

Cooperation and Norm Enforcement - The Individual-Level Perspective

Felix Albrecht^{†,‡}, Sebastian Kube^{‡,‡}, Christian Traxler^{*,‡}

June 20, 2018

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 modified 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 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 important for achieving cooperation. Incorporating information on punishment types explains large amounts of the between- and within-group variation in cooperation.

JEL-Classification: C9; D03

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

Acknowledgements: We would like to thank the editor, Tim Cason, two anonymous referees as well as Nick Bardsley, Christoph Engel, Louis Putterman, Christian Thöni and numerous seminar and workshop participants for helpful comments and suggestions. Financial support from DFG (Deutsche Forschungsgemeinschaft; grant 50130225) is gratefully acknowledged. The usual disclaimer applies.

† University of Marburg; ‡ University of Bonn; ‡ Max Planck Institute for Research on Collective Goods;
* Hertie School of Governance, Berlin.

1 Introduction

An extensive body of research documents cooperation among humans (e.g., Andreoni, 1988; Ledyard, 1994; Fischbacher and Gächter, 2010; Balliet et al., 2011; Chaudhuri, 2011, to name only a few), pointing out that cooperation problems can be mitigated by appropriate institutional settings (e.g., Ostrom et al., 1992; Kosfeld et al., 2009). Among these, the ubiquitous mechanism of peer punishment plays a prominent role in the literature (e.g., Fehr and Gächter, 2000, 2002; Carpenter, 2007; Reuben and Riedl, 2013). Even though peer punishment makes successful cooperation much more likely to occur, there are still groups who fail to use decentralized punishment in an effective and pro-social manner. This might be due to the fact that peer punishment constitutes a cooperation problem in itself (Yamagishi, 1986). A breakdown in cooperation that coincides with a failure of peer punishment could thus capture two sides of the same coin (see, e.g., Ones and Putterman, 2007; Peysakhovich et al., 2014). 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 dispositions do not coincide, which of the two is relatively more important in achieving cooperative outcomes under peer punishment?

We study these questions employing a classical workhorse in the literature on cooperation and punishment: a linear public-goods game (VCM) with decentralized 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 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).¹

¹An alternative approach, based on a conventional strategy-method together with a strongly restricted choice set, is implemented by Cheung (2014) and Kamei (2014), who offer interesting complementary findings on cooperation and punishment patterns, respectively. Beyond method and sample size, the present paper also differs from these studies in that we analyze the link between cooperation and punishment types as well as the role of the different types for achieving cooperative outcomes in a repeated game.

Using this strategy-method to elicit punishment patterns reveals substantial heterogeneity between individuals. In our sample of $N = 628$ 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 contributions, 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 (in the spirit of ‘anti-social punishment’; see, e.g., Herrmann et al., 2008). Moreover, we document that among pro-social punishment types, patterns are almost exclusively ‘self-centered’ around the own contribution level.

Linking individual punishment patterns to the corresponding individual dispositions to cooperate — which 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 behavioral archetypes. (i) For the majority of our subjects cooperation and punishment types are aligned: we find that 55% of conditional cooperators punish pro-socially and that 56% of free-riders are non-punishers. (ii) Consequently, 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 diverging — is a major benefit from combining our 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 in the presence of punishment opportunities. 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 that highlight the important role of conditional cooperators (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 in this richer environment: 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 both, subjects classified as free-riders and among conditional cooperators.

These findings underline that (at least in the context of peer punishment) 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 causal evidence on this link. The results complement recent studies that have hinted at the importance of individuals' inclination to punish. [Ones and Putterman \(2007\)](#) rank lab subjects according to a composite index, which is based on previous contribution and punishment decisions in a repeated VCM. Using the ranking to form homogenous groups of similar types, they find that subsequent cooperation is higher in groups with 'higher-ranked' subjects, i.e., among individuals that tend to be more cooperative and/or willing to engage in pro-social punishment.

Studying field data, [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.² 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 punishment patterns, whereas [Ones and Putterman \(2007\)](#) combine both patterns into a single index. By contrast, [Falk et al. \(2005\)](#) study both individual punishment and cooperation behavior in isolation, 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.

²The 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.

The classification of individuals along two dimensions offers additional insights 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 focused on variation in punishment and cooperation patterns at the *aggregate level*.³ Our analysis complements these studies of group-level heterogeneity and thus constitutes a potential micro-foundation that might prove useful for future studies.

It seems natural to explore several follow-up questions, especially (but by no means exclusively) in the context of decentralized sanctioning and norm enforcement. Knowledge about individuals' (punishment) types might, for instance, 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). Currently, this literature is strongly focusing on how different *contribution* types are affected by, or react to, the institutions at hand (e.g., Brekke et al., 2011). However, it might be worthwhile to extend this line of thinking to include *punishment* types as well — namely as soon as the institution at hand relies on some form of mutual monitoring, expression of preferences over certain norms, or other mechanisms that are likely to appeal differently to different kind of punishment types. Moreover, if institutions need to be adapted endogenously (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 Güreker et al., 2006), information about a population's type composition might allow to anticipate the support for an institution for a given population. Finally, knowledge about individuals' cooperation and punishment types might offer new solutions to optimal team composition problems (e.g., Burlando and Guala, 2005; Gächter and Thöni, 2005; Ones and Putterman, 2007).

The remainder of this paper is structured as follows. The next section discusses the design and implementation of the experiment. Section 3 presents the results from the classification of cooperation and punishment types. Section 4 shows how the presence of these different types influence group outcomes and individual behavior in a repeated game. Section 5 concludes.

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

³Consider, 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).

(*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 heterogeneous 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. In addition to these three games, subjects answered a brief questionnaire.

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 $\bar{g}_j \in \{0, 1, \dots, 20\}$ rounded to integers. 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 a maximum of 10 punishment points to the other group members $j \neq i$, $0 \leq d_{ij} \leq 10$. 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 again be 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 modified strategy-method at the punishment stage.⁴ 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). All 11 triples are then 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 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.⁵ 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

⁴This 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. For a related but different approach, see Cheung (2014) and Kamei (2014).

⁵One subject might see, for instance, $\{0, 0, 0\}$ for the combination $\{g^L, g^L, g^L\}$ and $\{0, 2, 8\}$ for $\{g^L, g^L, g^M\}$. A different subject might face $\{0, 2, 3\}$ for the former and $\{0, 2, 14\}$ for the latter. Balancing tests indicate that randomization at the individual level was successful.

neither know which one is the ‘real’ triple,⁶ 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 respective matching-group.⁷

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 4). 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. 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, 2015). To prevent strategic spillovers between games, subjects received the instructions for each subsequent part of the experiment once the previous part had been completed by all participants. Standard experimental procedures were followed.⁸ 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.

⁶Testing whether subjects punish the (unknown) real versus the hypothetical contributions differently, we find no significant differences whatsoever.

⁷The instructions under the stranger protocol explicitly informed subjects that groups are randomly re-shuffled each period, without stating the total number of matching groups per lab session. (One session typically consisted of 24 subjects spread over three matching-groups.) The implementation of multiple matching groups per session is a common practice that balances the benefits from reducing reputational concerns and total implementation costs. Moreover, since parts of our analyses exploit between matching-group variation in types, smaller matching-groups are important to obtain sufficient between group variation.

⁸The instructions and further details on the procedure are available in the Online Appendix.

3 Individual Patterns in Cooperation and Punishment

This section studies individual punishment (Subsection 3.2) and cooperation patterns (3.3). Section 3.4 analyzes if and how these two patterns are aligned. Before doing so, we first discuss behavioral predictions based on the related literature.

3.1 Behavioral Predictions

As the VCM with punishment is well studied in the literature, the baseline predictions (as well as their limited predictive power) are well-known: As already noted above, the subgame-perfect Nash equilibrium for selfish payoff-maximizing players is that nobody contributes to the public good at the first stage since players anticipate that nobody will engage in costly punishment at the second stage. In that case, we should observe contribution and punishment profiles that are both ‘flat’ at zero, i.e., free-riders that do not punish.

Concerning the contribution profiles, previous studies have found a significant number of conditional cooperators. The seminal paper by Fischbacher et al. (2001), for instance, classifies 50% of subjects in their sample as conditional cooperators and 30% as free-riders. Similarly, heterogeneity has also been observed with respect to punishment behavior. For example, the between-group comparison in Herrmann et al. (2008) reveals significant differences in both the level and the targeting of punishment. Correspondingly, heterogeneity in individual punishment behavior is observed in Falk et al. (2005), Kube and Traxler (2011), Cheung (2014), and Kosfeld and Rustagi (2015). Although these articles differ in their specific elicitation methods and underlying games, they all point to the existence of different punishment types: some subjects do not use punishment, while others do engage in costly punishment. Among the latter, some apply punishment in a pro-social manner (targeting free-riders or low cooperation levels) others punish anti-socially (e.g., targeting high cooperation levels).

In light of the previous evidence, we therefore expect to observe individual-level differences in contribution and punishment patterns, too. However, it still remains open if and how individual contribution and punishment patterns are related.

Conceptually, one might argue that these patterns should be closely aligned, since both peer punishment and voluntary contributions (in the C-game) constitute a cooperation problem. One should thus expect that free-riders do not punish (as long as no monetary gains are expected to arise from punishment, an argument already made by Oliver, 1980). Similarly, Fehr and Gächter (2000, p.984) postulate that conditional cooperators are willing to engage in the costly punishment of free-riders, arguing that this would be predicted by models of reciprocity and equity. In fact, Leibbrandt and López-Pérez (2012, p.762) find that a “combination of inequity-averse and selfish types can sufficiently capture ... punishment patterns”, as measured by conditional responses to a dictator’s choices in ten binary allocation decisions. Likewise, the results in Che-

ung (2014, p.130) on the determinants of peer punishment “are directionally consistent with the predictions of the Fehr and Schmidt (1999) model of inequality aversion.”⁹

Similar notions of a close alignment of cooperation and punishment are developed in many other contributions. The review by Gächter and Herrmann (2009) on human cooperation, for instance, refers to voluntary contributions as positive and to punishment as negative reciprocity. Likewise, Ones and Putterman (2007, p.498) implicitly assume that positive and negative reciprocity are “two sides of the same coin” when they explore the impact of punishment and contribution behavior on group outcomes. A very different view is offered by Peysakhovich et al. (2014), who use factor analyses to compare individual decisions across six different one-shot games.¹⁰ They conclude that “punishment and cooperation may be separate phenomena, rather than being driven by a common altruistic motivation” (p.2) and thus “may not be two sides of the same coin” (p.3). If their findings were to extend to cooperation and punishment behavior within a given situation, it might be that conditional cooperators do not necessarily engage in (pro-social) punishment. Likewise, it might be that free-riders do spend resources on (pro-social) punishment.

