Keeping up appearances: Experiments on cooperation in social dilemmas

van den Broek, E.M.F.

Publication date
2014

Chapter 3: Public goods and private aversions

3.1 Introduction[20]

In the previous chapters I studied situations in which people were confronted with someone to whom they could offer help. In everyday life, such confrontations are more than simply random events. Instead, people choose with whom they interact; moreover, acting cooperatively is not limited to one person but may extend to a group. In this chapter I shift the focus from dyadic helping to cooperation in groups and consider the relevance of partner choice. I do so in a setting of voluntary public good provision.

A public goods game, as described in Chapter 1, captures a social dilemma in which the individual interest of the group members conflicts with the collective interest. A general finding of laboratory studies of such games is that members' contributions to the public good decline as the dilemma is repeated. Many studies have been devoted to formal[21] and informal enforcement mechanisms that may sustain long-term efficiency. Two informal mechanisms stand out as particularly effective in solving the free-rider problem: costly punishment of fellow group members and the possibility to form groups with preferred partners (Chaudhuri, 2011).

In the world outside the laboratory, the options for punishment and endogenous group formation may often occur simultaneously. People choose with whom they prefer to interact and punish or exclude others. The interplay between partner choice on the one hand, and punishment on the other hand, is not straightforward. Groups may use exclusion as a substitute for or an extreme form of punishment (Ahn et al., 2008). Alternatively, the two informal mechanisms may enforce each other's effects: a group member who punishes fiercely may be valued either as a reliable partner (Nelissen, 2008) or, conversely, excluded out of fear for retaliation.

[20] This chapter is based on Van den Broek, Kocher and Schram (working paper, 2010).
[21] Formal enforcement mechanisms include for instance exclusion from club goods on the basis of formal contracts and agreements (Buchanan, 1965).

To date, endogenous grouping and punishment have been studied in isolation, however. An important part of the literature on endogenous group formation addresses the exclusion of group members. Granting players the power to exclude fellow players may enhance cooperation in two ways. Firstly, the threat of irreversible expulsion may be enough to discipline free-riders. In Cinyabuguma et al. (2005) participants could exclude fellow group members through a majority voting system. Low contributors avoided expulsion by increasing their contributions after receiving a high number of exclusion votes. Secondly, endogenous group formation through exclusion, but also through voluntary matching, may raise average contributions by assortment. For example, the formation of groups consisting of a majority of conditional contributors (who constitute around 50% of the population, Fischbacher et al., 2001) increases contributions (Gunnthorsdottir et al., 2007).

Our experiment adds to an emerging literature on endogenous grouping that started with Ehrhart and Keser (1999). Here, costly migration sustained high cooperation levels but created iterated chasing of cooperators by free riders. Other mechanisms of endogenous group formation also sustain high cooperation levels, like majority voting about entry and exit of group members with evolving group size (Charness and Yang 2010). Group reduction or ostracism, the inverse of group formation, is another effective institution (Maier-Rigaud et al. 2005). Subjects use exclusion to punish 'unfair' behavior (non-strategic reason) and expect changes in behavior in response to exclusions (strategic reason; Masclet et al., 2003).

To our knowledge, the only study that combines endogenous group formation with punishment is Page et al. (2005)[22]. In their setup subjects can rank each other's candidacy as a co-member and groups are formed that optimize agreement on candidacy. Ranks are based on contribution levels only, and subjects cannot exclude each other with certainty. Although ranking is costly, 80% of the subjects always use the ranking opportunity, showing that ranking alone is a potentially strong mechanism. There are two drawbacks to this combination of ranking with punishment, however. Outside of the laboratory, we not only prioritize between people but also exclude them; and secondly, a most important aspect of the interplay between punishment and partner preferences is the reputation effect of punishing. In Page et

[22] In Ones and Putterman (2007) low contributors are exogenously matched with punishers, which increased contributions.

al. (2005) it remains unclear how punishers are viewed; i.e., would they be excluded or preferred.[23]

This chapter deals with the empirical questions that arise from the interplay between endogenous group formation and punishment. Our main research questions are

(i) To what extent do the two mechanisms, punishment and endogenous group formation (including exclusion), increase contributions to a public good?

(ii) How efficient are these mechanisms?

(iii) Are they complements or substitutes?

We present a laboratory environment with fixed group size, where we allow both for ranking and exclusion of partners based on information on their previous contribution and punishment choices. We find that endogenous group formation with exclusion creates high cooperation and efficiency levels. In combination with punishment, exclusion yields even higher contributions.

The remainder of this chapter is organized as follows. The following section presents a brief overview of the theory applicable to the environment we study in our experiments. A description of the experimental procedures and design in Section 3 is followed by the results in Section 4. Section 5 concludes.

[23] In Rockenbach and Milinski (2011), a single observer may learn about contributions and punishment behavior in the first 15 periods and exclude a partner for the second remaining periods. Here, punishment did not have an effect on eligibility as a co-player.

3.2 Game description and theoretical predictions

The game consists of three stages. At stage 1, groups are either exogenously formed, or endogenously determined by individuals' preferences with respect to whom they would like to be matched with. At stage 2, individuals participate in a public goods game (see Chapter 1) where each member of a group decides whether or not to contribute to a public good. At stage 3, members of the group may decide to enforce costly punishment on any other member(s).

