Exp Econ
DOI 10.1007/s10683-010-9257-1

Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature

Ananish Chaudhuri

Received: 29 June 2009 / Accepted: 14 September 2010

Abstract I survey the literature post Ledyard (Handbook of Experimental Economics, ed. by J. Kagel, A. Roth, Chap. 2, Princeton, Princeton University Press, 1995) on three related issues in linear public goods experiments: (1) conditional cooperation; (2) the role of costly monetary punishments in sustaining cooperation and (3) the sustenance of cooperation via means other than such punishments. Many participants in laboratory public goods experiments are "conditional cooperators" whose contributions to the public good are positively correlated with their beliefs about the average group contribution. Conditional cooperators are often able to sustain high contributions to the public good through costly monetary punishment of free-riders but also by other mechanisms such as expressions of disapproval, advice giving and assortative matching.

Keywords Public goods · Conditional cooperation · Monetary punishments · Non-monetary punishments · Moral suasion · Sorting

JEL Classification C71 · C91 · C92

1 Introduction

I provide an overview of developments in the experimental literature on linear public goods games since the survey undertaken by Ledyard (1995).[1,2] I will structure my

[1] The dilemma inherent in this game is highlighted by Olson (1965). See Andreoni (1988), Isaac et al. (1984, 1985), Isaac and Walker (1988a, 1988b), Kim and Walker (1984) and Marwell and Ames (1979, 1980, 1981) for some of the early experimental studies in this area.

[2] Ledyard (1995, p. 112) provides a succinct description of how a generic linear continuous public goods game is implemented. Most readers will be well acquainted with this. Here I provide a brief description

A. Chaudhuri (✉)
Department of Economics, University of Auckland,
655 Owen G Glenn Building Level 6, 12 Grafton Road, Auckland 1142, New Zealand
e-mail: a.chaudhuri@auckland.ac.nz

Published online: 29 September 2010

survey around three distinct but related ideas. These are: (1) findings regarding the heterogeneity of social preferences among participants and in particular the concept of *conditional cooperation*; (2) the use of costly monetary punishments in sustaining cooperation in this game and (3) the fact that cooperation can also be sustained via mechanisms such as expressions of disapproval, advice giving and assortative matching.[3]

1.1 Where things stood circa 1995

Ledyard (1995) starts by providing an overview of the extant experimental literature on public goods games till the mid 1990s starting with the influential early work undertaken in the area by Bohm (1972, 1983) as well by Robyn Dawes and John Orbell and their colleagues (for instance Dawes 1980; Dawes et al. 1977, 1986; Orbell et al. 1990), by Gerald Marwell and Ruth Ames (for instance Marwell and Ames 1979, 1980) and by Mark Isaac and James Walker (such as Isaac and Walker 1988a, 1988b) and their colleagues as cited above. The findings of this line of work suggested that: (1) In one-shot versions of the public goods game, there is much more contribution than predicted in the Nash equilibrium of the game. Groups of participants on average contribute between 40% and 60% of the optimal level with wide variations in individual contributions ranging from 100% contribution by some to 0% by others; but (2) if the players interact repeatedly over a number of rounds then contributions often start out at between 40% and 60% of the social optimum and decline steadily over time as more and more players choose to "free ride."

Ledyard then goes on to identify a number of factors that enhance cooperation which include (1) communication, particularly that dealing directly with the problem at hand; (2) the inclusion of a threshold and/or provision point and (3) the magnitude of the MPCR; a higher MPCR increases contributions. He also discusses other factors that might be expected to play a role but have little effect overall such as gender or

mostly as a reminder of terms and concepts that I will use in the rest of the paper. In a one-shot play of the game, a group of n participants is told that each of them has an endowment of ω tokens. Each participant i must make a contribution decision C $(0 \leq C \leq \omega)$ on how many of those tokens she wants to contribute to a public account. Any remaining tokens are allocated to the private account. Contributions by all the participants in a group are made simultaneously, without any communication and typically in whole tokens. In addition to the tokens allocated to the private account, each participant i receives a fixed percentage (α) of the total group contribution to the public account, where $0 < \alpha < 1 < n\alpha$. The term α is referred to as the *marginal per capita return (MPCR)* from the public good. At the end of a round participants either get to see the individual contributions made by members of the group or the total (and therefore average) contributions to the public account without learning the identity of the group members. Each participant's personal earning is the sum of the tokens kept in the private account plus the return from the public account. If the game is repeated finitely then this one-shot game is played over a number of rounds where each successive round proceeds in the same manner, starting with a new endowment of ω for each participant.

[3] Even when coverage is restricted to the period after 1995, the literature is still vast. I will focus only on those papers that look at linear continuous public goods games. I will not be discussing papers in a number of related areas such as those that analyze public goods with a provision point, prisoner's dilemma games or common-pool resource usage games. I will mostly focus on published papers, but I will discuss a handful of unpublished papers because I thought that these papers were innovative enough to warrant discussion.

the size of the group; contrary to intuition larger groups are no worse—and may even be better—at providing the public good than smaller ones. And in an anticipation of the literature that followed he also presents models of behavior which incorporate reciprocal motivations and beliefs about others' contributions.

1.2 What have we learned since then?

At the expense of over-generalization I suggest that advances have been made on two broad fronts. The first of these is a greater understanding of the fact that there are distinct types of players in such games who differ in their social preferences and/or their beliefs about their peers, either one of which may be sufficient to generate behavior not commensurate with the simple game theoretic prediction of free riding in such situations. The most notable finding in the area is that many participants behave as "***conditional cooperators***", *whose contribution to the public good is positively correlated with their beliefs about the contributions to be made by their group members.*

This idea that there may be different types of players certainly did not arise *de novo* and was foreshadowed in Ledyard's concluding section and in other previous studies. But there had not been a systematic exploration of the motivations behind these types till that point. At the time Ledyard was writing his survey it was also not entirely clear as to what factors lay behind the usual pattern of decaying contributions. This was variably attributed to kindness on the part of some and confusion on the part of others (Andreoni 1995), the "warm glow" of giving (Andreoni 1990), a combination of learning to play the dominant strategy and strategic play by self-interested players (Andreoni 1988; Andreoni and Croson 2008) or decision errors of various types (Palfrey and Prisbrey 1997 or Anderson, Anderson et al. 1998). Ledyard ends this particular section of his survey by suggesting that more research is needed to understand the reasons behind contributions decay. We have now come to realize that the usual decaying pattern of contributions can be better understood by appealing to heterogeneity in the types of players interacting with one another.

Furthermore, recent experimental research has discovered that conditional cooperators are often willing to engage in punishment of free-riders even when such punishment is personally costly and confers no long-term benefits. They are also often successful in sustaining high contributions to the public good even without such costly punishments via other mechanisms which can broadly be categorized as *moral suasion* (to borrow a term from Ledyard 1995) and/or assortative matching.[4] See for instance Axelrod (1986), Fehr and Fischbacher (2004a, 2004b, 2005a, 2005b), Fehr and Gächter (2000, 2002), and Bowles and Gintis (2002) among others. Experimental economists have in turn used their understanding of such type heterogeneity to design better institutions that can help sustain cooperation. This is the part of the literature that I focus on in the rest of this article.

