Bühren, Christoph; Dannenberg, Astrid

Conference Paper

The Demand for Punishment to Promote Cooperation Among Like-Minded People

Beiträge zur Jahrestagung des Vereins für Socialpolitik 2021: Climate Economics

Provided in Cooperation with:
Verein für Socialpolitik / German Economic Association

Suggested Citation: Bühren, Christoph; Dannenberg, Astrid (2021) : The Demand for Punishment to Promote Cooperation Among Like-Minded People, Beiträge zur Jahrestagung des Vereins für Socialpolitik 2021: Climate Economics, ZBW - Leibniz Information Centre for Economics, Kiel, Hamburg

This Version is available at:
https://hdl.handle.net/10419/242427

The Demand for Punishment to Promote Cooperation Among Like-Minded People

Abstract

We use an experiment to test the hypothesis that groups consisting of like-minded cooperators are able to cooperate irrespective of punishment and therefore have a lower demand for a costly punishment institution than groups of like-minded free riders, who are unable to cooperate without punishment. We also predict that the difference in the demand for punishment is particularly large when members know about the composition of their group. The experimental results confirm these hypotheses. However, the information about the composition of the group turns out to be even more important than we expected. It helps cooperative groups to avoid wasting resources for an unneeded punishment institution. In uncooperative groups, it helps members to recognize the need for punishment early on and not to follow an uncooperative path that produces a persistently competitive attitude. These findings highlight the role of group composition and information for institution formation and that lessons learned by one group cannot be readily transferred to other groups.

Keywords: Institution formation; public goods game; cooperation; punishment; controlled group formation

JEL: C72, C91, H41

1. Introduction

Many people are initially reluctant to the idea of punishment but over time come to appreciate it as a way to enforce cooperation (Gürerk et al., 2006; Ertan et al., 2009). There are two possible explanations for the initial reluctance. First, people think the punishment is not needed and a waste of resources, especially when the punishment institution is costly. They resort to punishment only when they experience free riding and adjust their beliefs about the *need* for punishment. Second, people anticipate free riding but do not think that punishment will make a difference. In this case, they may learn that there is more cooperation than they initially expected and adjust their beliefs about the *effect* of punishment. Either way, people learn about others' behavior during the game and the usefulness of punishment. This social learning has become an important field of investigation in behavioral economics. It is based on the observations that people differ in their inclinations to cooperate and in their beliefs about others' cooperativeness. 'Conditional cooperators' are willing to cooperate as long as they know or believe that others cooperate, too, while 'free riders' do not cooperate irrespective of what they believe about others (Fischbacher et al., 2001). In randomly formed groups where conditional cooperators are likely to meet free riders, cooperation usually dwindles over time. By contrast, if conditional cooperators are separated from free riders by some sorting mechanism, they can maintain significantly higher cooperation levels than randomly formed groups and groups consisting only of free riders (Gächter and Thöni, 2005). These findings suggest that groups consisting of like-minded cooperators have a different need for punishment than groups of like-minded free riders.

In this paper, we use an experiment to investigate if different needs for punishment exist and if they translate into different demands for punishment. We use a one-shot public goods game to measure subjects' cooperativeness and form groups of like-minded cooperators and free riders. We call this game the *sorting game*. The sorted groups then play a repeated public goods game where they can choose among the following versions: (i) a standard public goods game without punishment, (ii) a public goods game in which not contributing to the public good is mildly punished and players pay a low institutional cost, and (iii) a public goods game in which not contributing to the public good is severely punished and players pay a high institutional cost. We call this game the *institution formation game*. The available punishment institution is a formal institution that, once in place, is automatically enforced. The institutional cost is borne by all players and increases with the severity of the punishment. As a treatment variable, we vary whether subjects receive information about the initial cooperativeness of their group before they choose the first time among the different versions of the public goods game.

Consistent with our hypothesis, we find that cooperative groups need less punishment than uncooperative groups. Yet, they only demand less punishment when they know about the cooperativeness of their group. The information about the composition of the group plays a more important role than we expected. It helps cooperative groups to recognize that punishment may not be needed and to avoid wasting resources on unneeded institutions. In uncooperative groups, it helps members to acknowledge the need for punishment early on and to avoid following an uncooperative path under a weak institution that fuels a competitive attitude of some people. This attitude prevents successful cooperation even when a strong punishment institution is finally

introduced. These findings highlight the importance of group composition and information about group composition for the formation of institutions.

An important implication of our research is that lessons learned by one group cannot be readily transferred to other groups. Large amounts of empirical research in economics and other social sciences have gone into the analysis of institutions. A lot of this research has considered small-scale societies which share a common pool resource (Ostrom, 1990; Cox et al., 2010), but there are also studies on how the legal systems of larger societies influence social outcomes (Ehrlich, 1977; Levitt and Miles, 2007; Devos, 2013; Dularif et al., 2019). Common to this research is the attempt to measure social outcomes, for instance by the condition of the shared resource or the crime rate, and to test if the outcome can be traced back to the institutional setting. A common finding is that groups characterized by a high willingness to invest in monitoring and imposing sanctions on offenders achieve better outcomes than other groups. Comparing successes and failures in different contexts seems a natural way to gain a better understanding of the functioning of institutional rules. Our experiment, however, cautions against this approach, which we discuss in more detail in the concluding section.

2. Related literature

Our study contributes to the experimental literature on the endogenous choice of institutions in social dilemma games (for a review, see Dannenberg and Gallier, 2020). Participants in these experiments have the opportunity to influence the rules of the game before they play the game, for instance, impose a costly fine on defection in a prisoners' dilemma game. The experiments show how individuals and groups choose between different versions of the game, how they perform after having made that choice, and how they adjust their behavior over time. The implementation of an institution is usually associated with better outcomes, but not all groups use the opportunity to implement an institution. Weak institutions that do not completely remove the incentives to free ride and institutions which cover only a subset of players are not very popular (e.g. Kosfeld et al., 2009; Markussen et al., 2014). Institutional costs also reduce the willingness to implement the institution (e.g. Barrett and Dannenberg, 2017; Dal Bó et al., 2018; Dannenberg et al., 2020). Learning is an important determinant of institution formation as subjects are more likely to vote in favor of the institution in later rounds of the game and when they receive information about other groups (e.g. Gürerk et al., 2006; 2014; Ertan et al., 2009).

Most relevant for our purpose are the studies of formal punishment institutions, which automatically impose a fine when players do not cooperate. Tyran and Feld (2006) compare the adoption of a deterrent and a non-deterrent punishment institution, each relative to a standard public goods game without punishment. The punishment institutions are costless to implement and they are automatically enforced if players contribute less than the full amount. They find that 75 percent of players vote for the game with punishment when it is deterrent and 50 percent vote for it when it is non-deterrent. In both cases, contributions and payoffs are significantly higher with punishment than without. Gallier (2020) provides similar results for a non-deterrent punishment scheme while Vollan et al. (2017) in an experiment with a Chinese sample find that implementing a non-deterrent scheme does not lead to higher payoffs than playing the game without punishment.

The groups in the experiment by Markussen et al. (2014) choose repeatedly between a standard public goods game, a public goods game with a costly formal punishment, and a public goods game with an informal peer punishment option. To avoid strategic voting, players only choose between two institutions at a time. The informal punishment option is surprisingly popular and the groups choosing it earn more on average than the groups choosing the standard game. The formal punishment scheme is only popular and associated with higher payoffs than the standard game when it is cheap. The expensive formal schemes yield higher contributions but not higher payoffs, and even the deterrent scheme is not very popular. Kamei et al. (2015) study the choice between an informal punishment option and a formal punishment scheme. The formal scheme may involve an institutional cost and its severity is determined by the players. Their experimental results show that the groups that implement the formal scheme usually choose the most severe sanction rate. Nevertheless, the costly formal scheme is unpopular and not profitable compared to the informal punishment scheme. Without institutional costs, the majority of groups favor the formal over the informal scheme and earn about the same profits as the groups that implement the informal punishment option.

In all these experiments, groups are formed randomly and then sort themselves into the different institutions, in most cases through majority voting. In our experiment, we first sort the participants into homogenous groups and then let them choose among institutions, which allows us to study the demand for punishment by different types of groups. As another difference to the previous literature, we make the deterrent punishment scheme more expensive than the non-deterrent punishment scheme. This design is arguably more realistic and presents a more difficult trade-off because subjects usually prefer a deterrent over a non-deterrent scheme and a cheap over an expensive scheme.

Our experiment also relates to the theoretical and experimental literature on social preferences and social learning. According to the theoretical literature on social preferences (e.g., Rabin, 1993; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), it should be possible to sort subjects into specific groups which have different needs for punishment to enforce cooperation. The inequality aversion model by Fehr and Schmidt (1999), for example, predicts that a group of subjects who are all highly averse to advantageous inequality may not need punishment, although it could perhaps help to coordinate towards full cooperation. A group of purely selfish subjects will need deterrent punishment while more social groups may do well with cheaper non-deterrent punishment. The experimental literature confirms that controlled group formation has a significant effect on cooperation and punishment. Sorted groups composed of cooperators contribute significantly more to the public good than groups composed of free riders and randomly formed groups (Ones and Putterman, 2007; Gunnthorsdottir et al., 2007; Burlando and Guala, 2015). The closest study to ours is the experiment by Gächter and Thöni (2005), in which randomly formed groups and groups whose members know that they are composed of cooperators or free riders play either a standard public goods game or a public goods game with an informal punishment option. Gächter and Thöni (2005) find that sorted cooperative groups contribute significantly more to the public good than the most cooperative randomly formed groups – they achieve almost efficient cooperation levels with and without the punishment option and rarely use the punishment option. Sorted uncooperative groups contribute more to the public good than the least cooperative randomly formed groups, probably because they know that they cannot rely on any cooperators in the group. The punishment

option does not increase the contributions in the uncooperative groups, but the option is used surprisingly often, mostly targeted at free riders but also at contributors.

A related literature compares cooperative and punitive behavior across different cultural groups. For example, Hermann et al. (2008) compare the behavior of 16 different student samples from around the world in public goods games with and without punishment option. They find that all subject pools punish low contributors but differ greatly in the use of anti-social punishment. Some pools do not punish high contributors, while other pools punish high contributors as much as low contributors and, by this, destroy the cooperation-enhancing effect of punishment. Further analyses of the same dataset by Gächter et al. (2010) indicate that cultural differences in cooperation and punishment exist in the sense that variation within cultures is smaller than across cultures. Henrich et al. (2006) let members of 15 diverse small-scale societies play ultimatum games and third-party punishment games. The authors observe large differences in the willingness to use costly punishment across populations, with some societies showing a very low willingness to punish, others revealing a high willingness, and yet others showing a willingness to punish both too selfish and too generous behavior. The meta-analysis by Balliet et al. (2011) shows that punishment has a large positive effect on cooperation in experiments run in Australia, Japan, Israel, and Switzerland, a medium effect in the Netherlands and the US, and no effect in Russia. In another meta-analysis based on observations from 18 countries, Balliet and Van Lange (2013) find that punishment more strongly promotes cooperation in societies with high levels of trust.

