Matching Technology and the Choice of Punishment Institutions in a Prisoner's Dilemma Game

Veronika Grimm, Friederike Mengel

Title: Matching Technology and the Choice of Punishment Institutions in a Prisoner's Dilemma Game

Authors: Veronika Grimm, Friederike Mengel

PII: S0167-2681(11)00048-5
DOI: doi:10.1016/j.jebo.2011.01.018
Reference: JEBO 2686

To appear in: *Journal of Economic Behavior & Organization*

Received date: 6-8-2009
Revised date: 15-12-2010
Accepted date: 26-1-2011

Please cite this article as: Grimm, V., Mengel, F., Matching Technology and the Choice of Punishment Institutions in a Prisoner's Dilemma Game, *Journal of Economic Behavior and Organization* (2008), doi:10.1016/j.jebo.2011.01.018

Matching Technology and the Choice of Punishment Institutions in a Prisoner's Dilemma Game.*

Veronika Grimm† University of Erlangen–Nuremberg

Friederike Mengel ‡ Maastricht University

December 15, 2010

Abstract

We experimentally investigate the effect of endogenous matching within a segmented population on peoples' willingness to cooperate as well as their attitudes towards cooperative norms. In the experiment participants can repeatedly choose between two groups, where in one of them a (local) punishment institution fosters cooperation. The degree of population viscosity (i.e. the extent to which matching is biased towards within–group interactions) is varied across treatments. We find that both, the share of participants that choose into the group with the punishment institution and the share of participants that cooperate, increase monotonically with the degree of population viscosity. Furthermore — with higher population viscosity — significantly more subjects claim to support a punishment institution in a post–experimental questionnaire.

Keywords: Experiments, Cooperation, Punishment Institutions, Evolution, Population Viscosity.
JEL classification: C70, C73, Z13, C92.

*We thank Dirk Engelmann, Axel Ockenfels, the associate editor, and an anonymous referee, as well as seminar participants at Amsterdam, Cologne, Magdeburg and Milan (EEA 2008) for helpful comments and suggestions, and Rene Cyranek, Felix Lamoroux and Michael Seebauer for excellent research assistance. Financial support by the *Deutsche Forschungsgemeinschaft,* the *Instituto Valenciano de Investigaciones Económicas (IVIE)* , the *Spanish Ministery of Education and Science* (grant SEJ 2004-02172) and the European Union (grant PIEF-2009-235973) is gratefully acknowledged.

†University of Erlangen–Nuremberg, Lehrstuhl für Volkswirtschaftslehre, insb. Wirtschaftstheorie, Lange Gasse 20, D-90403 Nürnberg, Germany, Tel. +49 (0)911 5302-224, Fax: +49 (0)911 5302-168, email: veronika.grimm@wiso.uni-erlangen.de

‡Maastricht University, Department of Economics (AE1), PO Box 616, 6200MD Maastricht, Netherlands. *e-mail*: F.Mengel@maastrichtuniversity.nl

Research Highlights

- In a segmented population people voluntarily participate in a local punishment institution if interaction with outsiders is unlikely enough
- Self selection into institutions can sustain cooperation if population viscosity is high enough
- A punishment institution that successfully implements cooperative behavior shifts the agents' attitudes towards norm enforcement

Matching Technology and the Choice of Punishment Institutions in a Prisoner's Dilemma Game.*

December 15, 2010

Abstract

We experimentally investigate the effect of endogenous matching within a segmented population on peoples' willingness to cooperate as well as their attitudes towards cooperative norms. In the experiment participants can repeatedly choose between two groups, where in one of them a (local) punishment institution fosters cooperation. The degree of population viscosity (i.e. the extent to which matching is biased towards within–group interactions) is varied across treatments. We find that both, the share of participants that choose into the group with the punishment institution and the share of participants that cooperate, increase monotonically with the degree of population viscosity. Furthermore — with higher population viscosity — significantly more subjects claim to support a punishment institution in a post–experimental questionnaire.

Keywords: Experiments, Cooperation, Punishment Institutions, Evolution, Population Viscosity.
JEL classification: C70, C73, Z13, C92.

*We thank Dirk Engelmann, Axel Ockenfels, the associate editor, and an anonymous referee, as well as seminar participants at Amsterdam, Cologne, Magdeburg and Milan (EEA 2008) for helpful comments and suggestions, and Rene Cyranek, Felix Lamoroux and Michael Seebauer for excellent research assistance. Financial support by the *Deutsche Forschungsgemeinschaft,* the *Instituto Valenciano de Investigaciones Económicas (IVIE)* , the *Spanish Ministery of Education and Science* (grant SEJ 2004-02172) and the European Union (grant PIEF-2009-235973) is gratefully acknowledged.

1 Introduction

Social structure is important in a wide range of interactions, including the buying and selling of various goods and services, the transmission of information, the decision to hold a honorary office or to engage in criminal activity, and informal insurance. In these social and economic environments people typically have some freedom to choose who to interact with, i.e. they can endogenously decide on their match or peer group. However, choosing a peer group often does not entirely prevent interactions with someone outside this peer group. Behavior in such "*viscous*" populations (where matching is biased towards within–group interactions) is the focus of our study. We investigate how the degree of *population viscosity* affects peoples' willingness to cooperate in a social dilemma situation as well as their attitudes towards cooperative norms.

In many of the environments mentioned above cooperation may be achieved when a group of individuals has established implicit or explicit institutions which punish uncooperative behavior either through social disapproval or even materially. Often participation in such a (punishment) institution cannot be enforced, but is voluntary and involves the adoption of certain norms. In this case the institution does not apply to the whole population but only to a certain subgroup. This is in particular characteristic of mechanisms of social disapproval. In this study we are interested in the competition of such different norms in large populations where agents will typically not know ex ante whether or not their match adheres to a certain social norm. Typical examples involve norms for cooperation in cultural subgroups in large anonymous interactions. Our design could also be applied to situations where agents interact in small groups (and know each other's preferences) but have to choose an action before they know who they are matched with. Examples could be interactions at the workplace where effort has to be put into background research for a project before it is known who one's team-members for the project will be. Typically, social disapproval is a *local* punishment mechanism, i.e. non–cooperative behavior towards group members is sanctioned, but not non–cooperative behavior towards non–group members.[1] We will say that punishment is *local* if sanctions are not effective when group members behave non–cooperatively towards outsiders, and that it is *internalized* if sanctions apply to all interactions of a group–member. Punishment institutions may, moreover, differ along a second dimension: In endogenous mechanisms punishment is decided upon case by case by the group members after behavior has been observed. Exogenous punishment, on the contrary, is automatically implemented upon defection.

In this study we ask under what conditions agents will voluntarily opt for a group where a *local and exogenous* punishment mechanism is at place if possibly not all agents participate. We focus in particular on the question to what extent *population viscosity* (i.e. a high degree of group separation in our context) is needed to sustain

[1]The reason is that (i) non–members do not engage in social disapproval towards others i.e. no punishment has to be expected from interactions with non–members and (ii) since non–members do not share the cultural norm they do not care about social disapproval by anyone, i.e. they cannot be punished.

cooperation.[2] Furthermore, we study whether the degree of *population viscosity* (and the success of the punishment mechanism) has feedback effects on the participants' attitudes towards cooperation. This is an important question, as in practice the survival of institutions in the long run may depend on whether they influence attitudes towards cooperation in a way that supports them.

In our experiment we consider the following situation. There are two groups of agents in a population. In one group (group A) a "norm for cooperation" is implemented that is enforced through a *local and exogenous* punishment institution. In the other group (group B) agents do not care about this norm and cannot be punished via the mechanism implemented among group A members.[3] Throughout the experiment subjects can repeatedly choose between those two groups. The matching technology (i.e. the degree of viscosity) is varied across treatments. In one extreme case, full separation, agents interact exclusively with agents of their own group. In the other extreme case, random matching, the probability to interact with someone from the other group corresponds to the share of agents in that group. The most realistic cases are probably intermediate cases where the probability to interact with someone of the other group is lower than the share of agents in that group, but not zero.