[4] For a shorter review of some of the issues discussed here—such as conditional cooperation and the notion of altruistic punishments—see Gächter and Thöni (2007).

The second major area of advance has been the development of theoretical models of behavior in such games based on experimental findings in the area.[5] These have in turn prompted further experiments designed to test the implications and applicability of these models to other games and situations. There are two distinct types of models that have been developed to deal with such issues: (1) models that focus on distributional concerns such as those in Fehr and Schmidt (1999) and Bolton and Ockenfels (2000); and (2) intentions based models that focus on participants' beliefs about each others' actions and a concern for reciprocity. The latter class of models include Rabin (1993) and Dufwenberg and Kirchsteiger (2004). Models that combine elements of both include Charness and Rabin (2002), Cox et al. (2007, 2008) and Falk and Fischbacher (2006). A recent theoretical paper that does not quite belong to either group is Ambrus and Pathak (2009).

Models incorporating distributional concerns such as Fehr and Schmidt (1999) assume that players are "inequity averse" in the sense that they experience "guilt" if they receive a payoff that is higher than others (advantageous inequity) while they experience "envy" if they receive a payoff that is smaller (disadvantageous inequity). Utility functions are linear in the difference between these payoffs with disadvantageous inequity leading to higher losses in utility than advantageous inequity. The Bolton and Ockenfels model also assumes that players care about relative payoff but allows for non-linearity in the utility function. However, the results presented in Charness and Rabin (2002) among others raise questions about the conclusions or broad applicability of such inequity aversion based models.

On the other hand, studies like Rabin (1993), Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) develop models that explicitly incorporate first and higher order beliefs—beliefs of player j about the strategies to be adopted by player k as well as beliefs of player j about the beliefs of player k about the strategies to be adopted by player j and so on—directly into the utility function.

A recent theoretical paper by Ambrus and Pathak (2009) extends the earlier work by Kreps et al. (1982) in explaining cooperation in finitely repeated prisoner's dilemma games. These authors assume that there are two kinds of players—purely self-interested and reciprocal. The actual proportion of each type is common knowledge and hence there is no asymmetric information about types. The main feature of the model is a reciprocity function whose arguments include past and current contributions of other group members. The authors show that as long as this reciprocity function obeys certain regularity conditions, then there is a unique sub-game perfect equilibrium in which contributions drop off over time. The model is able to explain a number of observed regularities in behavior including the "re-start effect".

In the interests of parsimony, I refrain from further elaborating on these theoretical models, which makes this survey less than exhaustive. The rest of this paper is structured in the following way. Section 2 discusses the issue of conditional cooperation. Section 3 examines the role of costly monetary punishments. Section 4 looks at how

[5] These theoretical advances quite possibly deserve a separate survey of their own at this point. The chapter on other-regarding preferences by Cooper and Kagel (2009) forthcoming in the second volume of *The Handbook of Experimental Economics* discusses some of these models.

mechanisms other than monetary punishments can also sustain cooperation among either sorted or non-sorted groups. Section 5 concludes.[6]

2 Conditional cooperation

In this section I focus on experimental studies that explore the phenomenon of "conditional cooperation".[7] If we assume standard preferences then free-riding is the dominant strategy equilibrium in one-shot plays of the game; free-riding is also the subgame perfect equilibrium in finitely repeated games and evolutionarily stable. However alternative models of social preferences that incorporate beliefs and/or inequity aversion show how conditionally cooperative behavior can emerge in this game.

Rabin (1993), building on the pioneering work of Geanakoplos et al. (1989) develops the concept of "fairness equilibrium" which explicitly incorporates the beliefs that players hold about others' actions. Rabin shows that once we allow for reciprocal motivations on the part of the players, then it is possible to think of the public goods game as a *coordination problem* with full contribution being an efficient equilibrium and full free-riding an inefficient equilibrium with other equilibria in between.

One drawback of Rabin's model is that it does not apply to sequential move games as in finitely repeated linear public goods games. Both Dufwenberg and Kirchsteiger (2004) as well as Falk and Fischbacher (2006) provide a generalization of Rabin's approach and develop the concept of "sequential reciprocity" which extends the notion of conditional cooperation to finitely repeated games. The difference between these papers lies in the fact that (1) Falk and Fischbacher incorporate both distributional concerns as well as beliefs about others' strategies and (2) Dufwenberg and Kirchsteigher allow for a more sophisticated Bayesian belief updating rule with players maximizing utility at each node of the game while in Falk and Fischbacher (2006) only initial beliefs enter into the utility function.

Conditional cooperation can also arise if participants are inequity averse as in Fehr and Schmidt (1999) or Bolton and Ockenfels (2000). A participant with such preferences will contribute more if he believes that his peers will contribute more due to his concern for equity in payoffs.

There are a number of studies that provide evidence of such conditional behavior using a variety of different means. One of the first direct experimental tests of this type of conditional cooperation is provided by Fischbacher et al. (2001), henceforth referred to as FGF.[8] Given that this is a paradigmatic and widely replicated study, I will dispense with a detailed discussion.

[6] Many of the papers discussed combine multiple mechanisms and could fit into more than one section. In such cases I have included those papers in the section that I deemed most appropriate.

[7] For another review of the phenomenon of conditional cooperation using both lab and field experiments, see Gächter (2007). This paper also discusses some policy implications of conditional cooperation in the areas of tax evasion, tax morale, contributions to charity, etc.

[8] Probably, the earliest study examining the concept of conditional cooperation is Kelley and Stahelski (1970) which focuses on the phenomenon in the context of prisoner's dilemma games. However, Bryan and Test (1967) can also be construed as a study that examines the phenomenon of conditional cooperation as embodied in the decision to help only when others are doing so, except Bryan and Test do not elicit any information regarding beliefs.

FGF find that 50% of their participants are conditional cooperators. However the slope of the contributions profile for these participants lies below the 45 degree line. This implies that while these participants are willing to match the average contribution expected of others in the group, they do not quite match the group average dollar for dollar, but exhibit a small amount of "self-serving" bias. FGF go on to argue that it is the heterogeneity in player types that provides a rationale for why contributions decay over time. They suggest that any given unsorted group of participants in an experiment consists of both conditional co-operators and free-riders. Conditional co-operators with optimistic beliefs regarding the contributions to be made by their peers will contribute to the public account. But over time as they begin to discover the heterogeneity in types and particularly the presence of free-riders they will reduce their contributions leading to the decaying pattern in contributions.

