



Working Papers

www.cesifo.org/wp

Cooperation and Punishment: The Individual-Level Perspective

Felix Albrecht
Sebastian Kube
Christian Traxler

CESIFO WORKING PAPER NO. 6284
CATEGORY 13: BEHAVIOURAL ECONOMICS
DECEMBER 2016

An electronic version of the paper may be downloaded

- *from the SSRN website:* www.SSRN.com
- *from the RePEc website:* www.RePEc.org
- *from the CESifo website:* www.CESifo-group.org/wp

ISSN 2364-1428

Cooperation and Punishment: The Individual-Level Perspective

Abstract

We explore the relationship between individuals' disposition to cooperate and their inclination to engage in peer punishment as well as their relative importance for mitigating social dilemmas. Using a novel strategy-method approach we identify *individual* punishment patterns and link them with *individual* cooperation patterns. Classifying $N = 628$ subjects along these two dimensions documents that cooperation and punishment patterns are intuitively aligned for most individuals. However, the data also reveal a sizable share of free-riders that punish pro-socially and conditional cooperators that do not engage in punishment. Analyzing the interplay between types in an additional experiment, we show that pro-social punishers are more crucial for achieving cooperation than conditional cooperators. Incorporating information on punishment types explains large amounts of the between and within group variation in cooperation.

JEL-Codes: C900, D030.

Keywords: strategy method, punishment patterns, type classification, conditional cooperation, public-goods game.

Felix Albrecht
Universities of Marburg & Bonn
Marburg / Bonn / Germany
felix.albrecht.uni@gmail.com

Sebastian Kube
University of Bonn & Max Planck Institute
for Research on Collective Goods
Bonn / Germany
kube@coll.mpg.de

Christian Traxler
Max Planck Institute for Research on Collective Goods
& Hertie School of Governance
Bonn / Berlin / Germany
traxler@hertie-school.org

We would like to thank Nick Bardsley, Christoph Engel, Christian Thöni, and numerous participants at seminars and workshops in Berlin (BBE), Bonn (MPI and University), Florence (2013 Workshop), Lausanne, Marburg, at FSU, Waseda University, the 2013 Public Economic Theory Meeting (Lisbon) and the 2014 North-American ESA meeting (Fort Lauderdale) for helpful comments and suggestions. Financial support from DFG (Deutsche Forschungsgemeinschaft; grant 50130225) is gratefully acknowledged. The usual disclaimer applies.

1 Introduction

An extensive body of research documents that humans commonly fail to solve cooperation problems (e.g., Yamagishi, 1988; Andreoni, 1988; Ledyard, 1994; Fischbacher and Gächter, 2010; Balliet et al., 2011; Chaudhuri, 2011, to name only a few). Societies nevertheless manage to escape the ‘tragedy of the commons’ by using appropriate institutional mechanisms (Ostrom et al., 1992; Kosfeld et al., 2009). One such mechanism is peer punishment, which makes successful cooperation much more likely to occur (e.g., Fehr and Gächter, 2000, 2002; Carpenter, 2007; Reuben and Riedl, 2013). Many groups, however, fail to use punishment in an effective and pro-social manner — which could be due to the fact that peer punishment constitutes a cooperation problem in itself (e.g., Yamagishi, 1986). A breakdown in cooperation that coincides with a failure of peer punishment might thus capture the same phenomenon. This conjecture raises two fundamental questions that we try to answer in this paper: Firstly, what is the relation between an *individual’s* disposition to cooperate (Fischbacher et al., 2001; Fischbacher and Gächter, 2010) and her *individual* inclination to engage in peer punishment? Secondly, if these two individual traits do not coincide, is one more important than the other for mitigating the basic cooperation problem?

We study these questions employing a classical workhorse in the literature on cooperation and punishment: a linear public-goods game (VCM) with decentralized peer punishment (Fehr and Gächter, 2002). Subjects first make a contribution decision and can then assign costly punishment points that reduce the other group members’ payoffs. Within this prominent paradigm, we introduce a novel variant of the strategy-method at the punishment stage of the game that allows identifying heterogeneity in peer punishment at the *individual* level.

When making her punishment decisions, each subject is confronted with a random sequence of ‘scenarios’, i.e., combinations of others’ contributions. One of these scenarios corresponds to the other group members’ actual contribution decisions. All other scenarios are randomly drawn contributions that systematically cover relevant parts of the strategy space. Only the punishment decisions for the scenario with the actual contributions become payoff-relevant. As subjects do not know which scenario is the ‘relevant’ one, we have an incentive compatible strategy-method that induces *exogenous* variation in others’ contributions to consistently estimate individual peer punishment patterns in a one-shot game (see Bardsley, 2000, for a related approach eliciting cooperation patterns).

Using this strategy-method to elicit punishment patterns reveals substantial heterogeneity between individuals. In our sample with $N = 628$ experimental participants two patterns dominate:

Almost every second subject (47.1%) is classified as a *pro-social punisher*. Their individual punishment patterns are all significantly decreasing in the other's contribution, i.e., they target their punishment towards those contributing nothing or little to the public good. The second-largest group (40.3%) are *non-punishers* ('second-stage free-riders'), i.e., subjects that do not at all engage in peer punishment. Beyond these two dominant types, there is only a small fraction of subjects that displays either an unsystematic pattern or a pattern that is increasing in the other's contribution (often termed 'anti-social punishment'; see, e.g., Herrmann et al., 2008).

Linking these individual punishment patterns to the corresponding individual dispositions to cooperate — that we obtain from a within-subject design using the measure of conditional cooperation introduced in Fischbacher et al. (2001) — yields a two-dimensional classification that reveals two interesting behavioral archetypes. (i) For the majority of our subjects cooperation and punishment types are aligned in an intuitive way: we find that 55% of conditional cooperators punish pro-socially and that 56% of free-riders are non-punishers. (ii) Strikingly, this also implies that a significant share of subjects have individual punishment- and cooperation-patterns that are diverging: 35% of conditional cooperators are non-punishers and 32% of free-riders do engage in pro-social punishment.

The ability to identify these two behavioral archetypes — individuals whose cooperation and punishment patterns are either aligned or non-aligned — is a major benefit of combining our novel approach to classify punishment patterns at the individual level with the conditional cooperation-measure from Fischbacher et al. (2001). Moreover, as the individuals' inclinations to cooperate and to punish are far from being perfectly correlated, we can assess their respective importance for mitigating a social dilemma. To do so, we use these individual type-classifications from two one-shot games to explain group outcomes in a third game: a finitely repeated public-goods game with peer punishment — both among stable groups where players interact repeatedly (partner design) and among steadily alternating groups where a group's type composition changes over time (stranger design).

In both conditions, we observe that groups with more conditional cooperators achieve higher average contributions, that are also more stable over time, than groups with fewer conditional cooperators. While these observations mirror previous findings (e.g., Gächter and Thöni, 2005), we also obtain a similar picture with respect to the group members' *punishment* types. In fact, variation in punishers' types seems to be crucial: keeping constant the fraction of conditional cooperators, average contributions are significantly higher in groups that contain more pro-social punishers. The presence of pro-social punishers induces higher contributions among subjects classified as free-riders and among conditional cooperators. In fact, regression analyses of the

repeated-game data suggest that information about the number of pro-social punishers in a group is more powerful in explaining the heterogeneity in cooperation levels between groups than information on the number of conditional cooperators.

These findings underline that group outcomes crucially depend on the presence of pro-social punishment types. To the best of our knowledge, our paper is the first to present strong causal evidence on this link. As such, it is complemented by two recent papers that have hinted at the importance of the individual inclination to punish. Rustagi et al. (2010) find a positive correlation between natural groups' success in managing forest commons and the number of conditional cooperators in the respective groups. They attribute this to the difference between conditional cooperators and selfish persons in their self-reported statements about time spent on forest patrols.¹ In a similar vein, the correlational analyses by Kosfeld and Rustagi (2015) suggest that these natural groups are also better at managing forest commons if the corresponding leader's third-party punishment behavior, as measured in a lab experiment, promotes equality and efficiency rather than being arbitrary.

Rustagi et al. (2010) and Kosfeld and Rustagi (2015) focus either on cooperation or on sanctioning patterns. By contrast, Falk et al. (2005) study both individual punishment and cooperation behavior, but without exploring the relative impact of subjects' types on mitigating a social dilemma. They employ a strategy method on the peer punishment-stage of a binary prisoner's dilemma-game between three persons, and relate the punishment pattern to the subject's actual cooperation decision in the prisoner's dilemma. While the fraction of people who cooperate and punish is similar to what we find, it differs for those who defect and punish. To some extent, this is driven by the marked amount of anti-social punishment in their data. In parts, though, this might also be due to the fact that they use the actual decision (cooperate or defect) rather than eliciting cooperation types via a strategy method. After all, a defector might either be a selfish individual or a conditional cooperator that expects the other person to defect. Our two-dimensional type classification suggests that this distinction makes a difference for pinning down the linkage between cooperation and punishment patterns.

The classification of individuals along two dimensions yields also interesting results on how the interplay of different behavioral types drives group outcomes. Accounting for the heterogeneity in punishment types significantly improves our ability to explain the large and persistent differences in cooperation across groups. Moreover, the identification of systematically different punishment patterns at the *individual level* provides a novel contribution to the literature which has mainly

¹The authors conclude that [...] "better forest management outcomes are not only a result of conditional cooperators being more likely to abide by the local rules of the group but also being more willing to enforce these rules at a personal cost" (p.964). The systematic causal evidence provided in this paper confirms this line of reasoning.

focussed on variation in punishment and cooperation patterns at the *aggregate* level.² Our analysis complements these studies of group-level heterogeneity and thus constitutes a potential micro-foundation that might prove useful for future studies, too.

Corresponding studies must not necessarily be located exclusively in the area of decentralized peer punishment. Knowledge about individuals' (punishment) types might help to better explain the effectiveness of other institutional arrangements aimed at sustaining cooperation (see, e.g., our work on centralized punishment in Kube and Traxler, 2011). Certain aspects of institutions could appeal to different cooperation and punishment types (e.g., Brekke et al., 2011), which would in turn inform the design of institutions that are successful in fostering cooperation. In this respect, our paper is also related to the growing literature on the endogenous adaption and implementation of institutions (e.g., via elections as in Kosfeld et al., 2009, Hamman et al., 2011, or Kube et al., 2015, or via voting by feet as in Gürer et al., 2006), since heterogeneity in the support for an institution might be traced back to underlying differences in individuals' cooperation and punishment types. Information about a population's type composition might allow to anticipate the support for an institution for a given population, which might also help to solve optimal group formation (e.g., Robbett, 2015) and optimal team composition problems (e.g., Burlando and Guala, 2005; Gächter and Thöni, 2005).