We find that a significant number of participants choose into group A if and only if the degree of population viscosity is high enough. Moreover, while agents in group A cooperate if and only if the degree of viscosity is high enough, agents in group B almost never cooperate in any treatment. Finally both, the share of participants that choose into group A and the share of participants that cooperate, rise monotonically with the degree of viscosity. We also investigate whether attitudes towards punishment or norm enforcement are influenced by the matching technology. To this end we elicit those attitudes via a post–experimental questionnaire. While we find no significant differences between cooperator types[4], in treatments with a high degree of population viscosity significantly more subjects claim to be in favor of a punishment institution. That is, a punishment institution that successfully implements cooperative behavior shifts the agents' attitudes towards norm–enforcement. We conclude that matching technology plays a crucial role in establishing cooperative outcomes, in the short run (via changed incentives), as well as in the long run (by influencing the agents' attitudes).

The paper is organized as follows. In Section 2 we review related literature. In Section 3 we present the theoretical model underlying our study and summarize several hypotheses derived from this model. Section 4 describes the experimental design. The results from the experiment on local punishment institutions are presented and discussed in Section 5, and Section 6 compares local and internalized punishment

[2]Kocher et al. (2009), for example, have shown in a field study that cooperative behavior can differ significantly across different cultural groups within the same town.

[3]The reason for this could be that these agents just do not care about social disapproval through members of group A.

[4]Fischbacher et al. (2001) classify roughly 50% of all subjects as conditional cooperators and 30% as flat defectors. This is consistent with what we find in our questionnaire. See also Fischbacher and Gächter (2006) or Brandts and Schram (2001).

under low population viscosity. Section 7 concludes.

2 Relation to the Literature

Experimental economics has started to focus on the relation between matching structure and cooperation only recently. One of the first studies where agents could choose between groups is Ehrhardt and Keser (1999). In their experiment, however, cooperation is unstable and overall rates of cooperation decline steadily until the end of the experiment. Other studies where participants could choose their interaction group include Riedl and Ule (2002), Bonet and Kübler (2005), Gürerk et al. (2006), Grimm and Mengel (2009), or Kosfeld et al. (2009). Also other types of endogenous matching than group selection have been studied in the literature. Ostrom et al. (1992), Brown et al. (2004), Coricelli et al. (2004), Page et al. (2005), Engelmann and Grimm (2006), Goette et al. (2006), Huck et al. (2007), or Cabrales et al. (2010) study situations where agents endogenously choose interaction partners.[5]

In a nutshell, all those studies on endogenous matching show that the possibility to choose interaction groups or partners may explain more cooperative behavior, but only if punishment is feasible[6] and sufficiently effective and, moreover, cooperative agents can avoid interaction with others that are not subject to punishment (with a sufficiently high probability). The existing literature also demonstrates that cooperative outcomes are extremely sensitive to the institutional frame individuals interact in. None of the studies mentioned above (other than Grimm and Mengel, 2009) have investigated population viscosity, i.e. the possibility to interact with individuals which are not in the peer group chosen by an individual.

The effect of punishment institutions on cooperation rates has first been investigated by Fehr and Gächter (2000, 2002) and has been reviewed by Kosfeld and Riedl (2004), among others. Typically, the literature on punishment in experimental economics analyzes situations where agents can individually choose to punish others, i.e. to decrease the others' payoff, at some cost (endogenous punishment). There is abundant evidence that individuals use this option if the punishment technology is sufficiently effective, even though punishment is typically not individually rational. In our paper we do not investigate individual punishment decisions. Rather, in one of the groups (group A) punishment is *institutionalized*, but not in the other group (group B). Instead of choosing whom to punish, subjects in our experiment decide whether to participate in the punishment mechanism or not. Such a structure is typical of many real life environments.

Two closely related studies are Gürerk et al. (2006) and Grimm and Mengel (2009). Gürerk et al. study a public good game and show that subjects learn to choose into a

[5]See also Ones and Putterman (2006), Brosig (2002), McCabe et al. (2007) or the literature on network experiments reviewed in Kosfeld (2003).

[6]This includes also indirect punishment as, for example, not choosing an interaction partner any more in the future.

group where an endogenous punishment mechanism is at place. In their paper punishment is decentralized as in Fehr and Gächter (2000, 2002), i.e. each individual can decide whether to punish others or not. Our paper, in contrast, studies a centralized punishment mechanism. Moreover, Gürerk et al. study only the case where the two groups are fully separated. The focus of our study, in contrast, is whether a punishment mechanism can foster cooperation in an environment where interaction with others (not participating in the mechanism) cannot be avoided, which is characteristic of virtually all real life mechanisms.

The second closely related paper is Grimm and Mengel (2009) (GM in the following), where we also analyze behavior under population viscosity, i.e. in environments where the two groups are not perfectly separated. In GM the two groups differ since in one group the payoff for defection is reduced, but in a way that incentives are still those of a prisoner's dilemma game. The main differences between the two papers are that in the present study (i) the punishment mechanism may change strategic incentives in group A depending on the degree of viscosity and group choices of the population (whereas in GM the payoff reduction in group A never changes incentives), (ii) punishment is "local", i.e. effective only if group A members meet group A members (while in GM punishment is internalized) and (iii) the focus of GM is on the identification of different cooperator types while in this study the focus is on punishment institutions and how they shape normative attitudes.

Many studies have shown that it makes a difference whether cooperation occurs on a voluntary basis or whether it is explicitly enforced by monetary incentives (see e.g. Gneezy and Rustichini (2000), among many others). Those studies often found that explicit incentives can crowd out intrinsic incentives. Hence it is not clear whether both setups (the mere possibility of signaling one's willingness to cooperate as in GM and a punishment institution that may induce a monetary incentive to cooperate as in this paper) will have the same effect on behavior.

The second difference mentioned above is that in GM "punishment" in group A is *internalized* (in the group with lower defector payoffs the payoff of defectors is reduced irrespective of whom they interact with), while in the present study the punishment is *local*. Agents in group A are *only* punished if they defect against other agents of group A. This means that the expected punishment for defection in group A depends on the degree of viscosity as well as on the relative group sizes. The property of *local* punishment is characteristic of social disapproval mechanisms, where punishment is confined to interactions with others that share the same social norm. For example, a nuclear energy lobbyist may be quite unimpressed by social disapproval received from an environmentalist.

The difference of the two punishment institutions is also clearly reflected in the experimental data. In the present study (local punishment) cooperation rates in group A are high under full separation and high viscosity (0.97 and 0.91, respectively), lower under low viscosity (0.64), and even lower (0.45) for random matching. In GM, on the contrary, cooperation rates in group A are approximately constant across treatments ($0.62, 0.67, 0.60$ for the various degrees of population viscosity). In Section 6

we present data from a control treatment that is identical (in terms of payoff parameters) to the low viscosity–treatment in the present paper except for the fact that punishment is internalized. In this control treatment we observe a cooperation rate of 0.93 (as compared to 0.64 under local punishment, as mentioned above).

Another reason "local" punishment (or social disapproval) mechanisms are particularly interesting to study is that it is unclear whether normative beliefs supporting these mechanisms can evolve if mechanisms are competing. In the present study we partly address this question using questionnaire data. Elinor Ostrom (2000) writes in an article in the Journal of Economic Perspectives: "It is possible that [...] policy initiatives to encourage collective action [...] may have been misdirected — and perhaps even crowded out the formation of social norms that might have enhanced cooperative behavior in their own way." The environment in this experiment with "local" punishment, where institutions compete and where interaction with outsiders cannot be avoided, provides a tough test for the more optimistic conjecture that institutions designed to enhance cooperation can shift attitudes in a direction supporting them.

3 The Model

3.1 The Basic Game

The game we study is a standard (symmetric) Prisoner's Dilemma game, in which agents can either cooperate (C) or defect (D). The payoffs are given by the payoff matrix in table 1, where $c > a > d > b$. The cooperative outcome is efficient whenever $a > \frac{b+c}{2}$. In the experiment we use the following parametrization

$$a = 400;\ b = 50;\ c = 550;\ d = 200. \tag{1}$$

	C	D
C	a	b
D	c	d

Table 1: Payoff Matrix of the Prisoner's Dilemma

If agents are randomly matched to play this game the unique prediction is mutual defection and thus a payoff of d (= 200) for both. What happens if there is a group that has managed to implement a norm for cooperation (to play C) through a local punishment mechanism? Clearly whether agents will choose such a group and choose to cooperate will depend on (i) how many others do so and (ii) how likely it is to interact with someone from one's own group or someone from another group.