However, a number of other studies also explore such conditionally cooperative behavior and were published within close proximity of each other and of FGF. In all likelihood the first authors to actually use the term "conditional cooperation" were Sonnemans et al. (1999). They extend the classic "partners" versus "strangers" paradigm introduced by Andreoni (1988). Participants play for 36 rounds in groups of four. Group composition remains constant for a minimum of three and a maximum of 12 rounds. Groups change only gradually, with at most one subject at a time leaving the group with replacement and the timing of departure is common knowledge. This design feature is crucial because a subject leaving a group is guaranteed not to interact with his former group members again and so there is no incentive to engage in strategic behavior in the last period before leaving the group, while strategic behavior is still possible in other periods. Hence, this design is a combination of a partners and a "real" strangers treatment. It enables a comparison "between subjects" by looking at participants who leave a group and those who stay and "within subjects" by analyzing how the same person behaves in the last period in an old group and the first period in a new group.

The authors find evidence of strategic behavior in that there is a sharp decline in contributions in the last period prior to leaving a group but they also find evidence of conditional behavior in that subjects who expect others to contribute also contributed themselves.

Keser and van Winden (2000) replicate Andreoni's (1988) study using the "partners" versus "strangers" paradigm. Their main contribution lies in recognizing that participants do not behave either as free-riders or as altruists but rather in an inherently conditional manner. Participants use information about the average group contributions as an anchor for their own future contributions. In both treatments, about 80% of the participants behave in a conditional manner; those who are above (below) the average in one round decrease (increase) their contribution in the following round.[9]

[9] Croson (2007) explores different motivations behind the desire to contribute to a public good. These include theories of commitment, altruism and reciprocity. Croson suggests that participants are primarily motivated by reciprocal tendencies. Furthermore, when participants are shown the distribution of the contributions made by other members of the group, the data suggests that the participants try to match the median or the mean contribution in the group, rather than the minimum or the maximum.

Brandts and Schram (2001) use a "contribution function" approach similar to the one developed by Palfrey and Prisbrey (1997). Here the returns from the private account and public accounts varies so that for some values of the MPCR the dominant strategy is to free-ride while for others it is to contribute the entire endowment.[10] Participants are expected to switch from the latter to the former at a particular value. Brandts and Schram argue, contrary to Palfrey and Prisbrey (1997), that decision errors cannot be the primary driving force behind cooperation and that some participants behave in a conditionally cooperative manner while some others seem to be driven by self-interest. It is the interaction between these two groups that is important in explaining the temporal pattern of contributions.

It should be noted that FGF's implementation of the public goods game is different from the usual approach, as in Sonnemans et al. (1999) or in Keser and van Winden (2000), in the sense that FGF rely on the *"strategy method"* (Selten 1967).[11] One way to think about the difference between these two approaches is the following. Keser and van Winden (2000) study reciprocal behavior by looking at the extent to which participants' behavior is conditional on the *past behavior* of their peers, with all decisions being payoff relevant. Fischbacher et al., on the other hand, look at how participants respond to their *own prior beliefs* about the behavior of their peers. In the latter case, some decisions will not affect the participants' payoff. There is controversy regarding the consistency between the "hot" responses elicited directly by observing responses to others' behavior as opposed to the "cold" responses elicited via hypothetical questions. Brandts and Charness (2009), following up on their earlier work in Brandts and Charness (2000), analyze a number of studies that compare such direct versus indirect elicitation of responses and conclude that by and large these two types of responses are consistent more often than not.

Fischbacher and Gächter (2009, 2010) extend the FGF study by asking: when a particular participant indicates that he would contribute more if his peers in the group did so too on questionnaires of the type used by FGF, do those responses match actual behavior when playing the same game for money? They analyze participants' preferences using two different experimental treatments. In the *P-experiment* participants first play a one-shot public goods game and then fill out a questionnaire stating how much they will contribute to the public good conditional on the average contribution of the other group members. In the *C-experiment* participants play 10 rounds of a linear public goods game with random re-matching. At the end of each round participants are asked to estimate the other group members' average contribution. In half of the sessions participants play the P-experiment and followed by the C-experiment (P-C treatment). This sequence is reversed in the remaining sessions (C-P treatment).

The authors find that 55% of participants are conditional cooperators while 23% are free-riders. The results show a positive and stable correlation between the beliefs and contributions for conditional cooperators in both the C- and P-experiments. Participants who are classified as conditional cooperators using the questionnaire approach in the P-experiment behave in the same way when asked to play for 10 rounds

[10] Brandts and Schram use the term "marginal rate of transformation" which is the inverse of the MPCR.

[11] The strategy method, introduced by Selten (1967) collects data by asking participants to respond to hypothetical questions. Not all of the participant responses may be payoff relevant. This has the advantage that it allows the experimenter to collect large quantities of data.

in the subsequent C-experiment. The distributions of beliefs in the P-C or the C-P treatments are not significantly different. Hence eliciting participants' beliefs after they have participated in the public goods game does not affect their preferences.

Fischbacher and Gächter go on to argue that it is the so-called "self-serving bias" in conditional cooperation that leads to contributions decay. Even if an entire group of participants consists of conditional cooperators, as long as each conditional cooperator tries to contribute a little less than others, this would lead to a fall in contributions over time. Using simulations that rely on elicited beliefs, actual contribution patterns and the belief updating rule discovered from the data (beliefs of subject j in period t are based on a convex combination of j's beliefs in period $t-1$ and the contribution of other group members in period $t-1$) they show that such "imperfect" conditional cooperation is at the heart of the decay in contributions over time.[12]

Like Fischbacher and Gächter (2009, 2010), Kurzban and Houser (2005) also look at the robustness of conditional cooperation using a more elaborate experimental design. They begin by first classifying participants into types according to their behavior in a linear public goods game. The authors then go on to observe if participants remain true to type in a different public goods game.

There are four session with 84 participants who take part in a sequence of one-shot linear public goods games in randomly formed groups of four. Each game starts with players making a decision regarding how to allocate their endowments between a private account and the public account. Each contribution decision is followed by a number of rounds each of which proceeds as follows: first, one player in each group is provided with the current aggregate contribution to the public account and is afforded an opportunity to change his allocation to the two accounts. Then the next player is given the same opportunity and so on. Payoffs to participants in each game are determined by the final allocation of tokens between the private and public accounts at the point where the game ends. The exact number of rounds that succeed each contribution decision is unknown to the participants. They only know that following the initial contribution decision each participant will get at least one chance to change her mind and that there is a 4% probability that the game will end after each subsequent decision.