Taken together, there is ample evidence that societies differ in both their need for punishment and their ability to use it to enforce cooperation. We contribute to the literature by showing under highly controlled conditions how different types of groups choose among costly punishment schemes and how they perform under the self-chosen regimes, which to our knowledge has not been studied before.

3. Experimental design

3.1 Implementation

We conducted the experiment online with undergraduate students recruited from the general student population of a German university. In total, 536 students participated in the online experiment with each one taking part in one treatment (between-subject design). We used the o-Tree software to run the experiment (Chen et al., 2016) and ORSEE for the random recruitment of participants for each treatment (Greiner, 2015).

Subjects were informed at the beginning that the experiment would consist of two games and that they would receive the instructions for the second game only when they have finished the first game. After reading the instructions of each game (see Appendix), participants had to answer several control questions. To provide an incentive to read the instructions carefully, participants were offered an additional payment of €1 for each game if they answered all control questions correctly at the first attempt. While this might create a small income effect (for which we can control), the advantage is a generally higher level of understanding, which was important for our purpose and arguably is a bigger issue in online experiments than in lab experiments. Throughout

the experiment, participants could ask questions through a private computer chat with the experimenter. At the end of the experiment, participants completed a short questionnaire on their personal background, including gender, age, field of study, final high school grade, knowledge of game theory and behavioral economics, previous participation in experiments, and their motivation to take part in this experiment.

During the games, earnings were displayed in tokens. Participants knew that payments would be calculated by summing up the number of tokens earned over all games and rounds and by applying an exchange rate of €0.5 per 100 tokens. Payments varied between €10.36 and €23.23 and were made directly after the session via Paypal. Sessions lasted between 70 and 100 minutes.

3.2 Pilot treatments

The task in the experiment involves choosing between and playing different public goods games. The choice is always between a standard public goods game without punishment and a public goods game in which not contributing the full endowment to the public good is automatically punished. If a group chooses the game with punishment, it has to decide further if it wants to implement a relatively cheap non-deterrent punishment scheme ('Mild punishment'), in which zero contribution is still the dominant strategy, or an expensive deterrent punishment scheme ('Severe punishment'), in which full contribution becomes the dominant strategy.

An important goal of our study is to test if cooperative groups choose and use the institutions differently than uncooperative groups. To obtain an indicator of cooperativeness, participants in all treatments start the experiment with a one-shot public goods game. Participants in a given session are randomly divided into groups of $n = 5$ members to play the game. The payoff to player i in the one-shot game is given by $\pi_i = 350 - g_i + 0.3 \sum_{j=1}^{n} g_j$, where g_i denotes player $i's$ contribution to the public good. The five players choose simultaneously how much to contribute to the public and they learn the outcome of this game only at the end of the experiment. We chose a somewhat unusual endowment of 350 to get sufficient variation in contributions and avoid a peak at half of the endowment. Directly after this decision, participants are asked to guess what the other players in their group have contributed on average to the public good, knowing that correct guesses (+/- 5 tokens) are rewarded with €1. Up to this point, the procedure is identical across all treatments.

We started the experiment with a series of pilot treatments.[1] In the first treatment, *Pilot-Random*, subjects are randomly reshuffled after the first game, again into groups of $n = 5$ members, to play the second game. Participants are informed that their co-players in the second game may but do not have to be the same people as in the first game. The second game is a finitely repeated public goods game in which the group stays together for all rounds (partner matching). There are six phases which consist of four rounds each, with the game being fixed within a phase. At the start of each phase, the group chooses the game they want to play, with a simple majority deciding. They

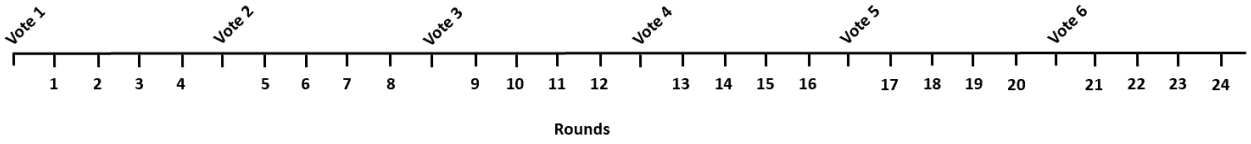
[1] The purpose of the pilot treatments was to test if our parameters elicit enough variation in the choice of institutions for a meaningful comparison between the different classes of groups and if the punishment institutions can potentially be welfare-enhancing. Since the second condition was not met, we decided to redesign the punishment institutions in the main treatments.

first decide if they want to play the game without punishment or with punishment. The instructions use neutral language. The public good is called 'a joint project' and the punishment scheme is called 'deduction system.' All members of the group simultaneously vote either for or against punishment. Abstentions are not allowed. For punishment to be selected, at least three out of the five members must vote in favor of it. Members are informed about the voting outcome, but not about the individual votes. If the group decides in favor of punishment, members choose between Mild and Severe punishment, again with a simple majority deciding and learning only the voting outcome but not the individual votes.

The payoff function of the game without punishment is given by $\pi_i = 100 - g_i + 0.3\sum_{j=1}^{n} g_j$, which is the same as in the first game except for the endowment. If a group decides to implement the Mild punishment scheme, the payoff function changes to $\pi_i = (100 - g_i)(1 - 0.5) + 0.3\sum_{j=1}^{n} g_j - 5$. This implies that half of every token not contributed to the public good is deducted, which can be interpreted as a tax or fine. Additionally, every player incurs a cost of 5 tokens to implement the scheme, which can be interpreted as the costs of monitoring and enforcement. The payoff function for the Severe punishment scheme is given by $\pi_i = (100 - g_i)(1 - 0.9) + 0.3\sum_{j=1}^{n} g_j - 20$. In this case, the fine deducts 90 percent of every token not contributed to the public good and every player incurs an institutional cost of 20 tokens. Markussen et al. (2014) and Kamai et al. (2015) used similar proportional punishment schemes, which deduct a certain proportion of the retained endowment.

After the choice of the game, the group plays the chosen game throughout the phase (four rounds). The chosen game (No punishment, Mild punishment, or Severe punishment) is shown to the players before they start playing and during the whole phase. After having completed a phase, players vote again to determine the game for the next phase. Figure 1 presents the timeline for the institution formation game. This timeline is used in all treatments.

Figure 1. Timeline in the institution formation game



Within each phase, players choose simultaneously in every round how much to contribute to the public good. If the group plays with punishment, deductions and institutional costs are automatically implemented according to the payoff function shown above. After each round, individual contributions and payoffs of all members are displayed on the screen in random order, so that it is not possible to track the contribution by other members over time. Each player is informed how his or her payoff is calculated from the own contribution, others' contributions, and, depending on the chosen game, the deduction and the cost of the punishment institution.

Standard economic theory, based on rational and self-interested actors, predicts that players will vote for and implement the Severe punishment scheme and then contribute their entire endowment to the public good. With this, they receive a payoff of 130, which is substantially higher than a payoff of 100, the predicted outcome of the standard game, and a payoff of 45, the predicted outcome for the game with Mild punishment. Not surprisingly, given the results of the previous literature, this is not what we observe for the 14 groups that took part in the *Pilot-Random* treatment. The Severe punishment scheme is chosen in most of the phases (46%), but also Mild punishment (27%) and No punishment (26%) are chosen frequently. Average contributions increase with the severity of punishment, from 33 tokens in the standard game to 60 tokens under Mild punishment and 86 tokens under Severe punishment. However, these higher contribution levels do not suffice to compensate for the costs of punishment. Average payoffs are highest in the standard game without punishment (117 tokens), followed by Severe punishment (111 tokens) and Mild punishment (105 tokens).

As the next step, we used the contributions made in the sorting game to form homogenous groups for the institution formation game. For this, the five subjects with the highest contribution in a given session form the first group, the five subjects with the second-highest contributions form the second group, and so on. Two or more subjects at the border who have chosen the same contribution are assigned randomly. A similar sorting mechanism was used by Gächter and Thöni (2005). We distinguish between two treatments, *Pilot-Sorted* and *Pilot-Sorted-Info*. Subjects in both treatments, like in the *Pilot-Random* treatment, are informed that their co-players in the second game may but do not have to be the same people as in the first game.[2] Subjects in the *Pilot-Sorted-Info* treatment get additional information about what their current co-players in the second game have contributed in the first game. Hence, unlike players in the *Pilot-Sorted* treatment, they know from the start whether they are in a cooperative or uncooperative group. After the sorting, and in *Pilot-Sorted-Info* the provision of information, subjects play the institution formation game as described above and illustrated in Figure 1. To summarize the behavior of the 15 groups in *Pilot-Sorted* and 13 groups in *Pilot-Sorted-Info*, we distinguish between "HIGH cooperators" who contribute more than half of the endowment in the sorting game and "LOW cooperators" who contribute half or less. With this classification, each category contains half of the observations in both treatments. In both treatments, LOW cooperators choose Severe punishment more often than HIGH cooperators, with the difference being greater in *Pilot-Sorted-Info* (43% vs. 28% of all phases) than in *Pilot-Sorted* (29% vs. 21%). This suggests that uncooperative groups have a higher demand for Severe punishment, especially when subjects are informed about the cooperativeness of their group. However, neither the uncooperative nor the cooperative groups benefit from Severe punishment. HIGH cooperators choose high contributions, irrespective of punishment, or whether

[2] Subjects were not informed about the sorting. As we did not use any false or misleading instructions, we still comply with the regulations of the laboratory that we used for the experiment. We also obtained ethical approval from the German Association of Experimental Economic Research (https://gfew.de/en). We are aware that other laboratories use stricter regulations, which we do not comply with in this experiment. Specifically, it is possible that, if subjects had known about the sorting, they would have decided differently in the sorting game. Omitting this information is the only way of getting a clean sorting and it is common practice in the experimental literature on assortative matching (e.g. Gächter and Thöni, 2005; Ones and Putterman, 2007; Gunnthorsdottir et al., 2007; Burlando and Guala, 2015). The analysis is impossible without the sorting mechanism because the probability of obtaining groups of like-minded cooperators by chance is too low. Relying on chance alone would require an unrealistically large subject pool and wasting a large proportion of the data.

they are informed about the cooperativeness of the group or not. In each version of the game, their average payoff is not far away from the social optimum, which means that they do not benefit from punishment because it is costly and they do not need it. LOW cooperators contribute only little when there is no punishment in place. Average contributions more than double under Severe punishment. However, a sizable minority of subjects deviate from the dominant strategy of contributing everything and choose a lower suboptimal contribution. In both treatments, the group average of LOW cooperators is below the efficient level in almost 80% of all rounds in which they play under Severe punishment. In comparison, this happens in groups of HIGH cooperators in less than 50% of all rounds. These suboptimal contributions together with the institutional cost make the Severe punishment scheme unprofitable for LOW cooperators compared to the game without punishment. A possible explanation for the suboptimal contributions under Severe punishment is that people misunderstand the punishment scheme and our data provide some evidence for this. Another possible explanation is that people want to earn more than others because a suboptimal contribution reduces the own payoff, compared to full contribution, but it reduces the other members' payoffs even more. We know from the literature on anti-social punishment that some people are very competitive and willing to pay for an advantageous position within the group (Hermann et al., 2008). Whether it is a lack of understanding or a desire to come out first, the punishment scheme may work better if the people who choose the optimal contributions had a better chance of earning more than the people who choose suboptimal contributions. This is what we aim at in the main treatments.

