, 93, 893-902.
- Balafoutas, L. and Nikiforakis, N. (2012) Norm Enforcement in the City: A Natural Field Experiment forthcoming *European Economic Review*.
- Baldassarri, D. and Grossman, G. (2011). Centralized sanctioning and legitimate authority promote cooperation in humans. *Proceedings of the National Academy of Sciences*, 108, 11023–11027.
- Bochet, O., Page, T. and Putterman, L. (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior and Organization*, 60, 11-26.
- Boehm, C. (2000) Conflict and the Evolution of Social Control, *Journal of Consciousness Studies* 7, 79-101.
- Bolton, G. E. and Ockenfels, A. (2006). ERC: a theory of equity, reciprocity, and competition. *American Economic Review*, 90, 166-93.
- Boyd, R., Gintis, H. and Bowles, S. (2010). Coordinated punishment of defectors sustains cooperation and can proliferate when rare. *Science*, 328, 617-620.
- Boyd, R., and Richerson, P. J. (1992). Punishment allows the evolution of cooperation (or anything Else) in sizeable groups. *Ethnology and Sociobiology*, 113, 171-195.

- Carpenter, J. P. (2007). The demand for punishment. *Journal of Economic Behavior and Organization*, 62, 522–542.
- Casari, M. (2007). On the design of peer punishment experiments. *Experimental Economics*, 8, 107–115.
- Casari, M. and Luini, L. (2009). Cooperation under alternative punishment institutions: an experiment. *Journal of Economic Behavior and Organization*, 71, 273–282.
- Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics*, 14, 47–83.
- Choi, J.-K. and Bowles, S. (2007). The coevolution of parochial altruism and war. *Science*, 318, 636–640.
- Cinyabuguma, M., Page, T. and Putterman, L. (2006). Can second-order punishment deter perverse punishment. *Experimental Economics*, 9, 265–279.
- Clutton-Brock, T. H. and Parker, G. A. (1995). Punishment in animal societies. *Nature*, 373, 209–216.
- Denant-Boemont, L., Masclet, D., Noussair, C. (2007) Punishment, Counterpunishment and Sanction Enforcement in a Social Dilemma Experiment, *Economic Theory* 33, 145–167.
- Dreber, A., Rand, D. G., Fudenberg, D. and Nowak, M. A. (2008). Winner don't punish. *Nature*, 452, 348–351.
- Egas, M., and Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences*, 275, 871–878.
- Ertan, A., Page, T. and Putterman, L. (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, 53, 495–511.
- Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114, 817–868.
- Fehr, E. and Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, 90, 980–994.
- Fehr, E. and Gächter, S. (2002). Altruistic punishment in humans. *Nature*, 415, 137–140.
- Fehr, E. and Rockenbach, B. (2003). Detrimental effects of sanctions on human altruism. *Nature*, 422, 137–140.
- Fischbacher, U. (2007). Z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10, 171–78.
- Fletcher, J.A. and Zwick, M. (2004). Strong altruism can evolve in randomly formed groups. *Journal of Theoretical Biology*, 228, 303–313.
- Fowler, J. H. (2005). Altruistic punishment and the origin of cooperation. *Proceedings of the National Academy of Sciences*, 102, 7047–7049.
- Frank, S. A. (1995). Mutual policing and repression of competition in the evolution of cooperative groups. *Nature*, 377, 520–522.
- Gächter, S., Renner, E. and Sefton, M. (2008). The long-run benefits of punishment. *Science*, 322, 1510–1512.
- Guala, F. (2012). Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate. *Behavioral and Brain Sciences*, 35, 1–59.
- Guillen, P., Schwieren, C. and Staffiero, G. (2006). Why feed the Leviathan? *Public Choice*, 130, 115–128.
- Gürerk, O., Irlenbusch, B. and Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, 312, 108–111.
- Henrich, J. (2006). Cooperation, punishment, and the evolution of human institutions. *Science*, 312: 60–61.

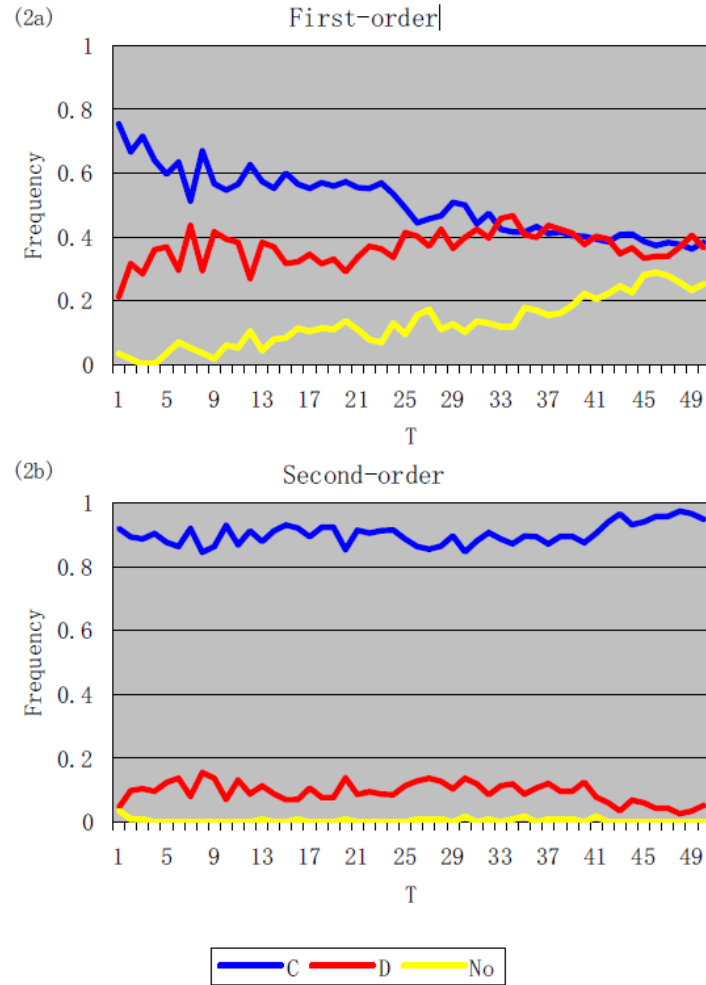
- Henrich, J., McElreath, R., Barr, A., Ensminger, J., Barrett, C., Bolyanat, A., Cardenas, J. C., Gurven, M., Gwako, E., Henrich, N., Lesorogol, C., Marlowe, F., Tracer, D. and Ziker, J. (2006). Costly punishment across human societies. *Science*, 312, 1767-1770.
- Herrmann, B., Thoni, C. and Gächter, S. (2008). Antisocial punishment across societies. *Science*, 319, 1362–1367.
- Kamei, K., Putterman, L., and Tyran, J-R. (2011). State or nature? Formal vs. informal sanctioning in the voluntary provision of public goods. Discussion Papers 11-05, University of Copenhagen. Department of Economics.
- Kiyonari, T. and Barclay, P. (2008). Free riding may be thwarted by second-order reward rather than by punishment. *Journal of Personality and Social Psychology*, 95, 826-842.
- Kosfeld, M., Okada, A. and Riedl, A. (2009). Institution formation in public goods games. *American Economic Review*, 99, 1335-1355.
- Maynard Smith, J. and Szathmary, E. (1995). *The Major Transitions in Evolution*, New York: Oxford University Press.
- Markussen, T., Putterman, L., and Tyran, J-R. (2011). Self-organization for collective action: an experimental study of voting on formal, informal, and no sanction regimes. Working Papers 2011-4, Brown University, Department of Economics.
- Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: can we really govern ourselves? *Journal of Public Economics*, 92, 91-112.
- Nikiforakis, N. and Engelmann, D (2011) Altruistic punishment and the threat of feuds. *Journal of Economic Behavior and Organization*, 78, 319-332.
- Oliver, P. (1980). Rewards and punishments as selective incentives for collective action: theoretical investigations. *American Journal of Sociology*, 85, 1356-1375.
- Olson, M. (1965). *The Logic of Collective Action: Public Goods and the Theory of Groups*, Harvard: Harvard University Press.
- Ostrom, E. (1990). *Governing the Commons: The Evolution of Institutions for Collective Action*, Cambridge: Cambridge University Press.
- Ostrom, E. (2005). *Understanding Institutional Diversity*, Princeton: Princeton University Press.
- Rockenbach, B. and Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature*, 444, 718–723.
- Sigmund, K. (2007). Punish or perish? Retaliation and cooperation among humans. *Trends in Ecology and Evolution*, 22, 593-600.
- Sigmund, K. (2010). *The Calculus of Selfishness*, Princeton: Princeton University Press.
- Sigmund, K., De Silva, H., Traulsen, A. and Hauert, C. (2010). Social learning promotes institutions for governing the commons. *Nature*, 466, 861-863.
- Sigmund, K., Hauert, C., Traulsen, A. and De Silva, H. (2011). Social control and the social contract: the emergence of sanctioning systems for collective action. *Dynamic Games and Applications*, 1, 149-171.
- Sugden, R. (1986). *The economics of rights, cooperation and welfare*, Oxford: Blackwell.
- Sutter, M., Haigner, S. and Kocher, M. G. (2010). Choosing the carrot or the stick? Endogenous institutional choice in social dilemma situations. *Review of Economic Studies*, 77, 1540-1566.
- Tiebout, C. (1956). A pure theory of local expenditures. *Journal of Political Economy*, 64, 416-424.