Each experimental session contains at least seven such games. This multiple elicitation of contribution responses is designed to see if there is attenuation in the tendency to behave in a conditional manner over time. Participants are classified as unconditional cooperators, free-riders and conditional cooperators on the basis of a plot

[12] I should point out that while all groups seem to contain some free riders, albeit a minority, their presence is not necessary for contributions to decay. For instance a self-serving conditional cooperator with optimistic beliefs who expects group members to contribute 8 tokens might contribute $(8-\varepsilon)$ tokens while a pessimistic conditional cooperator who expects group members to contribute 2 tokens might contribute $(2-\varepsilon)$ tokens but to the former the latter will appear as a free-rider leading to an eventual decay in contributions. In fact a recent paper by Chaudhuri and Paichayonvijit (2010b) suggest that the usual pattern of decay is caused by participants essentially realizing that while many of them behave in a conditional manner, the distribution of initial beliefs is very different. Those who have optimistic beliefs contribute more and those with pessimistic beliefs contribute less. Over time the optimists reduce their contribution while the pessimists actually increase their contributions but the contribution increases coming from the latter are too small to offset the reductions from the former leading to contributions decaying over time. This explanation is at odds with the one put forth by Fischbacher and Gächter (2009, 2010) or by Ambrus and Pathak (2009).

of the participant's contributions against the average contribution to the public account he observed before making his own contribution, along the lines of FGF. 63% are conditional cooperators, 20% are free-riders and 13% are unconditional cooperators.

The authors find that these classifications are stable by having the participants take part in three additional games and find that those classified as free-riders contribute less on average than their peers, cooperators more and conditional cooperators about the same as their group members. Furthermore groups that consist of more cooperators generate higher contributions on average.

Burlando and Guala (2005) also test for the robustness of conditional cooperation by having participants take part in four different tasks: (1) the "Strategy Method" used by FGF; (2) the "Decomposed Game Technique" used by Offerman et al. (1996);[13] (3) contributions to the public account in a repeated linear public goods game played for 20 rounds and (4) a questionnaire.[14] Overall, the outcomes of the four classification tasks are consistent. 35% are classified as conditional cooperators, 18% as unconditional cooperators, 32% as free-riders and the remaining 15% could not be classified.

Two studies replicate the FGF results with a similar experimental design but a more heterogeneous participant pool. Kocher et al. (2008) recruit 36 participants, in groups of three, from three different locations—North Carolina, USA, Innsbruck, Austria and Tokyo, Japan. The majority of the participants act as conditional cooperators except that there are more of them among the US participants (81%) than among those in Austria (44%) or Japan (42%). Hermann and Thöni (2008) recruit 160 participants at four separate universities spread across Russia, two in rural areas and two in small cities. They find that overall 56% of participants behave as conditional cooperators, and only 6% as free riders. The distribution of preferences does not differ significantly across rural or urban backgrounds and socio-economic conditions do not seem to have an impact on the preference for conditional cooperation. Brandts et al. (2004) also undertake a cross-cultural study with participants from Japan, Netherlands, Spain and USA, except, rather than using the FGF procedure they use the

[13] The Decomposed Game technique (Griesinger and Livingston 1973; Liebrand 1984) has been widely used by psychologists and economists in order to measure social values and attitudes towards cooperation. Participants are asked to make 24 choices between pairs of allocations. Each participant knows that she has been paired anonymously with another participant and the pairings remain unchanged for the entire duration of the session. Each allocation consists of a number of tokens paid to the player concerned and another sum paid to the other player. The token amounts can be positive or negative. A typical choice may involve, e.g., a combination $A = (75, -130)$ vs. $B = (39, -145)$, where one must choose between gaining either 75 or 39 tokens, with related losses on the other's part of either 130 or 145 tokens. There is no feedback concerning the other player's choices. The final payoff is obtained by combining the 24 choices of each participant with those of the other player. Players are then classified into one of five possible categories—altruistic, cooperative, individualistic, competitive and aggressive—depending on their choices in this allocation task. Offerman et al. (1996) use this decomposed game technique to classify participants into separate categories. In their study 61% of participants are classified as individualistic while 27% are deemed cooperative with relatively few people in the other categories. The authors then go on to investigate the behavior of the different types in a step-level public goods game.

[14] The questionnaire included questions such as (1) "What were you trying to do in the experiment (in other words: what were your goals or objectives)?" (2) "Did you achieve your objectives?" (3) "What were the other members of your group trying to do (what were their objectives)?" (4) "What was the scope of this experiment (in other words, what were the experimenters trying to discover)?"

contributions function approach pioneered in Brandts and Schram (2001). They report little differences in contributions across these locations and a recurring finding of conditionally cooperative behavior.

Given the plurality of conditional cooperators in these studies, one interesting question to ask is what happens when conditional cooperators learn about the presence of other conditional cooperators in the group. This is what Chaudhuri and Paichayontvijit (2006) examine. They implement a between subjects treatment, where in the control treatment participants fill out a conditional cooperation questionnaire similar to FGF and for some participants it is the response on this questionnaire that is relevant. This is followed by three other treatments, where participants are provided progressively more detailed information about the presence of conditional cooperators in the group. The authors find that 62% of participants are conditional cooperators. But more interestingly, there is an increase in contributions when participants are provided information about the presence of conditional cooperators and this increase is most pronounced for the conditional cooperators themselves.[15]

2.1 Concluding remarks

The evidence presented in this section suggests that the many participants in linear public goods games are conditional cooperators whose contributions to the public good are positively correlated either with their *ex ante* beliefs about the contributions to be made by their peers or to the actual contributions made by the same. The behavior of conditional cooperators deviates substantially from the game-theoretic prediction of free-riding in this context. Furthermore, the evidence suggests that conditional cooperation is a stable preference type that is quite prevalent among participants and the phenomenon is robust across cultures and the mode of elicitation of responses.

3 Sustaining cooperation by punishing free-riders

Fehr and Gächter (2000) study the efficacy of costly punishments using both "partners" and "strangers" protocol. Participants play for 20 rounds in groups of four—the first 10 rounds without any punishment possibility and then the next 10 rounds with punishment.[16] In each round a participant has an endowment of 20 tokens and the MPCR is 0.4. This study has been widely replicated and the experimental design is

[15] For the sake of completeness I should point out another issue that has been raised in this context. This has to do with the extent to which conditional cooperation is not a stable preference type but merely a form of conformism. The tendency to copy the most prevalent behavior among a group is a robust phenomenon, often because non-conformism can be psychologically painful. See for instance Asch (1951, 1955) and Moscovici (1985) among others. Carpenter (2004) and Bardsley and Sausgrüber (2005) suggest that conformism may partially explain conditionally cooperative tendencies. But the evidence presented in this section shows that there is clearly more to conditional cooperation than a mere desire to conform.

[16] Yamagishi (1986, 1988) and Ostrom et al. (1992) also look at the role of punishments in sustaining cooperation in social dilemmas. While these papers pre-date Fehr and Gächter, as mentioned above, I am going to focus on the period since 1995 and therefore I will forego a detailed discussion of these earlier studies.

Table 1 Punishment levels and associated costs for the punishing participant in Fehr and Gächter (2000)

Punishment Points	0	1	2	3	4	5	6	7	8	9	10
Cost of Punishment	0	1	2	4	6	9	12	16	20	25	30

Source: Fehr and Gächter (2000)

well-known. I will provide a brief summary only, primarily because I need to refer to aspects of this study when I discuss subsequent papers in this area below.