3.3 Main treatments

Our main treatments *Sorted* and *Sorted-Info* correspond exactly to the pilot treatments in that subjects first play the sorting game and then the institution formation game, in which they choose between the standard public goods game without punishment and the game with either Mild or Severe punishment. The groups in our main treatments are sorted according to their contributions in the sorting game before they play the institution formation game. Players in *Sorted-Info* additionally receive information about their co-players' contributions in the sorting game. During the institution formation game, groups in *Sorted* and *Sorted-Info* receive the same information. The only difference between the main treatments and the corresponding pilot treatments is the design of the punishment schemes. The payoff function of the standard game without punishment is the same as before and given by $\pi_i = 100 - g_i + 0.3\sum_{j=1}^{n} g_j$. If a group implements the Mild punishment scheme, the payoff function changes to

$$\pi_i = \begin{cases} 100 - g_i - 50 + 0.3\sum_{j=1}^{n} g_j - 5, \ if\ g_i < 100 \\ 100 - g_i + 0.3\sum_{j=1}^{n} g_j - 5, \ if\ g_i = 100 \end{cases}.$$

If a group implements the Severe punishment scheme, the payoff function changes to

$$\pi_i = \begin{cases} 100 - g_i - 90 + 0.3\sum_{j=1}^{n} g_j - 20, \ if\ g_i < 100 \\ 100 - g_i + 0.3\sum_{j=1}^{n} g_j - 20, \ if\ g_i = 100 \end{cases}.$$

The institutional costs, 5 tokens for Mild punishment and 20 tokens for Severe punishment, are the same as before. What is different is that the maximum penalty is triggered as soon as a player contributes less than the full endowment, even when the deviation is small. With this, punishment no longer 'fits the crime' but instead imposes an absolute penalty for all possible deviations from full contribution, which becomes very salient. A similar absolute punishment scheme was used by Feld and Tyran (2002) and Tyran and Feld (2006). The advantage, and the main reason for switching to this design, is that people, who contribute everything, have a better chance to earn more than people who choose lower contributions. Under the Mild punishment scheme, a player must contribute less than 50 tokens to earn more than a person who contributes everything. With Severe punishment, a player must contribute less than 10 tokens to earn more than a person who contributes everything. Theoretically, the absolute punishment and the proportional punishment are equivalent. A group of rational and self-interested individuals will, in either case, choose the Severe punishment scheme and then contribute the full endowment to the public good. However, a group in which some individuals choose their contributions more naively, based on who earns the most, may learn and perform better under the absolute punishment scheme.

3.4 Hypotheses for the main treatments

To organize and present the collected data, we divide the groups into 'HIGH cooperators' with an average contribution level of more than 250 tokens in the sorting game, 'MIDDLE cooperators' with an average contribution between 150 and 250 tokens, and 'LOW cooperators' with an average of less than 150 tokens. This classification makes sure that we have roughly a third of the observations in each class in both treatments.

Based on the existing evidence from the previous literature on controlled group formation (Gächter and Thöni, 2005; Ones and Putterman, 2007; Gunnthorsdottir et al., 2007; Burlando and Guala, 2015) and formal punishment schemes (Tyran and Feld, 2006; Markussen et al., 2014; Kamei et al., 2015), we formulate the following hypotheses:

Hypothesis 1

a) HIGH cooperators will choose high contributions with and without punishment.
b) HIGH cooperators will have a lower demand for punishment than MIDDLE and LOW cooperators.

Hypothesis 2

a) MIDDLE cooperators will contribute more with punishment than without punishment.
b) MIDDLE cooperators will have a higher demand for punishment than HIGH cooperators.

Hypothesis 3

a) LOW cooperators will contribute more with punishment than without punishment.

b) LOW cooperators will have a higher demand for punishment than MIDDLE or HIGH cooperators.

Hypothesis 4

The differences in the demand for punishment will be larger when subjects are informed about the composition of their group than when they are not.

Hypotheses 2a and 3a are based on the robust finding that formal punishment schemes increase contributions to the public good (Tyran and Feld, 2006; Markussen et al., 2014; Kamei et al., 2015). The potential for this increase in contributions of course is greater when groups contribute only little in the absence of punishment. According to the results by Gächter and Thöni (2005), we can expect sufficient room for an increase in contributions for the LOW groups and, perhaps to a lower extent, for the MIDDLE groups. In contrast, there may be no room for an increase in contributions for the HIGH groups, which is expressed in Hypothesis 1a. If these three hypotheses are true, there is a higher need for punishment in uncooperative groups than in cooperative groups and a larger potential for the punishment institutions to increase payoffs. Hypotheses 1b, 2b, and 3b then express the assumption that LOW groups will have the highest demand for punishment followed by MIDDLE groups and then HIGH groups. We refrain from forming hypotheses on which of the two punishment institutions will be used more frequently and more effectively because it is ex ante not clear if severity or cost will be the dominant factor. Finally, Hypothesis 4 predicts that the information provided in *Sorted-Info* helps players to recognize the need for punishment in their respective groups and adjust the demand accordingly. Nevertheless, as the information that is provided during the 24 rounds of the institution formation game is the same in *Sorted* and *Sorted-Info*, the differences between the two treatments may be pronounced at the beginning and level off over time.

4. Results

4.1 Contributions and demand for punishment by HIGH, MIDDLE, and LOW cooperators

For the analysis, we have 24 groups of LOW cooperators with an average contribution of 75 tokens in the sorting game, 24 groups of MIDDLE cooperators with an average contribution of 195 tokens, and 23 groups of HIGH cooperators with an average of 323 tokens.

Figure 2 shows average contributions and payoffs in the institution formation game across all rounds separated by treatment, class, and institution. Figures 3 and 4 illustrate how contributions and payoffs in each class and institution develop over time. Let's start with the HIGH cooperators shown on the right-hand side in the figures. They make very high contributions under all institutions and manage to sustain this high level throughout the game. Even without punishment, these groups contribute more than 80% of the endowment on average and show only a small last-round effect. These results confirm Hypothesis 1a. Because the HIGH cooperators always make high contributions and the punishment institutions are costly, they earn the highest payoffs in the game without punishment. Tables 1 and 2 present regression results on average contributions and payoffs separated by treatment and class. They show that the HIGH cooperators earn significantly less with punishment than without punishment. Contributions are not significantly affected by the

punishment institutions. It is interesting to see that average contributions by the HIGH cooperators under the different institutions are very similar in *Sorted* and *Sorted-Info*, even at the beginning of the game where the subjects in *Sorted-Info* have the advantage of knowing that they are in a cooperative group. This can be explained by our sorting mechanism. Recall that subjects choose their contributions in the sorting game without any information about their co-players and the subsequent sorting. HIGH cooperators are therefore people who cooperate even under strategic uncertainty. Nevertheless, the information provided in *Sorted-Info* does have an effect and this concerns the voting behavior. Figure 5 shows how often the institutions are implemented separated by treatment and class. The left panel shows the distribution of institutions in the first phase and the right panel shows the average distribution of institutions across all rounds. Most of the HIGH groups in *Sorted* implement a punishment institution when they decide the first time; only 8% choose the game without punishment. In contrast, 40% of the HIGH groups in *Sorted-Info* choose the game without punishment when they decide the first time and the remaining 60% choose the Mild punishment scheme. A Chi^2 test confirms that the distribution of institutions in the first phase is significantly different between *Sorted* and *Sorted-Info* ($p=0.055$). The information provided in *Sorted-Info* thus helps cooperative groups to realize that punishment may not be needed. Do the HIGH groups in *Sorted* learn this lesson over the course of the game? The answer is no. Across all rounds, the distribution of institutions is still different between *Sorted* and *Sorted-Info* ($p=0.004$). In *Sorted*, HIGH groups play the game without punishment only 24% of the time on average, which compares to 48% in *Sorted-Info*. Likewise, the Severe punishment scheme is chosen more often in *Sorted* than in *Sorted-Info* (22% vs. 7%). Learning over the course of the game is not equivalent to prior information about the cooperativeness of the group, arguably because the motivation behind the high contributions is more ambiguous. For instance, if a group starts playing with a punishment scheme in place, it is not clear whether high contributions are made because of punishment or because the members are cooperative. Even if a group starts playing without punishment, a high contribution level is not an unambiguous signal of cooperativeness because some people may contribute for strategic reasons at the beginning of the game. Due to the lower use of punishment in *Sorted-Info*, HIGH cooperators earn significantly higher payoffs in this treatment than in *Sorted* according to a Mann-Whitney-U (MWU) test ($p<0.01$).

Let's now turn to the MIDDLE cooperators. Figure 2 shows that the MIDDLE cooperators contribute more when a punishment scheme is in place than when they play without punishment, confirming Hypothesis 2a. The difference in contributions with and without punishment is more pronounced in *Sorted* than in *Sorted-Info*. The regression results in Table 1 indicate that, in *Sorted*, contributions under both Mild and Severe punishment are significantly higher than without punishment, while, in *Sorted-Info*, only the difference between Severe and No punishment is weakly significant. The coefficients of the punishment institutions in *Sorted* are roughly twice as large as in *Sorted-Info*. The main reason for this is that MIDDLE cooperators make larger contributions in the game without punishment when they are informed about the cooperativeness of their co-players and thus know that there are no strong free riders. This difference implies that the punishment institutions are only profitable in *Sorted* but not in *Sorted-Info*. MIDDLE groups in *Sorted* earn significantly higher payoffs under both Mild and Severe punishment compared to No punishment. MIDDLE groups in *Sorted-Info* earn significantly less under Severe punishment compared to No punishment. However, these differences in the profitability of the punishment

institutions are not reflected in the demand for punishment (see Figure 5). In both treatments, Mild punishment is the most popular institution, at the beginning of the game and across all rounds. The game without punishment is more popular than the Severe punishment at the beginning but similarly popular across all rounds. There are no significant differences in the distribution of institutions between *Sorted* and *Sorted-Info*, neither in the first phase (Chi^2 test, p=0.856) nor overall (p=0.561). The information about the cooperativeness of the co-players obviously does not change the perceived need for punishment, perhaps because MIDDLE groups come closest to what subjects expect when they do not have any information about their group.