Starting with stage 2, consider the standard linear public goods game as introduced in Chapter 1. Let $I = \{1, 2, \ldots, m\}$ denote a group of m subjects who interact in T periods. In each period $t \in \{1, 2, ..., T\}$ individual $i \in I$ receives an endowment ω which can be allocated either to a private good or a public good. The voluntary contribution of individual i to the public good in period t, $g_{i,t}$, is a binary choice, *i.e.* it must satisfy $g_{i,t} \in \{0, \omega\}$. The marginal per capita return from investing in the public good is denoted by α, and satisfies $0 < a < 1 < ma$, meaning that the self-interested choice and the socially optimal one are in conflict.
Consequently, i's payoff is

$$\pi_{i,t} = \omega - g_{i,t} + a\sum_{k=1}^{m} g_{k,t} \qquad (1)$$

At beginning of stage 3, each group member is informed about the individual contributions by the other group members and decides whether to punish other individuals in his group. Punishment is costly for the punisher as well as for the punished member. We implement punishment as a *binary* decision.[24] The effectiveness of punishment is captured by the variable c. Taking into account the monetary consequences of the second stage in each period yields the following payoff function for i.

$$\pi_{i,t} = \omega - g_{i,t} + a\sum_{k=1}^{m} g_{k,t} - c\sum_{k\neq i} p_{ik,t} - \sum_{h\neq i} p_{hi,t} \qquad (2)$$

[24]The reason to choose a simple binary punishment technology is that it creates relatively simple summary statistics of any particular individual's choices. These statistics may be used by others when deciding on with whom they wish to be matched in the endogenous group formation of stage 1.

where $p_{kh,t} = 1$ if member *h* has punished member *k* in period *t* and zero otherwise and the costs of punishing have been set to 1.

The novel element in our experiment takes place at stage 1. Individuals observe information about contributions, $g_{h,\_}$, and punishment decisions, $p_{kh,\_}$, by all other *h* in the population in all previous periods $\_ < t$. Subsequently, they can use this information to rank other individuals in terms of how much they would like to be in a group with them in period *t*. In doing so, they may exclude specific others altogether. A matching procedure is then used to form groups in accordance with these preferences. [25]

Assuming that subjects care only about their own monetary payoffs and assuming common knowledge of rationality, a risk-neutral decision maker will abstain from costly punishment. Since the game is finitely repeated, backward induction yields the standard result that the contribution decision will not be affected by the possibility of punishment. Consequently, the game will yield zero contributions (*i.e.* $g_{i,t} = 0$ for all *t*) because free-riding is a dominant strategy due to $a < 1$. Therefore common knowledge of selfishness and full rationality implies that all individuals will be indifferent with respect to whether they end up in a group or not, and whom they will be grouped with. The theoretical prediction in case of selfish preferences is therefore that any group may be formed at stage 1; there are no contributions at stage 2 and no punishment at stage 3.

Empirically, such reasoning proves to be of very limited use, because the slightest belief of somebody contributing a positive amount would induce even a purely selfish player to prefer being in a group with this person. As long as the threat of punishment is used rationally (in line with social preferences), there are equilibria (e.g., in Fehr-Schmidt models, see Chapter 1) in which everyone wants to join a group and contributes positive amounts, regardless of whether players are selfish (Kosfeld, 2009). The underlying intuition of such equilibria is that once punishment is a credible threat, everyone in a group will contribute. If all contribute,

[25] *Cf.* section 3.3 for details about the matching procedure used in our experiment.

everyone is welcome in all groups[26]. For cases where (some) individuals have other-regarding preferences, such results allow us to formulate the expected patterns per type. As for group formation, free riders prefer to team up with cooperators (whether punishment is possible or not); but cooperators would not reciprocate this preference. The latter holds even if cooperators have the option to punish free riders. This is because punishment costs are expected to be lower in groups without free riders. If punishment is possible, free riders try to avoid punishers; cooperators may be indifferent towards punishers. Finally, the predictions for contributions differ on whether punishment is possible. If it is, free riders may believe that the threat of punishment is so high that they actually contribute, if not, they will not.

It is beyond the scope of this chapter to develop a more detailed formal model of how preferences about group composition and exclusion are formed and how they develop based on observed choices. Ultimately, we aim for empirical evidence on the effects of endogenous group formation (with an exclusion possibility) and, particularly, on its interplay with punishment.

[26] In a different vein, Brekke et al. (2007) show that the fear of exclusion from a high-contribution team may lead to high cooperation levels. Hirshleifer and Rasmusen (1989) show that equilibrium with cooperation exists in a finitely repeated public goods game with costless expulsion.

3.3 Experimental design and procedures

A total of 90 subjects participated in 6 sessions of the experiment in March and April 2008 at the CREED laboratory in Amsterdam. Subjects were students with a variety of majors, including economics (33%) and psychology (24%). Each session lasted approximately 90 minutes and subjects earned on average € 24.06 including a show-up fee of € 7.

Subjects were brought into the laboratory and told that they would participate in two experiments. The first consisted of a value orientation test (Offerman *et al*. 1996) and served to obtain an independent measure of each participant's social value orientation (see Appendix A for further details).[27] After completion of this first experiment, subjects were informed that the second experiment would comprise two independent parts, that Part I would last 10 periods and that they would receive instructions for Part II after Part I had been completed. After the computerized instructions (see Appendix A) they completed a quiz to check for understanding. The experiment was computerized.[28] No participant was informed about the identity of others. Earnings in the experiment were in 'francs'. Aggregate earnings were exchanged for euros individually and privately after the experiment, at an exchange rate of € 0.03 for each franc.

Our experimental treatments in the second experiment are based on the public goods games introduced in the previous section. Part I consists of a standard ten-period game (*T = 10)* with or without punishment (depending on the treatment). In both cases, subjects are in groups of $m = 3$; the composition changes after each period within a matching group of nine. In each period, each subject is endowed with $\omega = 20$ francs (equivalent to € 0.60). They have a binary choice of either keeping the endowment or investing it in a public account (which we call a "group project" in the experiment). We use an MPCR of $a = 0.5$, which means that the sum of the group members' contributions in a period is multiplied by 1.5 and then equally divided amongst the three members. In the treatment with punishment, each subject is informed about the group members' decisions and can subsequently decide to allocate one "subtraction point"

[27] Details about the results of the first experiment are available upon request. The second experiment is the main focus of this chapter.
[28] We thank CREED programmer Jos Theelen for writing the Delphi program. It is available upon request.

to either or both of the other two in the group. This is again a binary decision, where allocation of a point costs the individual giving it 1 franc and the member receiving it $c = 3$ francs.