To wrap-up: on the basis of the existing evidence we clearly expect to observe heterogeneity in both contribution and punishment patterns. However, the literature offers competing conjectures regarding the outcome of the two-dimensional classification approach: behavioral patterns might be fully aligned or (at least partially) diverging. Our two-dimensional classification offers an explorative approach that seeks to clarify whether or not cooperation and punishment are indeed two sides of the same coin. While we can offer a novel take on this question, we do not offer a specific test of the underlying motivation. To convincingly sort out different channels emphasized in competing models (e.g., of other-regarding preferences), one would require much richer data on individuals’ beliefs and decisions under very different payoff functions (i.e., from different games and different parametrization).

3.2 Individual Peer-Punishment Patterns

3.2.1 Primary Classification of Punishment Types

In a first attempt 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 $j \neq i$):

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

⁹For the underlying intuition see footnote 22 below and, for a more formal analysis, the supplementary Online Appendix of Leibbrandt and López-Pérez (2012, p.A42).

¹⁰Among others, the authors observe an unconditional contribution decision in a VCM and a punishment decision in a binary prisoners’ dilemma.

The regressor in eq. (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).¹¹ Using the data from the strategy-method in the P-game (for the punishment decisions of the one-shot game), we *separately* estimate α_i and β_i for each of our 628 subjects. The estimated coefficients capture individual-level heterogeneity in punishment patterns.

It is important to realize that conventional observational data would 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. Similarly, in repeated games like our R-game, contributions shape punishment and punishment shapes contributions simultaneously.¹² Our strategy-method breaks this simultaneity by introducing 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.¹³

Running 628 regressions with $N_i = 30$, we obtain the estimates $\hat{\alpha}_i$ and $\hat{\beta}_i$ (along with robust standard errors) for each subject. Based on these estimates, we then classify the subjects' punishment patterns. Our classification distinguishes between 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 each case, 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 $p > 0.01$.¹⁴

¹¹Estimating a model with $d_{ij} = \alpha'_i + \beta'_i g_j + \varepsilon'_i$ would yield equivalent estimates with $\hat{\beta}_i = -\hat{\beta}'_i$.

¹²Due 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.5 in the Online Appendix).

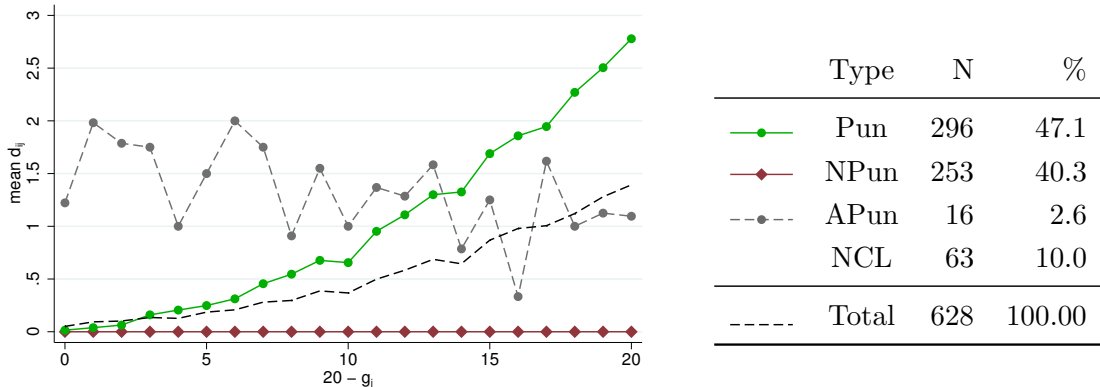
¹³Our results are insensitive to including the three punishment decisions for the real contribution triple.

¹⁴The 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). Our primary classification approach deviates from this self-centered notion of anti-social punishment as we do not consider punisher i 's own contribution g_i . Still, *APun*-types reflect patterns of punishment that are targeted towards high contributors.

Punishment patterns that cannot be assigned to one of these three types are summarized in a group of non-classified (*NCL*) patterns. The different types and their stylized punishment patterns are illustrated in Figure A1 in the Appendix.

The results from our classification approach are presented in Figure 1. 47.1% of our subjects are classified as *pro-social punishers*, 40.3% are *non-punishers*, 2.6% display an anti-social pattern, 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 A1). 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.¹⁵

Figure 1: Primary Punishment Types and Patterns



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 1, is therefore clearly increasing in $20 - g_j$. Note further that the slope of the punishment pattern is relatively steep. The average $\hat{\beta}_i$ among *Pun* types is 0.135 (the median is 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).

¹⁵These results are documented in the Online Appendix (see Figure S.1). It is further worth noting that we obtain very similar type distributions if we use Spearman's rank correlation to classify punishment patterns (see Tables S.1 and S.2).

3.2.2 Robustness and Classification of Self-Centered Punishment Types

How robust are our type classifications? Note first that the strategy-method introduces, by design, random variation in g_j . This renders the estimated coefficients from (3) fairly insensitive to adding further control variables (e.g., controls for contributions g_k and g_l , $k \neq l \neq j$).¹⁶ Obviously, this statement does *not* imply that eq. (3) is the ‘best model’ to describe individual punishment patterns. Moreover, it does *not* imply that we consider efficiency motives (to punish any deviation from the socially optimal behavior, i.e., from contributing the full endowment) as the primary driver of peer punishment. The point is simply that our primary classification approach yields quite robust results (see Section II in the Online Appendix).

Motivated by earlier findings in the literature, we nevertheless explore one alternative, more refined approach to classify patterns of peer-punishment. More specifically, we account for the fact that individual i ’s disposition to punish might be ‘self-centered’ around her own contribution (from the first stage of the P-game). We thus consider a model that allows punishment patterns to differ between the domain $g_j < g_i$ and $g_j \geq g_i$:

$$d_{ij} = \alpha'_i + \beta'_{i1} \max(g_j - g_i; 0) + \beta'_{i2} \max(g_i - g_j; 0) + \varepsilon'_i. \quad (4)$$

Based on the estimates from this model (in particular, $\hat{\beta}'_{i1}$ and $\hat{\beta}'_{i2}$) one can differentiate between numerous, refined patterns of punishment (see Figure A2 in the Appendix). Our refined classification will again focus on a limited set of types. (1) Subjects with a significant pro-social punishment slope in the domain $g_j < g_i$ (i.e., $\beta'_{i2} > 0$) but an insignificant slope coefficient for $g_j \geq g_i$ (i.e., β'_{i1} with $p > 0.01$) are classified as ‘Self-centered Punishers’ (*SPun'*-types). (2) All further subjects with statistically significant, positive slope coefficients (i.e., with two positive β -estimates or only $\beta'_{i1} > 0$) are subsumed in a class of (further) ‘Pro-social Punishers’ (*Pun'*-types). In addition, we distinguish between (3) ‘Non-Punishing’ (*NPun'*; defined as above) and (4) ‘Anti-socially Punishing’ (*APun'*) types. The residual category are again non-classified patterns (*NCL'*-types).¹⁷

Table 1 compares the results from this more refined type classification approach to our primary classification outcomes. Among the 296 subjects labeled as pro-socially punishing (*Pun*) according to our primary classification, 270 (more than 91%) do in-fact display a self-centered pattern of punishment (*SPun'*). Together with 11 further subjects (that were *NCL* according to our primary approach) we obtain at a total of 281 self-centered punishment patterns (44.7% of all 628 subjects). Beyond this large group, there are only six pro-social *Pun'*-types with patterns

¹⁶A classification that builds, for instance, on the estimates $\hat{\alpha}_i$ and $\hat{\beta}_{i1}$ from the equation $d_{ij} = \alpha_i + \beta_{i1}(20 - g_j) + \beta_{i2}(20 - g_k) + \beta_{i3}(20 - g_l) + \varepsilon_i$ differs for a mere 11 subjects (1.8% of our sample). Adding dummies that capture the sequence at which subject i faced a certain triple does not change this picture.

¹⁷In terms of the stylized illustration in Figure A2, the pattern from panel *b* would classify as *SPun'*-type; the pattern illustrated in panels *a*, *c*, and *d* would be all subsumed as *Pun'*-types; and, finally, the patterns from panels *a.i* to *d.i* would be classified as *APun'*-types.

that are not self-centered (as defined above). Hence, almost all pro-social punishment in our sample is self-centered and these self-centered patterns constitute the most frequent punishment type. We will return to these findings below.

Table 1: Primary and Refined (Self-Centered) Classification of Punishment Types

<i>Primary Classification</i> ↓	<i>Refined Type Classification</i>					<i>Sum (N)</i>
	Pun'	SPun'	NPun'	APun'	NCL'	
Pun	6	270	0	0	20	296
NPun	0	0	253	0	0	253
APun	0	0	0	5	11	16
NCL	0	11	0	0	52	63
Total	6	281	253	5	83	628

Notes: Row values display the outcome from our primary classification approach based on eq. (3). Column values depict classification results for the refined ('self-centered') approach that builds on equation (4).

Regarding the other types we observe only minor changes in the number of anti-social punishers and a modest increase in the number of non-classified patterns (from 63 in our primary to 83 in our refined classification approach). The latter observation is related to a conceptual limitation of the refined classification method: for very low or very high contributions g_i , there will be few observations with $g_j < g_i$ or $g_j \geq g_i$, respectively. This certainly reduces the scope to obtain precise estimates for both β coefficients of eq. (4).¹⁸

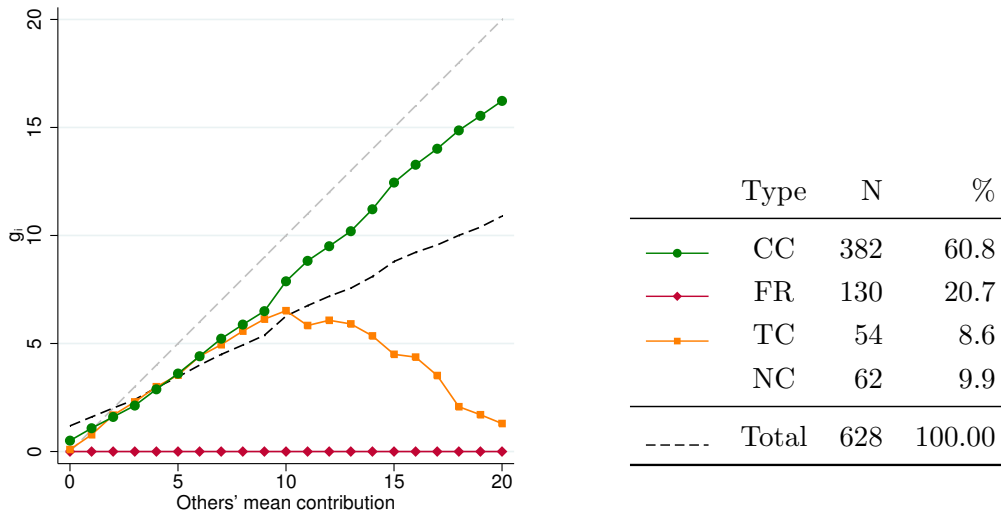
The methodological complications associated with the refined classification approach, together with the fact that pro-social punishment follows almost unanimously a self-centered pattern (which, in turn, nullifies the scope for distinguishing among different types of pro-social punishment), motivates us to work with the primary type classification throughout the remainder of the paper. Given the results from Table 1, however, it is unsurprising that the results from the subsequent analyses are qualitatively insensitive to using the type distribution from the self-centered classification approach.