- Traulsen, A., Röhl, T. and Milinski, M. (2012). An economic experiment reveals that humans prefer pool punishment to maintain the commons. *Proceedings of the Royal Society B: Biological Sciences*, 279, 3716-3721.
- Tyler, T. R. and Degoe, P. (1995). Collective restraint in social dilemmas: procedural justice and social identification effects on support for authorities. *Journal of Personality and Social Psychology*, 69, 482-497.
- Van Vugt, M., Henrich, J. and O'Gorman, R. (2009). Constraining free riding in public good games: designated solitary punishers can sustain human cooperation. *Proceedings of the Royal Society B: Biological Sciences*, 276, 323-329.
- Walker, J. M., Gardner, R., Herr, A. and Ostrom, E. (2000). Collective choice in the commons: experimental results on proposed allocation rules and votes. *The Economic Journal*, 110, 212-34.
- Wilson, D.S. (1975). A theory of group selection. *Proceedings of the National Academy of Sciences*, 72, 13-146.
- Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, 51, 110-116.



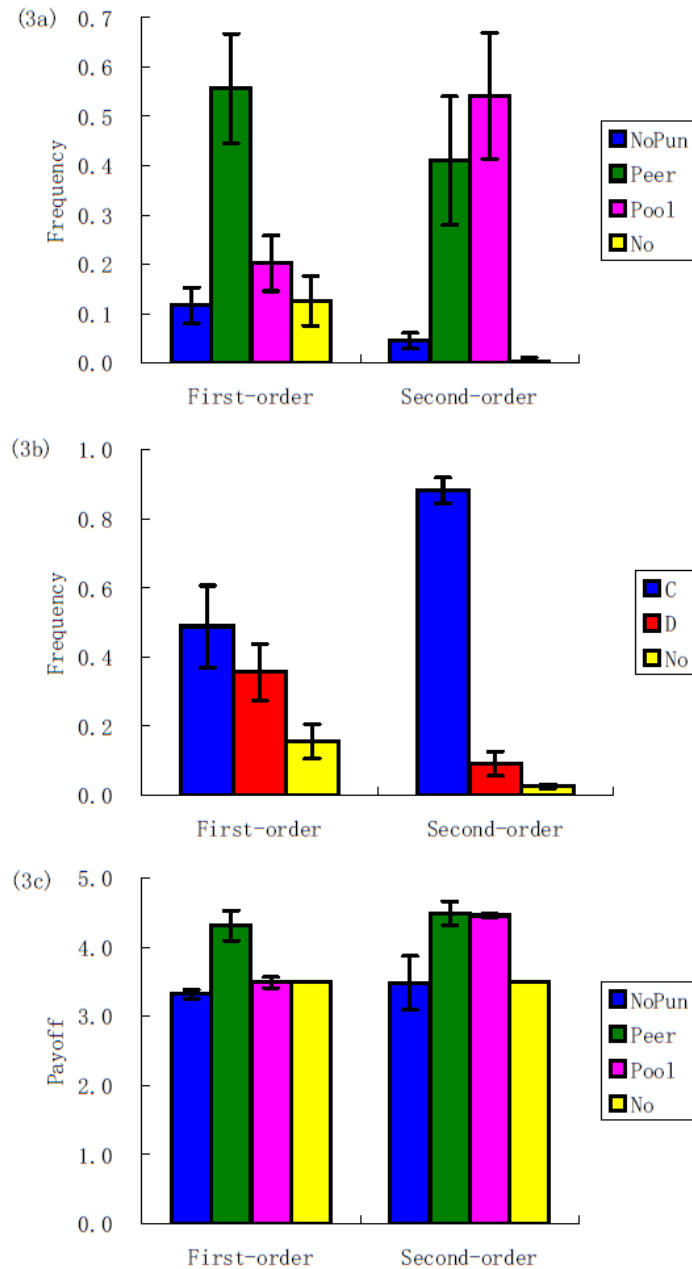
**Figure 1**

Time-evolution of the frequencies of players voting for the games (NoPun), (Peer), (Pool) or (No) in (a) the first-order treatment and (b) the second order treatment. In each case the numbers are obtained by summing over the nine corresponding groups. (For the frequencies in each group, we refer to Figure S2 of the Supplementary Online Resource).



**Figure 2**

Time-evolution of the frequencies of cooperation (C, blue), defection (D, red) and non-participation (No, yellow) over 50 rounds in the first- and the second-order treatments, summed over the nine corresponding groups. (2a) In the first-order treatment, defection was chosen by about one-third of the players in each round. The number of contributions declined in favor of non-participation. (2b) In the second-order treatment, almost all the players chose to contribute. This cooperative regime was stably sustained.



**Figure 3**

(3a) Frequencies of the decisions in favor of the different games, over 50 rounds, for the first- and the second-order treatments. In the first-order treatment, peer punishment is favored. In the second-order treatment, pool punishment is more frequent, but error bars overlap. (3b) Frequencies of the decisions to contribute to the Mutual Aid game, to defect (i.e., not to contribute) and to opt for non-participation, averaged over 50 rounds. Contribution is strongly promoted in the second-order treatment. (3c) Payoffs obtained for the different games (NoPun), (Peer), (Pool), averaged over fifty rounds, do not greatly differ. Nevertheless, in the first-order treatment, peer punishment games, and in the second-order treatment, both peer and pool punishment games provided the highest average payoff. Error bars indicate standard errors of the group means.



# Online Resource to “The evolution of sanctioning institutions: an experimental approach to the social contract”

Boyu Zhang, Cong Li, Hannelore De Silva,  
Peter Bednarik and Karl Sigmund \*

The experiment took place in a computer lab of the Vienna University of Economics and Business (WU) on six days. On three days, the first-order treatment was played, and on the other three days the second-order treatment. The lab has 50 computers and for each of the six sessions, some 40 students (3 groups) played together. The interactions were anonymous, and via PCs. Cardboard dividers ensured that the students could not see each other. Players were not allowed to communicate. They were also not allowed to ask questions during the experiment, but they could ask before the experiment (only four or five did).