The second set of ten rounds has two stages in each round. In the first stage, participants play a standard linear public goods game. This is followed by a second stage where players, having seen the contributions of others (but without learning their identities) can choose to punish the other group members. Each punishment point reduces the punished participant's payoff by 10%, so that if a participant receives 10 or more punishment points then this participant's earnings fall to zero. The punishment is costly to the punisher as well. Table 1 shows the cost associated with each punishment point in terms of tokens.[17]

Figure 1 provides an overview of average contributions in the two treatments without and with punishments. Across all rounds and the two treatments the average contribution to the public good is 19% without punishment and 58% with punishments. The average contribution in the last round without punishments is 10% but with punishments the average last round contribution is 62%. Fehr and Gächter (2000) also find that punishments are primarily aimed at those who contribute less than the group average in any round and the further below the group average is the participant's contribution, the greater is the magnitude of the punishment handed out to this participant.[18,19]

[17] Studies implementing a punishment mechanism fall into one of two categories. Either they use a "within subjects" design where the same participant takes part in two treatments, an experimental treatment with a punishment option and control treatment without punishment. Others use a "between subjects" treatment where some participants take part in treatments with punishment while others take part in a control treatment without. In what follows, every time I say that this is a "between subjects" treatment, it will imply that besides the experimental treatments, one set of participants always take part in a control treatment, where no punishment opportunity is available.

[18] Fehr and Schmidt's (1999) model of inequity aversion can justify the use of costly punishments. Given that free-riders are better off in payoff terms compared to cooperators, the latter may well be willing to sacrifice a further part of their payoffs to punish the free-riders especially if that reduces the inequity of payoffs between the two.

[19] Fehr and Gächter (2002) extend the results obtained by Fehr and Gächter (2000) using a similar within subjects design except with random re-matching in all sessions. In one treatment participants play 6 rounds without a punishment option followed by another 6 rounds with the punishment option. This sequence is reversed in the second treatment. Here each token of punishment reduces the earnings of the punished participant by 3 tokens. Since participants are randomly re-matched at the beginning of each round theories of reciprocity or costly signalling cannot explain cooperation and punishment in this environment. The authors suggest that punishment in this context is "altruistic" in the sense that this punishment is costly to the punisher but does not generate any immediate benefits for the current group members since groups are randomly re-constituted at the end of each round. The results show that 84% of participants punish at least once. Punishments appear to be triggered by negative emotions given that 74% of punishment is imposed by participants who contribute higher than the group average on those who contribute lower than the group average.

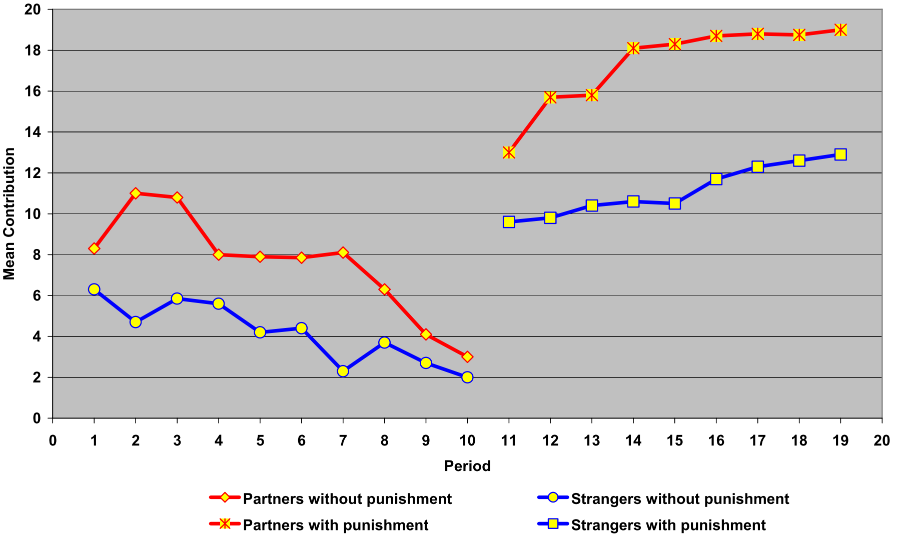


Fig. 1 Average contributions without and with punishments in Fehr and Gächter (2000). Source: Fehr and Gächter (2000)

While the papers by Fehr and Gächter show that costly punishments can indeed raise contributions to levels above those attainable in the absence of such punishments, in those studies the participants do not get to choose whether punishments are available or not. Gürerk et al. (2006) analyze contribution behavior in a public goods game where participants can *choose* to be in either a sanctioning environment (i.e. one which allows participants to punish their group members) or a sanction-free environment.

Each round of this experiment consists of multiple stages. In stage 1, participants have an opportunity to choose to be in either a *sanctioning* or a *sanction-free* institution. In stage 2 participants participate in a linear public goods game. The round ends here for participants who choose to be in the *sanction-free* institution. Participants who choose to be in the *sanctioning* institution continue to stage 3 where they can allocate either positive or negative sanction points to other members. A positive sanction (actually a reward) awards the recipient 1 token and costs the sender 1 token. A negative sanction (a punishment) costs the recipient 3 tokens and costs the sender 1 token. Participants receive feedback on earnings at the end of the round. The experiment consists of 30 rounds and participants are randomly re-matched at the end of each round. This is a between subjects design and once participants choose to be in a particular institution they do not learn the results of the other institutions.

In the first round of the game, a majority (63%) of the participants choose to be in the *sanction free institution* (SFI) rather than the *sanctioning institution* (SI). Participants who do choose to be in the SI, however, contribute on average 64% to the public good which is significantly higher than the 37% contributed by those who choose to join the SFI. Over time the SI becomes the predominant institution and eventually close to 100% of participants chooses the SI. By round 10 of the 30 round interactions, contributions in the SI increases to 90% and continues to go up from

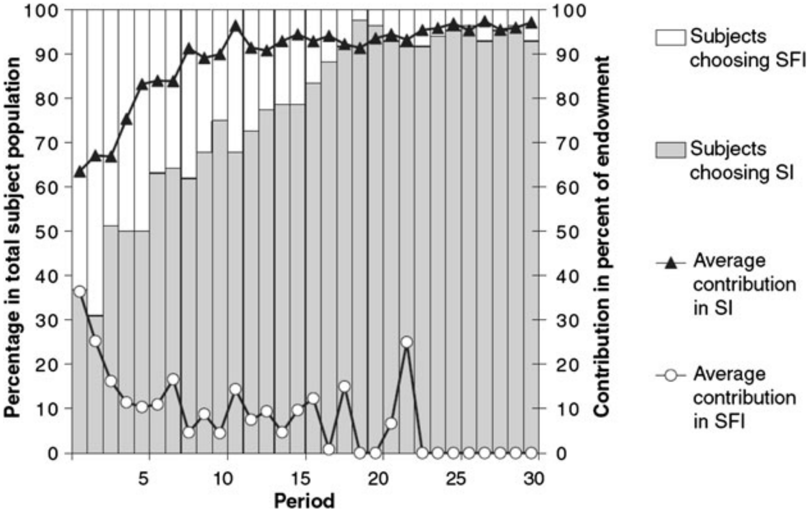


Fig. 2 Evolution of subjects' choice of institutions and contributions over time from Gürerk et al. (2006). Source: Gürerk et al. (2006)

there. In contrast, contributions in SFI decrease to zero. Averaged across all rounds, contributions in the SI are 91%, significantly higher than the 14% in the SFI.