In Part II the game is exactly the same as in Part I (either with or without punishment). The difference is that groups are no longer exogenously formed. In stage 1 of Part II, subjects are matched according to their own preferences as expressed in their willingness to be in a group with specific other subjects. This is organized as follows. First, subjects are, like in Part I, allocated to matching groups of nine that remain constant throughout Part II. In each of the ten periods, at most three groups of three are formed. Group formation takes place in three steps in each period. In step 1, three players are randomly selected to be *proposers* (called "type A" in the experiment). The remaining six players are *responders* ("type B"). Note that most players will be proposer in some periods and responder in others. A group always consists of one proposer and two responders.

In step 2, proposers are asked to submit a preference profile over all responders. Before doing so, they are informed about each responder's contributions to the group project and (if applicable) her allocation of punishment points in all previous periods.[29] A preference profile consists of a score for each of the six responders on an ordinal 6-point scale ranging from 1="highly preferred" to 6="least preferred". Ties are allowed. It is also possible to exclude one or more responders from the proposer's group by not giving them a score.[30] While the proposers are scoring the responders, the latter have to indicate for each of the proposers their willingness to be in a group with her. This is a binary decision and can be based on information we provide about the proposers' contributions and punishment decisions in all previous periods.[31]

[29] During the group formation stage, subjects can scroll through information about the contribution, number of punishment points assigned and the earnings of the other players in every previous period. See the instructions in Appendix A for an example. A "–" appears in the table if a subject had not participated in the public good game in a specific period. The same information remains visible in the subsequent contribution phase.

[30] For example, a proposer submitting scores 2,3,1,−,2,− for respondents 1…6, respectively, indicates that she refuses to be in a group with respondents 4 and 6, that she most prefers to be with 3 followed by a tie for 1 and 5 and that respondent 2 is the least preferred of the acceptable ones. Note that the ranking is ordinal, submitting 2,3,1,−,2,− indicates the same preference as 4,6,3,−,4,−.

[31] Numbers indicating specific proposers or respondents are scrambled in each period such that individuals cannot be identified across rounds.

In step 3 the preferences of proposers and responders are used to form groups in the following way: a proposer is randomly chosen and matched with her two most preferred responders.[32] If the preferred responders have agreed to join this proposer, a group is formed; otherwise, the next preferred responder that is willing to join that proposer will be chosen. If it is not possible to find two responders that are acceptable for the proposer and at the same time are willing to join her, this proposer is not allocated to a group in that period. This procedure is repeated for the other proposers with responders who have not yet been assigned. All players that have not been assigned to a group at the end of step 3 receive their endowment ω but cannot play the public goods game in the period concerned or allocate punishment points and receives no information during that period about the other players.

For each of the six responders coupled to a proposer, the proposer can either indicate a score on a six-point scale or no score if she wishes to exclude them. A responder can either agree or decline to play with each of the three proposers, in the latter case effectively excluding someone from her group. Each type of player can effectively withdraw from playing by excluding all players of the other type. Aside from giving us a procedure to develop groups based on preferences, this structure provides detailed information allowing us to carefully analyze social exclusion. We can do so by comparing per player the choice to exclude someone when the player is a responder, i.e. when the only other option is to agree to join, with the choice to exclude someone when the player is a proposer, i.e. when the alternative includes the possibility to assign a relatively low score[33].

In short, our design consists of a repeated public goods game and distinguishes between two treatments variables: (i) punishment opportunities versus no-punishment, and (ii) exogenous versus endogenous group formation. We varied (i) between subjects and (ii) within subjects. The reason for the latter choice is that it allows subjects to get acquainted with the game (in Part I) before they need to form preferences about group membership in Part II. Table 3.1

[32] In case of a tie, one is chosen at random.
[33] In the treatment without punishment, excluding is weakly dominated by assigning a relatively low score to someone, because a small chance that one of the group members contributes a positive amount makes it more profitable to be in a group than to end up alone.

summarizes our treatments and gives the number of independent observations (*i.e.* matching groups) in each treatment.

Table 3.1. Treatments

	Baseline treatment *(4 groups)*	**Punishment treatment** *(6 groups)*
Part I (10 periods)	Random grouping Contribution	Random grouping Contribution Punishment
Part II (10 periods)	Endogenous grouping • type allocation • preference elicitation • grouping Contribution	Endogenous grouping • type allocation • preference elicitation • grouping Contribution Punishment

Notes. Timeline of treatments. Matching groups will be used as the unit of observation for our statistical tests.

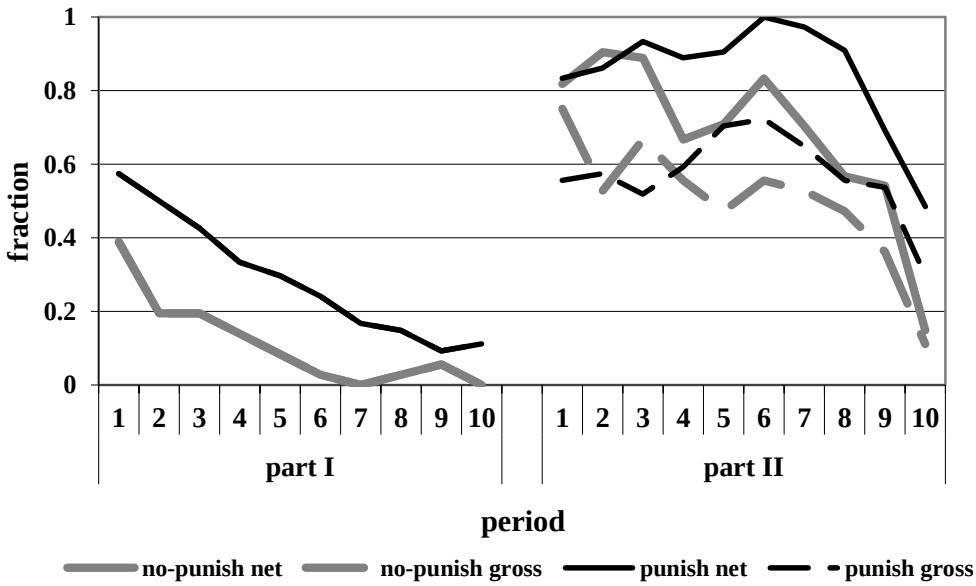
3.4 Experimental results

We start this section with a general overview of contributions and efficiency (section 4.1). The overview is followed by an analysis of punishment and its effect on contributions (section 4.2). In section 4.3, we analyze the choice to exclude and individual preferences with respect to the other group members. Finally, in section 4.4, we will compare between distinct mechanisms and analyze how our subjects choose what to use.