**Table S1**

## **Group size in the first-order and the second-order treatment**

Group sizes in the first-order treatment									
group 1	group 2	group 3	group 4	group 5	group 6	group 7	group 8	group 9	Total
13	13	13	13	13	13	14	14	14	120

Group sizes in the second-order treatment									
group10	group11	group12	group13	group14	group15	group16	group17	group18	Total
14	14	13	12	12	12	14	14	13	118

The practice rounds lasted about 45 min, almost for as long as the subsequent experiment (students knew that the sessions would last at most for two hours, but were not told the number of rounds, so as to avoid end round effects). All players were given the same instructions (in German, see screen shots). The groups were then re-shuffled before the actual experiment started, and remained unchanged for its entire duration. The translation of the instructions for the practice rounds and the experiment can be found at the end of the Online Resource. The average income was 19.6 euro (minimum 15.3, maximum 24.9). All steps were time-limited. Players knew that if they did not decide within 15 seconds, they would be allocated a random decision. Since the players had familiarized themselves with each game,

---

\* Author for correspondence: Telephone: +43 (0)1427750612; Fax: +43(0)142779506.  
E-mail: karl.sigmund@univie.ac.at

during the practice rounds, this happened only 9 times in 11900 decisions, and is omitted from the statistics.

In the groups 1-9, which offered the first-order treatment of pool punishment, peer punishment was preferred, as can be seen in Figure S1a and Table S2a. In the following tables S2a and S2b, the standard error is based on individual decisions (not on the groups).

**Table S2a**

**Decisions in the first-order treatment**

Groups 1-9: votes for the different games (including non-participation)

Decisions	Number of times	Percentage	Average payoff	Standard error
(No) non-participation	754	0.126	3.500	0
(NoPun) no-punishment game	701	0.117	3.342	0.0299
(Peer) peer punishment game	3330	0.556	4.299	0.0164
(Pool) pool punishment game	1208	0.202	3.492	0.0250
Total	5993	1	3.924	0.0126

After including among non-participants those players who found no partners

Decisions	Number of times	Percentage	Average payoff	Standard error
(No) non-participation	926	0.155	3.500	0
(NoPun) no-punishment game	618	0.103	3.32	0.0339
(Peer) peer punishment game	3300	0.551	4.31	0.0165
(Pool) pool punishment game	1149	0.192	3.49	0.0262

**Decisions within each game**

Decisions	Number of times	Percentage	Average payoff	Standard error
contribution in no-punishment game	99	0.017	2.601	0.1094
non-contribution in no-punishment game	519	0.087	3.458	0.0311
contribution, but no punishing, in peer punishment games	2049	0.342	4.636	0.0150
non-contribution in peer punishment games	859	0.143	3.614	0.0297
peer-punishment and contribution	392	0.065	4.100	0.0681
contribution, but no punishing, in pool punishment games	338	0.056	3.486	0.0522
non-contribution in pool punishment games	757	0.126	3.566	0.0295
pool-punishment and contribution	54	0.009	2.477	0.1189

In the first-order pool punishment games, neither contributions to the Mutual Aid nor to the sanctioning took off. Only a tiny fraction of the decisions in this group (54 out of 1149)

favored investing into the punishment pool. The large majority seems to have sensed that the punishment threat would not be carried out, and defected. Defection was the most profitable decision in the pool punishment game, but the average payoff (3.566 MU) was only slightly higher than what non-participants obtained. (This difference was not significant). Peer punishment was clearly preferred. The average payoff obtained by opting for the peer punishment game was 4.3 MU, higher (not significant) than for opting for a pool punishment game (3.49 MU, Mann-Whitney U-test,  $n_1=9$ ,  $n_2=9$ ,  $p=0.11$ ) or the game without punishment (3.34 MU, Mann-Whitney U-test,  $n_1=9$ ,  $n_2=9$ ,  $p=0.03$ ). Indeed, the average payoff values in the pool punishment or no-punishment games were lower than the non-participation payoff of 3.5 MU. A majority (62 percent) of players opting for the peer punishment game contributed to the Mutual Aid game, but did not choose the punishment option. All in all, 48.9 percent of all decisions were in favor of contributing to the Mutual Aid game, rather than defecting (35.6 percent) or abstaining from the game (15.5 percent). The frequency of cooperative decisions in the (Peer) game is significant higher than that of the (Pool) game (74% vs 34.1%, Mann-Whitney U-test,  $n_1=9$ ,  $n_2=9$ ,  $p<0.0001$ ) and the (NoPun) game (74% vs 16%, Mann-Whitney U-test,  $n_1=9$ ,  $n_2=9$ ,  $p<0.0001$ ). But as mentioned in the main text, the time evolution over the fifty rounds shows a clear decline in contributions over time (See Table S3b). We also note that free-riding was the most frequent and most successful behavior in the pool punishment game, but that the average payoff (3.566 MU) was only insignificantly higher than what non-participants obtained. Remarkably, the payoff for defecting in the games without punishment was almost the same (3.458 MU).

In the groups 10 to 18, pool-punishment was offered in the second-order treatment, i.e., it included punishing those who contributed to the Mutual Aid but not to the punishment pool. This time, pool punishment was preferred, as can be seen in Figure S1b and Table S2b. Almost all decisions in the (Pool) game are cooperative (contributions to both pools or only to the Mutual Aid), which is significant higher than for the (Peer) game (99.7% vs 83.5%, Mann-Whitney U-test,  $n_1=9$ ,  $n_2=9$ ,  $p<0.0001$ ) or the (NoPun) game (99.7% vs 23.8%, Mann-Whitney U-test,  $n_1=9$ ,  $n_2=9$ ,  $p<0.0001$ ). Only 4.5 percent of all decisions were in favor of game (NoPun). The free-riders, in that case, did about as poorly as in the (Peer) game (3.696 MU vs 3.689 MU), since they found only few to exploit. Very few decision was in favor of non-participation. In many more cases, non-participation was the unintended consequence of choosing a game that was not chosen by anyone else in the group. Second-order free-riding (i.e., to opt for the peer punishment game, and contribute, but not punish) achieved the highest payoff, 4.77 MU (see Figure 3c).

The time-evolution in the different groups is interesting (see Figures S1 and S2). In seven of the nine groups where pool punishment was offered in the first-order treatment, the initial majority voted for peer punishment and in the other two groups, the initial majority voted for pool punishment. Three groups (3, 4 and 6) quickly reached consensus on peer punishment but all other groups went to chaos. During fifty rounds, players persisted in switching from one game to another. We note that in the three groups leading to peer punishment, two-thirds of the players, in each round, decided not to actually punish. The threat of the remaining third sufficed to ensure co-operation, although that threat had rarely to be carried out.