Moreover, not only do more and more participants migrate from the SFI to the SI over time, migrating participants engage in high levels of cooperation and also very quickly adopt the prevailing SI norm of punishing low contributors. Over time as the amount of free-riding falls away to zero and the need for punishment diminishes, the difference in payoffs between high contributors who do not punish free-riders and high contributors who do punish becomes smaller, suggesting that "selection pressures" against strong reciprocators become weaker over time.

Figure 2 shows the evolution of subjects' choice of institutions and contributions over the 30 periods of interaction. The average contributions in both institutions over time are measured as the percentage of the endowment contributed to the public good.[20]

[20] Rockenbach and Milinski (2006) extend this line of investigation by analyzing the interaction between punishment and reputation formation. In their main experimental treatment (called PUN&IR for punishment and indirect reciprocity), participants are put into groups of eight and interact for 20 rounds. In each round there are four stages. In the first stage participants decide whether to allow for costly punishments or not following the public goods game. In the second stage, they play the public goods game. In the third stage, those who choose to allow for sanctions get to mete out and/or receive punishment points. Finally, in the last stage, each participant plays a gift exchange game along the lines of Berg et al. (1995), both as a sender and a receiver. But prior to sending money the sender learns about the receiver's "reputation" by receiving information about the receiver's contributions in the public goods game in the previous stage and about the amounts sent by the receiver in prior plays of the gift exchange game. There is a control treatment (called PUN) where participants get to choose whether to allow punishments or not following the public goods game but there is no gift exchange game after that. Their main conclusion is that both contributions and efficiency are higher in the treatment that allows for both punishment and indirect reciprocity. Efficiency in this treatment averages more than 90% of the social optimum and is close to 100% during the

Gächter et al. (2008) examine, using a between subjects design, whether the duration of interaction affects the efficacy of punishments by looking at two different punishment treatments, one which lasts ten rounds (treatment *P10*) while the other lasts fifty rounds (treatment *P50*). There are also two control treatments without any punishment opportunities, one lasting 10 rounds (treatment *N10*) and the other lasting 50 rounds (treatment *N50*). Here each punishment point costs the punisher one token but reduces the punished participant's payoff by three tokens. What is particularly striking in this study is that per period contributions in the *P50* treatment are 25% higher than those in the *P10* treatment and 50% higher than those in the *N50* treatment. Average net earnings are significantly higher in the *P50* treatment compared to the *N50* or *P10* treatments. Finally, towards the later stages of the *P50* treatment, cooperation seems to become stabilized without much actual punishment being meted out which results in punishment costs becoming negligible resulting in higher earnings in this treatment.

Walker and Halloran (2004) look at the efficacy of rewards and/or sanctions in a sequence of one-shot games. The main innovation of this study is that both the reward and the sanction can either be *certain* or *uncertain*. In the *certain* treatments, each dollar of punishment (reward) reduces (increases) the recipient's payoff by two dollars. In the *uncertain* treatments, with each dollar of punishment (reward), there is a 50% chance that the punished subject's payoff will be reduced (increased) by 4 dollars and a 50% chance that the payoff will remain unchanged. There are five treatments including certain and uncertain punishment, certain and uncertain reward and a control treatment with neither.

However neither rewards nor sanctions—whether certain or uncertain—have any significant impact on contributions compared to the control treatment, suggesting that repeated interactions and the consequent dynamics have an important role to play in sustaining cooperation with punishments. A one-off threat of punishment may not be as effective.

Sefton et al. (2007) explore the relative efficacy of punishments and rewards as well using a repeated public goods game and a between subjects design. They look at treatments that allow for only rewards or only punishments or both. Here each dollar given up in punishment (or reward) reduces (increases) the recipient's payoff by the same amount.

While the introduction of an opportunity to punish or reward group members has a salutary effect on cooperation in that contributions in those treatments are significantly higher than those in the control treatment, the effect of sanctions is much less pronounced compared to say the Fehr and Gächter (2000) results. In fact the treatment that generates the highest contribution allows for both rewards and sanctions. However, the relatively low cost-effectiveness of the punishment mechanism implemented in this study—it costs \$1 to punish another participant by \$1—may have something to do with its relative lack of success. As I argue in the next section, in order to have an impact on contributions, the cost-effectiveness of punishments must be sufficiently high.

last 10 rounds of interaction. In contrast when there is only punishment and no possibility to establish a reputation, efficiency averages only about 75% and these differences are significant at conventional levels.

3.1 On the cost-effectiveness of costly punishments

Nikiforakis and Normann (2008) suggest that the ability of costly punishments to sustain high contributions to the public good depends crucially on the effectiveness of that punishment, i.e., the factor by which each punishment point reduces the recipient's payoff.

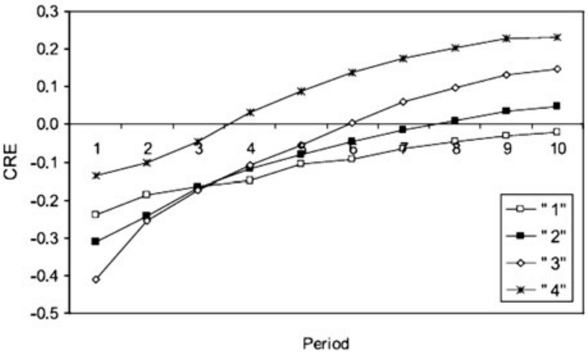
Nikiforakis and Normann look at four different experimental treatments where each punishment point meted out costs the punisher one experimental currency unit (ECU). But in treatment 1, each point reduces the punishment recipient's income by one unit, in treatment 2 by two units, in treatment 3 by three units and finally, in treatment 4 by four units. There is also a control treatment where no punishment mechanism is available. Participants are placed into groups of 4 and play for 10 rounds in a "partners" protocol. In order to prevent reputation formation, given fixed groups, participant identification numbers are changed randomly from one round to next.