3.4.1 Contributions

Figure 3.1 gives an overview of the fraction of subjects contributing their endowment to the public good. For part II, it distinguishes between gross fractions (the number of contributions divided by the total number of individuals) and net fractions (the number of contributions divided by the number of individuals that are allocated to groups). Treatments with and without punishment are shown separately.

Figure 3.1. Contributions to the public good over time



Notes. For each period, the graph shows the fraction of participants that contributed their endowment to the public good. Black (grey) lines show the fraction for treatments with(out) punishment. In part II solid lines show the 'net' contributions, i.e., as fraction of participants who have been allocated to groups and dashed lines show the 'gross' contributions, i.e., as a fraction of all participants.

Four things stand out in figure 3.1. First, far more participants contribute in part II than in part I. This holds for the whole population but even more so if we only look at participants that have been allocated to groups. The introduction of endogenous group formation increases average (gross) contribution levels from 0.11 to 0.5 without punishment and from 0.29 to 0.57 with punishment.[34] The increase is statistically significant in the treatments with and without punishment ($p < 0.01$, resp. $p = 0.02$; Wilcoxon-signed-rank tests[35]). Second, the possibility of punishment yields higher contribution levels in both parts I and II. In part I, participants contribute 11% of the time without and 29% with the opportunity to punish. This difference is statistically significant on the level of matching groups ($p = 0.05$; Mann-Whitney-U test). In part II, punishment yields an increase in (net) contributions from 43% to 57%. The difference, however, is not significant ($p = 0.16$; Mann-Whitney-U test). The opportunity to punish group members does not have significantly less of an effect if participants have a say in the formation of their groups ($p = 0.17$, $F(1,8) = 2.22$). Third, while contributions gradually diminish towards zero in part I, they remain at higher levels longer in part II.

Finally, some individuals are excluded from groups. This can be seen from the fact that the gross fraction is always lower than the net fraction (with one exception in the final round), meaning that the denominators in the net fractions (*i.e.*, the number of individuals in groups) must be lower than the corresponding denominators in the gross fractions (the number of individuals). We will closely analyze the pattern of exclusion below. At an aggregate level, the number of groups that is formed is an indication of the extent of exclusion. Without punishment, 74% of the possible groups are formed and with punishment the percentage is 67%. The difference is statistically not significant ($p = 0.32$; Mann-Whitney-U-test).

Results 1: *(a) Endogenous grouping significantly increases contributions levels, both in the treatments with and without punishment. (b) Punishment increases contributions. (c)*

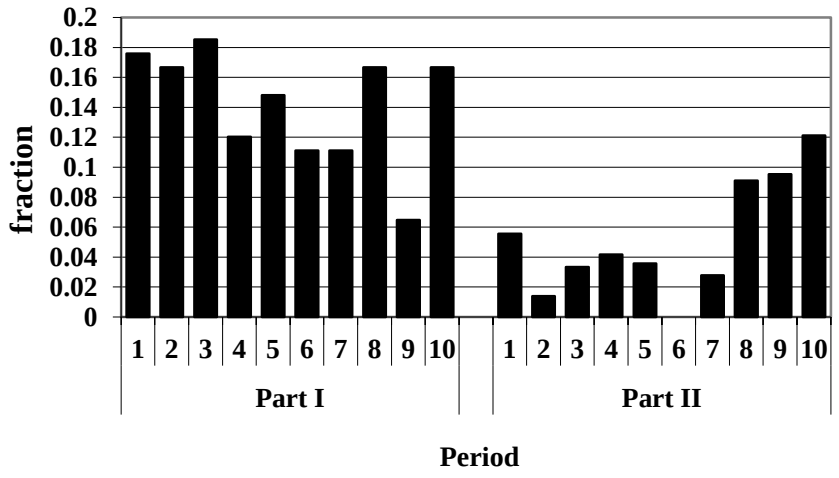
[34] We compare gross contribution levels because they are the more straightforward indicators of cooperation within the whole population.
[35] Unless indicated otherwise, tests reported in this section use matching groups consisting of nine subjects as the unit of observation.

Endogenous grouping leads to the exclusion of almost one out of every three participants. (d) High contributions through endogenous grouping break down in the final rounds of the game.

3.4.2 Punishment behavior

Figure 3.2 gives an overview of the frequency with which participants punished other group members as a fraction of the total number of times they could have punished them. For part II this fraction is calculated for the groups that were endogenously formed. The figure shows that punishment is lower across all rounds in part II (5 % of all punishment opportunities are used) than in part I (14%). The difference is statistically significant ($p = 0.05$; Wilcoxon-signed-rank test). Note, however, that the frequency of punishment increases in the final rounds of part II (but remains less than the average in part I).

Figure 3.2. Punishment behavior over time



Notes. For each period, bars show the fraction of times participants punished group members.

The fact that participants punish less in part II can be attributed to three possible causes. First, they may have less reason to punish because their group members are contributing more (*cf.* figure 3.1). Second, they may use punishment less because there are other mechanisms (group formation and exclusion) that they can use to enforce cooperation. Third, they may be excluded if they punish. To distinguish between these options, we ran random effects probit regressions where the decision to punish is explained by a number of variables, including the group member's contribution decision. We do so separately for parts I and II. If an individual

responds differently to identical situations in both parts, this is support for the second explanation for reduced punishment. Table 3.2 presents the results.