**Table S2b**

**Decisions in the second-order treatment**

Groups 10-18: votes for the different treatments (including non-participation)

Decisions	Number of times	Percentage	Average payoff	Standard error
(No) non-participation	23	0.004	3.500	0
(NoPun) no-punishment game	265	0.045	3.483	0.057
(Peer) peer punishment game	2421	0.410	4.490	0.018
(Pool) pool punishment game	3189	0.541	4.459	0.009
Total	5898	1	4.424	0.010

After including among the non-participants those players who found no partners:

Decisions	Number of times	Percentage	Average payoff	Standard error
(No) non-participation	154	0.026	3.500	0
(NoPun) no-punishment game	181	0.031	3.475	0.0836
(Peer) peer punishment game	2389	0.405	4.503	0.0178
(Pool) pool punishment game	3174	0.538	4.464	0.0094

Decisions in each treatment:

Decisions	Number of times	Percentage	Average payoff	Standard error
contribution in no-punishment games	43	0.007	2.767	0.1791
non-contribution in no-punishment games	138	0.023	3.696	0.0866
contribution, but no punishing, in peer punishment games	1781	0.302	4.770	0.0123
non-contribution in peer punishing games	393	0.067	3.689	0.0516
peer-punishment and contribution	215	0.036	3.776	0.0946
contribution, but no punishing, in pool punishment games	11	0.002	-0.955	1.3659
non-contribution in pool punishment games	8	0.001	0.313	1.8094

pool-punishment and contribution	3155	0.535	4.493	0.0017
----------------------------------	------	-------	-------	--------

There was not much switching in the groups where the second-order treatment of pool punishment was played. Despite the fact that in the first round, more players voted for (Peer) than for (Pool) punishment (65 vs. 43), pool-punishment emerged in six of the nine groups as consensus solution. In three groups (13, 17 and 18), the initial majority for peer punishers was large enough to ensure the fixation of peer punishment within a few rounds. However, group 17 collapsed eventually, since the threat of peer punishment was not actually carried out. The players then turned to the (Pool) game. A switch in the opposite direction occurred in group 15. After some initial oscillations, the pool-punishment game emerged as the majority choice, but it was never unanimous, and eventually became replaced by (Peer).

**Table S3**

Regression lines

Table S3a: Voting for different games

	Regression line (50 rounds)	R <sup>2</sup>	P-value
First-order peer game	$y = 0.6347 - 0.0031x$	0.4761	P-value<0.001
First-order pool game	$y = 0.2749 - 0.0029x$	0.4362	P-value<0.001
Second-order peer game	$y = 0.5220 - 0.0044x$	0.6671	P-value<0.001
second-order pool game	$y = 0.4325 + 0.0042x$	0.5939	P-value<0.001

Table S3b: Frequencies of C (contribute to the Mutual Aid) and D (defect)

	Regression line (50 rounds)	R <sup>2</sup>	P-value
First-order C	$y = 0.6689 - 0.0065x$	0.8812	P-value<0.001
First-order D	$y = 0.3292 - 0.0015x$	0.1761	P-value=0.024
Second-order C	$y = 0.8783 + 0.001x$	0.1779	P-value=0.023
second-order D	$y = 0.1164 - 0.001x$	0.1707	P-value=0.029

Table S3c: Voting for different games in the second-order treatment

	Regression line (first 20 rounds)	R <sup>2</sup>	P-value
Second-order peer game	$y = 0.6191 - 0.0136x$	0.8983	P-value<0.001
Second-order pool game	$y = 0.3262 + 0.0146x$	0.9167	P-value<0.001

Notes: y represents the frequency and x the round. R<sup>2</sup> is the coefficient of determination.

There are two related problems in establishing the statistics. One is that players opting for a game may end up with no partners, and thus become non-participants. Their decision was registered, and included in the statistics, but their payoff (3.5 MU) was not included in the average payoff for the game of their choice, since that game was cancelled. If we had added instead their 3.5 MU to the average, not much would have changed. The second

problem is how to count the decisions in favor of peer punishment in those peer punishment games where no defection took place. If a player sees that there is no one to punish, and then chooses ‘peer-punishment’, this can indicate an earnest commitment to uphold the sanctioning system to guarantee cooperation (Masclet et al. 2003), but it could just as well be a mere cost-free gesture. If conversely a player chooses ‘non-punishment’, this can either indicate a decision for second-order free riding, or merely mean that the player was aware that there was no need for sanctions anyway. There were 108 such rounds (out of 239). The average number of decisions for peer punishment in (Peer) games without defectors was higher than that in (Peer) games with defectors (1.02 vs 0.85), but the difference is not significant, see Table 4S. In computing average payoffs and frequencies, we decided to take the players statements at face value. But we also computed a ‘skeptical’ version (not shown here), where players who actually did not punish were counted as non-punishers, no matter whether they declared themselves to be peer-punishers or not. Frequencies and the average payoffs are different, but the main conclusions remain unaffected.

**Table S4**

Average number of peer punishers (in both treatments)

Number of defectors in the (Peer) game	0	1	2	3	4	5	6	7	8	9	>0
Number of (Peer) games	239	130	102	84	42	19	26	12	8	4	427
Average number of peer punishers	1.02	1.84	0.47	0.45	0.4	0.37	0.35	0.25	0	0	0.85

The experiment was motivated by a theoretical analysis (Sigmund et al, 2010). This analysis predicts that the emergence of pool punishment is possible only if second-order free-riders are also punished. This is confirmed in our experiment. On the other hand, we expected that peer punishment would be replaced, in that case, by pool punishment. As it turned out, we did not observe this anticipated ‘trading efficiency for stability’. Rather, we found examples for switches in both directions (groups 15 and 17, see Online Resource). A look at the time evolution in each group (see Online Resource, Figures S1 and S2) suggests that in both treatments, peer punishment offered a modicum of stability, but that when it failed, it gave way to asocial behavior (i.e., non-participation or defection) in the first-order treatment, and to pool punishment in the second-order treatment. As a consequence, contributions were stably sustained in the second-order treatment, at a very high level, whereas they declined, and were ultimately overtaken by defections, in the first-order

treatment (see Figure 2). This good performance of peer punishment may be due to the fact that retaliatory punishment was not possible in our design (Cinyabuguma et al. 2006; Nikiforakis 2008). Moreover, pool-punishers could not punish peer-punishers in our experiment. They belonged to different games. It is possible that ‘cross-punishment’ can change this outcome. (In Traulsen et al. 2012, this possibility was offered, but hardly ever used by the players.)