3.2 Two cultural groups with different norms

There are two cultural groups, A and B. Agents in group A share a norm to cooperate. This norm is institutionalized in group A through a local punishment institution. More precisely, whenever a member of group A defects in an interaction with another A member she incurs a payoff loss of γ (= 200). One can think of this as a social disapproval mechanism. While in the experiment punishment is material, in real life this term can also correspond to either a psychological payoff loss or to anticipation of a material loss in the future. Thus, whereas a member of group B faces the payoff matrix given in table 1, the relevant payoff matrix for a group A member is given by table 2, where δ_{AA} takes on the value $\delta_{AA} = 1$ if both interaction partners are

	C	D
C	a	b
D	$c - \gamma\delta_{AA}$	$d - \gamma\delta_{AA}$

Table 2: Payoff Matrix Group A

group A members and zero otherwise. Obviously, for a group B member $\delta_{AA} = 0$ independently of whom she interacts with. We chose this punishment technology as it best reflects the situation, where agents in one group (namely group A) have implemented a local punishment institution among themselves. Members of group B cannot be punished and do not punish (either because the institution requires signing a contract or — if we talk of social disapproval as punishment — because they do not care about the social disapproval defectors receive from members of group A).

In our model, group-membership defines an agent's type. At all times agents have incomplete information about the type of their match. When choosing an action in the bilateral game they have to estimate the type of their match from the distribution of types in the economy and from their knowledge about the matching technology described below. Clearly, for a group B–member defection is a dominant strategy. For a group A–member, whether cooperation or defection is optimal depends on the relative size of the two groups and on the degree of separation of the two groups (population viscosity).

	A	B
A	$1 - p_B x$	$p_B x$
B	$p_A x$	$1 - p_A x$

Table 3: Matching Probabilities

Matching takes place randomly in a viscous population, the latter meaning that individuals have a tendency to interact more often with individuals that are of the same type. The degree of viscosity is measured by the parameter $x \in [0, 1]$. $x = 1$

corresponds to the case of random matching. $x = 0$ means that the population is fully separated, implying that agents interact with probability 1 with agents of the same group and never with agents from another group. In a viscous population with parameter x, if p_A is the share of agents of type A (members of group A) the probability for any one of them to interact with a B type is $(1 - p_A)x = p_B x$ and the probability to interact with a member of group A is $(1 - (1 - p_A)x) = 1 - p_B x$. Obviously if the society is fully separated ($x = 0$), agents only interact with agents of their own cultural group. The matching probabilities are summarized in table 3.

3.3 Cultural Equilibria

We assume that materially successful groups attract agents and proliferate.[7] Denote by p_A the share of agents in group A and assume that p_A evolves as follows,

$$\dot{p}_A = p_A(1 - p_A)[\Pi_A - \Pi_B], \tag{2}$$

where Π_A and Π_B are the average payoffs of group A and group B members.

Let us call a *cultural equilibrium* a share p_A together with an action choice in the bilateral game, such that (i) the action choice is a Nash equilibrium given x and p_A and (ii) p_A is an asymptotically stable equilibrium of (2). Then, the theoretical prediction can be summarized as follows:

Proposition 1 (Cultural Equilibrium) *(i) If $x < \frac{1}{4}$ the globally stable cultural equilibrium has $p_A^* = 1$ with all players cooperating.*

(ii) If $x \in [\frac{1}{4}, \frac{4}{7}]$ there are two locally stable cultural equilibria: $p_A^* = 1$ with all players cooperating and $p_A^* = 0$ with all players defecting.

(iii) If $x > \frac{4}{7}$ the globally stable cultural equilibrium has $p_A^* = 0$ with all players defecting.[8]

Proof See appendix A. □

The intuition for this proposition is as follows. If $x < 1/4$ cooperation is a dominant strategy for agents in group A. Furthermore, as they interact among each other with very high probability their expected payoffs exceeds that of group B members. Consequently group A will proliferate. On the other hand if $x > \frac{4}{7}$ matching is close to random matching. In this case a group B member will often be able to exploit a group A member and enjoy higher payoffs. In the intermediate case, both equilibria are possible. In the experiment we implemented four treatments corresponding to the viscosity parameters $x = 0, \frac{1}{3}, \frac{2}{3}, 1$. The theoretical predictions for these treatments are summarized in the following section.

[7] See Boyd and Richerson (2005), Mitteldorf and Wilson (2000), Wilson and Sober (1994), Myerson et al. (1991) or Mengel (2007, 2008) among others.

[8] Note that whereas in the theory outlined above there is a continuum of agents this is obviously not the case in the experiment. Proposition 1 is derived for the discrete case of our experiment.

3.4 Hypotheses from the Theory

Group Choice In treatment $x = 0$ all subjects should join group A and cooperate. In treatments $x = \frac{2}{3}$ and $x = 1$ all subjects should join group B and defect. In treatment $x = \frac{1}{3}$ both outcomes are possible.

Cooperation Group B–members should always defect. Whether cooperation or defection is optimal for group A–members depends on (i) the relative size of the two groups, p_A, and on (ii) the degree of separation of the two groups, x. An equilibrium where subjects in group A cooperate exists if and only if the degree of population viscosity is sufficiently high (i.e. in treatments $x = 0$ and $x = \frac{1}{3}$).

Profits Average profits in the population should be equal to 400 in treatments $x = 0$ and equal to 200 in treatments $x = \frac{2}{3}$ and $x = 1$. Group A–members should have higher (lower) profits than group B–members in treatments $x = 0$ and $x = \frac{1}{3}$ ($x = \frac{2}{3}$ and $x = 1$).[9]

Rate of Convergence and Learning Dynamics Both approaches (the evolutionary model and the reinforcement learning model) predict that learning is fastest for $x = 0$, slower for $x = 1$ and slowest for the intermediate x–values.

4 The Experimental Design

The experiment was conducted in eight sessions in May, 2006, and in October, 2010. A total of 256 students (32 per session) were recruited among the student population of the University of Cologne — mainly undergraduate students with no (or very little) prior exposure to game theory.[10]

In order to answer our research questions we implemented four different treatments that differed in the degree of population viscosity, x, as defined in section 3. We chose the values $x \in \{0, \frac{1}{3}, \frac{2}{3}, 1\}$. One population consisted of 8 subjects. Each population constitutes an independent observation. We obtained six independent observations for each of the four treatments.

The members of a population were initially randomly assigned to groups A and B in equal proportions and played a Prisoner's Dilemma game for 100 rounds. In the first four rounds, each subject played the game described in section 3 with an interaction partner who was assigned randomly according to the matching technology. From round five on, each round had two stages. At the first stage, two of the eight subjects could decide to either join the other group, or to stay in their own group. (The reason we did not let all subjects switch groups at once is that we wanted to

[9] If this were not the case, an equilibrium where agents select into group A (B) in treatments $x = 0$ and $x = \frac{1}{3}$ ($x = \frac{2}{3}$ and $x = 1$) could not be stable under (2).

[10] Subjects were recruited using the Online Recruitment System by Greiner (2004).

create a more stable environment for learning.) As in total there were 8 subjects per matching group, each subject could make this decision every fourth round. At the second stage of each round, subjects played the (modified) prisoner's dilemma game as given by (2) with an interaction partner who was assigned randomly according to the matching technology. Prior to playing the game they were informed about (a) the percentage of subjects in group A and B, and (b) their individual probability to meet a group$A$– and group$B$–member, respectively.

Since in our experiment the population was necessarily finite, one–to–one matching was not feasible for matching technologies with $x \neq 1$ (i.e. in three out of four treatments). Instead, we first realized a random draw with the probabilities given in table 3 to decide whether a subject's "interaction partner" was from group A or B. Then the "interaction partner" played the actions "cooperate" or "defect" with probabilities that corresponded to the proportions with which those actions were played in the respective group (in that round). Note that what we call "interaction partner" (also in the experimental instructions) is *not* another participant of the experiment, but a draw from a distribution that corresponds to the actual distribution of action and group choices in the experiment. Of course, the general matching procedure we used was the same for all treatments. In the unlikely event that only one subject remained in a group (either A or B) and the first random draw determined that she had to play against a member of her own group, the subject's interaction partner was preprogrammed to play the equilibrium strategy.[11] After each of the 100 rounds, subjects were informed of whether their interaction partner belonged to group A or B, her action, and their own monetary payoffs.