Finally, note that our elicitation method could also be used to advance research in other areas. For example, the impact of non-cognitive skills on life outcomes has attracted growing attention among economists (e.g., Heckman and Rubinstein, 2001; Cunha and Heckman, 2007). Given the notion of non-cognitive skills (which includes not only personality factors but also items like motivation, socio-emotional regulation, time- and social-preferences), an individual's punishment type might be another interesting and important facet of an individual to add to this list. Another example would be the literature in law and economics that explores factors which affect the effectiveness of law-enforcement, of which individuals' inclinations to engage in peer punishment might be an important aspect to consider, too (compare the discussion in Falk et al., 2005, and references therein). Another instance where punishment types might matter is the domain of relational contracts (e.g., Baker et al., 2002). While the stability of relational contracts is usually based on concepts of trust and trustworthiness or related to the threat of terminating relationships, the punishment types of the individuals involved might be another important component to focus on (e.g., Chassang, 2010).

²Consider, for instance, Herrmann et al. (2008), who compare behavior in public-good games with peer punishment across 16 countries, or Henrich et al. (2006), who study third-party punishment in 15 diverse populations and observe at the *aggregate* level that "costly punishment positively covaries with altruistic behavior across populations" (p.1767).

The remainder of this paper is structured as follows. The next section introduces the design and explains the implementation of the experiment. Section 3 discusses our approach to classify punishment patterns and presents the results. Section 4 links punishment types to contribution types and reveals the existence of different behavioral archetypes. Section 5 shows how the presence of these individual types influence group level outcomes as well as other individuals' behavior in a repeated game. Section 6 concludes with a discussion of our findings and suggestions for follow-up studies.

2 Design and Procedures

Our experiment consists of three independent games: (1) a one-shot public-goods game without punishment (*C-game*), which allows us to identify individual cooperation patterns in the tradition of Fischbacher et al. (2001); (2) a one-shot public-goods game with peer punishment (*P-game*) that uses a strategy method at the punishment stage to elicit individual peer punishment patterns; and finally (3) a 10-period public-goods game with peer punishment (*R-game*). In the latter, random assignment produces heterogenous group compositions of cooperation and punishment types, as elicited from the *C-game* and *P-game*. We exploit this heterogeneity to analyze the interplay between the different types in the *R-game* and the impact on groups' abilities to overcome social dilemmas.

2.1 C-Game

The C-game is a standard one-shot linear public-goods game (VCM) with the strategy-method from Fischbacher et al. (2001). Subjects are randomly assigned into groups of four. Each subject $i \in \{1, \dots, 4\}$ is endowed with 20 tokens and decides how many tokens to contribute to the public good, g_i , and how many to keep for herself, $20 - g_i$. Each token allocated to the public good yields a marginal per capita return of 0.4. The payoff function is given by

$$\pi_i^C = 20 - g_i + 0.4 \sum_{j=1}^4 g_j. \quad (1)$$

Under the assumptions of rational payoff-maximizing behavior, contributing zero is the dominant strategy of the one-shot game. In contrast, the social optimum consists of all players contributing their entire endowment to the public good.

Following the procedure of Fischbacher et al. (2001), subjects are first asked to make an unconditional contribution decision, g_i . Using the strategy-method, subjects then make their

conditional contribution decisions. They have to indicate their contribution for all 21 possible whole numbers of average contributions among the other group members, $\bar{g}_j := \frac{1}{3} \sum_{j \neq i} g_j$, with $\bar{g}_j \in \{0, 1, \dots, 20\}$. After all decisions are made, one group member is randomly drawn. For this subject, the conditional contribution decision is implemented based on the average unconditional contributions of the other three group members. Contributions and payoffs are revealed to the subjects only at the end of the experiment.

2.2 P-Game

The P-game is a one-shot linear public-goods game with costly punishment (Fehr and Gächter, 2000, 2002). At the first stage of the game, subjects make their contribution decision, facing the same parameters as described above for the C-game. At the second stage of the P-game, each subject i can assign punishment points to the other group members $j \neq i$, $d_{ij} \geq 0$. Punishment is costly. Assigning one punishment point costs one token for the punisher and reduces the payoff of the punished subject by three tokens (Fehr and Gächter, 2002; Herrmann et al., 2008). The payoff function is

$$\pi_i^P = \underbrace{20 - g_i + 0.4 \sum_{j=1}^4 g_j}_{\text{VCM}} - \underbrace{1 \sum_{j \neq i} d_{ij}}_{\text{Pun. given}} - \underbrace{3 \sum_{j \neq i} d_{ji}}_{\text{Pun. received}} . \quad (2)$$

A fully rational, selfish agent would not engage in any punishment at the second stage of the game. Hence, contributing zero would be again the dominant strategy.

While Fehr and Gächter (2000, 2002) and the subsequent literature let subjects decide on the punishment levels for others' actual contributions, we implement a novel strategy method at the punishment stage.³ The strategy method confronts subjects with a sequence of contribution triples: each subject i faces 11 screens, where each screen s presents one triple $\{g_j^s, g_k^s, g_l^s\}$, with $j \neq k \neq l \neq i$ and $s \in \{1, \dots, 11\}$. One of the 11 triples comprises the actual contributions of the other group members. The other ten triples are hypothetical combinations of contributions, each being randomly drawn from a pre-defined set of combinations (see below) and presented in randomized order. For each triple, a subject has to decide how many punishment points (if any) to allocate to the other subjects.

As we aim at identifying punishment patterns at the individual level, we wanted to assure that subjects face combinations of contributions that cover different parts of the vast strategy

³This strategy method was first used in Kube and Traxler (2011). It can be seen as an instance of the 'Conditional Information Lottery' introduced by Bardsley (2000), who used it at the contribution stage of the game.

space (up to 21^3 potential triples). To do so, we partitioned contributions into three intervals: *low* (L), *intermediate* (M), and *high* (H) contributions with $g^L \in \{0, \dots, 4\}$, $g^M \in \{5, \dots, 15\}$, $g^H \in \{16, \dots, 20\}$. We then considered the ten resulting combinations of low, intermediate and high contributions:

$$\begin{array}{ccccc} \{g^L, g^L, g^L\} & \{g^L, g^L, g^M\} & \{g^L, g^L, g^H\} & \{g^L, g^M, g^M\} & \{g^L, g^M, g^H\} \\ \{g^L, g^H, g^H\} & \{g^M, g^M, g^M\} & \{g^M, g^M, g^H\} & \{g^M, g^H, g^H\} & \{g^H, g^H, g^H\} \end{array}$$

Within each of the ten contribution combinations, we randomly generated eight different triples (see Appendix A1 for further details). For all 10 contribution combinations, a subject would then face one of these triples.⁴ Following this protocol, we observe 3×11 punishment decisions for each subject.

It is common knowledge that ten out of the 11 triples are hypothetical and that only the punishment decisions for the real contribution triple become payoff relevant. However, subjects neither know which one is the ‘real’ triple,⁵ nor do they know the procedure to generate the hypothetical triples. Only at the end of the experiment, the actual contribution triple and punishment choices are revealed.

2.3 R-Game

The R-game is a public-goods game with costly peer punishment (Fehr and Gächter, 2000) that is played repeatedly for ten periods. The payoff function is equivalent to the one from the P-game, summarized in equation (2). Subjects play the R-game either under a *stranger* (R_s) or under a *partner* protocol (R_p). At the beginning of the R-game, players are randomly assigned into groups of four (partner protocol, with partners not identifiable between periods) or matching-groups of eight (stranger protocol) and remain in these groups for all 10 periods. In the stranger protocol, subjects are randomly re-matched each period within their matching-group.

2.4 Implementation

We evaluate data for 628 subjects that participated in 29 sessions. The large sample allows us to study the role of heterogenous group compositions for group outcomes (see Section 5). For each subject we observe 21 conditional contribution decisions in the C-game, 3×11 punishment decisions in the P-game as well as 10 contribution and 30 punishment decisions in the R-game.

⁴One subject might see, for instance, $\{0, 0, 0\}$ for the combination $\{g^L, g^L, g^L\}$ and $\{0, 2, 8\}$ for $\{g^L, g^L, g^M\}$, etc. A different subject might face $\{0, 2, 3\}$ for the former and $\{0, 2, 14\}$ for the latter.

⁵Testing whether subjects punish the (unknown) real versus the hypothetical contributions differently, we find no significant differences whatsoever.

452 subjects played the R-game under a partner protocol, 176 subjects under a stranger protocol. The experiments were conducted at the University of Bonn’s *BonnEconLab*, using the experimental software *zTree* (Fischbacher, 2007). Subjects were recruited online using Orsee (Greiner, 2004). Standard experimental procedures were followed.⁶ Results and payoffs from the C- and the P-game were only revealed at the end of the experiment. Results and payoffs from the R-game were revealed after each period. Including a follow-up questionnaire, a session lasted approximately 100 minutes. On average, subjects earned 19.88 Euro, including a 5 Euro show-up fee.

3 Individual Peer-Punishment Patterns

To classify individual peer-punishment patterns, we model punishment d_{ij} as a linear function of player j ’s contribution to the public good (with $j \neq i$):

$$d_{ij} = \alpha_i + \beta_i(20 - g_j) + \varepsilon_i. \quad (3)$$

The main regressor in equation (3), $20 - g_j$, is j ’s deviation from contributing the full endowment (20 tokens). This linear transformation will facilitate the interpretation of the coefficients (see below).⁷ Using the data from the strategy method in the P-game (for the punishment decisions of the one-shot game), we *separately* estimate β_i (and α_i) for each of our 628 subjects. In this vein, we can identify individual-level heterogeneity in punishment patterns.

To see the advantage of our approach, it is important to realize that conventional observational data do not allow for a proper identification of the coefficient β_i at the individual level. In one-shot public good games with peer punishment, one would only observe three punishment choices per subject. In addition, one might argue that g_j is shaped by the expectations about d_{ij} . Similarly, in repeated games like our R-game, contributions shape punishment and punishment shapes contributions simultaneously.⁸ Our strategy method breaks this simultaneity by introducing truly exogenous variation in g_j . Following this line of reasoning, we focus on the subjects’ punishment choices for the 10×3 exogenous contribution triples of the P-game, i.e., we exclude the triple with the actual contributions, leaving us with 30 observations per subject.⁹

⁶The instructions and further details on the procedure are available in the Online Appendix.