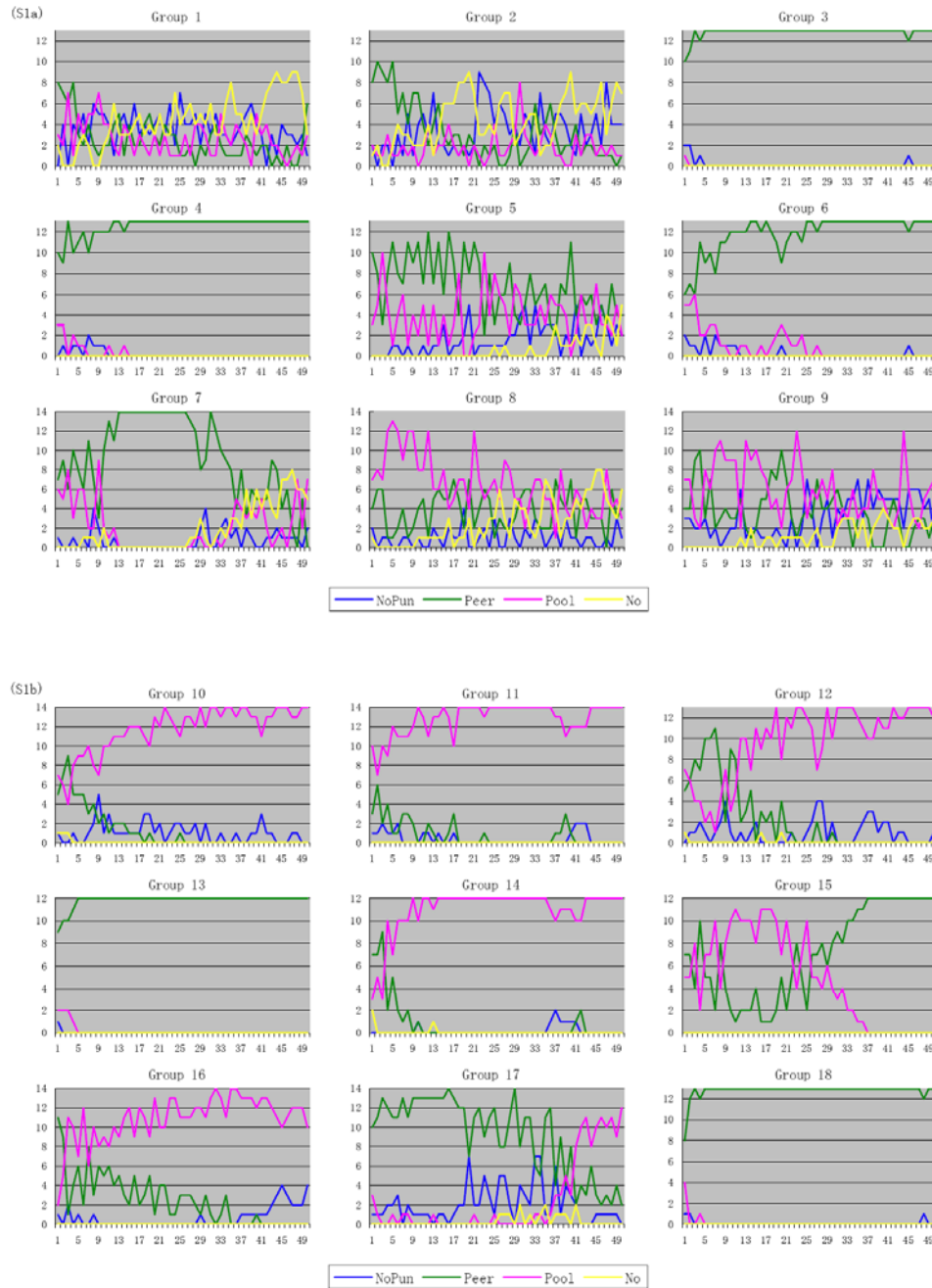
The initial phase of our experiment displayed a high rate of change in behavior in most groups. On average, more than one-fourth of the players switched to another decision between one round and the next, during the first twenty rounds. In the last ten rounds, the average switching rate was only 5.6 percent in the twelve groups that had settled on peer or pool punishment, but 50 percent in the others.

Another question that was not addressed here is whether the option to abstain from the game (‘non-participation’), which is crucial for the theoretical analysis (Sigmund et al, 2010), is also necessary for the experiment. For the theoretical analysis, it was assumed that innovative behavior (‘mutation’) is much rarer than copying behavior. In that case, non-participation is necessary as an escape from the homogeneous state of defection. Since actual human populations display high degrees of polymorphism (Traulsen et al. 2010), non-participation may not be needed. On the other hand, voluntary participation is likely to increase the perceived legitimacy of the sanctioning institution, and hence its efficiency (Tyler and Degoe 1995; Ertan et al. 2009).

## References

- Masclét, D., Noussair, C., Tucker, S. and Villeval, M-C. (2003). Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism. *American Economic Review*, 93, 366-380.
- Traulsen, A., Semmann, D., Sommerfeld, R. D., Krambeck, H-J. and Milinski, M. (2010). Human Strategy Updating in Evolutionary Games. *Proceedings of the National Academy of Sciences*, 107, 2962-2966.

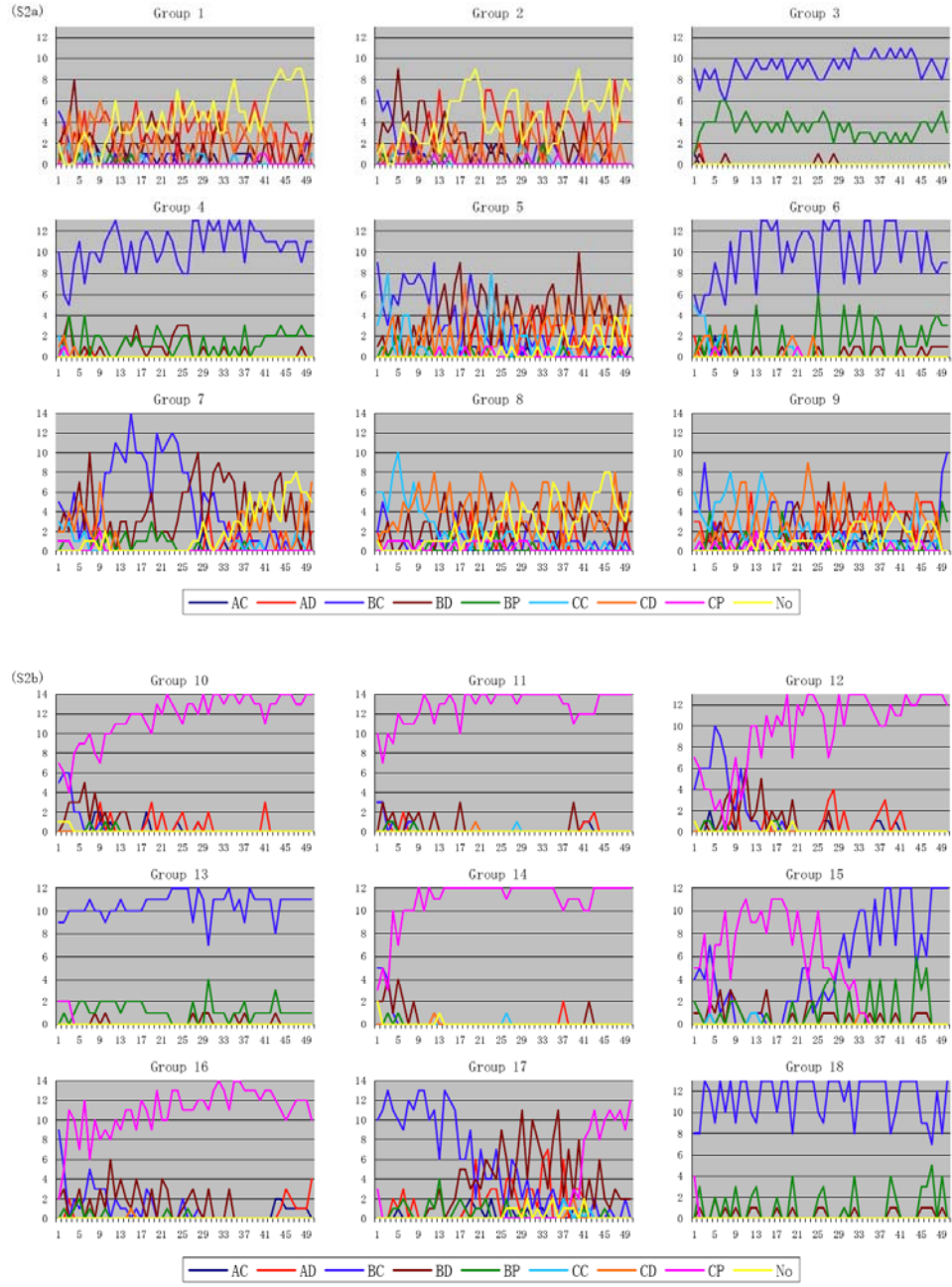
**Figure S1**



The time-evolution, over fifty rounds, of the frequencies of players voting for the games (NoPun), (Peer), (Pool) or (No). In Figure S1a (the first-order treatment), groups 3, 4 and 6 settled on the peer punishment game, (in the sense that during each of the last 10 rounds, more than half of the players opted for it). The six other groups remained undecided. In Figure S1b (second-order treatment), groups 10, 11, 12, 14, 16, 17 settled on the pool punishment game, and groups 13, 15, 18 settled on the peer punishment game.