Table 3.2. Determinants of the decision to punish

	Part I	**Part II**
Own contribution	0.65***	-0.068
Third party contribution	-0.61***	-0.94***
Negative deviation	1.37***	2.13***
Positive deviation	-0.86***	1.89***
Period	-0.23**	0.07
Period^2	0.02***	-0.00
# observations (groups)	1080 (4)	1080 (6)
Wald chi^2 (6)	201.41***	48.17***

Notes. The table presents the results of a random effects probit regression used to explain punishment of *j* by *i*. Formally, it gives the marginal effects at the means in $P_{jit} = \Phi\sum_i X'_{ijt}\beta + \mu_m)$ where Pr_{jit} is the probability that *i* punishes *j* in period *t*; Φ denotes the cumulative normal distribution and X_{ijt} is a vector of independent variables relating to *i* and *j* in *t* as described in the first column of the table. μ_m is a (white noise) matching-group-specific error that corrects for the dependencies within matching groups. The independent variables are defined as follows. "*Own Contribution*"= dummy variable equal to 1 if *i* contributed in t; "*Third party contribution*"= dummy variable equal to 1 if third member contributed in t; "*Negative deviation*"=dummy variable equal to 1 if j did not contribute and there was at least one contribution; "*Positive deviation*"=dummy variable equal to 1 if j contributed and there was at most one other contribution; "period"= period number; "*period*^{2}"=period number squared.*=statistically significant at the 10% level; **=statistically significant at the 5% level; *** = statistically significant at the 1% level.

The results show that individuals react strongly to a negative deviation of the other's contribution, even when controlling for own contribution and third party contribution. In the first and the second part, an individual is significantly more likely to punish a non-contributor if the other two group members (including the punisher) did contribute. The own contribution has a significant positive effect on punishment behavior in Part I and no effect in Part II. This may reflect a shift from punishment to exclusion; alternatively, it may result from a selection

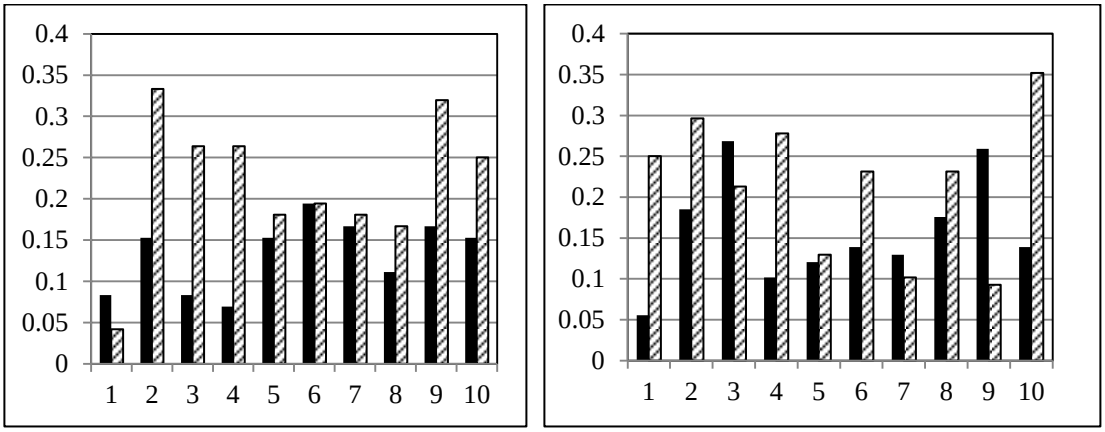
effect, since contributing participants are more likely to be included in a group. Third party contribution matters in both parts. The effect of a positive deviation of the punished person changes from a significantly negative to a significantly positive effect in Part II, probably because almost everyone contributes. In Part I, period and squared period have a significant effect; in Part II the effects disappear. Taken together, these results suggest that while multiple factors affect the decision to punish in part I, punishment in Part II is more focused towards negative deviators, as the increase in the negative deviation suggests. These differences show that the overall reduction in punishment is only partly explained by the higher contributions in the second part; participants resort to other methods than punishment. The fact that the dummy for negative deviation does not disappear in Part II suggests that fear of exclusion cannot explain the reduction in punishment completely.

Result 2: *The introduction of endogenous grouping partly crowds out punishment and directs punishment more strongly towards negative deviators.*

3.4.3 Exclusion and ranking

In the treatment without punishment, subjects excluded on average 18 % of the potential partners, compared to 19 % in the punishment treatment. Proposers excluded 15 % of the responders, while responders excluded 22 % of the proposers from interaction in the forthcoming period (MW, z=-1.82, p=0.07). The correlation between a subject's decisions to exclude as a proposer and as a responder is 0.27 ($p < 0.01$; Spearman rank correlation between the exclusion percentage of an individual in the role of responder and in the role of proposer). Figure 3.3 gives the pattern of exclusion over time; it shows no trend across periods.

Figure 3.3 Exclusion over time



Notes. For each period, bars show the fraction of potential partners that proposers (solid bars; fraction excluded per 6 options) and responders (striped bars; fraction they excluded per 3 options) excluded in the treatments without (left) and with punishment (right).

As a consequence of exclusion, on average 33 % (26 %) of all subjects did not participate in a group in the treatment with(out) punishment, either because they were deliberately excluded, or because they withdrew by excluding all others, or because all others they were willing to match with were already allocated to other groups. The difference between the two treatments is statistically not significant (p = 0.32; Mann-Whitney-U test) and does not reveal a clear pattern over time.

To further analyze the possible determinants of the decision to exclude others we ran random effects probit panel regressions. The binary decision by i whether to exclude j in period t ($exclude_{ijt}$) is explained by variables related to i's and j's history of contributing and punishing, as well as a dummy distinguishing between proposers and responders. Random effects on the matching group level correct for dependencies within the matching groups. We ran separate regressions for the punishment treatment and the baseline treatment and

distinguish between models with lagged variables (Model I and II) and models with averages of previous choices (Model III and IV). Table 3.3 presents the results[36].