⁷Estimating a model with $d_{ij} = \alpha'_i + \beta'_i g_j + \varepsilon'_i$ would yield equivalent estimates with $\hat{\beta}_i = -\hat{\beta}'_i$.

⁸Due to serial correlation in choices within subjects and (matching-)groups, one cannot easily avoid endogeneity problems (e.g., by using lagged values). In fact, our classification approach produces quite different results if we use the exogenous variation from our strategy method or the endogenous variation in the repeated game data (see Table S.7 in the Online Appendix).

⁹Our results are insensitive to including the three punishment decisions for the real contribution triple.

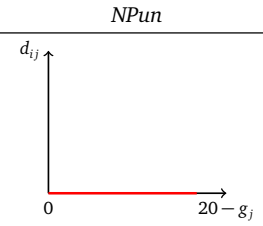
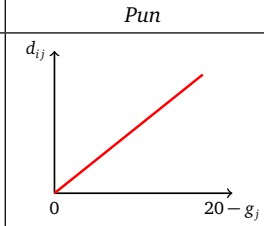
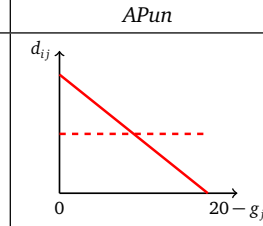
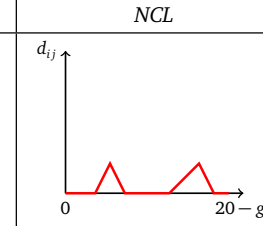
3.1 Punishment Types

Running 628 regressions with $N_i = 30$, we collect the estimates $\hat{\alpha}_i$ and $\hat{\beta}_i$ for each subject, along with robust standard errors. Based on these estimates, we then classify the subjects' punishment patterns. We distinguish subjects that do not punish, 'pro-social', and 'anti-social' punishers:

1. A subject is classified as a 'Non-Punisher' (*NPun*) if she assigns zero punishment points in all of the 30 punishment decisions, i.e., $d_{ij} = 0$ for all g_j . In equation (3), this is depicted by $\hat{\alpha}_i = \hat{\beta}_i = 0$.
2. Subjects that target their punishment towards those that contribute little or nothing to the public good have a punishment pattern that is upward sloping in $(20 - g_j)$. These subjects, with $\hat{\beta}_i > 0$ and $p \leq 0.01$, are classified as pro-social punishers (*Pun*).
3. Subjects are classified as anti-social punishers (*APun*), if their punishment is either increasing in the other's contribution g_j , i.e., if $\hat{\beta}_i < 0$ and $p \leq 0.01$, or if they display a significant positive but unsystematic level of punishment: $\hat{\alpha}_i > 0$ with $p \leq 0.01$ and an insignificant slope coefficient $\hat{\beta}_i$ with $p > 0.01$.¹⁰

Punishment patterns that cannot be assigned to one of these three types are summarized in a group of non-classified (*NCL*) patterns. The different punishment types and their punishment patterns are illustrated in Figure 1.

Figure 1: Stylized Illustration of Punishment Types

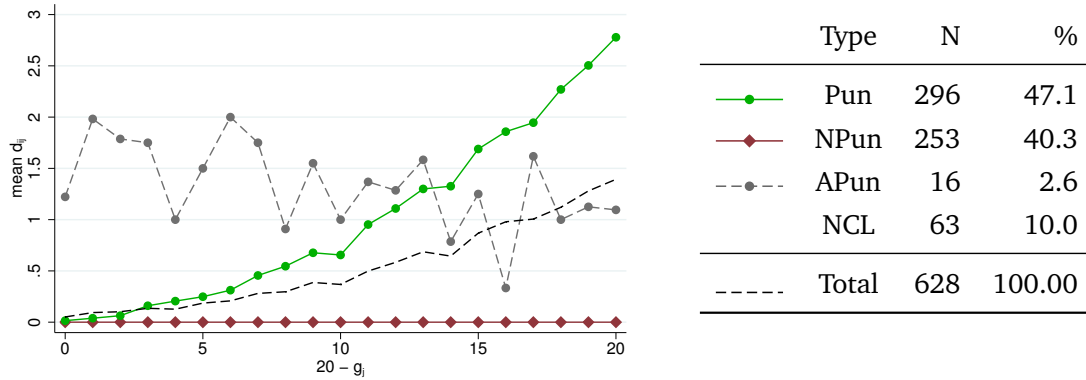
<i>NPun</i>	<i>Pun</i>	<i>APun</i>	<i>NCL</i>
			
$\hat{\alpha}_i = \hat{\beta}_i = 0$	$\hat{\beta}_i > 0$ with $p \leq 0.01$	$\hat{\beta}_i < 0$ with $p \leq 0.01$ or $\hat{\alpha}_i > 0$ ($p \leq 0.01$) & $\hat{\beta}_i$ insignif.	

The results from our classification approach are presented in Figure 2. 47.1% of our subjects are classified as *pro-social punishers*, 40.3% are *non-punishers*, 2.6% display an anti-social pattern,

¹⁰The literature typically defines anti-social punishment in reference to a subject's own contribution, i.e., if the punishment-receiving subject contributed a larger or equal amount to the public good compared to the punishing individual (e.g., Herrmann et al., 2008). Since our classification does not consider a punisher's own contribution g_i , it deviates from this self-centered notion of anti-social punishment. It nevertheless captures patterns of punishment that is targeted towards high contributors.

and 10.0% are in the residual group of non-classified patterns (*NCL*). Subjects from the latter group show very low levels of sporadic punishment (as illustrated in Figure 1). In fact, if we relax the strict definition of *NPun* to include also subjects with $\hat{\alpha}_i \approx \hat{\beta}_i \approx 0$, then every single *NCL* type would be re-classified as *NPun*. These (de-facto) non-punishers would then account for 50.3% of the sample.¹¹

Figure 2: Punishment Patterns and Punishment Types



Notes: Punishment type distribution and average punishment patterns (in the $20 - g_j$ -space) for the different types: pro-social punishers (*Pun*), non-punishers (*NPun*), anti-social punishers (*APun*), and non-classified punishment profiles (*NCL*). To ease illustration, the pattern for the latter is not plotted.

The results show that our sample is characterized by a high frequency of *Pun* types. The average punishment pattern, indicated by the dashed black line in Figure 2, is therefore clearly increasing in $20 - g_j$. Note further that the slope of the punishment pattern is relatively steep. The average [median] $\hat{\beta}_i$ among *Pun* types is 0.135 [0.124]. This suggests that a player j — who faces an average *Pun* type — receives around 0.14 punishment points for a one unit decline in her contribution g_j . If player j faces two [or even three] *Pun* types in her group, the marginal punishment increases to 0.28 [0.42] points. Given the parameters of the game (see equation 2) this translates into marginal costs of 0.84 [1.26] token — which weakly [strongly] dominates the marginal payoff gains from free-riding (0.6 token).

3.2 Robustness

How robust are our type classifications? Note first that the random variation in g_j induced by the strategy method renders the estimates of (3) fairly insensitive to adding further control variables

¹¹These results are documented in the Online Appendix, see Figure S.4.

(e.g., controls for contributions g_k and g_l , $k \neq l \neq j$).¹² Obviously, this does *not* imply that the simple equation from (3) is the ‘best model’ to describe individual punishment patterns. In a companion paper we explore richer models which capture, among others, self- and group-centered punishment (Albrecht and Traxler, 2016). There we show that (i) self-centered models outperform the simple model from equation (3) in terms of explanatory power (but not by much) and that (ii) the majority of the pro-social *Pun* types from above have a punishment pattern that is ‘kinked’ at the own contribution, g_i (see the Online Appendix). Studying more refined classifications of punishment patterns that can be derived from estimating more complex punishment equations is an interesting topic in itself, but it hardly illuminates the analysis that follows below. This paper therefore focuses on the simple type classification approach from above.

4 Cooperation Patterns and Two-Dimensional Classification

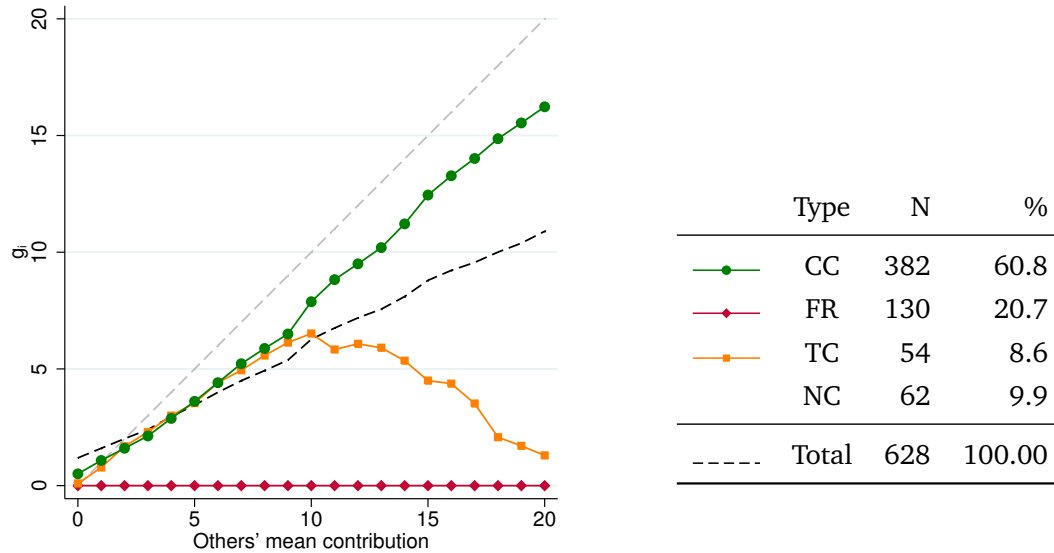
This section analyzes the strategy-method data from the C-game, where each subject states — conditional on all potential values for the others’ average contribution — how much to contribute to the public good (Fischbacher et al., 2001). Based on these data we will classify individual cooperation types. For each subject we will then derive a two-dimensional measure that links the individual punishment pattern with the individual cooperation type.