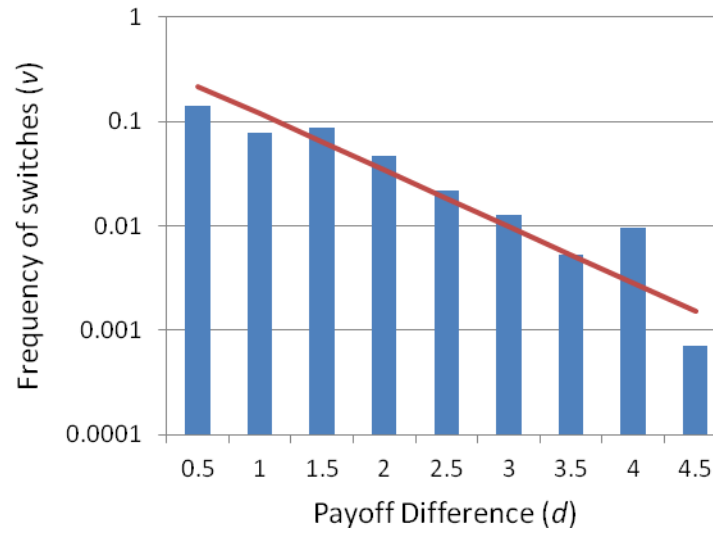


**Figure S2**



The time-evolution, over fifty rounds, of the frequencies of the strategies. Here AC, AD, BC, BD, CC and CD denote contribution resp. defection in (NoPun), (Peer) and (Pool), BP denotes peer-punishment, CP pool-punishment and No non-participation.

**Figure S3**



The frequency  $\nu$  of strategy switches failing to imitate the best is measured on a log scale.  $d$  denotes the payoff difference between the current optimum and the payoff currently achieved by the strategy which will be adopted after the switch. Blue bars represent the frequency of switches with payoff difference in half-open interval  $(d-0.5, d]$ . The red line corresponds to the regression curve  $\nu = 0.4058 \times 0.2888^d$ .

## **Instructions**

### **1. Instructions for the practice rounds (translated into English).**

Welcome and thank you for showing up.

Your minimal payoff will be 10 euros (guaranteed).

We first start with some practice games. These do not count towards your score. You can experiment.

### **COMMUNITY GAME**

In each round, you receive 3 MU and must decide whether or not to contribute 1 MU to your co-players' payoff.

I CONTRIBUTE means: you pay 1 MU and 3 MU will be distributed equally among all your co-players.

I DON'T CONTRIBUTE means: you keep 1 MU. This will not change your co-player's score.

You have 30 seconds for each round to decide and CONFIRM. If you do not decide in time, the computer will make a random decision.

After each round, you will see the scores.

### **EXAMPLE**

If all contribute, all end up with 5 MU.

If no one contributes, all end up with 3 MU.

In mixed groups, contributors always end up with less than the non-contributors.

### **DO YOU WANT TO CONTRIBUTE TO YOUR GROUP?**

YES

NO

*The round is played.*

*The scores are displayed.*

*This is repeated 5 times, with a reflection time of 30 seconds per round.*

### **COMMUNITY GAME WITH OPTION TO PUNISH**

This game consists of 2 stages. At the start of each round you receive 3 units.

The first stage is the community game, as above. You can decide whether or not to contribute 1 unit. You will then see the scores in your group, and how many contributed.

In the second stage, contributors can decide whether or not to punish all those who did not contribute.

If you punish, you have to pay 0.5 MU per non-contributor. Each non-contributor is then fined 1 MU.

You will then see the final score of the round.

### **EXAMPLE**

If 4 players punish a non-contributor, this costs each punisher 0.5 MU, and the punished player 4 MU.

If 3 players punish 2 non-contributors, this costs each punisher 1 MU and each punished player 3 MU.

If 2 players punish 3 non-contributors, this costs each punisher 1.5 MU and each punished player 2 MU.

### **DO YOU WANT TO CONTRIBUTE TO YOUR GROUP?**

YES

NO

x players out of y contributed.

### **DO YOU WANT TO PUNISH ALL NON-CONTRIBUTORS?**

YES

NO

*The round is played.*

*The scores are displayed.*

*This is repeated 5 times, with a reflection time of 30 seconds for each decision.*

## **COMMUNITY GAME WITH PUNISHMENT DEVICE**

At the start of each round you receive 3 MU. Again, you can decide to contribute 1 MU to the group or not. Contributors can additionally decide to pay for a punishment device. This costs the contributor 0.5 MU.

In the first-order treatment: Each punishment device will punish all non-contributors by 1 MU.

In the second-order treatment: Each punishment device will punish all non-punishers by 1 MU (irrespective of whether they contributed or not).

## **EXAMPLE FOR THE SECOND ORDER TREATMENT**

If 3 players chose a punishment mechanism, each pays 0.5 MU and 3 MU will be removed from the account of each player who did not chose the punishment mechanism.

Even if every player chooses the punishment mechanism and no-one will be punished, the costs for the punishment mechanism will have to be paid.

**DO YOU WANT TO CONTRIBUTE TO THE GROUP? DO YOU WANT A PUNISHMENT DEVICE?**

**JUST CONTRIBUTE TO THE GROUP**

**NEITHER, NOR**

**BOTH**

*The round is played.*

*The scores are displayed.*

*This is repeated 5 times, with 30 seconds per decision.*

## **2. Instructions for the full game with option to choose a game (still in the practice rounds)**

You will now have to decide, for each round, which game to play. You will receive 3 units for each round. You can choose to join

**A: COMMUNITY GAME WITH NO PUNISHMENT**

B: COMMUNITY GAME WITH OPTION TO PUNISH

C: COMMUNITY GAME WITH PUNISHMENT DEVICE

You can also decide not to play the game. In this case, you receive an additional 0.5 MU, but you cannot improve.

13 players participate in each round. But the sizes of the groups playing A, B or C are variable. If no co-player joins your group, you receive 0.5 MU and your game is cancelled.

At the end of each round, you will see the scores.

OPT FOR YOUR GAME:

A: COMMUNITY GAME WITH NO PUNISHMENT

B: COMMUNITY GAME WITH OPTION TO PUNISH

C: COMMUNITY GAME WITH PUNISHMENT DEVICE

D: NO GAME

*The round is played.*

*The scores are displayed.*

*This is repeated 10 times, with 30 seconds per decision.*

### **3. Instructions for the experiment (after the practice rounds)**

Now you will be paid according to your score (1 MU is 10 cents, so that 10 MU = 1 euro).

The average payoff will be around 20 euros.

OPT FOR YOUR GAME:

A: COMMUNITY GAME WITH NO PUNISHMENT

B: COMMUNITY GAME WITH OPTION TO PUNISH

C: COMMUNITY GAME WITH PUNISHMENT DEVICE

D: NO GAME


*The round is played*


*The scores are displayed.*

*Repeat this for 50 rounds, with 15 seconds per decision*

## Screen shots

### Login page

**universität  
wien**  
Fakultät für Mathematik

Department of Finance,  
Accounting and Statistics  
**WU**

Willkommen! Danke, dass Sie mitmachen.

**Ihr garantierter Mindestgewinn ist 10 euro.**

Login Daten:

Benutzername:

Passwort:

Experiment-Code:

Sprache:

Copyright ..... designed by Neil

### Practice rounds, instruction, game (NoPun)

**universität  
wien**  
Fakultät für Mathematik

Department of Finance,  
Accounting and Statistics  
**WU**

Willkommen! Dies ist p30 (logout) . Spiel startet um 10:00:00 .

Wir beginnen mit ein paar Übungsspielen. Diese werden Ihre Punktezah nicht beeinflussen. Sie können also experimentieren.