Table 3.3 Determinants of i's decision to exclude j

Model	**I**	**II**	**III**	**IV**
Treatment	**Baseline**	**Punishment**	**Baseline**	**Punishment**
Subject	-0.00*	0.00	-0.00	0.00
Responder (dummy)	0.31***	0.19***	0.47***	0.19**
Period	0.21***	0.13**	0.01**	-0.05
Period^2	-0.02***	-0.01**	0.02**	0.01
play i (t-1)	0.28**	0.23***		
play j (t-1)	0.90***	0.91***		
contribution i (t-1)	Coll	0.13		
contribution i (average)			0.51***	0.34***
punishment i (t-1)		0.07		
punishment i (average)				0.14
contribution j (t-1)	-2.70***	-2.24***		
contribution j (average)			-0.20	-0.77***
punishment j (t -1)		0.39**		
punishment j (average)				1.31***
# observations (groups)	1440 (4)	2160 (6)	1296 (4)	2160 (6)
Wald Chi^2	239.78***	393.05***	45.48***	148.48***

Notes. The table presents the results of four random effects probit regression used to explain exclusion of *j* by *i*. Formally, it gives the marginal effects at the means in $P_{jit} = \Phi\sum_i X'_{ijt}\beta + \mu_m)$ where Pr_{jit} is the probability that *i* excludes *j* in period *t*; Φ denotes the cumulative normal distribution and X_{ijt} is a vector of independent variables relating to *i* and *j* in *t* as described in the first column of the table. μ_m is a (white noise) matching-group-specific error that corrects for the dependencies within matching groups. Coll means the variable is collinear. Only data

[36] We ran a logit with the same specifications with similar results, in which no collinearity showed up. Results are available on request.

are used where *i* had to make a decision about *j*. The lagged contribution of the excluder is collinear with having played in the previous period.
Model I and II include only lagged variables; Model III and IV only averages over previous periods.

The results show again that responders exclude more than proposers, also in the baseline treatment where they could not punish free riders. Having played in the previous round increases the chances of excluding and being excluded. This could be due to two reasons. Either someone who did not play does not reveal any (potentially negative) information about previous behavior, and is therefore not excluded; or it could be due to a selection effect, since people who did not play may have been excluded in the previous round because of a history of non-contributing. This statistically increases the likelihood that someone who played in the previous period has a good track record.
Having contributed decreases the chance to be excluded, whereas punishing increases it. The models including averages confirm the picture. For free riders, it is intuitive that they exclude punishers; for contributors, the effect may be due to fear for antisocial punishment (Herrmann et al., 2008). Finally, the positive coefficients for the average contribution of the excluder show that on average subjects who contribute more often also exclude more often. This, too, is understandable: they have more to lose.

Although neither the own average punishment nor punishment in the previous period influences exclusion significantly, a difference between the treatments is the importance of the average contribution of the excluded. In the punishment treatment, this average contribution has a negative effect on exclusion, as expected; however, in the baseline treatment, this effect is not significant[37]. Overall, contributors exclude more often than free-riders, irrespective of whether they can punish or not. This result is in line with the predictions formulated in 3.2.

Results 3(a) *Low contributors and punishers are excluded more often.* ***(b)*** *Contributors exclude more than free-riders.*

Aside from exclusion, proposers can also reveal preferences about the responders they want to be grouped with by giving scores on a scale from 1 (high score) to 6 (low score). These scores

[37] A possible explanation is that contribution in the previous period is used as a cut-off rule for exclusion and average contribution is used for ranking.

allow us to investigate the factors that make a responder a relatively desirable group member. To do so, we ran an ordered probit regression explaining the scores given to a certain responder in round *t* by previous choices made by the involved proposer and responder.[38] Table 3.4 presents the results, separately for the cases with and without punishment and again distinguishing between a model with lagged variables and one with averages over previous rounds.

Table 3.4. Determinants of the proposer's score of responders

Model	**V**	**VI**	**VII**	**VIII**
Treatment	**Baseline**	**Punishment**	**Baseline**	**Punishment**
Subject	0.000	0.00**	0.00	0.00***
Period	-0.30	0.00	-0.01	-0.14
Period2	0.02	0.00	0.00**	0.01**
contribution i (t-1)	-0.21	-0.04		
contribution i (average)			0.10	0.12
punishment i (t-1)		0.06		
punishment i (average)				0.26
contribution j (t-1)	-0.99***	-1.01***		
contribution j (average)			-0.29*	-0.17
punishment j (t -1)		0.51***		
punishment j (average)				1.78***
# observations (groups)	805 (4)	1493 (6)	805 (4)	1493 (6)
Log Lkh	-513.51	-511.10	-539.21	-517.69

Notes. The table presents the results of four ordered probit regressions, used to explain the score of *j* (a responder) by *i* (a proposer). Ranking options ranged from high (1) to low (6) (exclusion (7) not included in this table). Formally, it gives the marginal effects at the mean of a vector of independent variables relating to *i* and *j*

[38] In these regressions, we disregard responders that were excluded by the proposers (*i.e.*, not given a score). Alternatively, one could consider exclusion as the lowest score (*e.g.*, a score of 7). We prefer not to do so, because we consider exclusion (as analysed in table 3) to be a qualitatively different decision than scoring. The alternative regressions are available upon request.

in *t* as described in the first column of the table. A (white noise) matching-group-specific error corrects for the dependencies within matching groups. Only data are used where *i* had to make a decision about *j*.
Model V and VI include only lagged variables; Model VII and VIII include averages over previous periods.

The results show that proposers prefer responders who contributed in the previous period. Consistent with their exclusion decisions, they dislike subjects who punished more in the previous period and prefer subjects who contributed more. In the models with average variables, responder's average punishment weighs more heavily than her average contribution. Remarkably, average contributions of the responder influence neither their rank nor whether they are excluded, whereas their contribution in the previous period does. The variables controlling for the proposer's own contribution and punishment have no significant effect on ranking.