Turning to the last class, the LOW cooperators, we find that they make significantly higher contributions with punishment than without (see Figure 2 and Table 1). The differences are highly significant in both *Sorted* and *Sorted-Info*. This confirms our Hypothesis 3a. In the game without punishment, average contributions are very low, below 20% of the endowment, and they remain at a low level throughout the game in both treatments (see Figure 3). What is interesting in this class is that contributions under both punishment schemes are higher in *Sorted-Info* than in *Sorted*. A MWU test shows significantly larger contributions for both Mild punishment (p=0.023) and Severe punishment (p=0.019) when subjects are informed about their co-players' cooperativeness. The regression results in Table 1 show substantially higher coefficients of the punishment institutions in *Sorted-Info* than in *Sorted*. This difference is especially puzzling for Severe punishment, where contributing the full endowment is the dominant strategy. We discuss this issue in greater detail further below. Because of this pattern in contribution behavior, the Severe punishment scheme leads to significantly higher payoffs compared to No punishment in *Sorted-Info* but not in *Sorted*. Mild punishment never leads to significantly higher payoffs compared to No punishment. Figure 5 shows that voting behavior also differs between treatments. LOW groups in *Sorted* start the game either without punishment (55%) or with Mild punishment (45%). No group implements the Severe punishment scheme in the first phase. Across all rounds, the Severe punishment scheme is implemented only 21% of the time, which is plausible once we know that it does not lead to higher payoffs. In *Sorted-Info*, 23% of LOW groups start the game with Severe punishment and, across all phases, it is implemented 42% of the time. A Chi^2 test indicates a significant difference in the distribution of institutions between *Sorted* and *Sorted-Info* across all phases (p=0.012), though not for the first phase (p=0.229). Due to the higher and more efficient use of the Severe punishment scheme in *Sorted-Info*, LOW cooperators earn significantly more in this treatment than in *Sorted* (MWU test, $p<0.01$).

4.2 Comparison of HIGH, MIDDLE, and LOW cooperators

We predicted that the LOW cooperators have a higher demand for punishment than the MIDDLE cooperators and that both of them have a higher demand than the HIGH cooperators. This is precisely what we observe in *Sorted-Info*. The distribution of institutions displayed in Figure 5 shows that the demand for punishment is highest for the LOW groups, somewhat lower for the MIDDLE groups, and lowest for the HIGH groups. A series of Fisher's exact tests shows that on average LOW and MIDDLE groups are more likely to implement a punishment scheme than HIGH groups ($p<0.05$ each) and they are more likely to implement the Severe punishment scheme ($p<0.05$ each). LOW groups are more likely to implement the Severe punishment scheme than MIDDLE

groups (p=0.093). In *Sorted*, however, we do not observe this ranking. If anything, the demand for punishment is lower for LOW groups than for HIGH and MIDDLE groups, though only the comparison of LOW and MIDDLE regarding their likelihood of implementing a punishment scheme reaches statistical significance (p=0.019). Using correlation tests, we find in the *Sorted* treatment a small positive correlation between subjects' contribution in the sorting game and their likelihood of voting in favor of punishment in the first phase of the institution formation game (rho=0.130, p=0.088). Cooperatively inclined people thus show a greater willingness to implement a punishment institution, which has also been found in experiments with randomly formed groups (Dal Bó et al., 2010; Ertan et al., 2009; Vollan et al., 2017; Gallier, 2020). In *Sorted-Info* the correlation is insignificant (rho=-0.025, p=0.740), which confirms that the information helps the subjects to better recognize the need for punishment.

Taken together, Hypotheses 1b, 2b, 3b are confirmed for *Sorted-Info* but not for *Sorted*. Hypothesis 4, which predicted larger differences in the demand for punishment in *Sorted-Info* than in *Sorted*, is overfulfilled because the differences in *Sorted* are not only smaller but go in the opposite direction.

So why do the different needs of punishment not translate into different demands for punishment in *Sorted*? We have argued above that it is difficult for groups of HIGH cooperators in *Sorted* to learn about their cooperativeness, especially when they start playing with punishment. This can explain why their demand for punishment is relatively high in *Sorted*. However, the LOW cooperators in *Sorted* should realize quickly that the contribution level without punishment is rather low. Why do they not implement the Severe punishment scheme more often? The answer is that the LOW cooperators do not use the punishment scheme wisely in this treatment. In 64% of all rounds in which the LOW groups play with the Severe punishment scheme in *Sorted*, they contribute less than the efficient amount. The same happens in *Sorted* for MIDDLE groups 20% of the time and for HIGH groups 26% of the time. In *Sorted-Info*, this happens 27% of the time for the LOW groups. This means that a non-negligible share of players deviates from the dominant, and socially optimal, strategy under Severe punishment and that this share is particularly large among the LOW cooperators in *Sorted*. Most of these deviations are zero contributions (70%, on average, and 77% among LOW cooperators in *Sorted*), implying that these people earn less than they could but more than the people who make positive contributions.

Table 3 presents regression results on the likelihood of contributing less than 100 tokens under the Severe punishment scheme, separated by class and treatment. With this explorative analysis we try to explain why suboptimal contributions under Severe punishment occur particularly often among LOW cooperators in *Sorted* and why our hypotheses on the demand for punishment fail there. The regression results provide three explanations for the deviation from the dominant strategy. The *first* explanation is incomplete comprehension of the game. Before the game started, subjects had to answer several control questions, out of which three questions were about the functioning of the Severe punishment scheme. Interestingly, the HIGH cooperators are more likely to answer the questions about the functioning of this punishment scheme correctly at the first attempt compared to MIDDLE and LOW cooperators (66% vs. 55% and 51%, Chi^2 test: p=0.051), while there are no significant differences when we consider all the other control questions. This suggests that cooperatively inclined people find it easier to recognize when cooperation is profitable and when

it is not. The regression results in Table 3 show that better comprehension leads to fewer deviations from the dominant strategy among the LOW and MIDDLE cooperators in *Sorted-Info*. We do not observe this in *Sorted*, where other factors appear to be more important.

The *second* explanation is that subjects who are outvoted and do not support the Severe punishment scheme are more likely to make <100 contributions than subjects who vote in favor of it. The negative sign of voting in favor of Severe punishment exists for all classes and treatments, but the effect is sizable and significant only for the LOW cooperators in *Sorted*; they are 22 percentage points more likely to deviate from full cooperation if they did not vote for the Severe punishment institution. Perhaps these subjects are disappointed about the voting outcome or they try to sabotage cooperation in the current phase to force the group to switch to another institution in the next phase. We do not observe this for the LOW groups in *Sorted-Info*, where the marginal effect is small and insignificant, perhaps because there is greater awareness of the need for Severe punishment even among those who vote against it.

The *third* explanation is the desire of some people to earn more than others. To shed some light on this motivation, we test if subjects who earned more than the group average in the previous round are more likely to make <100 contributions under Severe punishment in the next round. We find that this is indeed the case for LOW cooperators in both treatments, with the effect being larger in *Sorted* than in *Sorted-Info* (21 vs. 7 percentage points). These subjects seem to be willing to forego money in order to come out first in their group. The reason why this effect is larger in *Sorted* may lie in the dynamics of the game. All LOW groups in *Sorted* start the game without punishment or with Mild punishment; that is, with an institution under which people who make low contributions typically earn more than people who make high contributions. If a person wants to continue to earn more than the others under Severe punishment, he or she must choose a contribution close to zero. Subjects' experiences in *Sorted-Info* are different because more LOW groups implement Severe punishment early in the game. The regression results show that both starting the first phase with punishment and choosing Severe when playing the first time with punishment have a significant positive effect on making full contributions under Severe punishment, except for the HIGH cooperators. Also, LOW and MIDDLE cooperators in *Sorted-Info* are more likely to play the dominant strategy in the second half of the game than in the first half, confirming that this treatment is better suited for these groups to learn how to apply this punishment scheme.

Together, the results suggest that the information provided in *Sorted-Info* works through two channels in the LOW groups. First, it signals the need for punishment, also to those individuals whose first choice is not Severe punishment and who may otherwise retaliate. Second, the information allows the LOW groups to implement the Severe punishment scheme early on – on average in phase 2 compared to phase 4 in *Sorted* – and this facilitates the learning process towards successful cooperation. These differences explain why the Severe punishment scheme is profitable in *Sorted-Info* but not in *Sorted*.

The last question we want to investigate is whether we can use our data to predict which people will fall into the HIGH, MIDDLE, and LOW class. Through the ex-post questionnaire, we have information about our subjects' age, gender, final high school grade, field of study, number of semesters, previous participation in experiments, and prior knowledge in game theory and behavioral economics. Regression analysis shows that none of these variables has any predictive

power. This indicates that it would be difficult to replace the sorting game with another mechanism to sort people into the different classes.

5. Discussion and conclusion

The most important implication of our research is that the lessons learned by one group cannot be readily transferred to other groups. Strictly speaking, this holds for every single group. We know how the groups behaved under the chosen institutions, but we do not know how they would have behaved if they had chosen a different institution. In this sense, our classification into HIGH, MIDDLE, and LOW cooperators is merely an approximation. The chance that a group does well without punishment is higher in HIGH than in MIDDLE and LOW, even though there may exist HIGH groups that better play with punishment than without. Relying on this approximation, we can identify the specific challenges that the different classes of groups face. The HIGH groups do well under all institutions, but they need reliable information about their cooperativeness because this is difficult to infer in the institution formation game. The problem for the MIDDLE groups is that, even when they possess information about the composition of the group, this information does not offer clear advice about which institution is the best. It does not seem to be obvious to the members that the absence of strong free riders is sufficient to perform well without any punishment. Finally, the challenge for the LOW cooperators is to recognize and accept the need for Severe punishment and to implement it before a competitive attitude can spread in the group. The inability of some people in this class to adjust their behavior from a competitive strategy, that benefits only themselves, to a cooperative strategy under punishment, that benefits not only themselves but also the others, is one of the most interesting results of our experiment.

Our data also provide indications of what measures could help to select the right institutions for successful cooperation. Information about the composition of the own group appears to be a valuable no-harm measure because it helps the HIGH and LOW cooperators and does not harm the MIDDLE cooperators. On the other hand, information about the implementation and impact of institutions in *other* groups may be harmful, especially if these other groups differ in some unobservable ways. The more efficient use of Severe punishment in the main treatments as compared to the pilot treatments indicates that institutions work better when people who comply with the rules are better off than people who violate the rules. While this meant a drastic increase in the penalty level in our experiment, there are other possibilities in the real world, such as naming and shaming, loss of reputation, or social ostracism, that create differences between cooperators and defectors. It is also important that people understand the purpose, the need, and the functioning of the punishment institution so that enough people support the implementation and that those who do not actively support the implementation still accept it and do not retaliate. Although the LOW cooperators come closest to the prediction of standard economic theory, it would be too simple to equate these subjects with the rational and self-interested ‘homo economicus’. They are sensitive to information about the group and changes in the institutional setting. These results, together with the results on the HIGH and MIDDLE cooperators, provide further evidence that standard economic theory does not make reliable predictions for the formation and impact of institutions and that it needs to be supplemented by social preferences and social learning.

We believe that our experimental results are not only relevant to the theory of social preferences and learning but also have important implications for the empirical analysis of institutions. Much research in economics and the social sciences has gone into the analysis of institutions to understand why some groups succeed in solving collective action problems and other groups fail (Dietz et al., 2003; Faysse, 2005; Poteete et al., 2010). A relatively robust observation in the field is that successful groups are characterized by a high willingness to monitor the behavior of the group members and impose sanctions on the members who violate the rules (Ostrom, 1990; Cox et al., 2010). A tempting conclusion is that the implementation and enforcement of institutional rules are responsible for the success and that the comparison of successes and failures can help to gain an understanding of what works and what does not work to support collective action. Hilborn (2007), for example, argues that the approach of comparing successes and failures is akin to an approach in medicine, where people who are immune to an infectious disease are compared to those who are not in order to develop a vaccine.[3] Although he acknowledges that social systems are more complex, he believes that some general conclusions can be drawn from the comparison. Our findings caution against this approach because they show that it is very difficult to draw conclusions from the experience of one group to the possibilities of other groups. There is evidence that societies differ in their views on appropriate punishment and law compliance (Van Kesteren, 2009; Marien and Hooghe, 2011) and also that similar institutions can have different impacts in different societies. For example, anti-smoking laws had very different effects in Norway and Greece (Nyborg et al., 2016). The management of irrigation systems through strong government agencies and water markets worked well in some areas of the world but failed to live up to expectations in other areas because "the variability of local situations and the difficulty of transplanting institutions from one context to another were not taken into account" (Meinzen-Dick, 2007, p. 15200). Our results show that not only the local situation and context have to be taken into account but also the social preferences of the group, which may be difficult to measure. Of course, the necessary caution regarding the external validity of laboratory experiments also applies to our experiment. However, the usually criticized homogeneity of student samples makes our investigation rather conservative because one might expect even bigger differences in a more diverse sample.