4.1 Cooperation Types

Consistent with our approach from above we separately estimate for each subject i the linear model $g_i = a_i + b_i \bar{g}_j + e_i$. Applying the type classification proposed by Fischbacher and Gächter (2010) we distinguish between *Conditional Cooperators* (CC, with $\hat{b}_i > 0$ at $p \leq 0.01$), *Free-Riders* (FR, with $g_i = 0$ for all \bar{g}_j , i.e., $\hat{a}_i = \hat{b}_i = 0$), *Triangular Contributors* (TC), and *Non-classified* (NC) cooperation patterns. Figure 3 presents the distribution of these types among the 628 subjects from our sample. The observed type distribution, as well as the cooperation patterns, are remarkably similar to those reported in Fischbacher et al. (2001) and Fischbacher and Gächter

¹²The classification outcome that builds on the estimates $\hat{\alpha}_i$ and $\hat{\beta}_{i1}$ from the equation $d_{ijt} = \alpha_i + \beta_{i1}g_j + \beta_{i2}g_k + \beta_{i3}g_l + \sum_t \gamma_{it}D_{it} + \varepsilon_i$, where D_{it} are order dummies that capture the sequence at which subject i faces the different triples, differs for a mere 18 subjects (2.9% of our sample). It is also worth noting that we obtain fairly similar type distributions if we use Spearman’s rank correlation to classify punishment patterns. This point is documented in Table S.3 in the Online Appendix. Results from further refinements of the type classification approach are available from the authors upon request.

Figure 3: Cooperation Patterns and Contribution Types



Notes: The figure presents the distribution of contribution types, following Fischbacher et al. (2001) and Fischbacher and Gächter (2010), and the average cooperation patterns for the different types: *Conditional Cooperators (CC)*, *Free-Riders (FR)*, *Triangular Contributors (TC)*, and *Non-classified (NC)* cooperation patterns. To ease illustration, the pattern for the latter is not plotted.

(2010): 61% are conditional cooperators and 21% are free-riders. The remaining 18% display a triangular or a non-systematic contribution pattern.¹³

4.2 Two-Dimensional Type Distribution

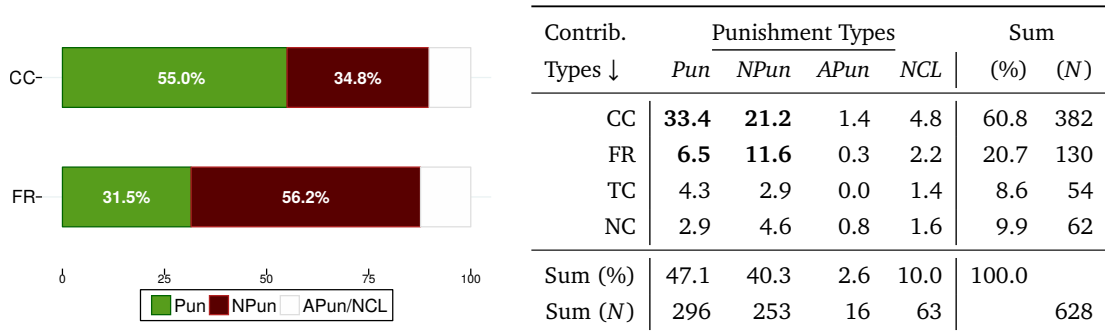
We now combine this latter classification with the classification of punishment types from the previous section to arrive at a two-dimensional type classification. In this vein, we can explore the relationship between individuals' disposition to cooperate and their inclination to punish their peers. The right panel in Figure 4 presents the results from the two-way classification.

The table reveals that, overall, a third of our sample (33.4%) are conditional cooperators with a pro-social peer punishment pattern ($CC \times Pun$). Almost 12% are free-riders in the C-game that do not punish in the P-game ($FR \times NPun$). In addition to these types with intuitively aligned patterns, we also observe a non-trivial fraction of subjects with diverging patterns: 21% of all subjects are conditional cooperators that do *not* punish at all ($CC \times NPun$) and more than 6% are free-riders with a pro-social punishment pattern ($FR \times Pun$).

¹³Classifications based on Spearman's rank correlation (as in Fischbacher et al., 2001) yield almost identical results. See the Online Appendix for details (Table S.5).

A different way of presenting the distribution of these four types — which cover almost three out of four subjects in our sample — is provided in the left panel of Figure 4. The bar graphs indicate that roughly every second conditional cooperator punishes pro-socially (55%) and that more than one out of two free-riders do not punish at all (56%). In addition to these types, whose cooperation and punishment patterns are aligned, there seems to be a second archetype of subjects with diverging patterns: every third (35%) conditional cooperator does not punish and, analogously, almost one in three (32%) free-riders punishes pro-socially.

Figure 4: Two-way Distribution: Contribution and Punishment Types



Notes: The left panel depicts the conditional frequency of *Pun* and *NPun* types among conditional cooperators (CC) and free-riders (FR), respectively. The table on the right shows the two-dimensional type distribution in our sample.

4.3 Supplementary Analyses

In a next step, we examine the distribution of the underlying coefficients of the type classifications (in particular, $\hat{\beta}_i$ and \hat{b}_i ; see Figure A.1 in the Online Appendix). The analysis reveals a positive correlation between \hat{b}_i and $\hat{\beta}_i$: ‘stronger’ conditional cooperators tend to have steeper punishment patterns. However, the correlation is far from perfect. Among *CC*×*Pun*-types, for instance, we observe an insignificant correlation coefficient of $\rho = 0.094$ ($p = 0.173$).¹⁴

We further studied whether individual characteristics, personality traits (big five, etc.) and attitudes (risk, trust, etc.) correlate with the contribution and punishment types (extensive margin variation) or patterns within types (intensive margin variation). Our analysis reveals three strong and robust predictors for the type assignments. First, using the questions from [Traxler and Winter \(2012\)](#) we find that subjects who express their willingness to impose social sanctions on norm violators (e.g., drunk drivers) among their peers are significantly more likely to be pro-social punishers (*Pun*-types). This observation suggests that the survey measure on norm enforcement

¹⁴The Spearman correlation is slightly stronger (0.128) and statistically significant ($p = 0.064$).

is consistent with the behavioral measure that builds on the observed pattern of peer punishment. Second, we find that subjects who report to be more reserved (see Rammstedt and John, 2007), are much more likely to be a *NPun*-type. Third, considering the different contribution types, we detect a strong gender effect: females have a much higher likelihood of being a conditional cooperator and, vice versa, a much lower probability of being a free-rider.¹⁵

To study intensive margin variation within types, we examined correlations of observables with the slopes of the subjects' contribution and punishment patterns ($\hat{\beta}_i$ and \hat{b}_i). Our analysis reveals that, among *Pun*-types, the slope of the punishment pattern is *lower* for females as well as for subjects with a high level of agreeableness in the big five (Rammstedt and John, 2007). For the cooperation patterns of *CC*-types, we find that those who express a high level of trust in others have a steeper contribution pattern: they are more likely to one-to-one match others' contributions.

Summing up, our two-dimensional classification reveals the existence of two interesting behavioral archetypes. First, for the majority of our subjects, there is a clear and intuitive overlap between cooperation and punishment types. This includes conditionally cooperative types that do invest in pro-social punishment (*CC* \times *Pun*) and free-riders that do not punish at all (*FR* \times *NPun*). Second, our analysis also identifies a significant share of individuals that are conditional cooperators which do not punish (*CC* \times *NPun*) as well as free-riders that are classified as pro-social punishers (*FR* \times *Pun*).

The identification of this second archetype, whose cooperation and punishment patterns are diverging, is interesting in itself.¹⁶ The finding further implies that individual inclinations to cooperate and to punish are far from being perfectly correlated in our sample. We can thus assess the interplay between the different types and their role for explaining outcomes in another independent situation: the R-game.

¹⁵Probit and LPM estimates underlying these results are available from the authors.

¹⁶One aspect that is beyond the scope of the present paper is the explanation of this second archetype based on existing theories of other-regarding preferences. Self-evident models to structure our data are based on theories of inequality aversion, in particular Fehr and Schmidt (1999) (F/S). (Obviously, we do not estimate coefficients from self-centered models of punishment. As pointed out in Section 3.2, the overall picture from our type classifications hardly changes for these more complex models.) Intuitively speaking, in F/S the decision to contribute is shaped by the aversion against advantageous inequality (i.e., the parameter β in F/S), whereas pro-social punishment is motivated by aversion against disadvantageous inequality (i.e., the parameter α). As such, F/S can easily accommodate the 'aligned' type combinations *CC* \times *Pun* (high α and β) and *FR* \times *NPun* (low α and β). Given the specific parameters of our experiment (4 players, MPCR of 0.4, and punishment technology of 1:3), also the less intuitive *CC* \times *NPun*-type is consistent with F/S-subjects with a sufficiently strong aversion against advantageous inequality but only a mild aversion against disadvantageous inequality. Yet, using F/S to explain the combination of free-riders that punish others with low-contributions (*FR* \times *Pun*) is not that straightforward and would require assumption regarding (players' expectations about) the distribution of the parameters α and β in the population.

5 Group Composition and Contributions in the Repeated Game

In this section we demonstrate the benefits from identifying heterogeneous punishment types for explaining group and individual level heterogeneity in repeated public goods games. To this end, we exploit the data from the 10 periods of the R_p - and the R_s -game (partner and stranger design, respectively).¹⁷ We analyze the influence of group compositions on group outcomes and individual behavior. Motivated by the results from Gächter and Thöni (2005), who document that grouping pro-social individuals leads to higher payoffs in a repeated VCM without punishment, we start out by computing the number of conditional cooperators (*CC*) and pro-social punishers (*Pun*) for each group (and for each matching group in the stranger design). Making use of the random assignment of subjects into groups — a point which is discussed in detail in the Appendix (see Table A.1) — we first evaluate the impact of having more or less *CC*- or *Pun*-types on a group's average contribution level.

5.1 Descriptive Evidence

A first glimpse at the results is provided by Figure 5. It depicts the average contribution per group over 10 rounds for different group compositions.¹⁸ Panel A [C] compares contributions for [matching-] groups with different numbers of *CC*-types. The figure shows a strong positive relationship between the number of *CC*-types and the average contribution level — an observation that is fully in line with the results from Gächter and Thöni (2005).