3.3 Individual Cooperation Patterns

Next we analyze 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

¹⁸In an attempt to cope with this limitation, we applied the classification only to individuals with $5 \leq g_i \leq 15$. For this range of contributions, we obtain a very similar pattern as the one displayed in Table 1. A further issue concerns the unbiasedness of the estimates: including g_i in the estimation model mechanically re-introduces endogeneity. To see this point, note that all unobserved individual factors ψ_i that shape punishment d_{ij} are absorbed in the error term ε'_i . As these unobserved factors ψ_i will also influence g_i , we obtain $\text{Cov}(g_i, \varepsilon'_i) \neq 0$. One can show, however, that this will mainly bias the estimates for α'_i .

Figure 2: 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.

the public good (Fischbacher et al., 2001). Based on these data we classify individual cooperation types.

Consistent with our approach from above we separately estimate for each subject i the linear model $g_i = a_i + b_i \bar{g}_j + e_i$ (with $\bar{g}_j := \frac{1}{3} \sum_{j \neq i} g_j$). 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 2 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 (2010): 61% are conditional cooperators and 21% are free-riders. The remaining 18% display a triangular or a non-systematic contribution pattern.¹⁹

3.4 Two-Dimensional Type Distribution

Finally, we combine the results from subsections 3.2 and 3.3 to arrive at a two-dimensional type classification, which links punishment and cooperation patterns at the individual level. In this vein, we can examine the relationship between individuals' disposition to cooperate and their inclination to engage in punishment. Table 2 presents the results from the two-way classification.

¹⁹Classifications based on Spearman's rank correlation (as in Fischbacher et al., 2001) yield almost identical results. (See Online Appendix, Table S.3.)

Table 2: Two-way Distribution: Contribution and Punishment Types

Contrib. Types ↓	Punishment Types				Sum	
	<i>Pun</i>	<i>NPun</i>	<i>APun</i>	<i>NCL</i>	(%)	(<i>N</i>)
CC	33.4	21.2	1.4	4.8	60.8	382
FR	6.5	11.6	0.3	2.2	20.7	130
TC	4.3	2.9	0.0	1.4	8.6	54
NC	2.9	4.6	0.8	1.6	9.9	62
Sum (%)	47.1	40.3	2.6	10.0	100.0	
Sum (<i>N</i>)	296	253	16	63		628

Notes: Within subject two-dimensional contribution and punishment type distribution in percent, for 628 subjects respectively. *N* shows the absolute type distribution per game.

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 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 ($FR \times Pun$).

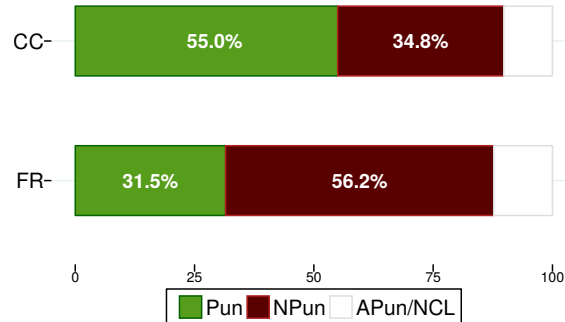
A different way of presenting the distribution of these four types — which cover almost three out of four subjects in our sample — is provided in Figure 3. The bar graphs indicate that roughly every second conditional cooperator punishes pro-socially (55%) and that more than one out of two free-riders 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. Hence, the overlap between conditionally cooperative and (pro-social) punishing individuals is far from perfect.

3.4.1 Further 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 A3 in the 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 again far from perfect. Among $CC \times Pun$ -types, for instance, we observe an insignificant correlation coefficient of $\rho = 0.094$ ($p = 0.173$).²⁰

²⁰The Spearman correlation is slightly stronger (0.128) and statistically significant ($p = 0.064$).

Figure 3: Conditional Distribution of Punishment-Types among *CC*- & *FR*-Types



Notes: The graph depicts the conditional frequency of *Pun*- and *NPun*-types among conditional cooperators (*CC*) and free-riders (*FR*), respectively.

Making use of the data from our questionnaire, 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, we find that subjects who express their willingness to impose social sanctions on norm violators among their peers (e.g., drunk drivers; see the survey questions in [Traxler and Winter, 2012](#)) are significantly more likely to be pro-social punishers (*Pun*-types). This observation suggests that the survey measure on norm enforcement is consistent with the behavioral measure that builds on the observed pattern of peer punishment. Second, we find that subjects who see themselves as 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.²¹

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.

²¹Probit and LPM estimates underlying these results are available from the authors.

3.4.2 Interim Summary

Summing up, our two-dimensional classification reveals the existence of two behavioral archetypes. First, the largest group of subjects is characterized by an overlap in their pro-social behavioral patterns. This group includes conditionally cooperative types that do engage in pro-social punishment ($CC \times Pun$) and free-riders that do not invest in punishment at all ($FR \times NPun$). For these types, cooperation and punishment are indeed “two sides of the same coin” (Ones and Putterman, 2007).

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 ($FR \times Pun$). About every third conditional cooperator and, analogously, every third free-rider displays such a divergence in cooperation and punishment patterns. The identification of this second archetype therefore suggests that cooperation and punishment may indeed be separate phenomena — at least for some individuals (see Peysakhovich et al., 2014).²² While the latter finding seems interesting in itself, it further implies that individual inclinations to cooperate and to punish are far from 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.

4 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 with peer punishment. 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). We analyze the influence of group compositions on group outcomes and individual behavior. Motivated by other studies which document the benefits from grouping pro-social individuals in a repeated VCM (e.g., Gächter and Thöni, 2005; Ones and Putterman, 2007), we start out by computing the number of conditional cooperators (CC) and pro-social punishers (Pun) for each group (and for each matching group of eight in

²²One 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 3.2.2, 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.

4.1 Descriptive Evidence

A first glimpse at the results is provided by Figure 4. It depicts the average contribution per group over 10 rounds for different group compositions.²³ Panel A [B] 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 C [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 C 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 B and D further shows that the differences among ‘top’ groups are less pronounced than in the R_p -game. 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.²⁴

4.2 Regression Analysis: Group Contributions

Figure 4 shows that the number of both *CC*- and *Pun*-types are important determinants of average contributions 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

$$\bar{g}_{\ell t} = \gamma_0 + \gamma_1 CC_{\ell}^{few} + \gamma_2 CC_{\ell}^{many} + \sum_t \delta_t D_t + \epsilon_{\ell t}, \quad (5)$$

where $\bar{g}_{\ell t} := \frac{1}{n} \sum_{i=1}^n g_{i\ell t}$ is the average contribution in group ℓ in period t . The explanatory variables are dummies indicating if there are few (one or two) or many (three or four) *CC*-types in a group.²⁵ In addition, the specification accounts for period-fixed effects. The results from linear random-effects estimations of equation (5) for the 113 groups in the partner design (R_p -game) are presented in column (1) of Table 3.²⁶

²³To ease exposition, the figure pools groups with similar type compositions. The raw data are illustrated in the Online Appendix (see Figure S.4).

²⁴Figure S.2 in the Online Appendix replicates Figure 4 for average group *payoffs* rather than contributions. This exercise delivers similar findings as those discussed above.

²⁵The reference category are groups with zero *CC*-types. In the interpretation of the point estimates discussed below, one should keep in mind that most groups are populated by at least some conditional cooperators (see Table A1 in the Appendix).

²⁶Tobit estimations yield almost identical results (see the Online Appendix, Table S.7).

Consistent with the graphical evidence from above, and again 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 (5), is only weakly significant. A Wald test further rejects $\gamma_1 = \gamma_2$ with $p = 0.003$.

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 (with the two estimates being significantly different from each other; $p = 0.000$). Note further that all information criteria reported in Table 3 indicate that the estimated model in column (2) clearly dominates the one from column (1): the R^2 strongly increases and the Akaike information criterion (AIC) declines, indicating a better model fit. This underlines the usefulness of information about the number of *Pun*-types in a group 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. These two dummies are certainly correlated; nevertheless, the significant share of subjects with diverging cooperation and punishment patterns (see above) in combination with our fairly large sample allows us to distinguish the role of *CC*- and *Pun*-types.

The results reported in column (3) show that the model clearly yields a better fit than the one using only information about *CC*-types (column 1); however, R^2 and AIC only improves modestly as compared to the specification from column (2). Put differently: once we account for a group's *Pun*-types, adding information about *CC*-types only weakly increases explanatory power. The results further indicate 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 $CC \times Pun$ -types (in the spirit of an interaction term). The outcome shows that, for a given number

Table 3: Group Composition and Average Contributions

	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
	<i>Partner Design</i>				<i>Stranger Design</i>			
	<i>Dependent variable: Average Group Contribution (\bar{g}_{it})</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

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.

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, the AIC 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.²⁷

In a next step, we consider the data from the stranger design. Columns (5)–(8) in Table 3 present the estimation output from an analogous set of regressions as those discussed above. The results are similar 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

²⁷Post-estimation tests following specifications (3) and (4) reject $\gamma_1 = \gamma_2$ ($p = 0.013$ and $p = 0.039$, respectively) and, analogously, the equality of the two Pun -dummies ($p = 0.000$). However, we cannot reject $CC^{few} = Pun^{few}$ and $CC^{many} = Pun^{many}$ ($p = 0.732$ and $p = 0.708$, respectively).

dummies for both types, only the ones on the *Pun*-types remain significant. In addition, the point estimates for the *CC*-dummies become much smaller now. In fact, post-estimation tests for specifications (7) and (8) both reject $CC^{\text{few}} = Pun^{\text{few}}$ ($p = 0.026$ and $p = 0.069$, respectively) as well as $CC^{\text{many}} = Pun^{\text{many}}$ ($p = 0.026$ and $p = 0.089$, respectively). The analysis therefore confirms the picture from above: in the presence of peer punishment, having more *Pun*-types in a group seems to be key for achieving high contribution levels, in particular, in the stranger design.