Acknowledgements. The work was financially supported by the European Union (EU) Horizon 2020 program, action ERC-2014-STG, Project HUCO, grant number 636746.

[3] Hilborn (2007) refers to the 18th-century physicist Edward Jenner who used the observation that milkmaids were generally immune to smallpox to develop vaccination based on cowpox, a similar but less virulent and less dangerous disease, which milkmaids got from cows.

References

Balliet, D., Mulder, L. B., & Van Lange, P. A. M. (2011). Reward, punishment, and cooperation: A meta-analysis. *Psychological Bulletin, 137*(4), 594–615.

Balliet, D., & Van Lange, P. A. M. (2013). Trust, punishment, and cooperation across 18 societies: A meta-analysis. *Perspectives on Psychological Science, 8*(4), 363-379.

Barrett, S., & Dannenberg, A. (2017). Tipping versus cooperating to supply a public good. *Journal of the European Economic Association, 15*(4), 910-941.

Bolton, G. E., & Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review, 90*(1), 166-193.

Burlando, R. M., & Guala, F. (2005). Heterogeneous agents in public goods experiments. *Experimental Economics, 8*(1), 35-54.

Chen, D. L., Schonger, M., & Wickens, C. (2016). oTree – An open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance, 9*, 88-97.

Cox, M., Arnold, G., & Tomás, S. V. (2010). A review of design principles for community-based natural resource management. *Ecology and Society, 15*(4): 38, 1-19.

Dal Bó, E., Dal Bó, P., & Eyster, E. (2018). The demand for bad policy when voters underappreciate equilibrium effects. *The Review of Economic Studies, 85*(2), 964-998.

Dal Bó, P., Foster, A., & Putterman, L. (2010). Institutions and behavior: Experimental evidence on the effects of democracy. *American Economic Review, 100*, 2205–2229.

Dannenberg, A., & Gallier, C. (2020). The choice of institutions to solve cooperation problems: A survey of experimental research. *Experimental Economics, 23*, 716–749.

Dannenberg, A., Haita-Falah, C., & Zitzelsberger, S. (2020). Voting on the threat of exclusion in a public goods experiment. *Experimental Economics, 23*, 84–109.

Devos, K. (2013). The role of sanctions and other factors in tackling international tax fraud. *Common Law World Review, 42*, 1–22.

Dietz, T., Ostrom, E., & Stern, P. C. (2003). The struggle to govern the commons. *Science, 302*(5652), 1907-1912.

Dularif, M., Sutrisno, T., & Saraswati, E. (2019). Is deterrence approach effective in combating tax evasion? A meta-analysis. *Problems and Perspectives in Management, 17*(2), 93-113.

Ehrlich, I. (1977). Capital punishment and deterrence: Some further thoughts and additional evidence. *Journal of Political Economy, 85*(4), 741-788.

Ertan, A., Page, T., & Putterman, L. (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review, 53*, 495–511.

Faysse, N. (2005). Coping with the tragedy of the commons: Game structure and design of rules. *Journal of Economic Surveys, 19*(2), 239-261.

Fehr, E., & Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *The Quarterly Journal of Economics*, *114*(3), 817-868.

Feld, L. P., & Tyran, J. R. (2002). Tax evasion and voting: An experimental analysis. *Kyklos*, *55*(2), 197-222.

Fischbacher, U., Gächter, S., & Fehr, E. (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics letters*, *71*(3), 397-404.

Gächter, S., Herrmann, B., & Thöni, C. (2010). Culture and cooperation. *Philosophical Transactions of the Royal Society B*, *365*, 2651–2661.

Gächter, S., & Thöni, C. (2005). Social learning and voluntary cooperation among like-minded people. *Journal of the European Economic Association*, *3*(2-3), 303-314.

Gallier, C. (2020). Democracy and compliance in public goods games. *European Economic Review*, *121*, 103346.

Greiner, B. (2015). Subject pool recruitment procedures: Organizing experiments with ORSEE. *Journal of the Economic Science Association*, *1*(1), 114-125.

Gunnthorsdottir, A., Houser, D., & McCabe, K. (2007). Disposition, history and contributions in public goods experiments. *Journal of Economic Behavior & Organization*, *62*(2), 304-315.

Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2006). The competitive advantage of sanctioning institutions. *Science*, *312*(5770), 108-111.

Gürerk, Ö., Irlenbusch, B., & Rockenbach, B. (2014). On cooperation in open communities. *Journal of Public Economics*, *120*, 220-230.

Henrich, Joseph, *et al.* (2006). Costly punishment across human societies. *Science*, *312*(5781), 1767-1770.

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

Hilborn, R. (2007). Managing fisheries is managing people: what has been learned? *Fish and Fisheries*, *8*, 285-296.

Kamei, K., Putterman, L., & Tyran, J. R. (2015). State or nature? Endogenous formal versus informal sanctions in the voluntary provision of public goods. *Experimental Economics*, *18*(1), 38-65.

Kosfeld, M., Okada, A., & Riedl, A. (2009). Institution formation in public goods games. *American Economic Review*, *99*(4), 1335-55.

Levitt, S. D., & Miles, T. J. (2007). Empirical study of criminal punishment. in: Polinsky, A. M., & Shavell, S. (eds.). *Handbook of Law and Economics*, North-Holland, Amsterdam, Volume 1.

Marien, S., & Hooghe, M. (2011). Does political trust matter? An empirical investigation into the relation between political trust and support for law compliance. *European Journal of Political Research, 50,* 267–291.

Markussen, T., Putterman, L., & Tyran, J. R. (2014). Self-organization for collective action: An experimental study of voting on sanction regimes. *Review of Economic Studies, 81*(1), 301-324.

Meinzen-Dick, R. (2007). Beyond panaceas in water institutions. *Proceedings of the National Academy of Sciences USA, 104*(39), 15200–15205.

Nyborg, K. et al. (2016). Social norms as solutions. *Science, 354*(6308), 42-43.

Ones, U., & Putterman, L. (2007). The ecology of collective action: A public goods and sanctions experiment with controlled group formation. *Journal of Economic Behavior & Organization, 62*(4), 495-521.

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

Poteete, A. R., Janssen, M. A., & Ostrom, E. (2010). *Working together: collective action, the commons, and multiple methods in practice*. Princeton University Press, Princeton.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review, 83*(5), 1281-1302.

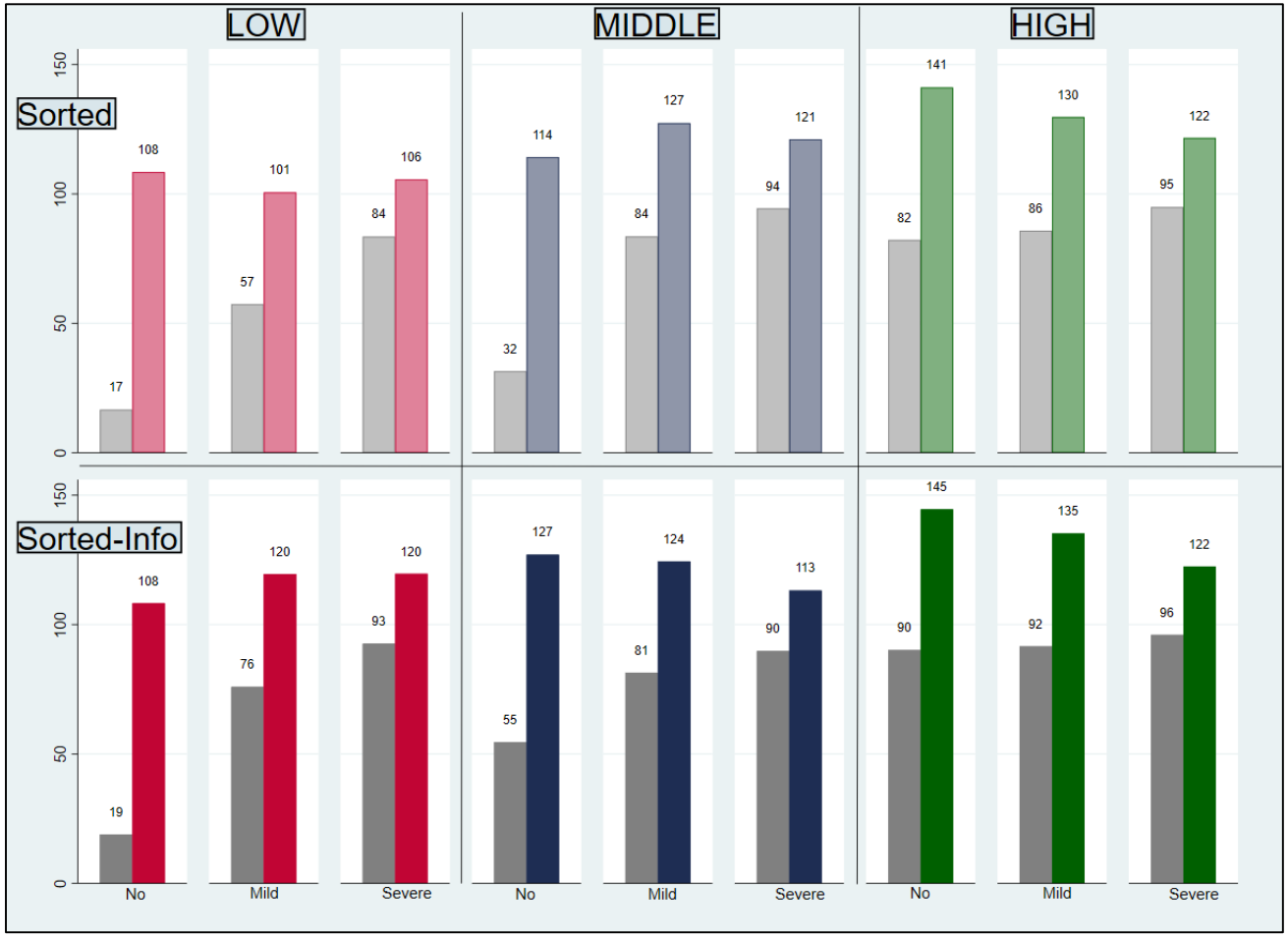
Tyran, J. R., & Feld, L. P. (2006). Achieving compliance when legal sanctions are non-deterrent. *Scandinavian Journal of Economics, 108*(1), 135-156.