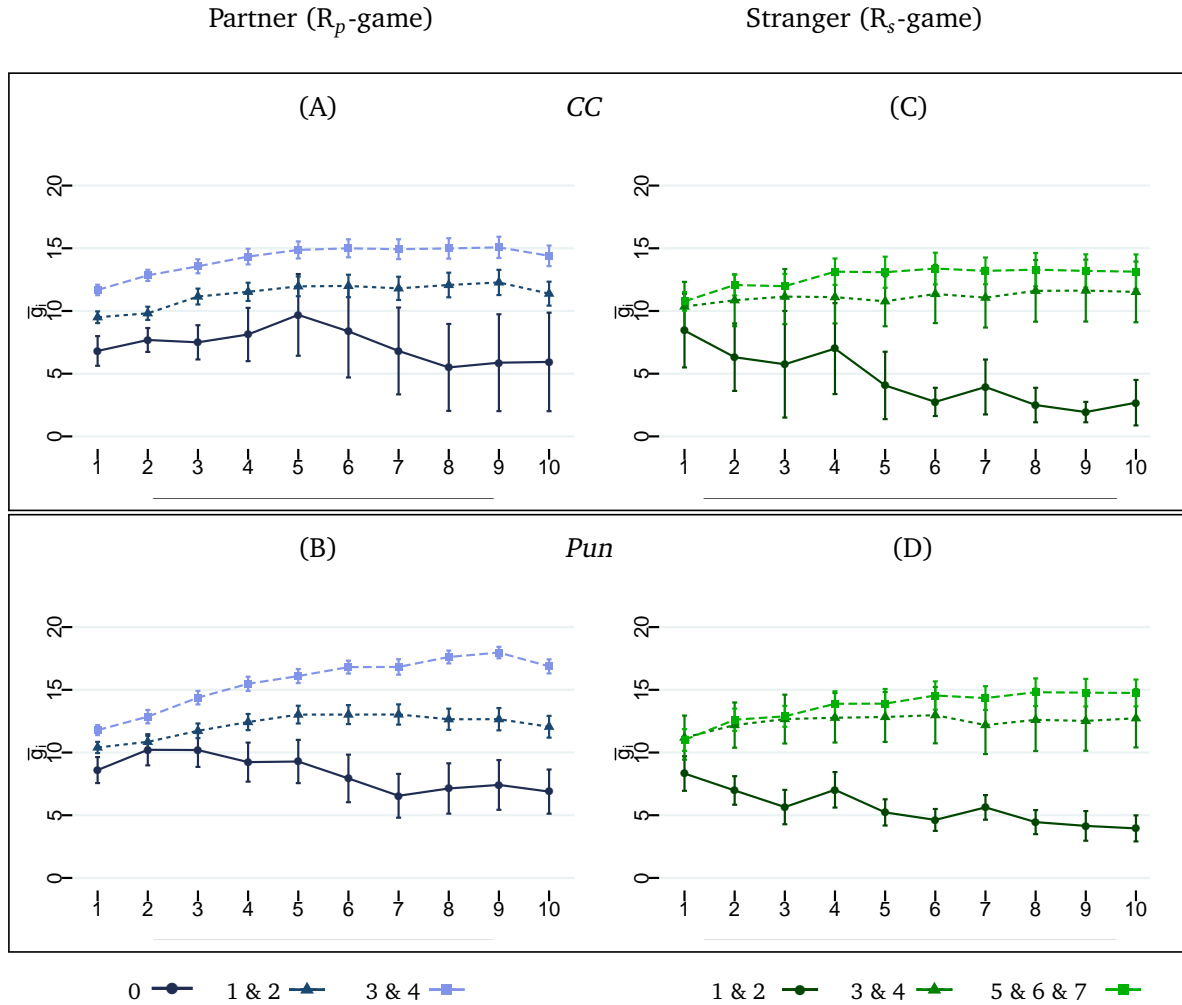
Panel B [and D] compares [matching-] groups with different numbers of *Pun*-types. Similar as above, we observe that contributions are higher in groups that contain more pro-social punishers. However, the standard errors are now smaller and, what is more important, average contribution in 'good' groups are higher in panel B as compared to panel A: During the last 5 periods of the R_p -game, groups with 3 or 4 *Pun*-types have an average contribution of 17.2 tokens. Groups with 3 or 4 *CC*-types 'only' reach 14.9 tokens on average. The difference is significant at the 5%-level ($p = 0.036$ in a two-sided t-test).

In the stranger design, we generally observe lower contribution levels. Comparing panel C and D further shows that the differences among 'top' groups are less pronounced than in the R_p -game.

¹⁷Recall that 452 subjects played the R-game in a partner design (R_p), i.e., in stable groups of four players, and 176 subjects in a stranger design (R_s) in stable matching groups of eight.

¹⁸To ease exposition, the figure pools groups with similar type compositions. The raw data are illustrated in the Online Appendix (see Figure S.6).

Figure 5: Average (Matching)-Group Contributions by Type Prevalence



Notes: Panels A and B [C and D] show the average contribution per period among the [matching]-groups for varying frequencies of CC- (panel A and C) and Pun-types (B and D). Panels A and B consider the groups of four subjects from the partner design (R_p), panel C and D are based on the eight-player matching groups from the stranger design (R_s). The underlying variation of types across (matching-)groups is presented in Table A.1 in the Appendix.

Matching-groups with either few *CC*- or few *Pun*-types show strongly declining contributions over time, a pattern well documented for repeated public goods games without punishment.¹⁹

5.2 Regression Analysis: Group Contributions

Figures 5 and S.2 show that both, the number of *CC*- and *Pun*-types are important determinants of average contributions and payoffs at the group level. To investigate the role of the different types in more detail, we conduct a regression analysis. We estimate models of the structure

$$\bar{g}_{\ell t} = \gamma_0 + \gamma_1 CC_{\ell}^{few} + \gamma_2 CC_{\ell}^{many} + \sum_t \delta_t D_t + \epsilon_{\ell t}, \quad (4)$$

where $\bar{g}_{\ell t} := \frac{1}{n} \sum_{i=1}^n g_{i\ell t}$ is the average contribution level in group ℓ in period t . The explanatory variables are dummies indicating if there are few or many *CC*-types in a (matching) group. In addition, the specification accounts for period-fixed effects. The results from linear random-effects estimations of equation (4) for the 113 groups in the partner design (R^P -game) are presented in column (1) of Table 1.²⁰

Consistent with the graphical evidence from above, and in line with Gächter and Thöni (2005), the estimates document that groups with a higher number of conditional cooperators achieve higher contributions. The point estimates indicate that groups with one or two *CC*-types reach contributions which are, on average, around 4 tokens higher than in groups with zero *CC* types. For groups with three or four *CC* types, this difference increases to 7 token. In economic terms, both coefficients are sizeable. Statistically speaking, however, the first coefficient, which corresponds to γ_1 from equation (4), is only weakly significant.

Column (2) reports the results for a model that uses dummies indicating groups with few or many *Pun*- (rather than *CC*-)types. The point estimates are of similar magnitude but the coefficients are more precisely estimated: on average, a group with one or two [three or four] *Pun*-types achieves contribution levels that are around 4 [7] tokens above those observed for groups with zero *Pun*-types. Both dummies are now significant at the 1% and 5% level, respectively. Note further that all information criteria reported in Table 1 show that the estimated model in column (2) clearly dominates the one from column (1): the R^2 strongly increases and the Akaike as well as the Bayesian information criteria (AIC and BIC) both decline, indicating a better model fit. This

¹⁹Figure S.2 in the Online Appendix replicates Figure 5 for average group *payoffs* rather than contributions. This exercise delivers similar findings as those discussed above.

²⁰Tobit estimations yield almost identical results (see the Online Appendix, Table S.8).

Table 1: Group Composition and Average Contributions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	<i>Partner Design</i>				<i>Stranger Design</i>			
CC^{few}	4.117*		3.114	3.168	6.602**		1.063	0.139
	(2.319)		(2.502)	(2.540)	(2.748)		(2.091)	(2.204)
CC^{many}	6.935***		5.246**	5.427**	8.187***		2.125	-0.118
	(2.286)		(2.507)	(2.652)	(2.042)		(1.874)	(3.365)
Pun^{few}		3.841**	3.148*	3.261*		6.867***	6.512***	5.622***
		(1.582)	(1.650)	(1.847)		(2.143)	(1.692)	(1.456)
Pun^{many}		7.323***	6.367***	6.662***		8.150***	7.184***	5.473***
		(1.507)	(1.614)	(2.049)		(1.258)	(1.057)	(1.546)
$CC \times Pun^{few}$				-0.193				2.704
				(1.375)				(2.958)
$CC \times Pun^{many}$				-0.523				5.190
				(1.875)				(3.862)
Obs.	1,130	1,130	1,130	1,130	220	220	220	220
R ²	0.117	0.187	0.235	0.236	0.229	0.446	0.459	0.501
AIC	7057	6964	6898	6902	1295	1222	1221	1207
BIC	7117	7025	6969	6983	1335	1263	1268	1261

Notes: Estimates from linear random-effects models for the R_p - (columns 1–4) and the R_s -game (columns 5–8). Dependent variable: average group contribution per period. The number of observations is $N = 1,130$ (113 groups of the partner design \times 10 periods) and $N = 220$ (22 matching-groups of the stranger design \times 10 periods), respectively. In the partner design, we use dummies for one or two (*few*) versus three or four (*many*) *CC*- or *Pun*-types. The omitted category pools groups with zero *CC*- or *Pun*-types. In the stranger design, we use dummies for matching groups with three or four (*few*) and five or more (*many*) *CC*- or *Pun*-types. The reference groups are then matching groups with two or less *CC*- or *Pun*-types. All specifications include a constant and a full set of period-fixed effects (coefficients not reported). Standard errors, clustered at the (matching-)group level, are in parentheses; *** / ** / * indicate significance at the 1%-, 5%-, and 10%-level, respectively.

suggests that information about the number of *Pun*-types in a group is more useful for explaining the heterogeneity in cooperation levels between groups.

The last point is further corroborated by the outcome reported in column (3). The specification includes both sets of dummies from before and thus directly assesses the relative importance of having more or less *CC*- or *Pun*-types in a group. Comparing AIC and BIC, we first note that the model from column (3) performs better than the specification from column (1); however, AIC and BIC only improve modestly (become smaller) as compared to the specification from column (2). Secondly, the results show that the estimated coefficients on the two *CC*-dummies shrink in magnitude while standard errors increase: one coefficient (γ_1) loses statistical significance, the other one (γ_2) remains significant at the 5% level. The precision of the two *Pun*-dummies decreases slightly, too; however, both coefficients remain significant at the 1% and 10% level, respectively.

The last specification, presented in column (4), adds dummies for the prevalence of *CC*×*Pun*-types (in the spirit of an interaction term). The outcome shows that, for a given number of *CC*- and *Pun*-types, having more or less of these two-way types does *not* matter for the groups' average contribution levels. In fact, AIC and BIC suggest that the simpler specification from column (3) dominates the one from (4). Concerning the other type dummies, it is reassuring to see that the estimates are almost unchanged — an observation that is consistent with the random assignment of subjects to groups.