4.3 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

$$g_{it} = \lambda_0 + \lambda_1 CC_{\ell}^{\text{few}} + \lambda_2 CC_{\ell}^{\text{many}} + \lambda_3 Pun_{\ell}^{\text{few}} + \lambda_4 Pun_{\ell}^{\text{many}} + \phi Pun_i + \sum_t \delta_t D_t + \epsilon_{it}, \quad (6)$$

The first set of dummies now captures whether individual i faces few or many *CC*- or *Pun*-types among the *other* players in her group ℓ .²⁸ The λ -coefficients thus reflect the impact from variation in the type composition among i 's peers on her contribution. 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. Finally, note that we estimate equation (6) *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. Moreover, any unconditional differences among these two contribution types will be reflected in different constants (λ_0).

The results from estimating eq. (6) for the partner design are presented in Table 4. Let us first focus on the estimates 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: post-estimation tests reject $\lambda_1 = \lambda_2$ ($p = 0.080$) and $\lambda_3 = \lambda_4$ ($p = 0.001$). In terms of statistical and economic significance, however, an increasing

²⁸More precisely, in the partners protocol, the dummies capture if there are few (one) or many (two or three) *CC*- or *Pun*-types among the other three players in the group. For the strangers design, the dummies with superscript few [many] indicate that two to four [five or more] subjects out of the seven other players in the matching group were classified as *CC*- or *Pun*-type, respectively.

Table 4: Group Composition and Individual Contributions (Partner Design)

	(1)	(2)	(3)	(4)	(5)	(6)
	<i>Conditional Cooperators (CC)</i>			<i>Free-Riders (FR)</i>		
	<i>Dependent variable: Individual Contribution (g_{it})</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

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.

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.²⁹

Estimations for the *CC*-types in the stranger design, which are presented in Table 5, show 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. In addition, we can reject $\lambda_1 = \lambda_3$ ($p = 0.008$) and find borderline evidence when testing $\lambda_2 = \lambda_4$ ($p = 0.1364$).

²⁹The Wald tests again reject $\lambda_3 = \lambda_4$ ($p = 0.001$). However, we do not detect a significant difference between the coefficients on CC^{many} and Pun^{many} ($p = 0.270$).

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 for obtaining high contribution levels among conditional cooperators.

Next we turn to the results for free-riders. Overall, the estimates from columns (4)–(6) in Tables 4 and 5 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 and somewhat under-powered. For the partner design, columns (4) and (5) of Table 4 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 is insignificant (but not statistically different from Pun^{few} ; testing $\lambda_3 = \lambda_4$ yields $p = 0.144$). Column (6), which presents the estimates for equation (6), 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.³⁰ 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 5 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_ℓ dummies lose significance whereas the Pun_ℓ dummies remain large and highly significant. Post-estimation tests reject $\lambda_3 = \lambda_4$ ($p = 0.008$) as well as $\lambda_1 = \lambda_3$ ($p = 0.008$) and $\lambda_2 = \lambda_4$ ($p = 0.033$).

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 point to a clear

³⁰In almost all cases when the Pun^{many} dummy is equal to one, CC^{many} is one, too.

Table 5: Group Composition and Individual Contributions (Stranger Design)

	(1)	(2)	(3)	(4)	(5)	(6)
	<i>Conditional Cooperators (CC)</i>			<i>Free-Riders (FR)</i>		
	<i>Dependent variable: Individual Contribution (g_{it})</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

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.

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 highlight 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)

A last point worth discussing is the fact that the estimates from Tables 4 and 5 allow for a comparison of the average contributions among the different types introduced in our type classification from above (see, e.g., Figure 3). To see this, one has to recognize that the constant term (λ_0 from equation 6) 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 heterogeneous 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.

5 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 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' dispositions to cooperate and their inclinations 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. Hence, for a majority of subjects cooperation and punishment indeed seem like two sides of the same coin. However, for a significant part of our sample, cooperation and punishment seem to be different behavioral traits.

The divergence between cooperation and punishment patterns further 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 with peer punishment. Our analyses provided strong, causal evidence on the relative importance of pro-social punishers for achieving and maintaining cooperation. While variation in cooperation types within a (matching) group explains large parts of the variance in group outcomes, similar variation in punishment types has a higher explanatory power.

The latter finding is relevant, 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- (e.g., Falk and Szech, 2013) and anti-social (e.g., Abbink and Serra, 2012) behavior.

The results and the methodologies from our study open several avenues for follow-up research. To advance our understanding of cross-cultural differences in cooperation (Henrich et al., 2006; Herrmann et al., 2008), one could readily apply our approach to examine the underlying variation in individual cooperation and punishment types. Exploring type variation in social dilemmas beyond linear public goods games (see, e.g., Cason and Gangadharan, 2015) 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). Depending on the parametrization of the game, self-centered models of inequality aversion might 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 a subject’s beliefs about others’ punishment — future research might address this 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.
- Andreoni, J. (1988). Why free ride?: strategies and learning in public goods experiments. *Journal of Public Economics* 37(3), 291–304.
- 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.
- Cason, T. and L. Gangadharan (2015). Promoting Cooperation in Nonlinear Social Dilemmas through Peer Punishment. *Experimental Economics* 18, 66–88.
- Chaudhuri, A. (2011). Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature. *Experimental Economics* 14(1), 47–83.
- Cheung, S. L. (2014). New insights into conditional cooperation and punishment from a strategy method experiment. *Experimental Economics* 17(1), 129–153.
- Dohmen, T., A. Falk, D. B. Huffman, and U. Sunde (2008). Representative Turst 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. and B. Herrmann (2009). Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* 364(1518), 791–806.
- 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. (2015). Subject pool recruitment procedures: organizing experiments with ORSEE. *Journal of the Economic Science Association* 1(1), 114–125.
- Güerker, 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.
- 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.
- Kamei, K. (2014). Conditional punishment. *Economics Letters* 124(2), 199–202.
- 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.
- Leibbrandt, A. and R. López-Pérez (2012). An exploration of third and second party punishment in ten simple games. *Journal of Economic Behavior & Organization* 84, 753–766.
- 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.
- Oliver, P. (1980). Rewards and punishments as selective incentives for collective action: Theoretical investigations. *American Journal of Sociology* 85(6), 1356–1375.

- Ones, U. and L. Putterman (2007). The ecology of collective action: A public goods and sanctions experiment with controlled group formation. *Journal of Economic Behavior & Organization* 62(4), 495–521.
- Ostrom, E., J. Walker, and R. Gardner (1992). Covenants With and Without a Sword: Self-Governance is Possible. *American Political Science Review* 86(2), 404.
- Peysakhovich, A., M. A. Nowak, and D. G. Rand (2014, 09). Humans display a ‘cooperative phenotype’ that is domain general and temporally stable. *Nature Communications* 5, 4939.
- 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.
- Rustagi, D., S. Engel, and M. Kosfeld (2010). Conditional cooperation and costly monitoring explain success in forest commons management. *Science* 330(6006), 961–5.
- 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.

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 A1 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 1 and 2, 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 2) we observe that the majority of groups are populated by two (35 groups) or three (48 groups) *CC*-types. In addition, there are 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

Table A1: 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>	

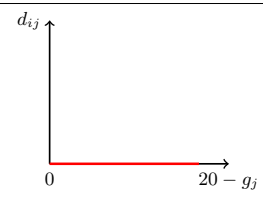
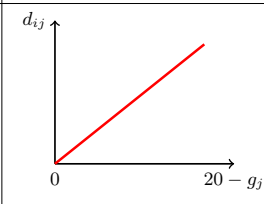
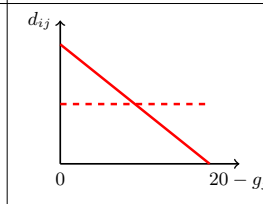
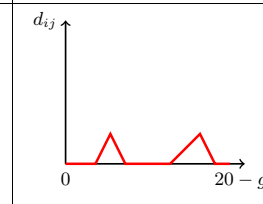
Notes: In the R_p -game subjects are counted at the *group level* (with 4 subjects per observational unit). In the R_s -game subjects are counted at the *matching-group level* (with 8 subjects per observational unit). The depicted distribution of subjects occurred from randomly assigning subjects to 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 1 and 2, respectively.

prevalence is close to one half (47.1 % of our sample, see Table 1). 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 A1 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.

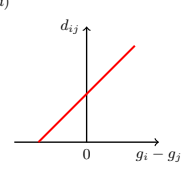
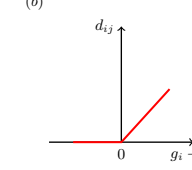
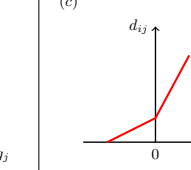
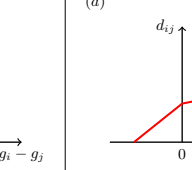
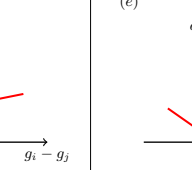
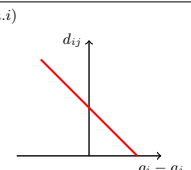
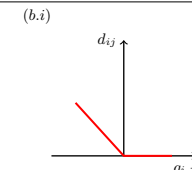
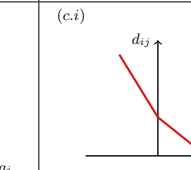
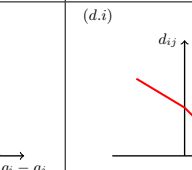
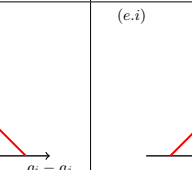
A3 Complementary Figures

Figure A1: Primary Classification Approach: 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.	

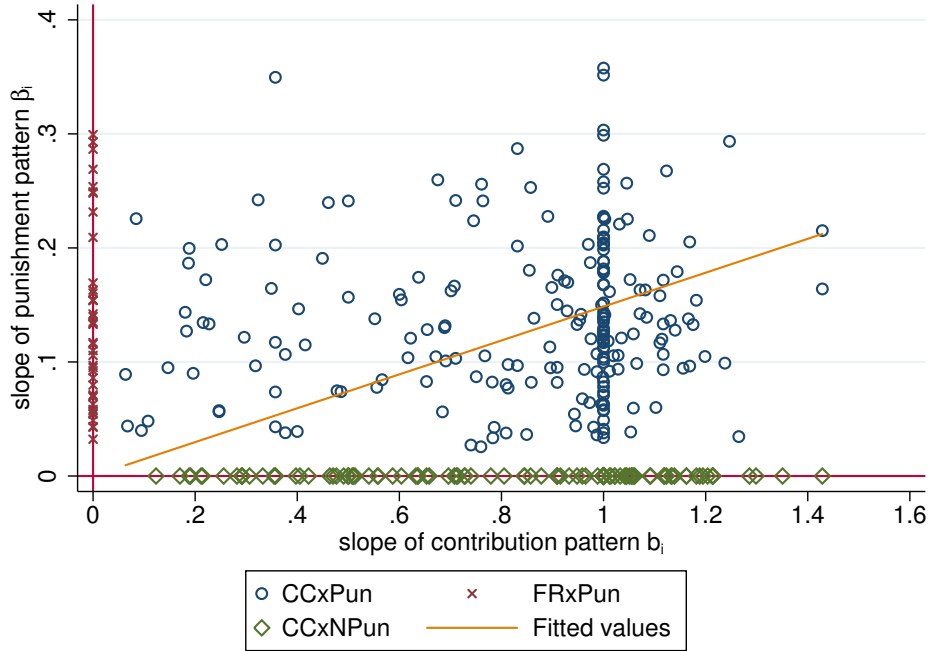
Notes: Type classifications based on $\hat{\alpha}$ and $\hat{\beta}$ obtained from estimating eq. (3).