Van Kesteren, J. (2009). Public attitudes and sentencing policies across the world. *European Journal on Criminal Policy and Research, 15,* 25–46.

Vollan, B., Landmann, A., Zhou, Y., Hu, B., & Herrmann-Pillath, C. (2017). Cooperation and authoritarian values: An experimental study in China. *European Economic Review, 93,* 90-105.

Figures and Tables

Figure 2: Average contribution and payoff levels by treatment, class, and institution



Notes: Grey bars: Average contribution. Colored bars: Average payoff. Contribution and payoff are measured by the average amount of tokens invested or earned in groups of LOW, MIDDLE, and HIGH cooperators by institution and treatment.

Figure 3: Average contribution levels over time by treatment, class, and institution

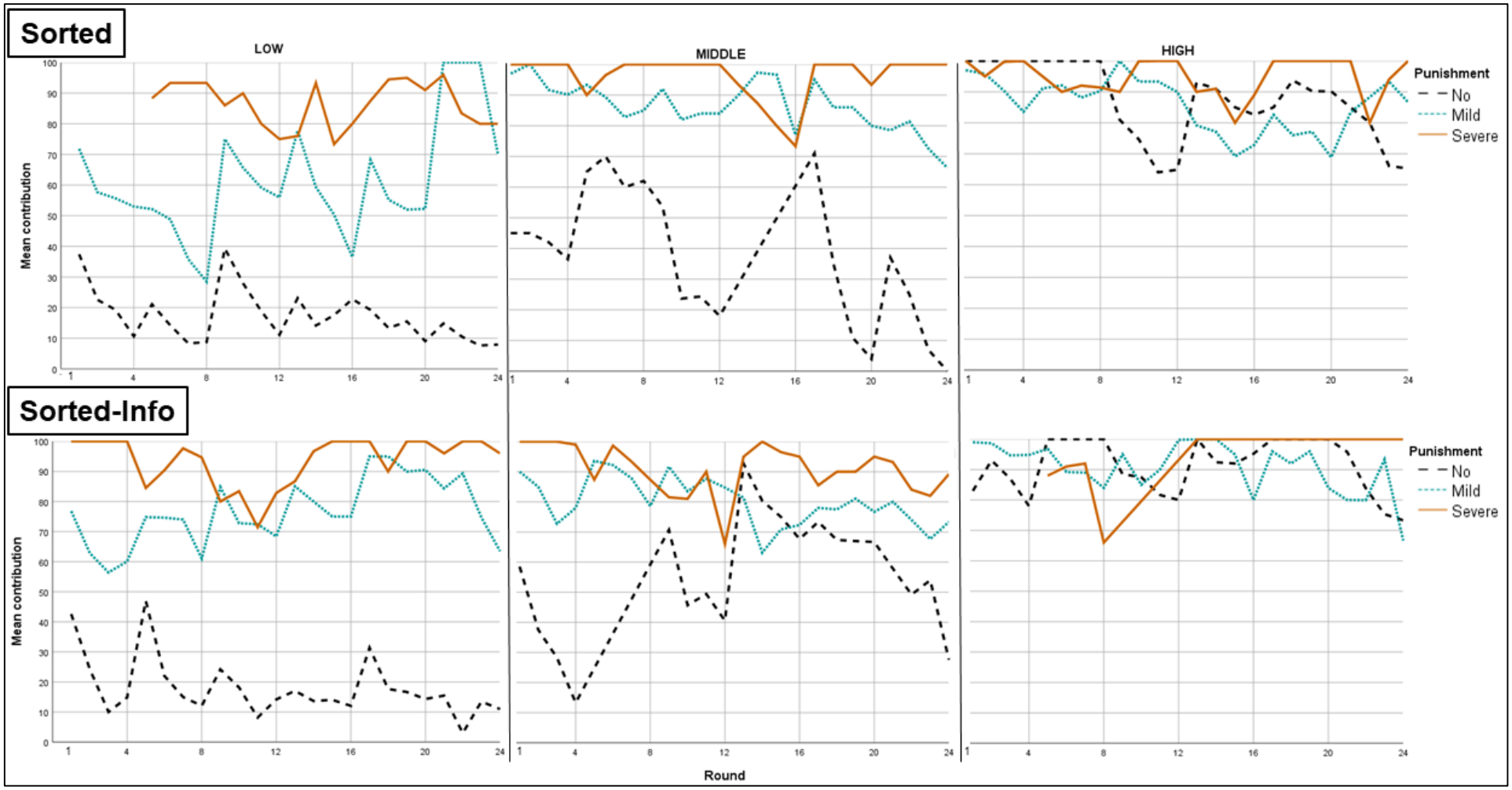


Figure 4: Average payoff levels over time by treatment, class, and institution

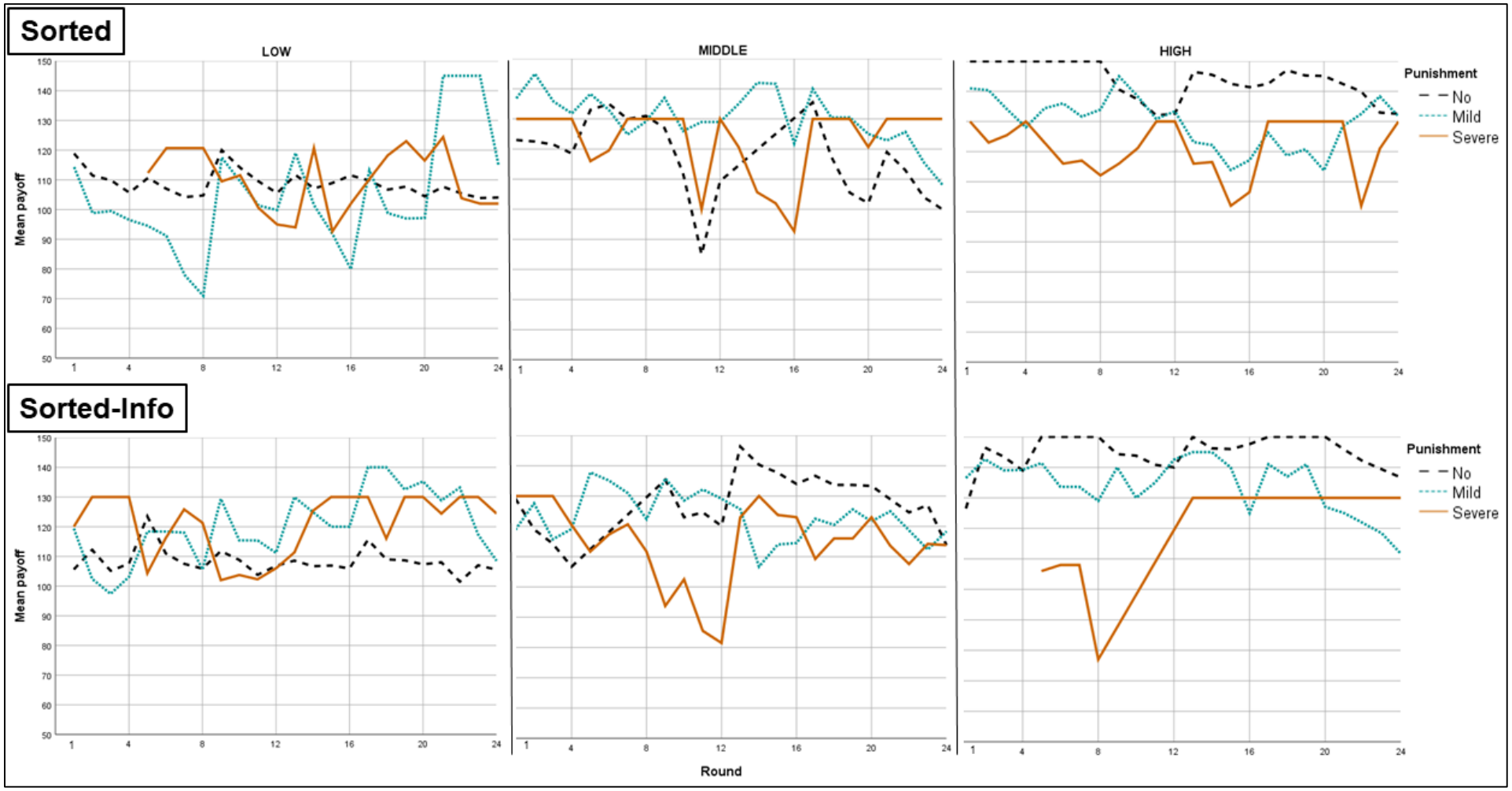
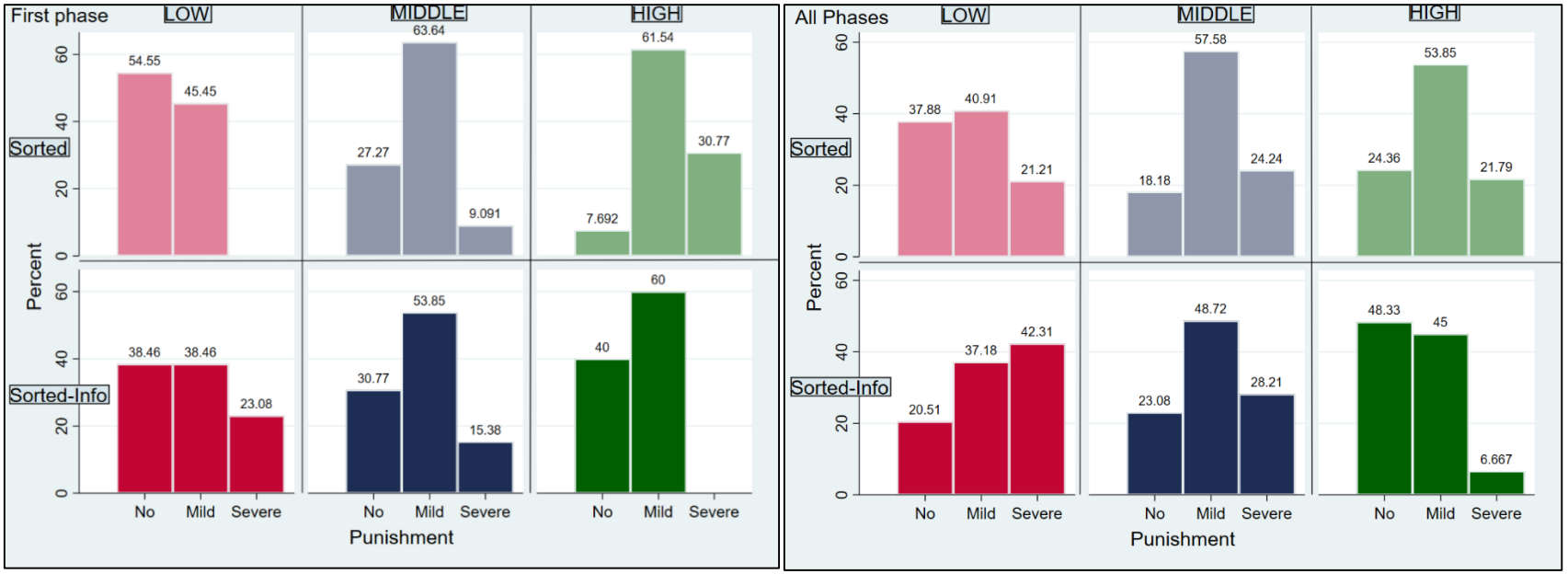


Figure 5: Distribution of institutions by treatment and class in the first phase (left panel) and across all phases (right panel)



Note: Distribution of institutions, No punishment, Mild punishment, and Severe punishment by treatment and class in the first phase (left panel) and across all phases (right panel).