In a next step, we consider the data from the stranger design. Columns (5)–(8) in Table 1 present the estimation output from an analogous set of regressions as those discussed above. The results are almost identical to those for the partner design. Again, we observe that a higher number of *CC*- or *Pun*-types within a matching group is associated with higher average contributions. Similar as above, specification (6), which controls for variation in the number of *Pun*-types, has a higher explanatory power and a better fit than specification (5). In column (7), when we add dummies for both types, only the ones on the *Pun*-types remain significant. The analysis therefore replicates the picture from above: having more *Pun*-types in a group seems to be of first-order importance to achieve high contribution levels.

5.3 Regression Analysis: Individual Contributions

Above we showed how variation in groups' type composition affects average group contributions. We now turn to the underlying individual behavior that is driving these results. To investigate the influence of the group composition on individual contribution decisions, we estimate the equation

$$g_{it} = \lambda_0 + \lambda_1 CC_\ell^{few} + \lambda_2 CC_\ell^{many} + \lambda_3 Pun_\ell^{few} + \lambda_4 Pun_\ell^{many} + \phi Pun_i + \sum_t \delta_t D_t + \epsilon_{it}, \quad (5)$$

where CC_{ℓ}^{few} and CC_{ℓ}^{many} are dummies indicating if individual i faces few or many conditional cooperators in her (matching-)group ℓ . The dummies for pro-social punishers are defined analogously. The λ -coefficients thus measure the effect of having different types in i 's (matching-)group on her contribution.²¹ The model further includes a dummy Pun_i , which indicates if i has been classified as a *Pun*-type herself. As an alternative, we will consider the dummy $NPun_i$, which indicates that she did not punish in the P-game. The coefficient ϕ then captures whether being a *Pun* (or *NPun*) type is correlated with higher or lower contributions.

Equation (5) is estimated separately for subjects classified as free-riders (*FR*) and conditional cooperators (*CC*). Considering these two groups separately allows for type-specific responses to variation in the group composition. Estimation outputs for the partner design are provided in Table 2.

Let us first focus on the results for conditional cooperators. Columns (1) and (2), which present specifications that separately include either the CC_{ℓ} or the Pun_{ℓ} dummies, suggest that a *CC*-type's contribution increases with the number of (other) conditional cooperators as well as with the number of pro-social punishers in the group. In terms of statistical and economic significance, however, an increasing number of *Pun*-types seems to exert a much stronger effect on contributions. This point is also documented in column (3), where the CC_{ℓ} dummies become statistically insignificant, whereas the coefficients on the effect from having few or many *Pun*-types in a group remain quantitatively large and significant at the 1%- and 5%-level, respectively.²²

Estimations for the *CC*-types in the stranger design, which are presented in Table 3, show very similar results. The CC_{ℓ} dummies are both insignificant (column 1), whereas the coefficients on the Pun_{ℓ} dummies are both large and relatively precisely estimated (column 2). When the two sets of dummies are combined, those for the prevalence of *Pun*-types in a matching-group remain highly significant.

Our results therefore show that — in the absence of group members who are willing to enforce a contribution norm — conditional cooperators per se do not necessarily perform well in coordinating on high contribution levels. Once pro-social punishers enter a group, conditional cooperators are much more willing to make higher contributions. The presence of *Pun*-types therefore seems to be essential to obtain high contribution levels among conditional cooperators.

²¹The effects are computed relative to the benchmark of groups with zero (in the partners protocol) or less than two (in the stranger protocol) *CC*- and *Pun*-types.

²²Post-estimation tests further show that the effect from having two or three *Pun*-types in a group is statistically different from having only one pro-social punisher ($p = 0.001$).

Table 2: Group Composition and Individual Contributions (Partner Design)

	(1)	(2)	(3)	(4)	(5)	(6)
	<i>Conditional Cooperators (CC)</i>			<i>Free-Riders (FR)</i>		
CC^{few}	1.464 (1.544)		0.551 (1.569)	1.673 (2.494)		1.925 (2.640)
CC^{many}	3.089** (1.482)		1.939 (1.498)	6.606*** (2.335)		6.648*** (2.483)
Pun^{few}		2.440** (0.975)	2.166** (1.040)		4.288*** (1.583)	3.643** (1.500)
Pun^{many}		4.469*** (0.988)	4.194*** (1.057)		2.500 (1.781)	1.225 (1.611)
Pun_i	2.982*** (0.599)	2.794*** (0.487)	2.641*** (0.516)			
$NPun_i$				-4.079*** (1.106)	-4.020*** (1.065)	-3.531*** (1.052)
Constant	7.906*** (1.332)	7.571*** (0.884)	6.376*** (1.410)	6.744*** (2.339)	8.860*** (1.606)	4.620 (2.853)
Obs.	2,790	2,790	2,790	950	950	950
R ²	0.094	0.137	0.147	0.220	0.149	0.253
AIC	18248	18111	18082	6403	6499	6364
BIC	18325	18188	18171	6466	6562	6437

Notes: Estimates from linear random-effects models for the R_p -game. Dependent variable: individual contribution per period. Dummies with superscript 'few' indicate that one, and dummies with 'many' indicate that two or three other subjects in the respective group are *CC*- or *Pun*-type. Columns (1)–(3) are based on the sample of conditional cooperators: $N = 2,790$ (279 *CC*-types over 10 periods); columns (4)–(6) use the sample of free-riders: $N = 950$ (95 *FR*-types over 10 periods). All specifications include a constant term and a full set of period-fixed dummies (coefficients not reported). Standard errors, clustered at the group level, are in parentheses; *** / ** / * indicate significance at the 1%-, 5%-, and 10%-level, respectively.

Table 3: Group Composition and Individual Contributions (Stranger Design)

	(1)	(2)	(3)	(4)	(5)	(6)
	<i>Conditional Cooperators (CC)</i>			<i>Free-Riders (FR)</i>		
CC^{few}	0.201 (2.398)		0.490 (2.313)	6.718*** (2.360)		-0.0456 (0.530)
CC^{many}	2.747 (2.269)		2.648 (2.227)	11.17*** (1.760)		0.489 (2.937)
Pun^{few}		7.584*** (1.386)	7.302*** (1.197)		6.538*** (1.535)	6.413*** (2.051)
Pun^{many}		7.622*** (1.501)	6.837*** (1.277)		13.24*** (1.392)	12.82*** (3.026)
Pun_i	2.568*** (0.868)	3.275*** (0.896)	3.053*** (0.928)			
$NPun_i$				-2.863* (1.665)	-4.079*** (1.334)	-4.014*** (1.324)
Constant	8.398*** (1.901)	2.330* (1.357)	1.387 (1.823)	1.216 (1.086)	1.691* (1.016)	1.677* (1.001)
Obs.	1,030	1,030	1,030	350	350	350
R ²	0.115	0.159	0.197	0.294	0.417	0.418
AIC	6428	6375	6332	2263	2184	2180
BIC	6492	6440	6406	2309	2234	2230

Notes: Estimates from linear random-effects models for the R_S -game. Dependent variable: individual contribution per period. Dummies with superscript ‘few’ indicate that two to four [three or four in columns 1–3], and dummies with ‘many’ indicate that five or more subjects in the respective matching group are *CC*- or *Pun*-type. (The pooling of dummies was based on the actual type allocation in the matching groups, with the objective to minimize loss of information.) Columns (1)–(3) are based on the sample of conditional cooperators: $N = 1,030$ (103 *CC*-types over 10 periods); columns (4)–(6) use the sample of free-riders: $N = 350$ (35 *FR*-types over 10 periods). All specifications include a constant term and a full set of period-fixed dummies (coefficients not reported). Standard errors, clustered at the group level, are in parentheses; *** / ** / * indicate significance at the 1%-, 5%-, and 10%-level, respectively.

Next we turn to the results for free-riders. Overall, the estimates from columns (4)–(6) in Tables 2 and 3 provide a similar picture. However, due to the limited number of observations (we only observe 95 free-riders in the R_p , and 35 in the R_s game), some of our findings are less instructive.

Let us first turn to the partner design. Columns (4) and (5) of Table 2 suggest that *FR*-types' contributions are, similar as those of *CC*-types, increasing in the number of *CC*- and *Pun*-types in their group. For the sample of free-riders, the coefficients on the CC^{many} dummy becomes larger and is now significant at the 1% level (despite a larger standard error as compared to column 1). Concerning the presence of pro-social punishers, we only find a large and statistically significant effect from having few (as compared to no) *Pun*-types. The Pun^{many} dummy, however, is insignificant. Column (6), which presents the estimates for equation (5), suggests that the largest effect comes from having many *CC*-types in a group. Having more *Pun*-types further increases the free-riders' contributions, but the effect is only statistically significant for one of the *Pun*-dummies.

From these estimates it appears tempting to conclude that the contributions of *FR*-types are more sensitive to the presence of conditional cooperators rather than pro-social punishers. However, a closer look at the data from the partner design shows an almost perfect overlap of *CC*- and *Pun*-types in the (few) groups of the free-riders.²³ Hence, the high correlation among types in this small sample impedes our ability to draw strong conclusions on the differential impact of the two different types on free-riders' behavior in the partner design.

For the stranger protocol (where the overlap of *CC*- and *Pun*-types in the matching groups is smaller), the results for the free-riders are much closer to those observed for the conditional cooperators. Columns (4) and (5) of Table 3 indicate that free-riders contribute significantly more, the more *CC*- and *Pun*-types are in their matching groups. For the model specification in column (6), the CC_i dummies lose significance whereas the Pun_i dummies remain large and highly significant.

To wrap-up, the estimates show that free-riders' contributions are influenced by both, the presence of *CC*- and *Pun*-types. While the data from the stranger protocol show a clear enforcement result — a higher share of pro-social punishers in a matching group pushes free-riders to contribute more to the public good — the data from the partner protocol point to the influence of conditional cooperators. While the latter observation is based on a small sample, it is consistent with the idea that (at least some) free-riders act strategically in the repeated game, playing high contributions that aim at encouraging reciprocal behavior of the *CC*-types (e.g. [Sonnemans et al., 1999](#); [Keser and van Winden, 2000](#); [Muller et al., 2008](#)).

²³In almost all cases when the Pun^{many} dummy is equal to one, CC^{many} is one, too.