Figure A2: Self-Centered Classification Approach: Stylized Illustration of Punishment Types

(a) 	(b) 	(c) 	(d) 	(e) 
$\hat{\beta}'_1 = \hat{\beta}'_2 > 0$ with $p \leq 0.01$	$\hat{\beta}'_1 \approx 0; \hat{\beta}'_2 > 0$ with $p \leq 0.01$	$\hat{\beta}'_2 > \hat{\beta}'_1 > 0$ with $p \leq 0.01$	$\hat{\beta}'_1 > \hat{\beta}'_2 > 0$ with $p \leq 0.01$	$\hat{\beta}'_1 < 0; \hat{\beta}'_2 > 0$ with $p \leq 0.01$
(a.i) 	(b.i) 	(c.i) 	(d.i) 	(e.i) 
$\hat{\beta}'_1 = \hat{\beta}'_2 < 0$ with $p \leq 0.01$	$\hat{\beta}'_1 < 0; \hat{\beta}'_2 \approx 0$ with $p \leq 0.01$	$\hat{\beta}'_1 < \hat{\beta}'_2 < 0$ with $p \leq 0.01$	$\hat{\beta}'_2 < \hat{\beta}'_1 < 0$ with $p \leq 0.01$	$\hat{\beta}'_1 > 0; \hat{\beta}'_2 < 0$ with $p \leq 0.01$

Notes: Different types based on $\hat{\beta}'_1$ and $\hat{\beta}'_2$ from estimating eq. (4). Panels *a* to *d* show different patterns of pro-socially punishing subjects. Panels *a.i* to *d.i* display anti-social patterns. The self-centered pattern *b* would be classified as *SPun'*, whereas pattern *a* would be labeled *Pun'*. As long as $\hat{\beta}'_1$ is not significantly positive, pattern *c* would be classified as *SPun'*, too. Empirically, we hardly observe patterns as those from panels *c* and *d*. Patterns *e* and *e.i* are mechanically possible but are not observed in our sample. (Patterns for *NPun'*- or *NCL'*-types are not illustrated here.)

Figure A3: 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* × *NPun*-type values are not plotted. The concentration of observations at $\hat{b}_i = 0$ and $\hat{b}_i = 1$ is due to ‘perfect’ free-riders and ‘perfect’ conditional cooperation, respectively. The former never contribute in the C-game, whereas the latter types perfectly (i.e., 1:1) match the average group contribution.

Supporting Online Material

Cooperation and Norm Enforcement - The Individual-Level Perspective

Felix Albrecht^{†,‡}, Sebastian Kube^{‡,‡}, Christian Traxler^{*,‡}

† University of Marburg; ‡ University of Bonn; ‡ Max Planck Institute for Research on Collective Goods;

* Hertie School of Governance

Content

I. Instructions

II. Sensitivity of Type Classification Approaches

III. Group Composition and Payoffs in the Repeated Game

IV. Complementary Tables

I Instructions

Below we provide the instructions, translated into English, as they were handed out and read aloud (in German) to the subjects. The first part of the instructions describes the C-game, the second the P- and the third part the R-game (R^P and R^S , respectively).

General Instructions for Participants

You are about to take part in an economic experiment. If you read the following instructions carefully, you can earn a considerable amount of money, depending on your decisions and the decisions of the other participants. It is therefore important that you read these instructions carefully and understand them well.

During the experiment, communication is absolutely forbidden. If you have any questions, please ask only us. Raise your hand and we will come to you. Disobeying this rule will lead to exclusion from the experiment and from all payments.

This experiment consists of several independent parts. You will be randomly matched into groups of four in each part. **The make-up of your group of four will change as each new part begins.** Participants cannot be identified beyond the individual parts, and you do *not* interact with the same participants in each part of the experiment.

For your participation today, you will initially receive a show-up fee of 5€. This amount increases by your earnings from the individual parts of the experiment. During the experiment, however, we will not speak of Euro, but of Token. Your total earnings are therefore initially calculated in Token. The total amount of Token you earn in the course of the experiment will be converted into Euro at the end and paid to you in cash. The exchange rate of Token to Euro will be told to you at the beginning of each part.

Now you will receive a description of the first part. You will receive the descriptions for the other parts later.

General Information on the First Part of the Experiment

In the first part of the experiment, the exchange rate from Euro to Token is: 5 Token = 1€.

The first part of the experiment consists of **one period only**. At the beginning of the first part, you will be assigned randomly to a group of four participants. Your group therefore consists of three other participants.

Each participant receives 20 Token. It is your task to decide how you will use your 20 Token. You can contribute all or part of your 20 Token to a **project**, or else put them in a **private account**. Each Token that you do not put towards to the project is automatically put in your private account by you. For instance, if your contribution to the project is 5 Token, then 15 Token remain in your private account.

Income from the private account:

For every Token that you put into your private account, you will earn exactly 1 Token. For example, if you put 20 Token in your private account (thus contributing nothing to the project), you will earn exactly 20 Token from the private account. If, for example, you contribute 12 Token to the project (thus putting 8 Token in your private account), you will earn 8 Token from your private account. *Nobody except you receives earnings from your private account.*

Income from the project:

For every Token that you or another participant from your group contributes to the project, *you and all other participants in your group* will earn 0.4 Token each. The income of each participant in your group from the project is therefore determined as follows:

$$\text{Income from the project} = \text{Sum of contributions to the project} * 0.4$$

Examples: If the sum of the contributions to the project by all participants from your group is 20 Token (e.g., if you and the three other participants each contribute 5 Token), you and all the other participants in your group receive $20 * 0.4 = 8$ Token from the project. If the sum of the contributions to the project is 10 Token in total, then you and all the other participants earn $10 * 0.4 = 4$ Token from the project.

Your income from the first part is the sum of your income from your private account and your earnings from the project. Therefore:

$$\begin{array}{l} \text{Income from your private account} (= 20 - \text{Contribution to the project}) \\ + \text{Income from the project} (= 0.4 * \text{Sum of contributions to the project}) \\ \hline \text{Income from the first part of the experiment} \end{array}$$

The calculations can be illustrated easily with an example:

You contribute 15 Token to the project, as do the other three participants. The total sum of contributions to the project is therefore $15 + 15 + 15 = 60$ Token. Your income in the example would be:

$$\underline{5 \text{ Token}} \text{ from your private account} + \underline{0.4 * 60 \text{ Token}} \text{ from the project} = 5 + 24 = \underline{29 \text{ Token}}.$$

However, if you contributed 0 Token to the project, for example, the total sum of contributions to the project would be $15 + 15 + 15 + 0 = 45$ Token. Your income would therefore be:

$$\underline{20 \text{ Token}} \text{ from your private account} + 0.4 * \underline{45 \text{ Token}} \text{ from the project} = 20 + 18 = \underline{38 \text{ Token}}.$$

The earnings for the other participants are calculated in the same way.

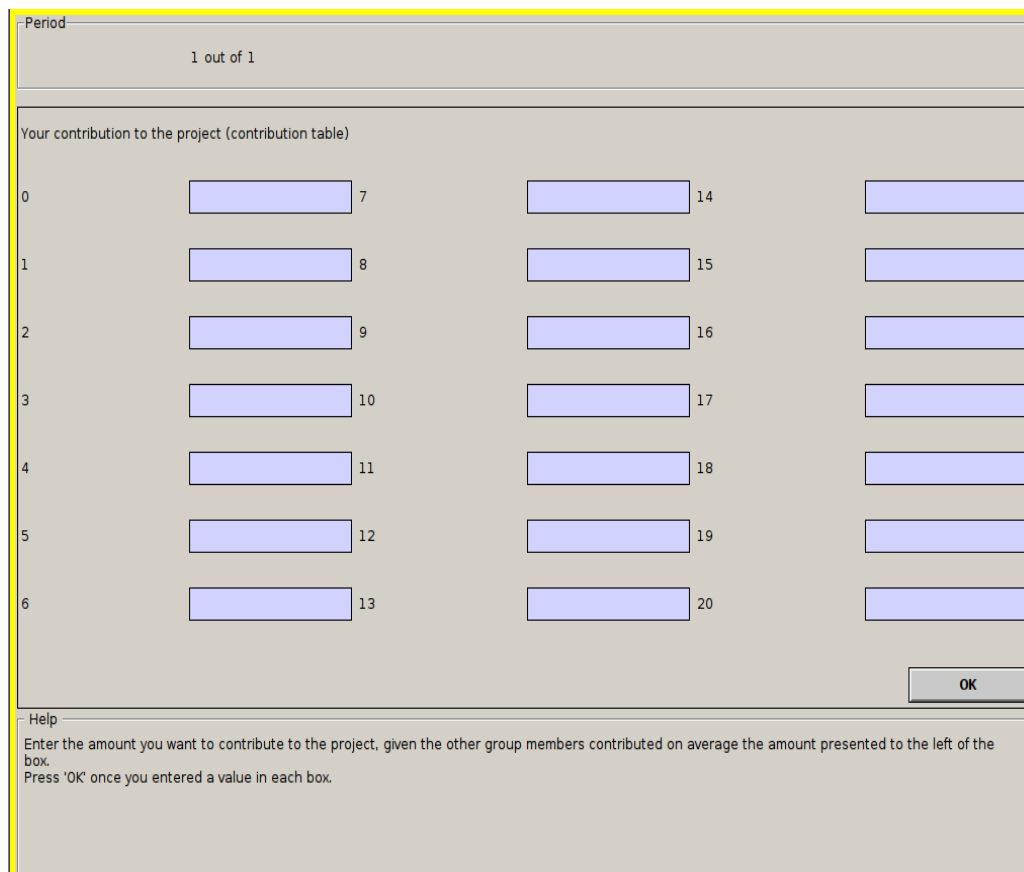
Do you have questions?

Additional Instructions for the First Part of the Experiment

You make your contribution decision as follows in the first part of the experiment:

First, you have to decide how many Token you wish to contribute to the project. From this point on, we will call this the **unconditional contribution**.