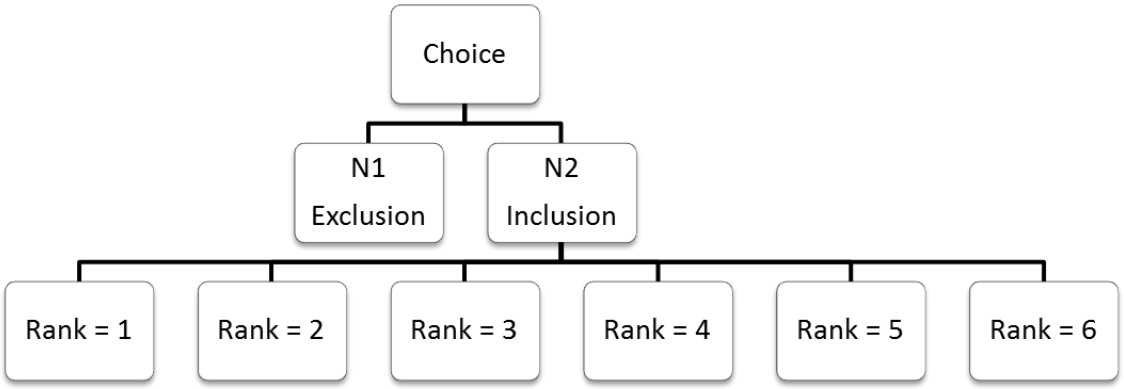
When we compare the above result to the determinants of exclusion (result 3a and 3b), it appears that exclusion and low ranks are spurred by similar circumstances. Both free riders and punishers are excluded and receive lower ranks. A participant's own contribution increases exclusion rates but not low ranking. Participants with a steady inclination to contribute (who contributed in previous rounds, too) may exclude co-players more readily and therefore end up in table 3.

Result 4: *Lower preferences are given to non-contributors and to punishers.*

The decision to rank a group member may be a two-stage decision, where a person first decides whether or not to exclude someone, and subsequently how to rank her. In such a nested model there are two nests (exclusion[39] or inclusion, N_1 and N_2) and six specifications in the second ranking level (rank = 1, .., 6) (see Figure 3.4).

[39] Exclusion is a degenerate nest, since it has only one alternative; so the conditional probability P (r=0|N=1) = 1.

Figure 3.4. Choice set



Notes. In the nested model there are two nests (exclusion or inclusion, N_1 and N_2) and six specifications in the second ranking level (rank = 1, .., 6).

The independence of irrelevant alternatives (IIA) characteristic of a multinomial logit model means that the probability to choose one alternative is independent of the nature of the alternatives; in this case, that the choice to exclude someone is made simultaneously with the choice to attribute a certain rank to someone.

The "inclusive value" I_{ij} (see Schram 1990) of person *i* ranking person *j* can be defined as

$$I_{ij} = \log\left\{\sum_{k=1}^{M} \exp(X'_i\beta_k)\right\}$$

with *X'* being the parameter vector and β_k the coefficients per rank. The probability to give *any* rank at all to a person (that is, the chance to not exclude a person) is then

$$Pr_{ij} = 1/[1 + \exp(X'_i\beta_0 - I_i)]$$

If we relax the IIA and allow for a different coefficient for the inclusive values we obtain the more general framework

$$Pr_{it} = 1/[1 + \exp(X'_i\beta_0 - (1 - \sigma)I_i]$$

If σ differs significantly from 0, the simultanous decision model can be rejected in favour of the sequential model and the IIA does not hold for the decision to exclude versus rank a person. If σ differs significantly from 1, we can reject the sequential model.

To derive the sigma, we estimated the individual coefficients of the ranks (without exclusion) by means of a multinomial logit. In the baseline treatment the decision to rank was hypothesized to depend on the previous contribution of the other and the period (based on Table 3.4). In the punishment treatment the decision to exclude and the ranking decision were explained by the punishment of the other in the previous round. With these coefficients we determine the I_i; with these inclusive values included as a parameter, we then estimate the chance of inclusion (or ranking) given all decisions (see Appendix B for the results). In the punishment treatment, we can reject the hypothesis that exclusion and ranking are a sequential decision (Wald X^2=6.95, $p<0.01$); but we cannot reject the hypothesis that exclusion and ranking are a simultaneous decisions (Wald X^2 = 2.71; $p=0.10$). In the baseline treatment we have to reject the sequential model (Wald X^2 = 8.94, $p<0.01$) and the simultaneous model (Wald X^2 =7.33, $p<0.01$).

Result 5: *The decision to include and rank someone is not a sequential decision in the punishment treatment.*

3.4.4 Comparing enforcement mechanisms

In part II of the treatment with punishment, proposers can choose between various mechanisms to enforce cooperation by others: punishment, ranking, or exclusion. Responders can choose between punishment and exclusion. Note that exclusion may be seen as a specific kind of punishment, since it denies the excluded participant future benefits from cooperation. Like punishing, exclusion is costly for the executer, because it increases the probability that she will not be included in any group. In this section, we compare the two enforcement mechanisms exclusion and punishment in terms of their effect on individual earnings and group efficiency. Obviously, we can only do so for the treatment with punishment.

Having established that both mechanisms are used in reaction to low contributions, we first check whether subjects exhibit distinct preferences for either. Consider reactions to a potential partner who played, but did not contribute in the previous round. For such a non-contributor, 80 % of the subjects decided to exclude at least once, but never punish and 20 % punished at least once, but never excluded such a non-contributor. 15 % neither punished nor excluded the defector and only 15 % of the subjects employed each option at least once[40].

Result 6: *Participants exhibit a preference for either punishing or excluding; many prefer exclusion to punishment.*

Ex ante it is unclear which of the two mechanisms yields higher earnings to the individual applying it. Obviously, the costs related to excluding (stemming from a higher chance of remaining groupless in future rounds) decrease as the number of remaining periods declines, but this may also be the case for the hidden costs of punishing. Subjects who excluded at least one fellow player in the first round[41] earned on average less than subjects who did not (239 vs 254 francs, MW z =2.45, p=0.01), which suggests that not playing may be costly. Punishing in the first round did not significantly decrease earnings in the long run (248 (first round punishers) vs 254 francs, MW z=0.19, p=0.85)). Considering all rounds, subjects who excluded at least one fellow player who did not contribute in the previous round earned on average the same as subjects who excluded no free riders (252 francs). Subjects who punished at least one free rider earned on average the same as subjects who punished no free riders (248 vs 254 francs, MW, z=0.02, p=0.98).