A last point worth discussing is the fact that the estimates from Tables 2 and 3 allow for a comparison of the average contributions among the different types introduced in our type classification from above (see, e.g., Table 4). To see this, one has to recognize that the constant term (λ_0 from equation 5) captures a type’s average contribution. Focusing on the partner design, the estimates from column (3) therefore suggest that an average conditional cooperator, who is *not* classified as pro-social punisher ($\text{Pun}_i = 0$), contributes 6.4 tokens (in the first period and with zero *CC*- and *Pun*-types among the group members). A *CC* \times *Pun*-type, in contrast, contributes significantly more: 9.0 tokens ($\lambda_0 + \phi$, based on column 3). From column (6) we further learn that an average free-rider, who is *not* classified as non-punisher ($\text{NPun}_i = 0$), makes a contribution of 4.6 tokens. A *FR* \times *NPun*-type would, *cet. par.*, contribute significantly less: 1.1 tokens. The different cooperation patterns from the one-shot C-game as well as the heterogenous punishment patterns from the one-shot P-game (which are used to classify these different types) are therefore strong predictors of the sizeable differences in individual contribution levels that are observed for the repeated game.

6 Concluding Discussion

Using a parsimonious strategy-method approach, we presented systematic evidence on the heterogeneity of punishment patterns at the individual level. We linked our novel classification of punishment-types to the popular cooperation-type classification from Fischbacher et al. (2001). This allowed for an individual-level analysis of the relationship between subjects’ disposition to cooperate and their inclination to enforce cooperation via peer punishment. The resulting two-dimensional classification suggested the existence of two distinct behavioral archetypes. On the one hand, we identified many subjects whose punishment and cooperation patterns are aligned. On the other hand, our analysis uncovered a non-trivial fraction of subjects whose cooperation and punishment patterns diverged: free-riders that punished pro-socially and conditional cooperators that did not punish.

The divergence between cooperation and punishment patterns allowed us to assess the role of the two-dimensional variation in types — which we identified in two independent one-shot games — for explaining group outcomes and individual behavior in a third, repeated game. Our analyses provided strong, causal evidence on the relative importance of pro-social punishers for achieving and maintaining cooperation. Exogenous variation in the number of punishment types within a (matching) group had a much higher explanatory power than similar variation in cooperation types.

The latter finding is particularly intriguing, since previous work has predominantly hinted at the importance of conditional cooperators for a group's success (e.g., Gächter and Thöni, 2005; Burlando and Guala, 2005). Except for Rustagi et al. (2010), however, the corresponding inferences are usually drawn from situations that do not entail elements of punishment. Given that the absence of sanctioning opportunities in natural environments is likely to be the exception rather than the rule, actual group outcomes might not be determined by individuals' cooperation types per se, but rather by the concomitant inclination to engage in pro-social punishment. Our results, in particular the identification of a behavioral archetype with diverging punishment and cooperation patterns, underline that this distinction indeed matters. It will be interesting to see in future studies if a similar differentiation also applies to other forms of pro-social (e.g., Falk and Szech, 2013) or anti-social (e.g., Abbink and Serra, 2012) behavior.

Our strategy method not only allows for a causal analysis of individual punishment patterns, it may also serve as a powerful tool to isolate the impact of different institutions on peer punishment, informal social sanctions (Maslet et al., 2003) or rewards (Sefton et al., 2007). The issue here is that almost any variation in the strategic environment — e.g., if subjects interact once or repeatedly — simultaneously influences behavior at the first (e.g., cooperation in our application) as well as the second stage of the game (punishment). By inducing a controlled level of exogenous variation at the first stage, our strategy method allows to distinguish the overall impact of an institutional change from its *ceteris paribus* effect at the second stage.

Finally, the results and the methodologies from our study open new avenues for follow-up research on cooperation and punishment. The approach adopted in this paper offers a rich set of opportunities to advance our understanding of differences in cooperation across cultures and societies (Henrich et al., 2006; Herrmann et al., 2008), e.g., by examining the underlying variation in individual cooperation and punishment types. Observing different punishment (and cooperation) patterns will also help to reassess the underlying motivations of peer punishment. If, for instance, people solely punish to reduce inequality in payoffs (in a self-centered way, e.g., following Fehr and Schmidt, 1999) this could intuitively explain the aligned behavioral archetype (pro-socially punishing conditional cooperators as well as individuals who free-ride in both stages of the game). However, self-centered models of inequality aversion would not be easily reconcilable with free-riders that are pro-social punishers or with conditional cooperators that do not punish. These diverging types would also be incompatible with a notion of strong reciprocity, assuming cooperation and punishment to be responses that are triggered by positive and negative reciprocity, respectively (Dohmen et al., 2008). Building on our design — e.g., by augmenting our strategy method to account for subject's beliefs about others' punishment — future research might address

these point and disentangle the influence of rational motives (Casari and Luini, 2012), emotions (Falk et al., 2005; Reuben and van Winden, 2008; Hopfensitz and Reuben, 2009) or inconsistency (Blanco et al., 2011) in explaining the different archetypes and their punishment patterns.

References

- Abbink, K. and D. Serra (2012). Anticorruption Policies: Lessons from the Lab. In *New advances in experimental research on corruption*, pp. 77–115.
- Albrecht, F. and C. Traxler (2016). Patterns of Norm Enforcement. Mimeo.
- Andreoni, J. (1988). Why free ride?: strategies and learning in public goods experiments. *Journal of Public Economics* 37(3), 291–304.
- Baker, G., R. Gibbons, and K. J. Murphy (2002). Relational Contracts and the Theory of the Firm. *Quarterly Journal of Economics* 117(1), 39–84.
- Balliet, D., L. B. Mulder, and P. A. M. Van Lange (2011). Reward, Punishment, and Cooperation: A Meta-Analysis. *Psychological Bulletin* 137(4), 594–615.
- Bardsley, N. (2000). Control Without Deception: Individual Behaviour in Free-Riding Experiments Revisited. *Experimental Economics* 3, 215–240.
- Blanco, M., D. Engelmann, and H. T. Normann (2011). A within-subject analysis of other-regarding preferences. *Games and Economic Behavior* 72(2), 321–338.
- Brekke, K. A., K. E. Hauge, J. T. Lind, and K. Nyborg (2011). Playing with the good guys. A public good game with endogenous group formation. *Journal of Public Economics* 95(9-10), 1111–1118.
- Burlando, R. M. and F. Guala (2005). Heterogeneous agents in public goods experiments. *Experimental Economics* 8(1), 35–54.
- Carpenter, J. P. (2007). Punishing free-riders: How group size affects mutual monitoring and the provision of public goods. *Games and Economic Behavior* 60(1), 31–51.
- Casari, M. and L. Luini (2012). Peer punishment in teams: expressive or instrumental choice? *Experimental Economics* 15(2), 241–259.
- Chassang, S. (2010). Building routines: Learning, cooperation, and the dynamics of incomplete relational contracts. *American Economic Review* 100(1), 448–465.
- Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics* 14(1), 47–83.
- Cunha, F. and J. Heckman (2007). The Technology of Skill Formation. *American Economic Review* 97(2), 31–47.
- Dohmen, T., A. Falk, D. B. Huffman, and U. Sunde (2008). Representative Trust and Reciprocity: Prevalence and Determinants. *Economic Inquiry* 46(1), 84–90.
- Falk, A., E. Fehr, and U. Fischbacher (2005). Driving Forces Behind Informal Sanctions. *Econometrica* 73(6), 2017–2030.
- Falk, A. and N. Szech (2013). Morals and markets. *Science* 340(6133), 707–11.
- Fehr, E. and S. Gächter (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90(4), 980–994.
- Fehr, E. and S. Gächter (2002). Altruistic punishment in humans. *Nature* 415(6868), 137–40.
- Fehr, E. and K. M. Schmidt (1999). A Theory of Fairness, Competition and Cooperation. *Quarterly Journal of Economics* 114(3), 817–868.
- Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-made Economic Experiments. *Experimental Economics* 10(2), 171–178.

- Fischbacher, U. and S. Gächter (2010). Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments. *American Economic Review* 100(1), 541–556.
- Fischbacher, U., S. Gächter, and E. Fehr (2001). Are people conditionally cooperative? Evidence from a public goods experiment. *Economics Letters* 71(3), 397–404.
- Gächter, S., E. Renner, and M. Sefton (2008). The long-run benefits of punishment. *Science* 322(5907), 1510.
- Gächter, S. and C. Thöni (2005). Social Learning and Voluntary Cooperation Among Like-Minded People. *Journal of the European Economic Association* 3(2), 303–314.
- Greiner, B. (2004). An Online Recruitment System for Economic Experiments. Technical report, Ges. für Wiss. Datenverarbeitung, Göttingen.
- Gürerk, O., B. Irlenbusch, and B. Rockenbach (2006). The competitive advantage of sanctioning institutions. *Science* 312(5770), 108–111.
- Hamman, J. R., R. A. Weber, and J. Woon (2011). An Experimental Investigation of Electoral Delegation and the Provision of Public Goods. *American Journal of Political Science* 55(4), 738–752.
- Heckman, J. J. and Y. Rubinstein (2001). The importance of noncognitive skills: Lessons from the GED testing program. *American Economic Review* 91(2), 145–149.
- Henrich, J., R. McElreath, A. Barr, J. Ensminger, C. Barrett, A. Bolyanatz, J. C. Cardenas, M. Gurven, E. Gwako, N. Henrich, C. Lesorogol, F. Marlowe, D. Tracer, and J. Ziker (2006). Costly punishment across human societies. *Science* 312(5781), 1767–70.
- Herrmann, B., C. Thöni, and S. Gächter (2008). Antisocial punishment across societies. *Science* 319(5868), 1362–7.
- Hopfensitz, A. and E. Reuben (2009). The Importance of Emotions for the Effectiveness of Social Punishment. *Economic Journal* 119(540), 1534–1559.
- Keser, C. and F. van Winden (2000). Conditional Cooperation and Voluntary Contributions to Public Goods. *Scandinavian Journal of Economics* 102(1), 23 – 39.
- Kosfeld, M., A. Okada, and A. Riedl (2009). Institution Formation in Public Goods Games. *American Economic Review* 99(4), 1335–1355.
- Kosfeld, M. and D. Rustagi (2015). Leader Punishment and Cooperation in Groups: Experimental Field Evidence from Commons Management in Ethiopia. *American Economic Review* 105(2), 747–783.
- Kube, S., S. Schaube, H. Schildberg-Hörisch, and E. Khachatryan (2015). Institution formation and cooperation with heterogeneous agents. *European Economic Review* 78, 248–268.
- Kube, S. and C. Traxler (2011). The Interaction of Legal and Social Norm Enforcement. *Journal of Public Economic Theory* 13(5), 639–660.
- Ledyard, J. O. (1994). Public goods: A survey of experimental research. In *The Handbook of Experimental Economics*, pp. 111–194.
- Masclot, D., C. Noussair, S. Tucker, and M.-C. Villeval (2003). Monetary and Nonmonetary Punishment in the Voluntary Contributions Mechanism. *American Economic Review* 93(1), 366–380.
- Muller, L., M. Sefton, R. Steinberg, and L. Vesterlund (2008). Strategic behavior and learning in repeated voluntary contribution experiments. *Journal of Economic Behavior & Organization* 67(3-4), 782–793.