At the end of the experiment (after all 100 rounds were finished) we had the participants answer a questionnaire designed to elicit their attitudes towards cooperation, the normative principles their decisions were guided by, and their attitudes towards norm enforcement.

We also ran a control treatment for a low degree of population viscosity ($x = \frac{2}{3}$) with "internalized" instead of "local" punishment in group A. In this additional treatment, defection in group A always led to a payoff loss of 200 ECU (not only in case of matching with another group A–member). Everything else was exactly as in the treatment described above for $x = \frac{2}{3}$. We have eight independent observations (i.e. populations) of this additional control treatment.

All experimental sessions were computerized.[12] Written instructions were distributed at the beginning of the experiment.[13] Each session took approximately 120 minutes (including reading the instructions, answering a post-experimental questionnaire and receiving payments). Subjects participating in the experiment received 2.50 Euros just to show up. On average subjects earned approximately 15 Euros.

[11]The subjects were informed that the interaction partner would play optimally given the situation in this case. This occurred in less than 0.1% of all cases.

[12]The experiment was programmed and conducted with the software z-Tree (Fischbacher 2007).

[13]The instructions for $x = 1/3$, translated into English, can be found in Appendix C. Instructions for the remaining treatments are available upon request.

5 Results

5.1 Prologue: Attitudes Towards Cooperation

Let us first report some results from the post–experimental questionnaire that we will later on refer to in our analysis of the data. In particular we want to report the answers to the following questions:

Suppose you played game 1 exactly once against a randomly drawn interaction partner.

Question 1 *If you knew that 0 (25, 50, 75, 100) % of all others are choosing C which action would you choose C or D?*

Question 2 *Do you think participants that choose D should get a deduction from their payoffs?*

We also asked subjects whether they think that participants that choose C should get a deduction or whether either participants choosing C or D should get a bonus. We did this to check whether subjects actually answered yes to question 2, because they think defection is "wrong", or because they just think everyone or no one should get a deduction. These additional questions also served as a consistency check. With respect to these questions we found that almost all subjects are consistent in their answers.[14]

Cooperation Type	Treatment				Overall
	$x = 0$	$x = \frac{1}{3}$	$x = \frac{2}{3}$	$x = 1$	
flat defectors	.19	.39	.37	.22	.29
altruists	.04	.02	.00	.02	.02
conditional cooperators	.33	.33	.35	.43	.36
hump shaped	.33	.23	.21	.19	.24
none of the others	.10	.02	.06	.12	.07

Table 4: Cooperation Types.

Question 1 (Cooperation Types) From the answers to Question 1 we identify four "cooperator types": (1) flat defectors (who answer defect at all shares 0 (25, 50, 75, 100) %), (2) altruists (who always answer cooperate), (3) conditional cooperators (who answer cooperate if and only if the share of cooperators is sufficiently high) and (4) hump shaped (who answer cooperate if and only if the share of cooperators

[14] We also asked the participants some questions about normative criteria, but here (just as with Question 1) treatment differences were not significant.

is intermediate and defect otherwise). Table 4 reports the results. We compared the entries in Table 4 pairwise using a Mann-Whitney Test with each individual as an independent unit of observation. There are no treatment differences in the share of altruists, conditional cooperators or hump shaped types ($p > 0.1551, p > 0.2969, p > 0.1054$). There is a significant difference between the share of defectors across treatments $x = 0$ and $x = \frac{1}{3}$ ($x = \frac{2}{3}$) ($p = 0.0421$), but no significant difference in any other pairwise treatment comparison ($p > 0.1217$).[15]

We take this as evidence that the basic attitude towards cooperation of the participants was not influenced by the experience in the experiment. The types we found in the questionnaire roughly correspond to what Fischbacher, Fehr and Gächter (2001) find. Just as in their study most participants are classified as defectors, conditional cooperators.

Question 2 (Norm Enforcement) In the following we will call a participant who answered "Yes" to Question 2 a "norm-enforcer". The reason is that we see the punishment institution in group A as implicitly defining a "norm for cooperation". This is why someone who supports punishment implicitly supports the norm. (Note that with our consistency questions described above, we checked whether participants gave this answer because they have some norm in mind).[16] In treatments with a high degree of group separation ($x = 0$ and $x = \frac{1}{3}$), the majority of subjects was in favor of punishment, while in those treatments where group separation was low ($x = \frac{2}{3}$ and $x = 1$), the majority was against punishment of defectors. The difference between treatments $x = \frac{1}{3}(x = 0)$ and treatment $x = 1$ is highly significant (Mann-Whitney Test, $p = 0.0043$ ($p = 0.0240$)). All other pairwise differences are not significant ($p > 0.1040$).[17]

Norm Enforcer?	Treatment $x = 0$	$x = \frac{1}{3}$	$x = \frac{2}{3}$	$x = 1$	Overall
no	.46	.39	.57	.69	.53
yes	.54	.61	.43	.31	.47

Table 5: Attitudes Towards Norm Enforcement.

We conclude that, while this was not the case for Question 1, here there seem to be feedback effects from the treatment variable (x) to the subjects' attitudes towards norm enforcement. In particular, the success of the punishment institution in sustaining cooperation in the experiment, had a positive feedback effect on the participants

[15] Logit regressions reveal no significant differences at all. See the regression tables in Appendix B.

[16] Of course participants could be motivated to answer yes to this question for a variety of reasons. They could be motivated by efficiency concerns, inequality aversion, fairness, etc. We make no claim about which of these motivations apply, but simply state that a participant supports enforcement of a norm to cooperate, when she answers yes. The term "norm" is thus left unspecified on purpose.

[17] Logit regressions reported in Appendix B confirm these results.

support of this institution. Recently Galbiati and Vertova (2010) have found that laws and regulations can be complementary to incentives in sustaining cooperation. One possible channel for the feedback effects identified could be that the punishment institution does not only affect incentives, but also implicitly defines an obligation to cooperate. This could at least partly explain the positive feedback effects identified. We also find that participants that claim to be conditional cooperators in the questionnaire are in favor of norm enforcement significantly more often (Spearman test, $\rho = 0.1228^{***}$).

5.2 Group Choice

Figure 1 illustrates the effect our treatment variable has on group choice. While for perfectly separated groups (treatment $x = 0$) almost all subjects join group A, the share of subjects that are in group A decreases as the degree of population viscosity decreases.

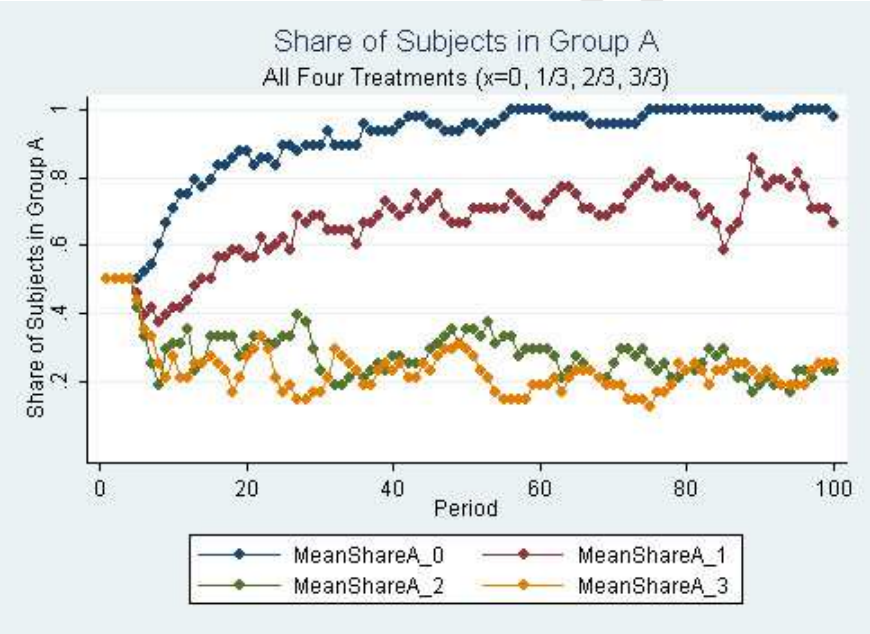

Figure 1: The Share of Subjects in Group A per Treatment ($MeanShareA_j$ denotes the average share of agents in group A in treatment $x = \frac{j}{3}$).