Afterwards you have to fill in a **contribution table**. In the contribution table, you have to **enter how many Token you want to contribute to the project for every possible (rounded) average contribution of the other participants in your group**. So you can decide how many Token you want to contribute, depending on how many Token the others have contributed on average. To understand this better, please take a look at the following screen. It will appear directly after you have made you unconditional contribution decision.



Other participants' average contribution	Total contribution to project	Your contribution
0	7	<input type="text"/>
1	8	<input type="text"/>
2	9	<input type="text"/>
3	10	<input type="text"/>
4	11	<input type="text"/>
5	12	<input type="text"/>
6	13	<input type="text"/>
	14	<input type="text"/>
	15	<input type="text"/>
	16	<input type="text"/>
	17	<input type="text"/>
	18	<input type="text"/>
	19	<input type="text"/>
	20	<input type="text"/>

Help
Enter the amount you want to contribute to the project, given the other group members contributed on average the amount presented to the left of the box.
Press 'OK' once you entered a value in each box.

The numbers to the left of the blue input fields on the screen present the potential rounded average contributions to the project by the **other** participants in your group. Now, simply enter, into every input field, how many Token **you** wish to contribute to the project – assuming that the other participants have contributed on average the depicted amount. **You have to make an entry into each input field**. You have to enter, for example, how many Token you contribute to the project if the other participants of your group contribute on average 0 Token to the project; how many Token you contribute if the others contribute on average 1, 2, or 3 Token, and so forth. You can enter **all** whole numbers from 0 to 20 into each field.

After all participants have made their unconditional contribution decision and filled in the contribution table, one participant is chosen at random from each group. For the **randomly chosen participant** only the completed **contribution table** is relevant for both the decision and the payoff. For the **other three participants** in your group, who have not been chosen randomly, only the **unconditional contribution** is relevant for both the decision and the payoff. An example illustrates this:

Example: Assume that **you have been chosen randomly, so that your contribution table is relevant for your decision**. Thus, for the other three participants, only the unconditional contribution decision is relevant. Assume this is given by 0, 2, and 4. The average contribution of these three participants is therefore 2.

If you have stated in your contribution table that you will contribute 1 in case the others contribute 2 on average, then the total group contribution to the project is $0 + 2 + 4 + 1 = 7$. All participants in your group thus earn $0.4 * 7 = 2.8$ Token from the project, plus the individual earnings from each private account.

On the other hand, if you stated in your contribution table that you would contribute 19, in case the others contribute 2 on average, the total contribution to the project is $0 + 2 + 4 + 19 = 25$. All participants in your group thus earn $0.4 * 25 = 10$ Token from the project, plus the individual earnings from each private account.

Only at the end of the third part of the experiment do you learn whether the contribution table or the unconditional contribution was relevant for you and how high your payoff is from this first part.

Do you have questions? If you do, please raise your hand now.

General Information on the Second Part of the Experiment

In the second part of the experiment, the exchange rate from Euro to Token is: 5 Token = 1€.

The second part of the experiment consists of **one period only**. At the beginning of the second part, you are once again assigned randomly to a group of four participants.

The decision situation is similar to the situation in part one; however, **in this part, an additional stage is introduced**. The process is now as follows:

In the first stage, as before, you have to decide how many Token you wish to contribute to a project.

Your income at the end of stage one is the sum of your income from the private account and your income from the project. Therefore:

$$\begin{array}{l} \text{Income from your private account (= 20 - Contribution to the project)} \\ + \text{Income from the project (= 0.4 * Sum of contributions to the project)} \\ \hline \text{Income from the first part of the experiment} \\ \hline \hline \end{array}$$

STAGE 2

At the beginning of stage 2, you will learn how many Token the other participants in your group have contributed to the project. You will then have the opportunity to **reduce** the stage 1 earnings of **each one** of the other participants in your group. The other participants can similarly reduce **your** earnings, if they choose to.

To reduce the earnings of a specific participant, you can assign so-called **points** to this participant.

For each point that you assign to a participant in your group, you reduce the earnings of this participant by 3 Token. Thus, if you assign 1 point to a participant, you reduce the earnings of this participant by 3 Token. If you assign 2 points to this participant, you reduce this participant's earnings by 6 Token, etc. If you do not want to reduce the earnings of a participant, assign 0 points to this participant.

The more points you assign to a participant, the larger is the reduction in the earnings of this participant. However, your own earnings are also reduced with every point that you assign to a participant. For each point that you assign, your earnings are reduced by 1 Token. For example, if you assign 2 points to a participant, you will incur costs of 2 Token; if you assign 4 points to a participant, you will incur costs of 4 Token; if you assign 0 points to a participant, you will incur no costs for this.

You decide for **each** participant in your group by how many Token you want to reduce his earnings. You may assign a maximum of 10 points to each participant.

If and how many Token in total are deducted from a participant's earnings depends not only on how many points you assigned to this participant, but also the other participants' points. For example, if a participant receives 1, 0, and 2 points, respectively, from the other three participants in the group, his earnings are reduced by $(1 + 0 + 2) * 3 = 9$ Token. Simultaneously, the earnings of the other participants are reduced because of the costs incurred by assigning the points by 1, 0, and 2 Token.

YOUR PAYOFF

Your payoff is thus determined as follows:

<p><i>Earnings from stage 1</i></p> <ul style="list-style-type: none">- $3x$ (The number of points from stage 2 that have been assigned to you)- The number of points from stage 2 that you have assigned to others <hr/> <p><u><i>Payoff</i></u></p>

Additional Instructions for the Second Part of the Experiment

You make your decisions in the second part of the experiment as follows:

First, you decide *once* how many Token you want to contribute to the project in the first stage.

In stage 2, you are confronted with a number of decision situations. In each decision situation, a combination of **possible** contributions by the other participants in your group is presented. Above, we pointed out that you will learn the precise contributions of the other three participants in your group in stage 2 – and after that you can assign points to each participant. However, in this part of the experiment, the three presented contributions might possibly be fictitious and do not represent the actual contributions of the other three participant.

After you decided about the assignment of points to the presented contributions, you will be presented with another (possibly fictitious) combination of contributions by the other participants in your group. For this decision situation, you also have to decide how many points you want to assign to each participant.

In total, you will be presented with **eleven** decision situations. Ten of these eleven decision situations are fictitious. In **exactly one** situation, you will be presented with the **actual** contributions of the other three participants in your group. How many points you assign to the other three participants in your group, and how large your payoff will be, will only be determined by the decisions in this one decision situation. The chosen assignment of points in the fictitious situations has no influence on your payoff or on that of the other participants. When deciding on the assignment of points in a decision situation, you will not know if the presented contributions are the actual contributions. Therefore you have to consider your assignment of points in every decision situation, as every situation might be relevant for you.

You will learn which situation was the actual situation and how big your earnings are from the second part of the experiment at the end of the third part of the experiment.

Do you have questions? If so, please raise your hand now.

General Information for the Third Part of the Experiment

In the third part of the experiment, the exchange rate from Euro to Token is: 50 Token = 1€.

The third part of the experiment consists of **ten periods**. At the beginning of the third part, you will again be assigned randomly to a group of four participants. The composition in all ten periods stays the same, which means **you will interact with the same participants in each of the 10 periods**.

The general decision situation is the same as in the second part of the experiment in each period, i.e., you will decide, in stage 1, how many Token you wish to contribute to a project; in the second stage, you can assign points to the other participants in your group. For each point that you assign to a participant, you reduce the earnings of this participant by 3 Token, and your own earnings by 1 Token.

In the **first period**, you will be confronted with eleven decision situations in stage 2. Ten out of the eleven decision situations are made-up. In exactly one situation, you will be presented with the actual contributions of the other three participants. Your payoff from the first period will only be determined by the decisions in this one decision situation (you know this already from the previous part of the experiment).

At the end of the first period, all participants in your group will learn how many points they have received from the other participants.

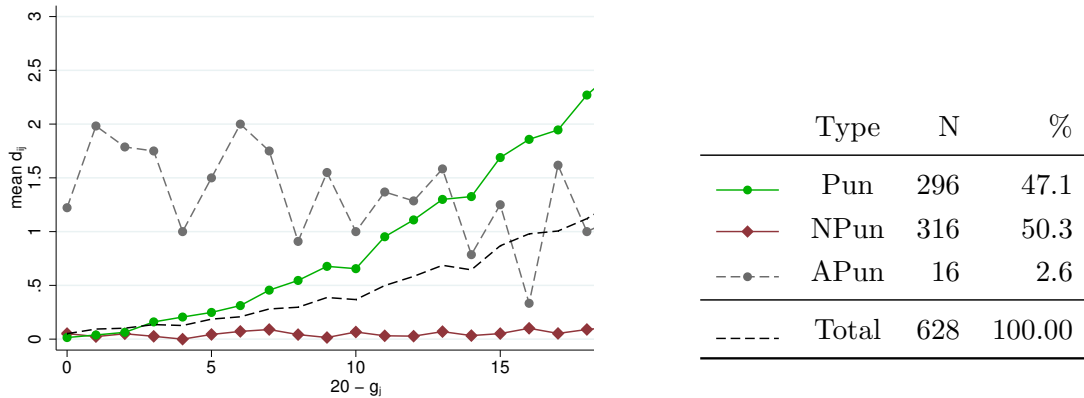
In the **subsequent nine periods**, and this is different from the previous part of the experiment, you will interact another nine times with the **same** participants. However, in the subsequent nine periods, you will **only be confronted with the actual** decision situation in stage 2.

After each of the ten periods, each participant will learn how many points he has received from the other participants in the group. Further, he will learn his payoff from this period. After this, each participant is given a new, randomly drawn number. You are therefore always with the same participants in one group, but cannot identify the individual participants from round to round.

Do you have questions? If so, please raise your hand now.

II Sensitivity of Type Classification Approaches

Figure S.1: Punishment Patterns and Punishment Types – Sensitivity Analysis



Notes: Primary type distribution and average punishment patterns (in the $20 - g_j$ -space) for different punishment types: pro-social punishers (*Pun*), non-punishers (*NPun*), anti-social punishers (*APun*). In contrast to our main approach, *NPun* is here classified as those with insignificant (rather than exact zero-) coefficients: $\hat{\alpha}_i \approx \hat{\beta}_i \approx 0$. With this alternative definition of *NPun*, all subjects that were classified as *NCL* in Figure 1 now fall into the extended *NPun* definition.