The authors find that there is a monotonic relation between the effectiveness of the punishment and mean contributions; as the effectiveness goes up, so does the mean contribution. Average contributions range from 9% in the control treatment to 33% in treatment 1, 57% in treatment 2, 87% in treatment 3 and 90% in treatment 4. However the effectiveness of the punishment matters, in that it is only in the two "high" punishment treatments, where each punishment point costs the recipient either three or four currency units that contributions actually increase over time as in the original Fehr and Gächter (2000) study. Contributions in the other less effective punishment treatments show the familiar pattern of decay.

Furthermore, higher contributions sustained on the basis of punishments do not necessarily translate into higher efficiency. Compared to the control treatment, it is only in treatment 4 (where the punishment inflicts the maximum penalty on the recipient) that average earnings are consistently higher. So it appears that the mere existence of punishment may not always be sufficient to enhance cooperation. In order for punishment to truly make a difference, it must inflict a penalty that is substantially higher than the cost of meting out that punishment.

Figure 3 summarizes that the role of punishment effectiveness on efficiency. It is only when the punishment has maximum effectiveness, depicted by the line with stars labeled "4", that it leads to significant gains in efficiency compared to the control treatment.

Fig. 3 Cumulative relative earnings across treatments in Nikiforakis and Normann (2008). Source: Nikiforakis and Normann (2008)



Egas and Riedl (2008) also report that the only punishment treatment, which succeeds in sustaining cooperation over time is the *low cost-high impact* treatment where each punishment point costs the punisher one token but reduces the recipient's payoff by three tokens. There are two innovative features of this study. The first one is the large sample size with 846 participants. The second is the non-standard pool of participants with any Dutch-speaking person eligible to participate, making the participants much more representative of the population as a whole. Participants were recruited via newspaper advertisements and the experiments were conducted over the Internet. Another interesting finding is that while contributions to the public good are only weakly influenced by the age of the participant (given the non-student participant pool in this study, the average age is 35 years with a range of 12–80 years), older males are significantly more likely to punish, controlling for the punished participant's level of contribution.[21]

Carpenter (2007a) approaches the issue of punishment effectiveness by asking what happens to monitoring of free riders and punishment as the group size becomes bigger.[22] On the one hand, as groups grow, it becomes harder for each individual to monitor others and as a result free riding might become more prevalent as punishment becomes less of a deterrent. On the other hand, in larger groups there are more people monitoring each free-rider, so that free-riders might actually receive more punishment in total compared to smaller groups.

Carpenter looks at both the effect of the *group size*, as well as the *monitoring fraction*, which refers to the fraction of the group each agent can monitor. There are two group sizes, 5 and 10, two values of the MPCR, 0.375 and 0.75 and four possible types of monitoring: (1) *no monitoring*; (2) *full monitoring*; (3) *half monitoring* and (4) *single monitoring*. In the *no monitoring* treatment agents see what others contributed but cannot punish them. In *full monitoring*, agents can punish any and all other group members as long as they have the resources to do so. In *half monitoring*, agents can punish that half of the group which is located closest to them on a circle and finally in *single monitoring* they can punish only one other group member.

The parameters of the punishment technology are the same as that in Fehr and Gächter (2000). For either value of the MPCR, contributions in the 5 person groups with either full of half monitoring start at about 55% and remain stable around that mark for the duration of the session. For the 10 person groups, with either full or half monitoring, contributions start at the same level but then show a clearly increasing profile reaching almost 100% contribution in the last three rounds of interaction.

[21] Nikiforakis (2010) reports that the efficacy of punishments also depends on the nature of the feedback provided to participants. In all prior studies looking at punishments, participants get feedback about individual contributions at the end of each round. These studies report that the availability of punishment opportunities leads to an increase in contributions. Nikiforakis shows that if participants are shown information about *individual earnings* at the end of each round, then this leads to significantly lower contributions and lower efficiency compared to a treatment where participants get information about *individual contributions*.

[22] Isaac and Walker (1988b) and Isaac et al. (1994) examine patterns of contributions in linear public goods games with groups of 5, 10, 40 and 100 players. Comparing groups of 5 with groups of 10, they find that contributions are not significantly different with a MPCR of 0.75 while larger groups achieve higher contributions with a lower MPCR of 0.375. When they look at groups of 40 and 100, they find that contributions are actually higher compared to groups of 5 or 10 and this is independent of the MPCR.

In both five and ten person groups, especially with MPCR = 0.75, there is a clear bifurcation with contributions showing an increasing profile with either full or half monitoring on the one hand and a decaying pattern with single or no monitoring on the other.

3.2 Do punishments obey the law of demand?

Anderson and Putterman (2006) explore the price responsiveness of the demand for punishment by systematically varying the cost of each punishment point. In each round, participants are divided into groups of three using a "perfect stranger" protocol and take part in a public goods game for 5 rounds with the cost of a punishment point changing from one round to the next. There are three treatments. In each treatment there are five different costs for buying a dollar's worth of punishment. (For instance, in one treatment the cost of a dollar of punishment can take of one of five values—0, 30, 60, 90 or 120 cents.) In each treatment, a participant faced each one of the five costs in random order over the five rounds.

What they find is that, across all treatments, the law of demand does hold for punishments. Even after controlling for a number of variables including whether a particular participant contributed more or less than the group average, the amount of punishment to this participant decreases when the price of punishment goes up. Assuming rational self-interest as the primary motivating factor, the act of punishment seems to be non-rational, seeing as it yields no strategic benefits especially with random re-matching of participants at the end of rounds. But there clearly seems to be a fundamentally rational element of price-responsiveness built into the decision to punish.

In Carpenter (2007b) participants play in groups of four using a "strangers" protocol for fifteen rounds. This study also has five possible per unit cost of punishment—0.25, 0.5, 1, 2 and 4. So if the cost is 0.25 (4), for instance, that implies it costs $1/4$ of a token (4 tokens) to buy one token of punishment. Participants face these costs either in ascending or descending order and each value stays unchanged for three consecutive periods. Participants are aware of this sequence. Another major design difference is that each participant can punish only one other member of her current group.

While Anderson and Putterman (2006) suggest that the demand for punishments may be elastic, Carpenter concludes that, after controlling for issues such as how far below the group average a particular participant's contribution is, the demand for punishment is inelastic: a 10% increase in the price of punishment leads to about 8% reduction in the quantity demanded. Furthermore, the quantity demanded is not responsive to changes in income, at least not significantly so.[23]

[23] I have chosen to discuss the Anderson and Putterman (2006) and Carpenter (2007b) papers in a different section than Nikiforakis and Normann (2008) or Egas and Riedl (2008) for the following reason. In the two latter papers the cost of buying each punishment point is always constant; what differs is the cost inflicted on the recipient which can vary from a low to a high value. In the two papers in this section, each punishment point meted out imposes the same cost on the recipient but the cost of buying each of those punishment points varies. The net effect is of course the same which is to vary the cost-effectiveness of punishments. But it is possible that participants may perceive these two mechanisms differently which might lead to differences in behavior.