All treatment differences (except the difference between $x = \frac{2}{3}$ and $x = 1$) are significant at the 1% level according to a random effects panel data logit regression of group choice on treatment dummies irrespective of whether we use individuals or matching groups as independent unit of observation.[18] Furthermore, participants

[18] See the regression tables in Appendix B.

classified as "conditional cooperators" in the questionnaire are more often in group A in treatments $x = 0$ and $x = \frac{1}{3}$ compared to all other types (Spearman test, $\rho = 0.0603^{***}$, $\rho = 0.0559^{***}$) and are less often found in group A in treatment $x = 1$ ($\rho = -0.0668^{***}$). Participants supporting norm enforcement are more often found in group A than those that do not in all treatments. (Spearman test, $\rho = 0.0749^{***}$, $\rho = 0.0490^{***}$, $\rho = 0.0774^{***}$, $\rho = 0.0981^{***}$).[19]

Result 1 (Group Choice) *(i) The share of subjects in group A is highest in treatment $x = 0$, followed by $x = \frac{1}{3}$ and lowest in treatments $x = \frac{2}{3}$ and $x = 1$.*

(ii) "Conditional cooperators" are more likely to be in group A than other types in treatments $x = 0$ and $x = \frac{1}{3}$ and more likely to be in group B in treatment $x = 1$.

(iii) "Norm enforcers" are more often found in group A in all treatments.

These results have very intuitive interpretations. The local punishment institution survives only if participants opting for the institution are sufficiently separated from others. Otherwise (almost) all participants prefer to opt out of the institution. Participants classified as "conditional cooperators" cooperate whenever matched with high probability with other cooperators. Consequently in treatments $x = 0$ and $x = \frac{1}{3}$ they cooperate and choose group A, whereas in treatment $x = 1$ they defect as the environment is characterized by defection. But then it is optimal to join group B. Remember that we did not find significant differences in the share of conditional cooperators across treatments. It also makes perfect sense that norm enforcers choose group A, where the norm is enforced. On the other hand, experiencing "successful" norm enforcement in treatments $x = 0$ and $x = \frac{1}{3}$ (where most subjects are in group A) possibly leads agents to support norm enforcement. These are the feedback effects already mentioned in the previous section.

5.3 Cooperation

As Figure 2 illustrates, the shares of cooperating subjects in the population evolves in line with the share of subjects in group A (compare figure 1). Analyzing cooperation shares separately for the two different groups reveals that in all treatments (except $x = 1$) the majority of subjects in group A cooperate, while almost no group B–member does (see table 6).

Table 6 shows that (a) subjects cooperate much more in treatments with higher population viscosity and (b) subjects cooperate much more in group A than in group B. The treatment differences are significant at the 1% level (see the regression table in Appendix B). Also the fact that cooperation is strongly correlated with being in group A is highly significant in all treatments (Spearman test $\rho = 0.7571^{***}$, $\rho = 0.8090^{***}$, $\rho = 0.6776^{***}$, $\rho = 0.4187^{***}$). It can also be seen from Table 6 that under

[19] We report the results of a Spearman correlation test rather than e.g. panel data logit regressions whenever there are possible endogeneity problems with the latter.

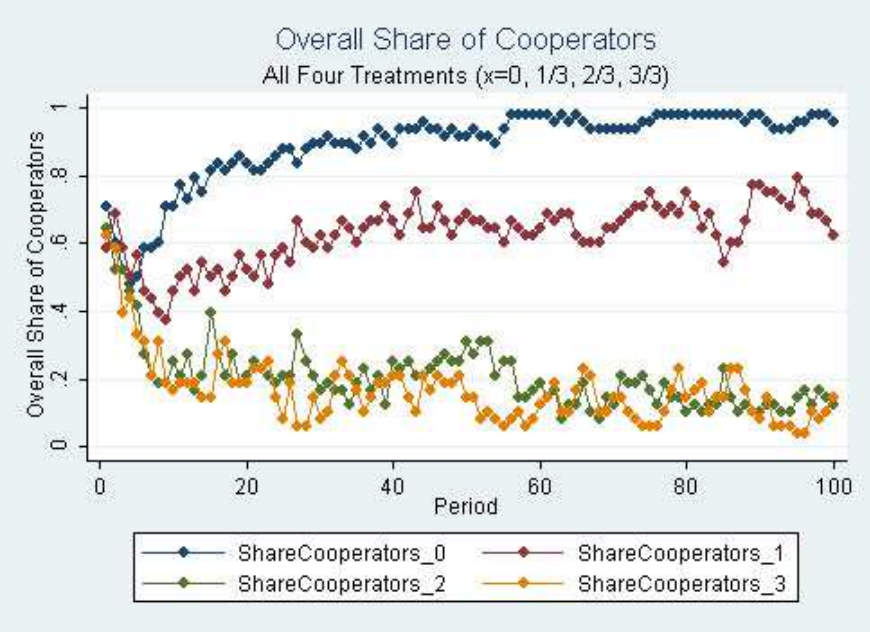

Figure 2: Shares of Cooperators.

the local punishment institution the rate of cooperation crucially depends on the degree of population viscosity. Even participants opting for the punishment institution decide to defect quite often if viscosity is low. We find in addition that participants classified in the questionnaire as "norm-enforcers" tend to cooperate more than others in all treatments ($\rho = 0.0428^{***}, \rho = 0.0980^{***}, \rho = 0.0951^{***}, \rho = 0.2234^{***}$). The same holds true for "conditional cooperators" with the exception of treatment $x = \frac{1}{3}$ where there is no significant correlation. ($\rho = 0.0929^{***}, \rho = 0.0052, \rho = 0.0738^{***}, \rho = 0.0499^{***}$).

Treatment	Group		Overall
	A	B	
$x = 0$	.97	.16	.89
$x = \frac{1}{3}$	.91	.08	.63
$x = \frac{2}{3}$	.64	.03	.20
$x = 1$	.45	.08	.17

Table 6: Average Cooperation Rates

Result 2 (Cooperation) *(i) Subjects in group A cooperate much more than subjects in group B.*

(ii) As x increases the overall rate of cooperation decreases.

(iii) "Norm Enforcers" are more likely to cooperate in all treatments.

(iv) "Conditional Cooperators" cooperate significantly more than all others in all treatments (except for $x = \frac{1}{3}$).

The interpretation of the results is intuitive. Defection is being punished in group A whenever one is matched with another group A member. The higher p_A, the more likely an agent is matched with a group A–member and in these cases there is more cooperation in group A. This effect is particularly strong in treatments $x = \frac{2}{3}$ and $x = 1$, where few agents tend to be in group A. (iii) and (iv) show that behavior in the experiment is roughly consistent with attitudes expressed in the questionnaire. Norm enforcers (who state that defection should be punished) are more likely to cooperate. Similarly those that we categorized to be altruists or conditional cooperators according to their answers in the questionnaire did indeed cooperate more in the experiment. Interestingly we also find that the difference in overall cooperation rates between treatments $x = \frac{2}{3}$ and $x = 1$ is mainly driven through the agents classified as norm enforcers, who cooperate significantly more in treatment $x = \frac{2}{3}$ compared to treatment $x = 1$.

5.4 Profits

The observed behavior (concerning group choice and cooperation) had clear consequences on profits. Recall that overall rates of cooperation were the higher, the more viscous the population was. Consequently, payoffs were highest in treatment $x = 0$, lowest (and close to the payoffs from mutual defection) for $x = 1$ and in between for the remaining treatments with intermediate degrees of population viscosity. Treatment differences are highly significant according to a panel data OLS regression except for the difference between $x = 1$ and $x = \frac{2}{3}$.[20] Members of group A have a higher payoff than members of group B in treatments $x = 0$ and $x = \frac{1}{3}$ (Spearman test, $\rho = 0.5745^{***}, \rho = 0.2750^{***}$) and vice versa in treatments $x = \frac{2}{3}$ and $x = 1$ ($\rho = -0.0929^{***}, \rho = -0.1651^{***}$). We summarize our results in table 7.