Table S.1: Primary Punishment Classification using Spearman's ρ

Type	N	%
Pun	307	48.9
NPun	253	40.3
APun	6	0.9
NCL	62	9.9
Total	628	100.00

Notes: Primary punishment type classification applying Spearman's rank correlation in line with the conditional cooperation classification proposed by Fischbacher et al. (2001). *Pun* are classified as positive ρ with $p \leq 0.01$. Subjects are *NPun* if all peer punishment decisions $d_{ij} = 0$. *APun* are classified as negative ρ with $p \leq 0.01$. All subjects that are not classified as one of the previous three, are classified as *NCL*.

Table S.2: Comparing Primary Punishment Classifications: Spearman’s ρ vs. OLS estimates

Spear. $\rho \downarrow$	OLS				Sum (N)
	Pun	$NPun$	$APun$	NCL	
Pun	293	0	0	14	307
$NPun$	0	253	0	0	253
$APun$	0	0	6	0	6
NCL	3	0	10	49	62
Total	296	253	16	63	628

Notes: Comparison of our primary peer punishment classification using Spearman’s ρ and the OLS estimates. The classification based on Spearman’s ρ shows only minor differences to our classification approach from the main text.

Table S.3: Contribution Type Classification using Spearman’s ρ

Type	N	%
CC	388	61.8
FR	130	20.7
TC	54	8.6
NC	56	8.9
Total	628	100.00

Notes: Contribution type classification applying Spearman’s ρ as proposed by Fischbacher et al. (2001). CC are classified as positive ρ with $p \leq 0.01$. Subjects are FR if all 21 conditional contribution decisions $g_i = 0$. TC -types initially show a positive relation to the average contributions and a decreasing slope in the latter part of the graph and are classified via eyeballing.

Table S.4: Comparing Contributor Classification Spearman’s ρ and OLS

Spear. $\rho \downarrow$	OLS				Sum
	<i>CC</i>	<i>FR</i>	<i>TC</i>	<i>NC</i>	(<i>N</i>)
CC	381	0	0	7	388
FR	0	130	0	0	130
TC	0	0	54	0	54
NC	1	0	0	55	56
Total	382	130	54	62	628

Notes: Comparison of individual conditional cooperation classification using Spearman’s ρ and our approach (based on OLS estimates). The Spearman’s ρ based classification shows only minor differences compared to our classification approach from the main text.

Table S.5: Comparing P- & R-game Punishment Classification

P-game \downarrow	R-game				Sum
	Pun	NPun	APun	NCL	(<i>N</i>)
Pun	177	16	4	99	296
NPun	45	105	5	98	253
APun	2	2	5	7	16
NCL	18	12	3	30	63
Total	242	135	17	234	628

Notes: Within subject comparison of primary punishment type classifications using P- vs. R-game data. Here we apply our classification approach (based on estimating the equation $d_{ij} = \alpha_i + \beta_i(20 - g_j) + \varepsilon_i$) to the observational data on cooperation and punishment in the repeated game (R-game). The two approaches show strong deviations in classification outcomes. Based on the R-game data, 234 subjects (more than a third of all) remain unclassified and get labeled as *NCL*.

III Group Composition and Payoffs in the Repeated Game

Figure S.2 replicates Figure 4 for the average group *payoffs*. The exercise delivers similar patterns as those discussed above. It is remarkable to note that the gains from higher contributions that groups with many *Pun*-types manage to achieve, are hardly offset by the costs from having more punishment. These findings concur with those in Fehr and Gächter (2000). In groups with 3 or 4 *Pun*-types the average payoff over all periods is 27.7 tokens, with an average of 29.2 during the last five periods. (Keep in mind that the maximum achievable group payoff is 32 tokens.) Groups with 3 or 4 *CC*-types, in contrast, end up with an average payoff of 26.8, and 27.7 during the last five periods. The payoff differences are significant at the 10 percent level (two-sided t-tests).

Next we run regressions to explore the role of *Pun*-types for a group’s average payoff. We build on equation (5) and use a (matching-)group’s average payoff, $\bar{\pi}_{it}$, as an alternative dependent variable. Estimation results for the partner and the stranger design are presented in Panel A of Table S.6. Consistent with the positive effect of *CC*-types on group contributions, column (1) shows positive and highly significant coefficients. In column (2), we find still positive but smaller coefficients for the *Pun*-type dummies. Only the dummy for 3 or 4 *Pun*-types is statistically significant. The point estimates from columns (1) and (2) imply that a group with 3 or 4 *CC*- [*Pun*-] types achieves a payoff per period that is on average 5.3 [3.7] tokens higher than in a group with zero *CC*- [*Pun*-] types. The estimates thus offer a slightly different picture than the one discussed in the main text: regarding average contributions, we have seen that having more subjects that punish pro-socially was unambiguously ‘better’ than having more conditional cooperators. For achieving higher payoffs, however, the positive role of *Pun*-types is limited by the fact that their (stronger) inclination to punish — which is instrumental for reaching high contribution levels — is costly and *cet.par.* lowers average group payoffs. This is why *CC*-types have a stronger positive effect on average payoffs, an observation that is further supported by the results from column (3).

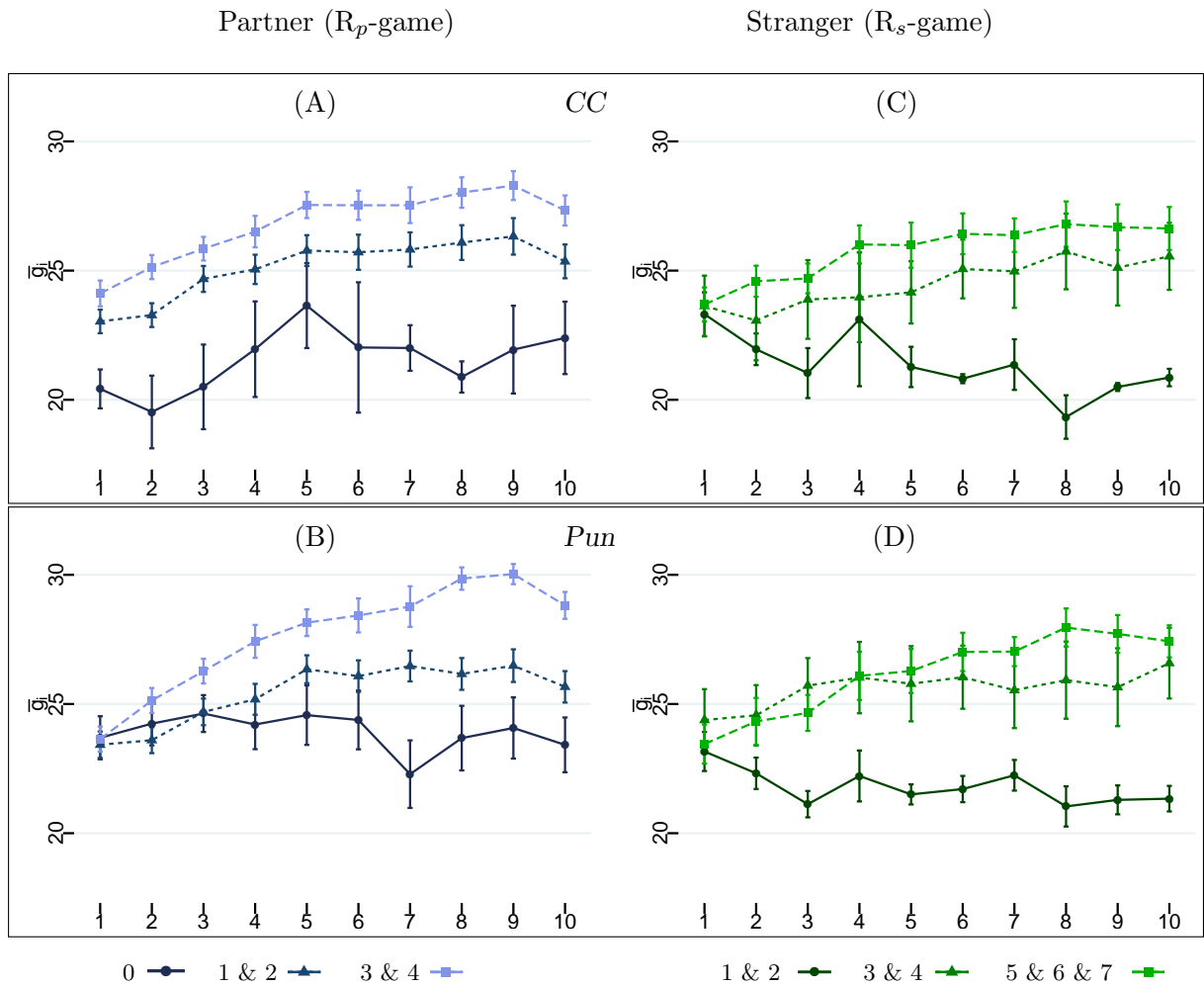
Three things are worth noting. First, the results for the stranger design, which are presented in columns (4)–(6) of Panel A in Table S.6, show again a much more positive impact of *Pun*- as compared to *CC*- types. In fact, column (6) reveals statistically significant coefficients for the two *Pun* but not for the *CC* dummies. The relative costs from having more *Pun*-types therefore seem to be higher in stable groups of four (partner design) as compared to matching groups of eight — a point we will return to below. Second, similar as above, the specification from column (2) outperforms the one from column (1), and the one from column (5) does better than the one from (4) in terms of explanatory power and information criteria. Thus, knowledge about the prevalence of *Pun*-types is still crucial for explaining variation in payoffs across heterogeneous groups. Thirdly, when we replicate the estimates from column (1)–(3) for the last five periods of the R_p -game, we observe again larger and more precisely estimated positive coefficients on the *Pun* dummies for the partner design. (These estimates are presented in Panel B of Table S.6.) Hence, excluding the early periods — where most punishment occurs (see Figure S.3) — the regressions capture again the beneficial, cooperation inducing effect of having more pro-social punishers in a group.