- Ostrom, E., J. Walker, and R. Gardner (1992). Covenants With and Without a Sword: Self-Governance is Possible. *American Political Science Review* 86(2), 404.
- Rammstedt, B. and O. P. John (2007). Measuring personality in one minute or less: A 10-item short version of the Big Five Inventory in English and German. *Journal of Research in Personality* 41(1), 203–212.
- Reuben, E. and A. Riedl (2013). Enforcement of contribution norms in public good games with heterogeneous populations. *Games and Economic Behavior* 77(1), 122–137.
- Reuben, E. and F. van Winden (2008). Social ties and coordination on negative reciprocity: The role of affect. *Journal of Public Economics* 92(1-2), 34–53.
- Robbett, A. (2015). Voting with hands and feet: the requirements for optimal group formation. *Experimental Economics* 18(3), 522–541.
- Rustagi, D., S. Engel, and M. Kosfeld (2010). Conditional cooperation and costly monitoring explain success in forest commons management. *Science* 330(6006), 961–5.
- Sefton, M., R. Shupp, and J. M. Walker (2007). The Effect of Rewards and Sanctions in Provision of Public Goods. *Economic Inquiry* 45(4), 671–690.
- Sonnemans, J., A. Schram, and T. Offerman (1999). Strategic behavior in public good games: when partners drift apart. *Economics Letters* 62(1), 35–41.
- Traxler, C. and J. Winter (2012). Survey evidence on conditional norm enforcement. *European Journal of Political Economy* 28(3), 390–398.
- Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology* 51(1), 110–116.
- Yamagishi, T. (1988). The provision of a sanctioning system in the United States and Japan. *Social Psychology Quarterly* 51(3), 265–271.

Appendix

A1 Contribution Triples

Below we list the hypothetical contribution triples that were used within each of the ten combinations of g^L , g^M and g^H (see Section 2.1). Before the experiment, these 10×8 triples were randomly generated by sampling with replacement from the corresponding sets g^L , g^M , g^H . Each player then faced a randomly selected triple within each combination. If the selected triple would by chance correspond to the real triple, the subject would *not* face this situation; instead another one of the pre-defined contribution triples for the corresponding combination would be drawn.

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(g^L, g^L, g^L) :	(0,0,0)	(0,2,3)	(1,1,3)	(1,2,2)	(1,2,3)	(1,2,4)	(1,3,3)	(1,3,4)
(g^L, g^L, g^M) :	(0,1,5)	(0,2,8)	(0,2,14)	(1,2,10)	(1,2,12)	(1,3,14)	(2,2,6)	(2,3,12)
(g^L, g^L, g^H) :	(0,3,18)	(1,2,20)	(1,3,19)	(1,4,20)	(2,2,18)	(2,2,19)	(3,3,18)	(4,4,17)
(g^L, g^M, g^M) :	(0,9,11)	(0,5,12)	(0,13,14)	(1,10,15)	(2,6,8)	(2,9,11)	(2,10,15)	(3,13,14)
(g^L, g^M, g^H) :	(0,6,19)	(0,14,17)	(2,6,17)	(2,8,20)	(2,11,19)	(3,7,18)	(4,8,17)	(4,10,20)
(g^L, g^H, g^H) :	(0,18,19)	(1,19,19)	(2,18,19)	(2,18,20)	(2,19,19)	(3,18,20)	(3,19,19)	(4,19,20)
(g^M, g^M, g^M) :	(5,7,12)	(5,14,15)	(6,6,9)	(6,10,10)	(7,8,9)	(7,10,13)	(7,14,15)	(8,9,11)
(g^M, g^M, g^H) :	(5,5,17)	(5,8,16)	(6,11,20)	(8,15,17)	(9,12,18)	(9,15,18)	(11,15,19)	(12,15,19)
(g^M, g^H, g^H) :	(5,18,20)	(7,18,19)	(9,18,20)	(11,17,17)	(12,17,18)	(12,18,18)	(14,17,20)	(15,17,19)
(g^H, g^H, g^H) :	(17,17,19)	(17,18,19)	(17,18,20)	(17,19,19)	(17,19,20)	(18,18,19)	(18,18,20)	(20,20,20)

A2 Type Distribution among (Matching-) Groups

Table A.1 illustrates the group composition that emerged from the random assignment of subjects into different [matching-] groups. In addition, the table presents the expected distribution (numbers in italics) based on the population frequencies of *CC*- and *Pun*-types as reported in Tables 2 and 3, respectively. The chance, for instance, of having four *CC*-types in one group is given by 0.608^4 . Among 113 groups, one should thus expect 15.4 groups with this composition. Stated differently: the numbers in italics form the ‘perfect randomization’ benchmark. The actual outcome is in fact very close to this benchmark.

The top part of the table illustrates the variation in the different types among the 113 four-player groups in the partner protocol (R_p -game). Consistent with the high population frequency of conditional cooperators (60.8 % of our sample, see Table 3) we observe that the majority of groups are populated by two (35 groups) or three (48 groups) *CC*-types. In addition, there are

Table A.1: Type Distribution per (Matching) Group

Number of subjects:		0	1	2	3	4	5	6	7	8	Sum
R_p -game	CC	4	13	35	48	13					113
		<i>2.7</i>	<i>16.6</i>	<i>38.5</i>	<i>39.8</i>	<i>15.4</i>					
	Pun	15	30	34	30	4					113
		<i>8.8</i>	<i>31.5</i>	<i>42.1</i>	<i>25.0</i>	<i>5.6</i>					
R_s -game	CC	-	1	1	4	2	5	8	1	-	22
		<i>0.0</i>	<i>0.2</i>	<i>0.8</i>	<i>2.6</i>	<i>5.0</i>	<i>6.2</i>	<i>4.8</i>	<i>2.1</i>	<i>0.4</i>	
	Pun	-	1	4	4	2	4	6	1	-	22
		<i>0.1</i>	<i>1.0</i>	<i>3.0</i>	<i>5.3</i>	<i>5.9</i>	<i>4.2</i>	<i>1.9</i>	<i>0.5</i>	<i>0.1</i>	

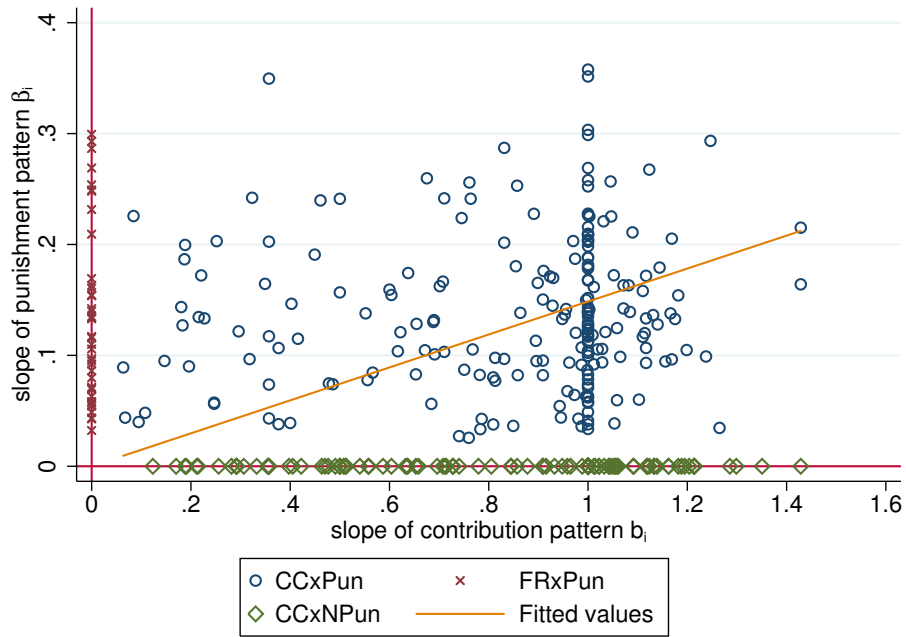
Notes: In the R_p -game subjects are counted at the *group level* (with 4 subjects per observational unit). In the R_s -game subjects are counted at the *matching-group level* (with 8 subjects per observational unit). The depicted distribution of subjects occurred from randomly assigning subjects to groups (matching-groups) at the beginning of the R-game. The numbers in italics present the expected distribution based on the population frequencies of *CC*- and *Pun*-types as reported in Tables 2 and 3, respectively.

several groups with no (4), one (13) or even four *CC*-types (13 groups). A slightly more symmetric distribution is observed for *Pun*-types — reflecting the fact that the population prevalence is close to one half (47.1 % of our sample, see Table 2). There are between 30 to 34 groups, each with either one, two, or three *Pun*-types. In addition, there are 15 groups with zero and four groups with four *Pun*-types. We use two-sided Fisher’s exact tests to assess the hypothesis that the observed and the predicted distribution of groups with different type-compositions stem from the same distribution. Consistent with random group assignment, this H_0 cannot be rejected ($p = 0.812$ for the distribution of *CC*-types, and $p = 0.539$ for the distribution of *Pun*-types).

The lower part of Table A.1 captures the variation in group compositions between the 22 matching groups (each with eight subjects) from the stranger protocol (R_s -game). Similar as above, the data indicate quite some variation in the type composition across groups. Given the limited number of matching groups, there appear to be larger deviations from the expected number of groups with different compositions. However, the actual distribution is again not different from the expected random distribution: the p-values from two-sided Fisher’s exact tests are, exactly as above, $p = 0.812$ for the *CC*- and $p = 0.539$ for the *Pun*-types.

A3 Complementary Figures and Tables

Figure A.1: Distribution of $\hat{\beta}_i$ and \hat{b}_i



Notes: Scatter plot for individual level peer punishment pattern slope $\hat{\beta}_i$ and contribution pattern slope \hat{b}_i for the four most prevalent types, i.e., *CC*, *FR*, *Pun*, and *NPun*. The estimated correlation between the respective $\hat{\beta}_i$ and \hat{b}_i is depicted as a yellow line. To ease illustration *FR* \times *NPun*-type values are not plotted. The apparent lump of observations in \hat{b}_i is attributed to 'perfect' conditional cooperation in the C-game, i.e., subjects match the shown average group contribution perfectly. Observations along the axes belong to off-diagonal types in figure 4.