Treatment	Group		Overall
	A	B	
$x = 0$	388	227	373
$x = \frac{1}{3}$	345	273	314
$x = \frac{2}{3}$	201	241	230
$x = 1$	192	241	230

Table 7: Profits.

[20] See the regression table in Appendix B.

Result 3 (Profits) *(i) Average profits in the population are highest in treatment* $x = 0$*, followed by* $x = \frac{1}{3}$ *and treatments* $x = \frac{2}{3}$ *and* $x = 1$*.*

(ii) The profit of a group A–member is higher than the profit of a group B–member in treatments $x = 0$ *and* $x = \frac{1}{3}$ *and vice versa in treatments* $x = \frac{2}{3}$ *and* $x = 1$*.*

5.5 Rate of Convergence and Learning Dynamics

How often do the participants switch groups during the experiment? Table 6 reveals that participants switch most often in the treatments with intermediate degrees of population viscosity, and least often if groups are perfectly separated. While in treatment $x = 0$ most of the "group switching" takes place during the first quarter (Q1, the first 25 rounds) of the experiment, in the treatments with intermediate degrees of separation there is still a substantial number of switches even in the last quarter (Q4). Consistently with theory (either the evolutionary or the reinforcement model) convergence to equilibrium is fastest in the $x = 0$ treatment and slowest for the intermediate treatments. The higher payoff differences between the two groups in the $x = 0$ and $x = 1$ treatments effectively seem to speed up learning, as the reinforcement model predicts. The observed behavior could also reflect a higher transparency of the economic incentives in these two treatments, though.

Treatment	Q1	Q2	Q3	Q4	Overall
$x = 0$	8.6	3.4	1.6	0.6	14.2
$x = \frac{1}{3}$	12.5	10.8	10.0	9.4	42.7
$x = \frac{2}{3}$	10.7	9.8	10.4	9.0	39.9
$x = 1$	8.8	8.9	7.7	8.8	34.2

Table 8: Average Number of Switches per Participant (per Quarter and Overall).

Result 4 (Convergence) *Play converges fastest in treatment* $x = 0$ *and slowest in treatments* $x = \frac{1}{3}$ *and* $x = \frac{2}{3}$*.*

6 Local versus Internalized Punishment

In this subsection we report the results of an additional treatment where we investigate the differences between local and internalized punishment institutions in viscous populations. Recall that under perfect separation of groups local and internalized punishment is the same and hence, this question could not arise in previous studies.

Theoretically there is a fundamental difference between the two mechanisms as we have argued in Section 2. Under local punishment it depends on the proportion of subjects that have chosen into group A and on the degree of population viscosity whether cooperation or defection is in the interest of a group A–member. Internalized

punishment implies that in group A cooperation is a dominant strategy if punishment is high enough (i.e. for the parameters used in this experiment).

In order to compare both punishment modes we have run one more treatment where punishment in group A is internalized (i.e. where defection always leads to a deduction of 200 ECU) for the case of a low degree of population viscosity ($x = \frac{2}{3}$). In this additional treatment everything else was exactly as in our treatment with local punishment and low population viscosity ($x = \frac{2}{3}$).

Treat-ment	ShareA***	Switches**	Cooperation Rates			Payoffs			Norm Enforcer*
			A***	B***	All***	A***	B	All***	
local	.28	39.9	.64	.03	.20	201	241	230	.43
internalized	.19	29.7	.93	.09	.25	228	247	243	.28

Table 9: Share of Players in group A, Average Number of Switches per Participant, Cooperation Rates, and Payoffs under Local and Global Punishment, $x = \frac{2}{3}$. The stars indicate whether the difference between the two treatments is significant according to a two-sided Mann Whitney test ($^{***} = 1\%, ^{**} = 5\%, ^{*} = 10\%$).

As expected, under internalized punishment a lower proportion of subjects chooses into group A. Moreover, we observe less experimentation (incentives seem to be more transparent) under internalized punishment than under local punishment. We also find that cooperation rates in group A are much higher under internalized punishment. Overall, high cooperation rates in group A outweigh the effect of low group A–size. That is, the overall cooperation rate and profits are higher under internalized than under local punishment. All differences, except for the payoff difference of group B–members, are highly significant, as we have indicated in table 9.[21]

We conclude that, while the two punishment modes induce clearly different behavior with respect to cooperation decisions in group A and group choice, overall cooperation rates are — though significantly different — qualitatively close. We find some evidence that internalized punishment tends to foster cooperation somewhat more than local punishment. Whereas less participants choose into group A, those who do so constantly cooperate. In total, cooperation rates and profits are therefore higher under internalized punishment.

In practice whether internalized punishment can be implemented will depend on (i) whether or not behavior of group members towards non-members can be observed and/or (ii) whether — in case punishment is realized via social disapproval — non group members will disapprove of "defection". If only one cultural group shares a social norm which is sustained via social disapproval, then the punishment mechanism will likely evolve to be local.

[21] While a non-parametric test seems the right approach to us here, we realize that a Mann-Whitney test does not take into account dependencies caused by repeated observations of the same individual. Hence we provide regression results in Appendix B that do so. The results are very consistent.

7 Conclusion

In this paper we experimentally investigated whether agents will opt for a local punishment institution when participation in the mechanism is voluntary. Participants in our experiment could repeatedly choose between two groups, where in one of them a punishment institution was in place. The degree of population viscosity was varied across treatments.

We found that the share of participants that choose into the group where a punishment mechanism is at place increases with the degree of population viscosity. Participants cooperated more than a model based only on monetary incentives would predict. Consequently, intrinsic incentives to cooperate are not crowded out in our environment, as it may be suggested by the literature on crowding out of intrinsic motivation by material incentives (see e.g. Gneezy and Rustichini, 2000).[22] We even find evidence for feedback effects of the interaction structure on the subjects' attitudes towards sanctioning mechanisms in a post-experimental questionnaire. Participants in treatments characterized by high viscosity (who experience the power of the institution to sustain cooperation) tend to be more in favor of norm enforcement. In short, matching structure seems a powerful and important factor for sustaining cooperation. It plays a crucial role in establishing cooperative outcomes, in the short run (by changing incentives), as well as in the long run (by influencing the agents' attitudes). To understand the way it acts on economic incentives, both extrinsic and intrinsic, gives a rich potential for further theoretical and experimental research.

While we have investigated the effect of endogenous matching on cooperation levels, there is also a small literature on endogenous institution formation (see, e.g., Sutter et al. (2009) or Kosfeld et al. (2009)). It is thus natural to ask how groups would endogenously form, establish institutions, and to establish boundaries towards outsiders. The answers to those questions would shed light on phenomena that are well known from interaction between cultural groups and from the practice of certain religious groups to isolate their members from outsiders.[23] A further interesting research question is how the possibility to identify individuals, their group membership and thus, their intentions would affect cooperative behavior and group choice in environments with endogenous matching. As it has been shown, for example, by Engelmann and Grimm (2006), the imputation of bad intentions can have tremendous negative effects on the willingness to cooperate.

[22]Note that endogenous matching does not always lead to significantly higher degrees of cooperation. In Ehrhardt and Keser (1999), for example, groups did not differ in their payoff structure and therefore, subjects that tried to establish "cooperative groups" were quickly followed and exploited by "defectors". In Bohnet and Kübler (2005) subjects selected their group membership once prior to multiple round interaction in different PD games. Cooperation could not establish in the group that selected the PD with the lower defection payoff.

[23]In the examples, a high degree of separation comes at a certain cost to the group members, which is not accounted for in our setup.

References

[1] Brandts, J. and A. Schram (2001). Cooperation and noise in public goods experiments: applying the contribution function approach, *Journal of Public Economics* 79, 399–427.

[2] Bohnet, I. and D. Kübler (2005). Compensating the Cooperators: Is Sorting in the Prisoner's Dilemma Possible? *Journal of Economic Behaviour and Organization* 56, 61–76.