Spende-Spiel

In jeder Runde bekommen Sie 3 E und können entscheiden, ob Sie davon 1 E Ihren Mitspielern spenden.

**ICH SPENDE** heißt, Sie zahlen 1 E, und 3 E werden unter Ihren Spielpartnern gleichmäßig aufgeteilt.

**ICH SPENDE NICHT** heißt, Sie zahlen nichts ein. Die Konten Ihrer Spielpartner werden nicht verändert.

In jeder Runde haben Sie 30 Sekunden Zeit, sich zu entscheiden und **BESTÄTIGUNG** zu drücken. Wenn Sie das nicht zeitgerecht tun, wählt der Computer zufällig eine der beiden Alternativen.

Nach jeder Runde sehen Sie den Punktestand.

Beispiele für das Spende-Spiel


Wenn es keine Spender gibt, bekommt jeder Nichtspender 3E.

Wenn alle spenden, bekommt jeder 5E.

In gemischten Gruppen erhalten die Nicht-Spender immer mehr als die Spender.


Copyright ..... designed by Neil

## Practice rounds, game (NoPun)



**universität  
wien**  
Fakultät für Mathematik

Department of Finance,  
Accounting and Statistics

**WU** 

Willkommen! Dies ist test! (logout) .  
Dies ist die **2. Runde**, Ihre Auszahlung ist **3**  
Bitte treffen Sie Ihre Entscheidung in den nächsten **8 Sekunden**

Wählen Sie Ihr Spiel!

Ihre Wahl: **Spende-Spiel ohne Bestrafung**,

An Ihrem Spiel nehmen 3 Spieler teil.

Wollen Sie für Ihre Gruppe spenden?

☒ JA  
☐ NEIN

**Resultate für Runde 1**

Sie wählten: Spende-Spiel mit Strafmechanismus  
Sie wählten: Spenden und bestrafen  
und erhielten **3 E**

	Anzahl der Spieler	Auszahlung
<b>Ohne Bestrafung</b>	0	
☐ Spenden	0	N/A
☐ Nicht spenden	0	N/A
<b>Mit Strafoption</b>	0	
☐ Nur spenden, nicht bestrafen	0	N/A
☐ Weder spenden noch bestrafen	0	N/A
☐ Spenden und bestrafen	0	N/A
<b>Mit Strafmechanismus</b>	3	
☐ Nur spenden, nicht bestrafen	1	2.5
☐ Weder spenden noch bestrafen	1	5
☐ Spenden und bestrafen	1	3
<b>Kein Spiel</b>		
☐ Nicht-Teilnehmer	0	N/A

Copyright ..... designed by Neil

## Practice rounds, instruction, game (Peer)



**universität  
wien**  
Fakultät für Mathematik

Department of Finance,  
Accounting and Statistics

**WU** 

Willkommen! Dies ist p30 (logout) . Spiel startet um **10:00:00** .

Spende-Spiel mit Strafoption

Dieses Spiel ist zweistufig. Am Anfang jeder Runde erhalten Sie wieder **3 E**.

**Stufe Eins** besteht aus dem Spende-Spiel, wie vorher. Sie können **1 E** spenden oder nicht. Dann sehen Sie die Auszahlungen in Ihrer Gruppe, und die Anzahl der Spender.

In **Stufe Zwei** kann jeder Spender entscheiden, ob er die Nichtspender bestrafen will oder nicht.

Wer bestrafen will, muss pro Nichtspender **0.5 E** zahlen. Jedem Nichtspender werden dann **1 E** abgezogen.

Dann sehen Sie die Auszahlung in dieser Runde.

Beispiele für die Strafoption:

Wenn 4 Spieler einen Nichtspender bestrafen, kostet das jeden Bestrafer 0.5E und den Bestraften 4E.

Wenn 3 Spieler 2 Nichtspender bestrafen, kostet das jeden Bestrafer 1E und jeden Bestraften 3E.


Wenn 2 Spieler 3 Nichtspender bestrafen, kostet das jeden Bestrafer 1.5E und jeden Bestraften 2E.

Spender werden nicht bestraft.

Copyright ..... designed by Neil





## Practice rounds, game (Peer)



**universität  
wien**  
Fakultät für Mathematik

Department of Finance,  
Accounting and Statistics

Willkommen! Dies ist test1 (logout) .  
Dies ist die **2. Runde**, Ihre Auszahlung ist **3**  
Bitte treffen Sie Ihre Entscheidung in den nächsten **8 Sekunden**

Wählen Sie Ihr Spiel!

Ihre Wahl: **Spende-Spiel mit Strafoption**,

An Ihrem Spiel nehmen **3 Spieler** teil.

Wollen sie für ihre Gruppe spenden?

Sie haben **1E** gespendet.

An Ihrem Spiel nehmen **2 Spender**, und **1 Nichtspender** teil.

Möchten Sie die Nichtspender bestrafen?

☒ JA  
☐ NEIN

**Resultate für Runde 1**

Sie wählten: **Spende-Spiel mit Strafmeehanismus**  
 Sie wählten: **Spenden und bestrafen**  
 und erhielten **3 E**

	Anzahl der Spieler	Auszahlung
<b>Ohne Bestrafung</b>	0	
<input type="radio"/> Spenden	0	N/A
<input type="radio"/> Nicht spenden	0	N/A
<b>Mit Strafoption</b>	0	
<input type="radio"/> Nur spenden, nicht bestrafen	0	N/A
<input type="radio"/> Weder spenden noch bestrafen	0	N/A
<input type="radio"/> Spenden und bestrafen	0	N/A
<b>Mit Strafmeehanismus</b>	3	
<input type="radio"/> Nur spenden, nicht bestrafen	1	2.5
<input type="radio"/> Weder spenden noch bestrafen	1	5
<input type="radio"/> Spenden und bestrafen	1	3
<b>Kein Spiel</b>		
<input type="radio"/> Nicht-Teilnehmer	0	N/A

Copyright ..... designed by Neil

## Practice rounds, instruction, game (Pool), first-order variant



**universität  
wien**  
Fakultät für Mathematik

Department of Finance,  
Accounting and Statistics




Willkommen! Dies ist p30 (logout) . Spiel startet um **10:00:00** .

Spende-Spiel mit Strafmeehanismus

Wieder erhalten Sie zu Beginn jeder Runde **3 E**, und können entscheiden, ob Sie Spender von **1 E** sein wollen oder nicht. Spender können zusätzlich einen Strafmeehanismus kaufen. Die Zusatzkosten sind **0.5 E**. Jeder Strafmeehanismus zieht dann **1 E** vom Konto jedes Nichtspenders ab.

Beispiel für den Strafmeehanismus,

Wenn sich 2 Spieler für einen Strafmeehanismus entscheiden, zahlt jeder **0.5 E** und jedem Nichtspender werden **2 E** abgezogen.

Wenn sich 3 Spieler für einen Strafmeehanismus entscheiden, zahlt jeder **0.5 E** und jedem Nichtspender werden **3 E** abgezogen.

Die Kosten für den Strafmeehanismus sind unabhängig von der Anzahl der Nichtspender. Auch wenn niemand bestraft wird (weil jeder spendet), müssen die Kosten des Strafmeehanismus getragen werden.

Copyright ..... designed by Neil

## Practice rounds, instruction, game (Pool), second-order variant



**universität  
wien**  
Fakultät für Mathematik

Department of Finance,  
Accounting and Statistics

**WU** 

Willkommen! Dies ist p30 (logout) . Spiel startet um **10:00:00** .

Spende-Spiel mit Strafmechanismus

Wieder erhalten Sie zu Beginn jeder Runde 3 E, und können entscheiden, ob Sie Spender von 1 E sein wollen oder nicht. Spender können zusätzlich einen Strafmechanismus kaufen. Die Zusatzkosten sind 0.5 E. Jeder Strafmechanismus zieht dann 1 E vom Konto jener Spieler ab, die nicht den Strafmechanismus gewählt haben (egal, ob sie nun gespendet haben oder nicht).