Table 1: OLS regressions on groups' average contributions by treatment and class

	(1) Contribution *Sorted*, LOW		(2) Contribution *Sorted*, MIDDLE		(3) Contribution *Sorted*, HIGH		(4) Contribution *Sorted-Info*, LOW		(5) Contribution *Sorted-Info*, MIDDLE		(6) Contribution *Sorted-Info*, HIGH	
	Coef.	**Std. Err.**	**Coef.**	**Std. Err.**	**Coef.**	**Std. Err.**	**Coef.**	**Std. Err.**	**Coef.**	**Std. Err.**	**Coef.**	**Std. Err.**
Punishment												
Mild	41.39**	13.20	52.95***	8.19	1.63	9.19	58.08***	9.03	23.99	14.82	0.19	6.71
Severe	66.71***	3.90	65.10***	5.68	9.95	9.65	74.72***	3.19	33.38*	15.47	4.74	8.54
Phase												
2	-6.40	8.96	-9.29	5.99	-2.80	2.87	-0.35	4.52	11.01*	5.64	0.67	1.30
3	3.43	7.52	-10.20	9.17	-6.89	5.94	-6.80	8.09	3.44	7.78	-4.51	5.13
4	-0.83	5.98	-11.45	7.53	-12.16**	5.43	2.46	7.54	8.07	9.68	2.34	6.49
5	3.05	8.62	-12.42	8.63	-8.81	6.17	8.82	5.31	6.64	9.12	3.36	6.62
6	2.51	7.49	-17.41**	6.59	-8.51*	4.37	2.85	6.33	-0.47	10.99	-9.35	6.27
Constant	16.15**	6.90	40.62***	8.85	90.41***	9.70	16.94**	6.06	51.72***	14.90	92.06***	8.47
N	66		66		78		78		78		60	
R^2	0.560		0.607		0.093		0.706		0.224		0.074	
Adj. R^2	0.507		0.559		0.002		0.676		0.146		-0.051	
F, Prob>F	78.36, p<0.01		42.23, p<0.01		3.81, p=0.02		264.39, p<0.01		12.59, p<0.01		3.30, p=0.05	

Notes: *: p<0.1, **: p<0.05, ***: p<0.01. Contribution refers to the average contribution by group and phase in the institution formation game. N is the number of phases. Robust standard errors clustered at the group level. No Punishment and Phase 1 are used as baselines.

Table 2: OLS regressions on groups' average payoffs by treatment and class

	(7) Payoff *Sorted*, LOW		(8) Payoff *Sorted*, MIDDLE		(9) Payoff *Sorted*, HIGH		(10) Payoff *Sorted-Info*, LOW		(11) Payoff *Sorted-Info*, MIDDLE		(12) Payoff *Sorted-Info*, HIGH	
	Coef.	**Std. Err.**	**Coef.**	**Std. Err.**	**Coef.**	**Std. Err.**	**Coef.**	**Std. Err.**	**Coef.**	**Std. Err.**	**Coef.**	**Std. Err.**
Punishment												
Mild	-6.90	13.24	14.09*	6.56	-12.89*	5.99	12.23	9.29	-5.13	8.97	-10.08*	5.00
Severe	-3.67	4.07	9.13**	3.55	-21.81***	5.82	12.09***	3.44	-15.31*	7.98	-22.60**	9.62
Phase												
2	-3.61	9.32	-8.83	5.61	-4.25	3.43	1.24	4.17	9.67	6.15	-0.72	3.45
3	2.75	7.58	-10.07	8.56	-5.13	3.99	-5.79	8.49	-0.94	8.88	-1.07	6.10
4	-1.13	6.20	-11.74	7.47	-13.22**	5.61	6.80	7.86	5.27	10.31	3.32	5.61
5	5.76	8.43	-9.61	7.74	-9.46	6.17	12.67**	4.79	4.72	9.23	3.36	6.81
6	5.81	7.48	-14.34**	6.16	-6.65*	3.55	7.22	5.91	-0.47	11.74	-6.58	8.04
Constant	106.65***	5.70	122.23***	5.61	148.90***	5.57	103.93***	5.26	125.63***	9.94	145.20***	5.91
N	66		66		78		78		78		60	
R^2	0.041		0.119		0.195		0.135		0.073		0.238	
Adj. R^2	-0.074		0.012		0.114		0.049		-0.020		0.135	
F, Prob>F	2.21, p=0.12		9.92, p<0.01		15.29, p<0.01		14.04, p<0.01		4.12, p<0.02		2.30, p<0.12	

Notes: *: p<0.1, **: p<0.05, ***: p<0.01. Payoff refers to the average payoff by group and phase in the institution formation game. N is the number of phases. Robust standard errors clustered at the group level. No Punishment and Phase 1 are used as baselines.

Table 3: Marginal effects after Probit regressions on individuals' contributions<100 under Severe punishment

	(13) Contribution<100 *Sorted*, LOW		(14) Contribution<100 *Sorted*, MIDDLE		(15) Contribution<100 *Sorted*, HIGH		(16) Contribution<100 *Sorted-Info*, LOW		(17) Contribution<100 *Sorted-Info*, MIDDLE		(18) Contribution<100 *Sorted-Info*, HIGH	
	dy/dx	**Std. Err.**	**dy/dx**	**Std. Err.**	**dy/dx**	**Std. Err.**	**dy/dx**	**Std. Err.**	**dy/dx**	**Std. Err.**	**dy/dx**	**Std. Err.**
Cooperation	-.0001	.0006	.0001***	.0003	-.0011**	.0005	-.0002	.0002	-.0003	.0004	-	-
Belief	-.0001	.0003	-.0001	.0002	.0002	.0002	-.0002	.0002	.0008***	.0003	-.0002	.0003
Severe correct	.0351	.0510	-.0331	.0261	-.0468	.0290	-.0484***	.0158	-.0485*	.0255	-.1702	.1636
1st phase punish	.0147	.0692	-.1971**	.0877	-	-	-.0811***	.0311	-.2169***	.0455	-	-
1st punish Severe	-.0943**	.0467	-.1043***	.0302	.0419	.0366	-.0433**	.0179	-.0676**	.0261	-	-
Second half	.0620	.0416	.0197	.0240	-.0049	.0321	-.0605***	.0189	-.1385***	.0360	-	-
Vote Severe	-.2158***	.0701	-.0319	.0259	-.0220	.0325	-.0229	.0150	-.0438	.0342	-	-
Earn more previous round	.2060***	.0989	-.0088	.0224	.0538	.0301	.0747*	.0398	.0587	.0419	.0236	.0317
N	250		259		320		645		430		80	
Pseudo R^2	0.169		0.219		0.058		0.259		0.273		0.230	
Chi^2	33.61		55.74		15.71		80.06		63.33		3.12	
Prob>Chi^2	p<0.01		p<0.01		p=0.03		p<0.01		p<0.01		p=0.374	

Notes: *: p<0.1, **: p<0.05, ***: p<0.01. Marginal effects at sample averages. Robust standard errors clustered at the individual level. 'Contribution<100' takes the value 1 if a player contributes less than 100 tokens under the Severe punishment scheme and 0 otherwise. 'Cooperation' is an individual's contribution in the sorting game. 'Belief' is an individual's belief about the other group members' average contribution level in the sorting game. We paid an additional €1 if the belief was correct (+/- 5 tokens). 'Severe correct' takes the value 1 if an individual answers all three control questions about the Severe punishment scheme correctly on the first attempt and 0 otherwise. '1st phase punish' takes the value 1 if the group implements a punishment scheme (either Mild or Severe) in the first phase of the institution formation game and 0 otherwise (No punishment). '1st punish Severe' takes the value 1 if the group's first implemented punishment scheme is Severe and 0 otherwise. 'Second half' takes the value 1 for the rounds 13 to 24 in the institution formation game and 0 otherwise. 'Vote Severe' takes the value 1 if an individual votes for the Severe punishment scheme in the respective phase and 0 otherwise. 'Earn more previous round' takes the value 1 if an individual earned more than the average group payoff in the previous round and 0 otherwise. N is the number of individuals in all rounds under Severe punishment. For HIGH cooperators, we just observe 22 out of 320 contributions that are smaller than 100 under Severe punishment in *Sorted* and only 5 out of 80 contributions in *Sorted-Info.* That is why we cannot use all the independent variables used for the other classes in the regressions including only HIGH cooperators, especially in *Sorted-Info.*

Appendix

Instructions

Welcome to this experiment!

General guidelines

You can earn money in this experiment. Your earnings depend on the decisions you and your fellow players make. Please do not leave your seat for the duration of the experiment and do not engage in other activities. Also, it is vital that you make all decisions yourself without consulting with others. We will inform you when the experiment is over and when you can leave your seat.

You will play for "points" in the experiment. These points will be converted into euro. You receive 0.50 euro per 100 points earned in the experiment (**200 points = 1 euro**). The more points you earn, the more money you will make. It is essential that all participants finish the game and do not drop out. We are only able to calculate and pay-out your earnings if you finish the game. The pay-out will be made via PayPal.

Please read the following rules of the game **carefully**. You can contact the experimenter via chat for any questions.

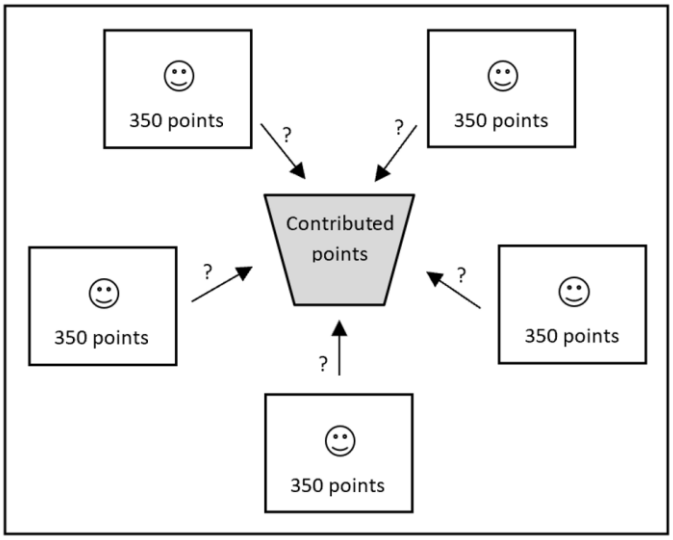
This experiment consists of two games. These are the rules of the first game which you will play subsequently. After finishing the first game you will be instructed on the second game which you will play subsequently. In the end, you will be paid-out the points you earned during both games.

After the instructions of each game, you will answer a few control questions on the rules. Should you get all the control questions right at the first attempt, you will in each case get **1 euro** on top.

Rules for the first game

There are five players, namely you and four other players. Every player is faced with the same decision problem.