[3] Boyd, R. and P. Richerson (2005). The Origin and Evolution of Cultures (Evolution and Cognition), University of Chicago Press.

[4] Brosig, J. (2002). Identifying Cooperative Behavior: Some Experimental Results in a Prisoner's Dilemma Game, *Journal of Economic Behavior and Organization* 47 (3), 275–290.

[5] Brown, m., A. Falk, and E. Fehr (2004). Relational Contracts and the Nature of Market Interactions, *Econometrica* 72 (3), 747–780.

[6] Cabrales, A., G. Charness, and M.–C. Villeval (2010). Hidden Information, Bargaining Power, and Efficiency: An Experiment. *Experimental Economics*, forthcoming.

[7] Coricelli, G., D. Fehr and G. Fellner (2004). Partner Selection in Public Goods Experiments, *Journal of Conflict Resolution* 48, 356–378.

[8] Ehrhardt, K.M. and C. Keser (1999). Mobility and Cooperation: On the Run, CIRANO working papers 99s-24.

[9] Engelmann, D. and V. Grimm (2006). Overcoming Incentive Constraints - the (In)effectiveness of Social Interaction, University of Cologne Working Paper.

[10] Fehr, E. and S. Gächter (2000). Cooperation and Punishment in Public Goods Experiments, *American Economic Review* 90 (4), 980–994.

[11] Fehr, E. and S. Gächter (2002). Altruistic Punishment in Humans, *Nature* 415, 137–140.

[12] Fischbacher, U. (2007). Z-tree Zurich Toolbox for Readymade Economic Experiments, *Experimental Economics* 10(2), 171–178.

[13] Fischbacher, U., Gächter, S. and, E. Fehr (2001), Are People Conditionally Cooperative? Evidence from a Public Goods Experiment, *Economics Letters* 71, 397–404.

[14] Fischbacher, U. and S. Gächter (2006). Heterogenous Social Preferences and the Dynamics of Free-Riding in Public Goods, CeDEx Discussion Paper.

[15] Galbiati, R. and P. Vertova (2010), How Law Affects Behavior: Obligations, incentives and cooperative behaviour, mimeo.

[16] Gneezy, U. and Rustichini A. (2000). A Fine is a Price. *Journal of Legal Studies*, 29(1), 1–17.

[17] Goette, L., D. Huffman, and S. Meier (2006). The Impact of Group Membership on Cooperation and Norm Enforcement: Evidence Using Random Assignment to Real Social Groups, Working Paper 06-07, Federal Reserve Bank of Boston.

[18] Greiner, B. (2004). An Online Recruitment System for Economic Experiments, in: K. Kremer and V. Macho (eds.), Forschung und wissenschaftliches Rechnen 2003, GWDG Bericht 63, Ges. für Wiss. Datenverarbeitung, Göttingen, Germany, 79–93.

[19] Grimm, V. and F. Mengel, (2009). Cooperation in Viscous Populations - Experimental Evidence, *Games and Economic Behavior* 61(1), 202–220.

[20] Gürerk, Ö., B. Irlenbusch, and B. Rockenbach (2006). The Competitive Advantage of Sanctioning Institutions, *Science* 312 (5770), 108–111.

[21] Huck, S., G. Lünser, and J.–R. Tyran (2007). Consumer Networks and Firm Reputation: A First Experimental Investigation, ELSE Working Paper No. 291.

[22] Kocher, M., P. Martinsson, and M. Visser (2009). Social Background, Cooperative Behavior, and Norm Enforcement, Working Paper No. 385, University of Gothenbburg.

[23] Kosfeld, M. (2003). Network Experiments, IERE Working Paper No. 152, University of Zurich.

[24] Kosfeld, M., A. Okada, and A. Riedl (2009). Institution Formation in Public Goods Games, *American Economic Review* 99 (4), 1335–1355.

[25] Kosfeld, M. and A. Riedl (2004). The Design of Decentralized Punishment Institutions for Sustaining Cooperation, mimeo, University of Zurich.

[26] Lindbeck, A., S. Nyberg and J. Weibull (1999). Social Norms and Economic Incentives in the Welfare State, *Quarterly Journal of Economics* 114, 1–35.

[27] McCabe, K., M. Rigdon and, V. Smith (2007). Sustaining Cooperation in Trust Games, *Economic Journal* 117, 991–1007.

[28] Mengel, F. (2007). The Evolution of Function-Valued Traits for Conditional Cooperation, *Journal of Theoretical Biology* 245, 564–575.

[29] Mengel, F. (2008). Matching Structure and the Cultural Transmission of Social Norms, *Journal of Economic Behavior and Organization* 67, 608–623.

[30] Mitteldorf,J. and D.S. Wilson (2000). Population Viscosity and the Evolution of Altruism, *Journal of Theoretical Biology* 204, 481–496.

[31] Myerson R.B., G.B. Pollock, and J.M. Swinkels (1991). Viscous Population Equilibria, *Games and Economic Behaviour* 3, 101–109.

[32] Ones, U. and L. Puttermann (2006). The Ecology of Collective Action: A Public Goods and Sanction Experiment with Controlled Group Formation, *Journal of Economic Behavior and Organization* 62, 304–315.

[33] Ostrom, E., J. Walker and R. Gardner (1992). Covenants With and Without a Sword: Self-Governance is Possible, *American Political Science Review* 86, 404–417.

[34] Page, T., L. Putterman and B. Unel (2005). Voluntary Association in Public Goods Experiments: Reciprocity, Mimicry and Efficiency, *The Economic Journal* 115, 1032–1053.

[35] Richerson, P., R. Boyd and J. Henrich (2003). Cultural Evolution of Human Cooperation, in: P.Hammerstein (ed.), Genetic and Cultural Evolution of Cooperation, MIT-Press.

[36] Riedl, A. and A. Ule (2002). Exclusion and Cooperation in Social Network Experiments, Working Paper.

[37] Sutter, M., S. Haigner, and M. Kocher (2010). Choosing the Stick or the Carrot? Endogenous Institutional Choice in Social Dilemma Situations. *Review of Economic Studies*, forthcoming.

[38] Vega-Redondo, F. (1996). Evolution, Games and Economic Behaviour, Oxford University Press.

[39] Weibull, J. (1995). Evolutionary Game Theory, Cambridge: MIT-Press.

[40] Wilson, D.S. and E. Sober (1994). Re-Introducing Group Selection to the Human Behavioural Sciences, *Behavioral and Brain Science* 17, 585–654.

A Proof of Proposition 1

First note that whereas in the theory outlined in Section 2 there is a continuum of agents this is obviously not the case in the experiment. In the following we provide a proof of the proposition for the discrete case.[24] Denote thus the number of agents in group A (B) by n_A (n_B) and the total number of agents by n.

Furthermore note that agents in group B will always defect (it is a dominant strategy to do so in this group independently of the number of subjects in groups

[24]The proof works analogously for the continuous case.

A and B). Assuming that all agents in group A cooperate, the payoff of an agent in group A from cooperating given that all agents in group A cooperate (denoted $\Pi_A(C|C)$) is given by

$$\Pi_A(C|C) = \left(1 - \frac{n_B}{n-1}x\right) a + \left(\frac{n_B}{n-1}x\right) b$$

and the payoff from defection is

$$\Pi_A(D|C) = \left(1 - \frac{n_B}{n-1}x\right) (c-\gamma) + \left(\frac{n_B}{n-1}x\right) d,$$

where $\frac{n_B}{n-1}x$ is the probability for an agent from group A to interact with an agent from group B.[25] An agent in group A has incentives to deviate from cooperation and to defect in group A whenever $\Pi_A(D|C) > \Pi_A(C|C)$ or, equivalently, whenever

$$\begin{aligned} \left(\frac{n-1-n_B}{n-1}\right) &< \frac{x(d-b)-(1-x)(a-(c-\gamma))}{x(d-(c-\gamma)-b+a)} \\ &= \frac{200x-50}{200x}, \end{aligned} \tag{3}$$

where the last equality follows from substituting in the parameter values from our experiment. Only if $\left(\frac{n-1-n_B}{n-1}\right) = \frac{n_A-1}{n-1} \geq \frac{200x-50}{200x}$ an equilibrium where agents in group A cooperate can exist.