Table S.6: Group Composition and Average Payoffs

Panel A. All 10 Periods							
	<i>Partner Design</i>			<i>Stranger Design</i>			
	(1)	(2)	(3)	(4)	(5)	(6)	
CC^{few}	3.580*** (1.159)		2.989** (1.351)	CC^{few}	3.158** (1.274)	-0.112 (1.137)	
CC^{many}	5.254*** (1.137)		4.367*** (1.333)	CC^{many}	4.431*** (0.818)	1.150 (0.882)	
Pun^{few}		1.493 (1.011)	1.026 (1.074)	Pun^{few}		3.826*** (1.221)	
Pun^{many}		3.734*** (0.965)	3.051*** (1.054)	Pun^{many}		3.954*** (1.200)	
					4.401*** (0.716)	3.803*** (0.900)	
Obs.	1,130	1,130	1,130	Obs.	220	220	220
R^2	0.135	0.146	0.189	R^2	0.224	0.371	0.394
AIC	6472	6457	6402	AIC	1103	1057	1052
Panel B. Last 5 Periods							
	(1)	(2)	(3)	(4)	(5)	(6)	
CC^{few}	4.008*** (1.305)		3.166* (1.624)	CC^{few}	4.714*** (1.234)	1.205 (1.122)	
CC^{many}	5.889*** (1.265)		4.537*** (1.590)	CC^{many}	6.006*** (0.791)	1.769** (0.734)	
Pun^{few}		2.605** (1.237)	2.137 (1.338)	Pun^{few}		4.423*** (1.363)	
Pun^{many}		5.612*** (1.184)	4.920*** (1.298)	Pun^{many}		3.979*** (1.381)	
					5.906*** (0.794)	5.137*** (0.904)	
Obs.	565	565	565	Obs.	110	110	110
R^2	0.0795	0.149	0.190	R^2	0.268	0.499	0.515
AIC	3339	3294	3270	AIC	556	514	514

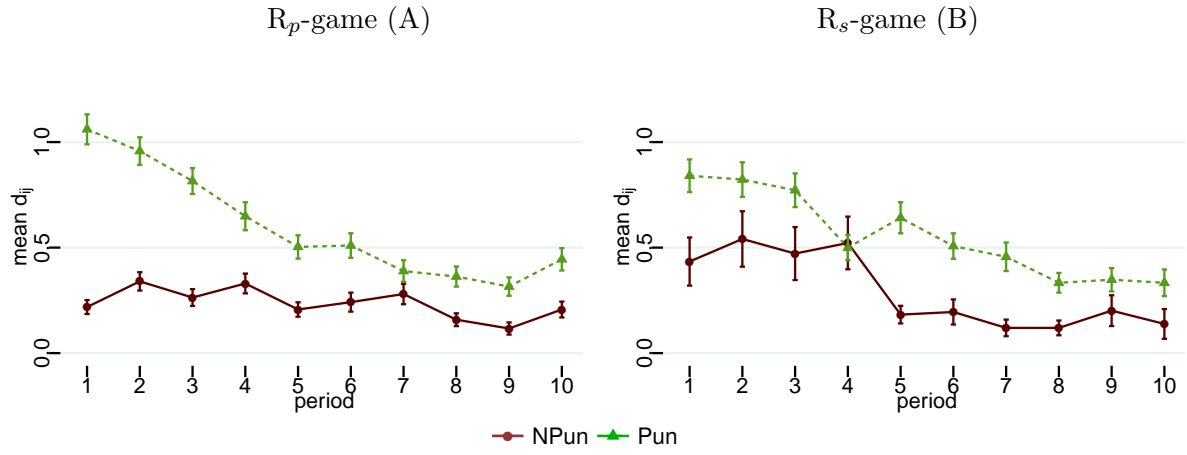
Notes: Estimates from linear random-effects models. Dependent variable: average (matching-)group payoff per period. Columns (1)–(3) for the R_p -game, columns (4)–(6) for the R_s -game. Panel A considers all 10 periods, the lower Panel B only the last 5 periods of the game. Dummies with superscript ‘*few*’ indicate that one or two [three or four], and dummies with ‘*many*’ indicate that three or four [five, six, or seven] subjects in the respective [matching]-group in models 1 to 3 [4 to 6] are *CC*- or *Pun*-type. In columns (1)–(3) the number of observations is $N = 1,130$ (Panel A) and $N = 565$ (Panel B; 113 groups of the partner design \times 10 and 5 periods, respectively). In columns (4)–(6) it is $N = 220$ and $N = 110$ (22 matching-groups of the stranger design \times 10 or 5 periods, respectively). 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.

Figure S.2: Average (Matching)-Group Payoffs by Type Prevalence



Notes: Panels A and B [C and D] show the average payoff 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 A1.

Figure S.3: Mean Observed Punishment per Period for Pun- & NPun- Types

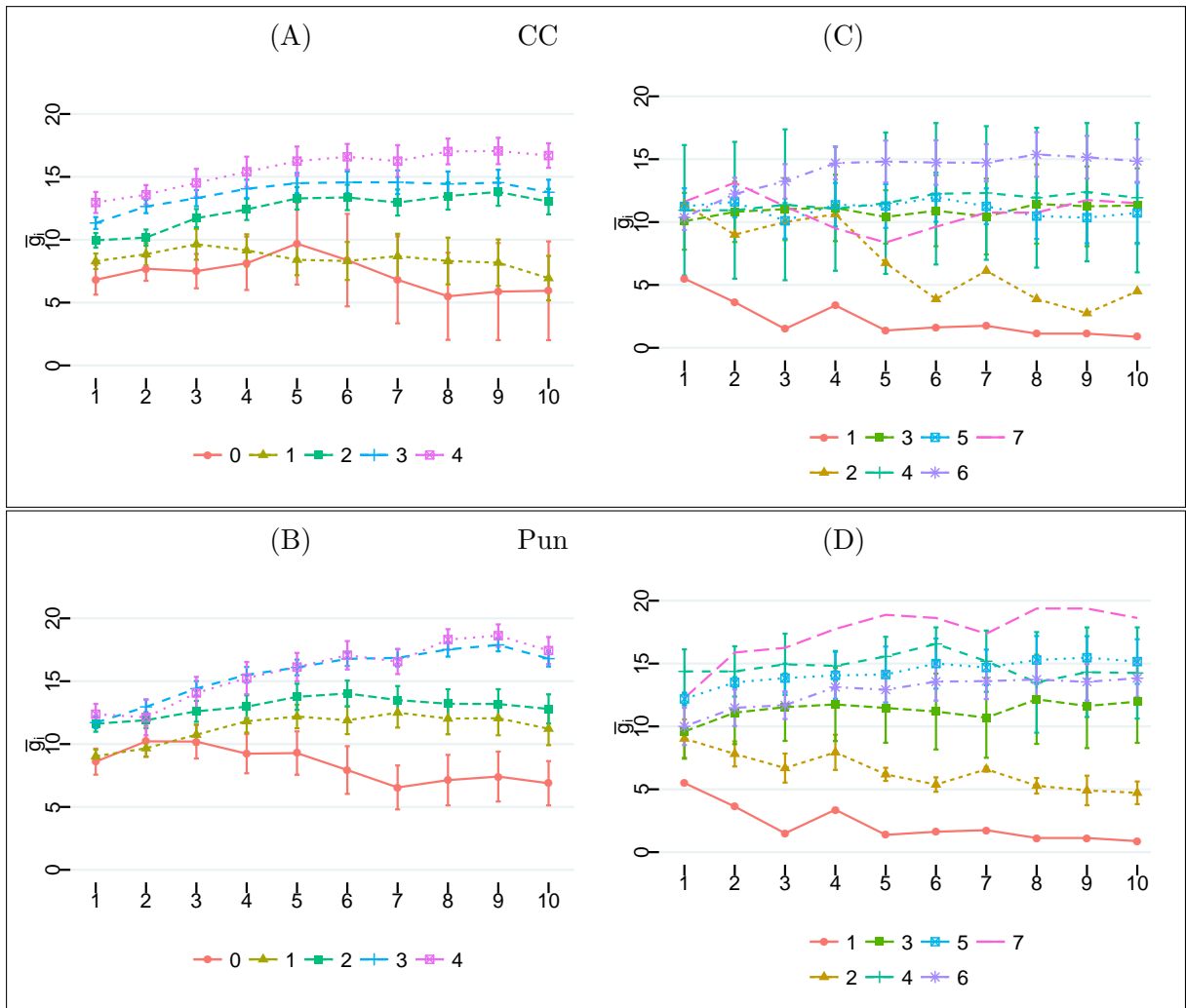


Notes: Mean observed punishment by Pun and NPun over the 10 periods of the R-game.

Figure S.4: Average (Matching)-Group Contributions by Type Prevalence

Partner (R_p -game)

Stranger (R_s -game)



Notes: Graphs depict the average individual contribution per period averaged over all (matching)-groups containing the corresponding number of subjects of a respective type (CC; Pun). Given the distribution of (matching) group compositions (See table A1.) the graph contains average contribution levels of single groups.

IV Complementary Tables

Table S.7: Group Composition and Group Contributions: Tobit Estimates (Partner Design)

	(1)	(2)	(3)	(4)
CC^{few}	4.395 (2.829)		3.353 (2.620)	3.384 (2.749)
CC^{many}	7.645*** (2.806)		5.853** (2.614)	6.044** (2.915)
Pun^{few}		4.356*** (1.498)	3.551** (1.466)	3.646* (1.876)
Pun^{many}		7.937*** (1.618)	6.838*** (1.594)	7.184*** (2.339)
$Pun \times CC^{few}$				-0.183 (1.690)
$Pun \times CC^{many}$				-0.662 (2.711)
AIC	5361	5352	5347	5350

Notes: Estimates from random-effects Tobit models (with a lower bound at zero (23 obs.) and an upper bound at 20 (186 obs.)). Dependent variable: average group contribution per period. The observational unit is a group (of 4 subjects) per period. Dummies with superscripts ‘*few*’ indicate one or two, with superscripts ‘*many*’ three or four CC , Pun , and $Pun \times CC$ type subjects per respective group. Number of observations: $N = 1,130$ (113 groups of the partner design \times 10 periods). All specifications include a constant and a full set of period-fixed effects (coefficients not reported). Standard errors in parentheses; *** / ** / * indicate significance at the 1%-, 5%-, and 10%-level, respectively.

Table S.8: Group Composition and Group Contributions: Tobit Estimates (Stranger Design)

	(1)	(2)	(3)	(4)
CC^{few}	6.602** (3.318)		1.053 (3.270)	0.134 (3.364)
CC^{many}	8.221*** (3.072)		2.130 (3.074)	-0.101 (3.719)
Pun^{few}		6.867*** (2.024)	6.517*** (2.416)	5.631** (2.447)
Pun^{many}		8.193*** (1.803)	7.224*** (2.196)	5.520** (2.377)
$Pun \times CC^{few}$				2.690 (2.926)
$Pun \times CC^{many}$				5.166 (3.646)
AIC	972	963	967	968

Notes: Estimates from random-effects Tobit models (with a lower bound at zero (0 obs.) and an upper bound at 20 (5 obs.)). Dependent variable: average matching-group contribution per period. Dummies with superscript ‘*few*’ indicate that three or four, and dummies with ‘*many*’ indicate that five, six, or seven subjects in the respective matching-group are *CC*- or *Pun*-type. For two-dimensional types superscripts ‘*few*’ indicate two or four and ‘*many*’ five or six $Pun \times CC$ per respective matching-group. The observational unit is a matching group (8 subjects) per period. Number of observations: $N = 220$ (22 matching-groups of the stranger design \times 10 periods). All specifications include a constant and a full set of period-fixed effects (coefficients not reported). Standard errors in parentheses; *** / ** / * indicate significance at the 1%-, 5%-, and 10%-level, respectively.