Every player gets 350 points. The players then decide whether to keep these points or to contribute them to a shared project. The kept points only benefit the player himself. The points contributed to the shared project benefit all players. The contributed points are then multiplied by 1.5 and equally distributed among all five players. Thus, every player benefits from the contributed points regardless of how much they contributed themselves. The earnings of a player, therefore, consist of the sum of the kept points and the points from the shared project.



The players decide simultaneously how many points they each contribute to the shared project. Contributions of 0 to 350 points are possible. All points which are not contributed to the shared project are thus kept by the player. After all players have made their contributions to the shared project, the game ends. The contributions and earnings of all players are announced only at the end of the experiment.

Examples:

- If all five players contribute 60 points each to the shared project and keep 290 points, each player will receive 380 points (=290+1.5*300/5).
- If all five players contribute 300 points each to the shared project and keep 50 points, each player will receive 500 points (=50+1.5*1500/5).
- If four players contribute 350 points each to the shared project and one player contributes nothing, the four players will each receive 420 points (=0+1.5*1400/5) while the one player will earn 770 points (= 350+1,5*1400/5).

Control questions for the first game

Please answer the following control questions now. Once you completed all questions, please click "Continue". The computer will check your answers. Should the answers to one or more questions be wrong, you will be asked to review the question.

Should you get all the control questions right on the first attempt, you will get **1 euro** on top.

True or false? There are a total of five players in the game.

O True O False

Assume you have contributed 100 points to the shared project and the other players have contributed a total of 400 points together. Would your earnings be higher, lower, or the same if the

other players contributed **more** than 400 points together? (Tip: There would be more points in the shared project.)

O Higher O Lower O The same

Assume you have contributed 100 points to the shared project and the other players have contributed a total of 400 points together. Would your earnings be higher, lower, or the same if you contributed **more** than 100 points? (Tip: You would keep less points and there would be more points in the shared project. Consider the stronger influence.)

O Higher O Lower O The same

Assume all players have contributed a total of 1,000 points to the shared project with your contribution being 200 points (thus keeping 150 points). How high are your earnings (in points)? (Tip: Please feel free to use a calculator.)

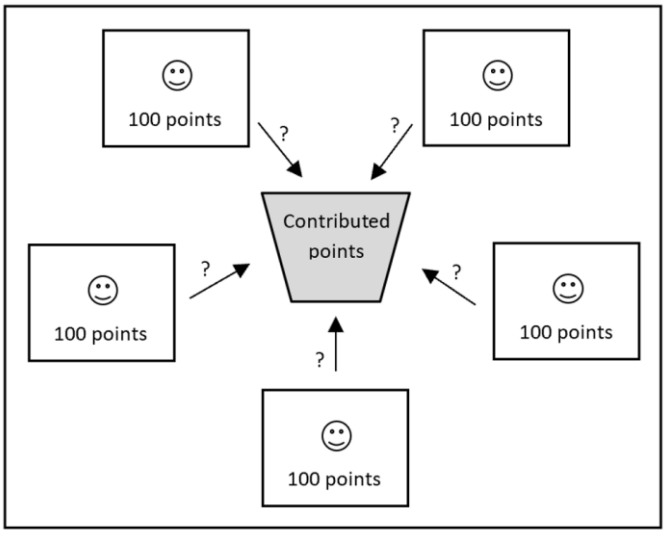
O 350 O 400 O 450 O 500 O 550 O 600

Rules for the second game

The second game is similar to the first game.

There are <u>five</u> players, namely you and four other players. These four players are not necessarily your fellow players from the first game. Every player is faced with the same decision problem.

Every player gets 100 points. The players then decide whether to <u>keep</u> these points or to <u>contribute</u> them to a shared project. The kept points only benefit the player himself. The points contributed to the shared project benefit all players. The contributed points are then multiplied by 1.5 and equally distributed among all five players. Thus, every player benefits from the contributed points regardless of how much they contributed themselves. The earnings of a player, therefore, consist of the sum of the kept points and the points from the shared project.



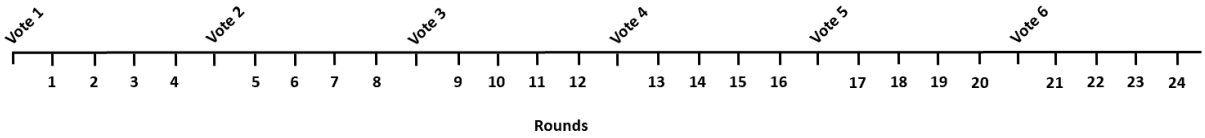
All players decide simultaneously how many points they want to contribute to the shared project. Contributions of 0 to 100 points are possible. All points which are not contributed to the shared project are thus kept by the player. This part of the game is similar to the first game. However, there are a few important differences which we will explain in the following.

Before the game, the group decides if they want to introduce a deduction system to the game. With the deduction system, points are deducted automatically from each player who contributes less than 100 points to the shared project. For the introduction of the deduction system, each player has to pay a fixed price. Taking a vote on the introduction of the deduction system is a two-step process. First, the players decide on the introduction of a deduction system. If the majority votes against the introduction (at least 3 of 5 players), the deduction system will not be introduced. In this case, the game is played as afore-mentioned. If the majority votes for the introduction of a deduction system, it will thus be introduced. Then, the group will decide which of two possible deduction systems should be implemented:

- The low deduction system means: Should a player contribute less than 100 points to the shared project, 50 of his or her points will automatically be deducted. No points will be deducted if the player contributes 100 points to the shared project. The fixed price for the introduction of this deduction system is 5 points per player.
- The high deduction system means: Should a player contribute less than 100 points to the shared project, 90 of his or her points will automatically be deducted. No points will be deducted if the player contributes 100 points to the shared project. The fixed price for the introduction of this deduction system is 20 points per player.

If the majority votes for the low deduction system, the low deduction system will be introduced. If the majority votes for the high deduction system, the high deduction system will be introduced.

The game consists of several rounds. The group does not change during these rounds. After choosing a deduction system, the group plays the chosen form of the game for 4 successive rounds. The course of these 4 rounds is identical. This means that every player gets 100 points each round which can either be kept or contributed. In case a deduction system was introduced, every player pays the fixed price each round. After 4 rounds, the group takes another vote on the introduction of the deduction system (yes or no, if yes: high or low) and the chosen game is played for another 4 successive rounds. In total, the group will take their vote six times on whether a deduction system should be introduced and will then play the chosen form of the game in four successive rounds. So, there are 24 rounds in total:



After every vote, the outcome of the vote will be announced but not the exact vote distribution. The group members will learn before the game whether a deduction system was chosen and if so, which.

After every round, it will be shown on the monitor how many points each player has kept or contributed. If the group introduced a deduction system, the individual deductions will also be shown. The amounts, deductions, and earnings of the players will be displayed randomly and anonymously. The randomization changes in every round. Your own decisions are displayed separately.

[The text in the box below is shown only in the Sorted-Info treatment]

Important: Before the players take their first vote on the introduction of the deduction system, they learn how many points their fellow players contributed to the shared project in the first game.

The points earned in all rounds make up the earnings of a player in the second game.

Examples:

Assume the group decided to implement the low deduction system. Remember: The fixed price is 5 points per player. This system leads to the automatic deduction of 50 points should a player contribute less than 100 points to the shared project.

- If all five players contribute 20 points each to the shared project and keep 80 points, each player will receive 55 points (=80-50+1,5*100/5-5).
- If all five players contribute 100 points each to the shared project and keep no points, each player will receive 145 points (=1.5*500/5-5).
- If four players contribute 80 points each to the shared project and one player contributes nothing, the four players will each receive 61 points (=20-50+1.5*320/5-5) while the one player will receive 141 points (=100-50+1.5*320/5-5).

Assume the group decided to implement the high deduction system. Remember: The fixed price is 20 points per player. This system leads to the automatic deduction of 90 points should a player contribute less than 100 points to the shared project.

- If all five players contribute 20 points each to the shared project and keep 80 points, each player will receive 0 points (=80-90+1,5*100/5-20).
- If all five players contribute 100 points each to the shared project and keep no points, each player will receive 130 points (=1.5*500/5-20).
- If four players contribute 80 points each to the shared project and one player contributes nothing, the four players will each receive 6 points (=20-90+1.5*320/5-20) while the one player will receive 86 points (=100-90+1.5*320/5-20).

Control questions for the second game

Please answer the following control questions now. Once you completed all questions, please click "Continue". The computer will check your answers. Should the answers to one or more questions be wrong, you will be asked to review the question.

Should you get all the control questions right on the first attempt, you will get **1 euro** on top.

True or false? There are a total of five players in the game.

O True O False

True or false? The game is played for 24 rounds. Your fellow players will be the same for the 24 rounds.

O True O False

True or false? At the beginning and then after every four rounds, the group takes a vote on the introduction of a deduction system. The group first decides whether to introduce a deduction system at all (yes or no) and, if yes, which deduction system to introduce (high or low). The majority of votes counts.

O True O False

True or false? If the majority of players vote against the introduction of a deduction system, the group will play the subsequent four rounds without a deduction system.

O True O False

What is the fixed price per player for the **low** deduction system (in points)?

O 5 O 10 O 15 O 20 O 25 O 30

With the **low** deduction system, how many points are automatically deducted if a player contributes less than 100 points to the shared project?

O 10 O 30 O 50 O 70 O 90

What is the fixed price per player for the **high** deduction system (in points)?

O 5 O 10 O 15 O 20 O 25 O 30

With the **high** deduction system, how many points are automatically deducted if a player contributes less than 100 points to the shared project?

O 10 O 30 O 50 O 70 O 90

Assume the group introduced the **low** deduction system. You have contributed 20 points to the shared project (while keeping 80 points). Would your earnings be higher, lower, or the same if you contributed 100 instead of 20 points? (Tip: You would keep less points and there would be more points in the shared project. Consider the stronger influence and whether the deduction would change.)

O Higher O Lower O The same

Assume the group introduced the **low** deduction system. You have contributed 90 points to the shared project (while keeping 10 points). Would your earnings be higher, smaller, or the same if you contributed 100 instead of 90 points? (Tip: You would keep less points and there would be more points in the shared project. Consider the stronger influence and whether the deduction would change.)

O Higher O Smaller O The same

Assume the group introduced the **high** deduction system. You have contributed 20 points to the shared project (while keeping 80 points). Would your earnings be higher, lower, or the same if you contributed 100 instead of 20 points? (Tip: You would keep less points and there would be more points in the shared project. Consider the stronger influence and whether the deduction would change.)

O Higher O Lower O The same

Assume the group introduced the **high** deduction system. You have contributed 90 points to the shared project (while keeping 10 points). Would your earnings be higher, lower, or the same if you contributed 100 instead of 90 points? (Tip: You would keep less points and there would be more points in the shared project. Consider the stronger influence and whether the deduction would change.)

O Higher O Smaller O The same

Assume the group introduced the **high** deduction system. All players have contributed a total of 350 points to the shared project. You have contributed 70 points to the shared project (while keeping 30 points). How high are your earnings in this round (in points)? (Tip: Please feel free to use a calculator.)

O 25 O 62 O 66 O 74 O 88 O 117