Analogously it can be shown that an equilibrium in which members of group A defect can exist only if

$$\frac{n_A-1}{n-1} < \frac{150-(1-x)200}{x}. \tag{4}$$

Now recall that the payoff of a group A member if all agents in group A cooperate is given by

$$\Pi_A(C|C) = \left(1 - \frac{n_B}{n-1}x\right) a + \left(\frac{n_B}{n-1}x\right) b.$$

If all group A members defect they receive

$$\Pi_A(D|D) = \left(1 - \frac{n_B}{n-1}x\right) (d-\gamma) + \left(\frac{n_B}{n-1}x\right) d.$$

Agents in group B always defect. If agents in group A cooperate they receive[26]

$$\Pi_B(D|C) = \left(\frac{n_A}{n-1}x\right) c + \left(1 - \frac{n_A}{n-1}x\right) d.$$

If agents in group A defect their payoff is

$$\Pi_B(D|D) = d.$$

Now we are in the position to prove proposition 1:

[25] Note that, for the matching probabilities, the number of *other* subjects in groups A and B matter (exclusive of the subject under consideration). Thus, in the discrete case, the equivalent to p_B in the matching probability is $\frac{n_B}{n-1}$.

[26] $\Pi_B(D|C)$ ($\Pi_B(D|D)$) denote the payoff of a group B member if all members of *group A* cooperate (defect).

Case (i) If $x < \frac{1}{4}$ agents in group A will always (independently of n_A) cooperate as can be read from (3) and (4). But note that if this is the case and $x < 1/4$, we have that $\Pi_A(C|C) > \Pi_B(D|C)$ and the dynamic equation implies that $\dot{p}_A > 0 \; \forall p_A \in [0, 1]$. All agents will thus end up in group A ($n_A^* = n$).

Case (ii) Now consider the interval $x \in [\frac{1}{4}, \frac{4}{7}]$. Note that in this case an equilibrium where agents in group A cooperate exists only if (3) holds (if there are sufficiently many agents in group A), whereas an equilibrium where agents in group A defect exists if and only if (4) holds. Furthermore, $\Pi_A(C|C) > \Pi_B(D|C)$ for high n_A and $\Pi_A(D|D) < \Pi_B(D|D) \; \forall n_A \neq 0$. Consequently in this parameter range both equilibria $n_A^* = n$ and $n_A^* = 0$ coexist.

Case (iii) Finally consider the interval $x > \frac{4}{7}$. Note that in this case $\Pi_A(C|C) < \Pi_B(D|C) \; \forall n_A$ and $\Pi_A(D|D) < \Pi_B(D|D) \; \forall n_A \neq 0$. Consequently, independently of whether agents in group A cooperate or defect (and independently of how many they are) agents in group B will always receive a higher payoff than agents in group A and thus, $\dot{p}_A < 0 \; \forall p_A \in [0, 1]$. Group B will proliferate. The unique equilibrium will have $n_A^* = 0$.

B Regression tables

Choice of Group A	(1)	(2)
constant	-1.8247^{***}	-1.2148^{***}
	(0.1469)	(0.0828)
treatment 0	4.6529^{***}	3.5190^{***}
	(0.2740)	(0.1543)
treatment 1	2.5821^{***}	1.9740^{***}
	(0.3134)	(0.0920)
treatment 2	0.4084	0.1305
	(0.2637)	(0.0877)
ρ	0.5080	0.0838

Table 10: Random Effects Panel Data Logit Regression. (1) using individuals as independent unit of observation. (2) using matching group. $^{***}1\%,^{**}5\%,^{*}10\%$. ($(\Pr > \mathcal{X}^2) < 0.0001$)

cooperation	(1)	(2)
constant	-2.2512^{***}	-1.6517^{***}
	(0.1611)	(0.0431)
treatment 0	4.9873^{***}	4.0914^{***}
	(0.2514)	(0.0733)
treatment 1	2.9586^{***}	2.2056^{***}
	(0.2513)	(0.0545)
treatment 2	0.3556	0.1640^{***}
	(0.2513)	(0.0565)
ρ	0.4692	0.0800

Table 11: Random Effects Panel Data Logit Regression. (1) using individuals as independent unit of observation. (2) using matching group. $^{***}1\%,^{**}5\%,^{*}10\%$. ($(\Pr > \mathcal{X}^2) < 0.0001$)

Profits	(1)	(2)
constant	229.81***	229.81***
	(5.51)	(8.75)
treatment 0	142.91***	142.91***
	(7.80)	(14.73)
treatment 1	84.11***	84.11***
	(7.80)	(18.59)
treatment 2	0.1458	0.1458
	(7.80)	(13.56)
ρ	0.0834	0.0834

Table 12: Panel Data OLS Regression (1) without (2) with standard errors clustered by matching group ***1%,**5%,*10%. ((Pr > $\mathcal{X}^2$) < 0.0001)

	(Defector)	(Altruist)	(CondCoop)	(Hump)	(Norm-Enforcer)
constant	−1.2130***	−3.8501***	−0.2513	−1.4663***	−0.7884***
	(0.3434)	(1.0105)	(0.2909)	(0.3698)	(0.3113)
treatment 0	−0.2533	0.7146	−0.4418	0.7731	0.9555**
	(0.5046)	(1.2421)	(0.4223)	(0.4801)	(0.4253)
treatment 1	0.7021	0.0000	−0.4418	0.3677	1.2113***
	(0.4547)	(1.4291)	(0.4223)	(0.4978)	(0.4290)
treatment 2	0.7022	dropped	−0.3494	0.1313	0.5371
	(0.4547)	()	(0.4192)	(0.5129)	(0.4261)
Pseudo R2	0.0292	0.0133	0.0060	0.0148	0.1060

Table 13: Logit Regression on Questionnaire Data. ***1%,**5%,*10%. ((Pr > $\mathcal{X}^2$) < 0.0001)

	(Group A)	(Cooperation)	(Cooperation A)	(Cooperation B)
constant	-1.4136^{***}	-1.8492^{***}	0.8384^{***}	-4.1893^{***}
	(0.2123)	(0.19931)	(0.2905)	(0.2653)
internalized	-0.5814^{**}	0.4039	2.6795^{***}	1.2695^{***}
	(02610)	(0.2726)	(0.3721)	(0.3253)
ρ	0.4853	0.4182	0.5126	0.4170
	(Profits)	(Profits A)	(Profits B)	(Norm Enforcer)
constant	229.95^{***}	193.18^{***}	240.27^{***}	-0.2513
	(4.75)	(7.54)	(4.90)	(0.2909)
internalized	13.30^{**}	26.89^{***}	7.41	-0.6109^{*}
	(6.28)	(10.50)	(6.43)	(0.3993)
ρ	0.0468	0.0441	0.0546	0.0161 (pR2)

Table 14: Using Regressions (on data from Treatments with $x = \frac{2}{3}$) for differences between local and internalized punishment. Panel Data Random Effects Logit Regressions on binary variables (Group A, Cooperation), Simple Logit for variable Norm Enforcer and panel data OLS regression on profis. $^{***}1\%, ^{**}5\%, ^{*}10\%$.

C Instructions for Treatment $x = 1/3$

Welcome and thank you for participating in this experiment. Please read these instructions carefully. They are identical for all participants with whom you will interact during this experiment.

If you have any questions please raise your hand. One of the experimenters will come to you and answer your questions. From now on communication with other participants is forbidden. If you do not conform to these rules we have to exclude you from the experiment. Please switch off your mobile phone at this moment.

For your participation you will receive 2,50 Euro. During the experiment you can earn more money. How much more depends on your behavior and the behavior of the other participants. During the experiment we will use ECU (Experimental Currency Units). At the end we will pay you in Euros according to the exchange rate 1 Euro = 2500 ECU. All your decisions will be treated confidentially.

The Experiment

At the beginning of the experiment we split you and the other participants equally into two groups — **group A and group B**. In each round of the experiment you play a game with a "representative member" either from group A or group B that we will call your **interaction partner** in the following. At the beginning of the experiment you play at least four rounds as a member of the group that you have been assigned to originally. In each of these four rounds you play a game that we describe in the next section. Starting with round five each round has two phases:

- **Phase 1**: Each round some of the participants can decide whether to change