Result 7: *Excluding someone in the first round has hidden costs, punishing does not.*

Next, we consider the efficiency of the choices made in the various treatments. Maximum efficiency is achieved in a period if all participants are in groups and every member contributes her endowment to the public good. Earnings are then 270 francs per matching

[40]Note that the categories show overlap.
[41] Subjects who excluded another subject in the first round had no information to base their decision on.

group of nine, *i.e.*, 30 francs per person.[42] In the treatment without punishment, lowest possible welfare is obtained if no participant contributes, irrespective of whether they are in a group. Everyone then earns 20 francs (*i.e.*, 10 less than with maximum efficiency, and 200 francs over 10 periods). Since not ending up in a group is the default option, we take this as reference point for the efficiency calculations. For any given level of contributions, punishment decreases welfare and may lead to negative efficiency in a period.

We therefore measure efficiency in the 10 period game is as $\left(\sum_{t=1}^{10}\pi_{i,t} - 200\right)$ / (270–200), with payoff defined after subtracting punishment points.

The observed efficiency in part I of our experiment is 0.16 without punishment and 0.26 with punishment. This difference is statistically insignificant (MW, p = 0.29). In part II, efficiency was higher, to wit, 0.71 without and 0.77 with punishment. This difference is statistically insignificant (MW, p = 0.20). We conclude that there is no evidence of an efficiency enhancing effect of punishment (in fact, evidence on exogenously formed groups shows that punishment in previous studies usually reduces efficiency in the short run[43]). The difference between Part I and Part II is marginally significant without punishment (Wilcoxon signed rank test, p = 0.07) and significantly so with punishment (p = 0.03). Hence, the endogenous formation of groups increases efficiency levels when combined with punishment, even when considering the welfare loss due to punishment per se. Finally, the efficiency observed in part I with punishment is significantly lower than in part II without (MW, p = 0.01). This is an indication that exclusion and voluntary group formation are, together, a more efficient way to support cooperation in social dilemmas than punishment in isolation.

Result 8. *Endogenous group formation increases efficiency.*

Finally, we observed in Figure 3.2 that the option to punish is used much less often in Part II, when there is also an opportunity to exclude and express preferences for group formation. This large difference in exerted punishment (153 vs 32 punishment points in total, or 14% vs 3% of all opportunities to punish someone) is an indication that 80 % of these acts of punishment are

[42] Each matching group consists of 3 groups. In each group the public good yields at most 90 francs to be split amongst the members.
[43] See also Gächter et al. (2008) for other papers on efficiency-reducing punishment.

replaced by exclusion in Part II (in 103 cases, or 6 % of all 1560 opportunities a participant was excluded). It seems that at least some of the subjects treat exclusion as a substitute for punishment; however, the efficiency results at the group level show that endogenous group formation and punishment are most effective in combination.

Result 9. *Endogenous group formation is used as a substitute for punishment, but works as a complement at the group efficiency level.*

3.5 Discussion

In this chapter we have presented experimental results on the interplay between endogenous grouping and punishment in a public goods game. Participants ranked and excluded their group members based on behavior in previous rounds, including previous contributions to the public good and punishment of other group members. Standard theory predicts that group formation is irrelevant, since no rational player contributes to the public good and thus everyone is indifferent about being in a group or not.

Our results show that (i) endogenous grouping increases contributions more than the punishment option does. The combination of the two mechanisms increases contributions even further. Subjects use the possibility to rank and exclude others and therefore punish less. Exclusion and low ranking are directed towards non-contributors as well as to punishers and we show that the decision whether or not to exclude someone precedes the allocation of a rank. (ii) Furthermore, we observe that efficiency increases significantly only when endogenous grouping is combined with punishment. (iii) Under the endogenous grouping regime, higher contributions are achieved with less punishment.

The results support the proposition that heterogeneity in types and the threat of exclusion motivate subjects to signal their (good) type through contributions (Ones and Putterman, 2007; Van Vugt and Hardy, 2010). The readiness with which exclusion is used (in spite of the fact that for any belief about types, exclusion is dominated by allocating a low rank) shows that this fear is justified. In early periods, punishment could be used as a signal of being an altruistic punisher, but the upward trend in punishment contradicts this explanation. Given that many subjects contribute and punish even in the last period, their types seem to be intrinsic instead of strategic[44].

The effect of punishment depends to a large extent on the composition of groups. Although the addition of punishment to endogenous grouping increases the efficiency at the group level,

[44] See also Fudenberg and Pathak (2010) on unobserved punishment.

individual punishers are less preferred as partners[45]. The positive influence of punishment at the group level therefore is not reflected by higher attractiveness of individual punishers during the group formation phase. Free riders could be expected to assign lower ranks to punishers than to contributors; but among contributors this aversion for punishers can only be attributed to fear of antisocial punishment[46]. Irrespective of the reason for the aversion however, if punishers expected their attractiveness as a partner (for contributors) to decrease by inflicting punishment, they may have refrained from punishment and turned to the seemingly cheaper alternative: the exclusion of free riders. Whether this effect would be strong enough to eliminate punishment in the long run – and thereby the competitive advantage of punishment and the preference for it- remains to be assessed[47]. One way to tease out the effects of punishment types on group formation (and vice versa, the effect of group formation on punishment types) would be by explicitly introducing heterogeneity in the punishment costs or inflictions.

Altogether the lower punishment rates account for a large part of the higher efficiency of the combination of the two regimes. Having a say about one's interaction partner reduces the need for punishment, but preserves the threat. This complementary effect of punishment and endogenous grouping may well account for the level of cooperation we observe in real institutions.

[45] See Rockenbach and Milinski (2011) who, on the contrary, find no effect of punishment in assessing eligibility for later periods.
[46] See Herrmann et al., 2008.
[47] Gürerk et al. (2006) show that when given the choice between a regime with or without punishment, participants ultimately migrate to the former.