Beispiel für den Strafmechanismus:


Wenn sich 2 Spieler für einen Strafmechanismus entscheiden, zahlt jeder 0.5 E und jedem Spieler, der nicht den Strafmechanismus gewählt hat, werden 2 E abgezogen.

Wenn sich 3 Spieler für einen Strafmechanismus entscheiden, zahlt jeder 0.5 E und jedem Spieler, der nicht den Strafmechanismus gewählt hat, werden 3 E abgezogen.

Auch wenn jeder den Strafmechanismus bezahlt, und daher niemand bestraft wird, müssen die Kosten des Strafmechanismus getragen werden.


Copyright ..... designed by Neil

## Practice rounds, game (Pool)



**universität  
wien**  
Fakultät für Mathematik

Department of Finance,  
Accounting and Statistics

**WU** 

Willkommen! Dies ist test1 (logout) .  
Dies ist die 2. Runde, Ihre Auszahlung ist 3  
Bitte treffen Sie Ihre Entscheidung in den nächsten **9 Sekunden**

Wählen Sie Ihr Spiel!

Ihre Wahl: **Spende-Spiel mit Strafmechanismus**,

An Ihrem Spiel nehmen 3 Spieler teil.

Wollen Sie Spender sein? Wollen Sie einen Strafmechanismus?

☒ Nur Spenden  
☒ Weder, noch  
☐ Beides

**Resultate für Runde 1**

Sie wählten: Spende-Spiel mit Strafmechanismus  
Sie wählten: Spenden und bestrafen  
und erhielten 3 E

	Anzahl der Spieler	Auszahlung
<b>Ohne Bestrafung</b>	0	
<input type="radio"/> Spenden	0	N/A
<input type="radio"/> Nicht spenden	0	N/A
<b>Mit Strafoption</b>	0	
<input type="radio"/> Nur spenden, nicht bestrafen	0	N/A
<input type="radio"/> Weder spenden noch bestrafen	0	N/A
<input type="radio"/> Spenden und bestrafen	0	N/A
<b>Mit Strafmechanismus</b>	3	
<input type="radio"/> Nur spenden, nicht bestrafen	1	2.5
<input type="radio"/> Weder spenden noch bestrafen	1	5
<input type="radio"/> Spenden und bestrafen	1	3
<b>Kein Spiel</b>		
<input type="radio"/> Nicht-Teilnehmer	0	N/A

Copyright ..... designed by Neil

## Practice rounds, instruction, full game with option to choose a game

**universität  
wien**  
Fakultät für Mathematik

Department of Finance,  
Accounting and Statistics  
**WU**

Willkommen! Dies ist p30 (logout) . Spiel startet um **10:00:00** .

Experiment

Sie können sich nun in jeder Runde entscheiden, an welchem Spiel Sie teilnehmen wollen. Sie erhalten **3 E** pro Runde. Sie können wählen:

A: Das Spende-Spiel ohne Bestrafung  
B: Das Spende-Spiel mit Strafoption  
C: Das Spende-Spiel mit Strafmechanismus

Sie können auch beschließen, an keinem Spiel teilzunehmen. In diesem Fall erhalten Sie **0.5 E** zusätzlich, aber können sich nicht verbessern.

**15** Spieler nehmen an jeder Runde teil. Aber die Größe der Gruppen, die A, B oder C spielen, ist veränderlich.

Wenn kein weiterer Spieler ihr Spiel wählt, findet es nicht statt, und Sie erhalten **0.5 E** Entschädigung.

Nach jeder Runde sehen Sie die Ergebnisse.

Copyright ..... designed by Neil

## Practice rounds, full game with option to choose a game

**universität  
wien**  
Fakultät für Mathematik

Department of Finance,  
Accounting and Statistics  
**WU**

Willkommen! Dies ist test1 (logout) .  
Dies ist die **2.** Runde, Ihre Auszahlung ist **3**  
Bitte treffen Sie Ihre Entscheidung in den nächsten **23 Sekunden**

Wählen Sie Ihr Spiel

☒ Spende-Spiel ohne Bestrafung  
☐ Spende-Spiel mit Strafoption  
☐ Spende-Spiel mit Strafmechanismus  
☐ Kein Spiel


**Resultate für Runde 1**

Sie wählten: Spende-Spiel mit Strafmechanismus  
Sie wählten: Spenden und bestrafen  
und erhielten **3 E**

	Anzahl der Spieler	Auszahlung
<b>Ohne Bestrafung</b>	0	
☐ Spenden	0	N/A
☐ Nicht spenden	0	N/A
<b>Mit Strafoption</b>	0	
☐ Nur spenden, nicht bestrafen	0	N/A
☐ Weder spenden noch bestrafen	0	N/A
☐ Spenden und bestrafen	0	N/A
<b>Mit Strafmechanismus</b>	3	
☐ Nur spenden, nicht bestrafen	1	2.5
☐ Weder spenden noch bestrafen	1	5
☐ Spenden und bestrafen	1	3
<b>Kein Spiel</b>		
☐ Nicht-Teilnehmer	0	N/A

Copyright ..... designed by Neil


## Experiment, instruction



**universität  
wien**  
Fakultät für Mathematik

Department of Finance,  
Accounting and Statistics

**WU**




Willkommen! Dies ist p30 (logout) . Spiel startet um 10:00:00 .

Jetzt werden Sie gemäß Ihrer Punktezahl ausbezahlt (1 Einheit ist zehn Cent wert, also 10E=1euro). Der mittlere Gewinn beträgt etwa 20 euro.

Copyright ..... designed by Neil


## Experiment, resulting page



**universität  
wien**  
Fakultät für Mathematik

Department of Finance,  
Accounting and Statistics

**WU**



Willkommen! Dies ist p30 (logout) .

Spiel beendet, Ihre Auszahlung beträgt **227.87**

	Anzahl der Teilnehmer	Mittlerer Gewinn
<b>Ohne Bestrafung</b>	0	N/A
☐ Spenden	0	N/A
☐ Nicht spenden	0	N/A
<b>Mit Strafoption</b>	356	4.51
☐ Nur spenden, nicht bestrafen	265	4.81
☐ Weder-noch	35	3.48
☐ Spenden und bestrafen	56	3.74
<b>Mit Strafmechanismus</b>	238	4.41
☐ Nur spenden, nicht bestrafen	3	-1.33
☐ Weder-noch	2	5
☐ Spenden und bestrafen	233	4.47
<b>Kein Spiel</b>		
☐ Nicht-Teilnehmer	0	N/A

Sie haben **22.8** Euro gewonnen

**Resultate für Runde 50**

Sie wählten: Spende-Spiel mit Strafoption  
Sie wählten: Nur spenden, nicht bestrafen  
und erhielten **5 E**

	Anzahl der Spieler	Auszahlung
<b>Ohne Bestrafung</b>	0	
☐ Spenden	0	N/A
☐ Nicht spenden	0	N/A
<b>Mit Strafoption</b>	12	
☐ Nur spenden, nicht bestrafen	12	5
☐ Weder spenden noch bestrafen	0	N/A
☐ Spenden und bestrafen	0	N/A
<b>Mit Strafmechanismus</b>	0	
☐ Nur spenden, nicht bestrafen	0	N/A
☐ Weder spenden noch bestrafen	0	N/A
☐ Spenden und bestrafen	0	N/A
<b>Kein Spiel</b>		
☐ Nicht-Teilnehmer	0	N/A

Copyright ..... designed by Neil