3.3 The possibility of "perverse" punishments

The above discussion suggests that costly monetary punishments of free-riders can sustain high levels of contribution to the public good. Nikiforakis (2008), however, sounds a note of caution about the salutary effects of punishments. He shows that if one allows the possibility of counter-punishments by punished free-riders, cooperators are less willing to punish. He finds that punishments are often "perverse" or "anti-social" in nature in the sense that those who free-ride often punish those who cooperate and such punishments are driven partly by strategic considerations and partly by a desire to avenge the punishments meted out by others. In what follows I will stick to the terms "anti-social" punishment to indicate punishment of high contributors and "pro-social" punishment to indicate the more usual punishment of free-riders by cooperators.

Nikiforakis looks at two different punishment mechanisms. The *punishment* treatment is identical to that implemented in Fehr and Gächter (2000) discussed above with the same punishment costs as shown in Table 1. The other treatment, which is the primary focus of this study, adds a third *counter-punishment* stage following the second punishment stage to each round. At the beginning of this third stage each participant is informed about the number of punishment points assigned to him by his group members and is given an opportunity to assign counter-punishment points to those participants in turn. The punishment costs are the same as in stage 2 (shown in Table 1). These costs work cumulatively. A participant's final earnings in a round in this third treatment is his earnings from the public goods game in stage 1 minus all the income reductions caused by the punishment points assigned to this participant by others and those assigned by the participant to his peers over the two stages of punishment and counter-punishment.

A crucial feature of this study is that only those participants who are actually punished in stage 2 are allowed to engage in counter-punishment and they can only punish those who punished them in the first place; moreover, a participant must have a positive payoff in order to carry out any counter-punishment and these must be carried out in the stage immediately following the punishment.

This is a within subjects design with both fixed groups and random re-matching. Participants interact for 20 rounds in two blocks which are counter-balanced. In one of those blocks participants play the standard public goods game while in the other block they have the opportunity to either engage in punishment or both punishment and counter-punishment.

Nikiforakis finds that individuals in the punishment treatment are significantly more likely to contribute to the public account compared to those in the counter-punishment treatment, where contributions show the familiar pattern of decay. The counter-punishment treatment also leads to lower average earnings compared to *both* the punishment treatment and the control treatment.

In looking at why the counter-punishment treatment does worse than the punishment treatment, the author finds that participants in this treatment engage in substantial amounts of "anti-social" punishment which can be attributed to one of two factors or a combination of those. One is the anticipation by some free-riders of the forthcoming punishment from cooperators and their willingness to retaliate those sanctions.

Here participants use counter-punishments strategically to signal that future sanctions will not be tolerated and this is especially true in fixed groups, which affords scope for such signaling. The second factor is the desire to avenge sanctions meted out to them in previous periods. In fact, participants in the counter-punishment treatment are 15% less likely to punish free-riding compared to the punishment treatment mostly because cooperators anticipate that this might, in turn, lead to "anti-social" punishment and wish to avoid the same.

However, Cinyabuguma et al. (2005) suggest that the specific mechanism for implementing counter-punishments might make a big difference to their eventual impact. In the Cinyabuguma et al. study, participants do not learn the identity of those who punished them. This makes it impossible to engage in targeted revenge. Instead each participant is told the pattern of punishing high, average and low contributors in the group in the first stage and then that participant can decide who to counter-punish, i.e. whether to engage in counter-punishment of "pro-social" or "anti-social" punishers.

Contrary to Nikiforakis, who found that contributions and earnings in the treatment with counter-punishment were lower than the punishment treatment, Cinyabuguma et al. find that the availability of counter-punishment does not reduce either contributions or earnings to levels lower than those achieved in the punishment only treatment; at least not significantly so. It is conceivable that this is due mostly to the design differences between the two studies. The opportunity to engage in counter-punishment then seems to have very different impact depending on its exact implementation.

Ertan et al. (2009) allow their participants to vote on who should be punished—those who contribute less than, equal to or greater than the group average contribution. The authors find that there are no groups where a majority voted to allow punishment of participants who contribute higher than the group average contribution. This ruled out the possibility of "anti-social" punishments. In initial votes there is a tendency to vote for no punishment. However over time, there is a clear evolution towards allowing punishment of low contributors while still prohibiting the punishment of high contributors.

Hermann et al. (2008) provide even more compelling evidence of such "anti-social" punishments in an ambitious cross-cultural experiment that compares the behavior of under-graduate students across 16 different locations using a between subjects protocol.[24] The punishment parameters are the same as in Fehr and Gächter (2000). The incidence of "anti-social" punishments is the lowest among the participants from western industrialized nations where the bulk of the previous experimental data comes from. This, in turn, suggests that the evidence in favor of the cooperation enhancing role of punishments that comes from prior experiments run in western societies may actually be over-estimating the efficacy of such punishments and in

[24] These locations are: Athens, Bonn, Boston, Chengdu, Copenhagen, Dnipropetrovs'k, Istanbul, Melbourne, Minsk, Muscat, Nottingham, Riyadh, Samara, Seoul, St. Gallen and Zurich. Henrich et al. (2010) point out that most people are not WEIRD meaning Western, Educated, Industrialized, Rich and Democratic. Yet the participants in the vast majority of studies including those cited in this article come from that somewhat unusual pool. Hence, the extent to which these results are applicable to those who are not WEIRD is open to debate.

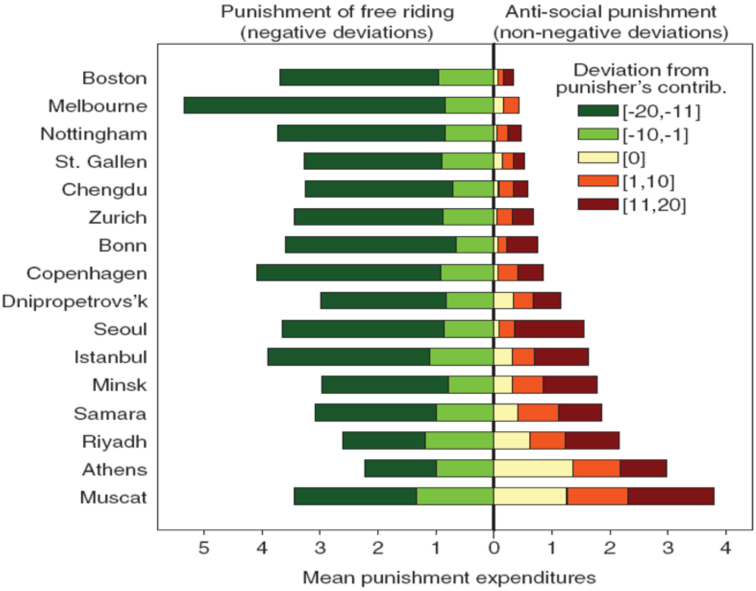


Fig. 4 Mean punishment expenditures across different locations in Hermann et al. (2008). Source: Hermann et al. (2008)

other societies the presence of "anti-social" punishments may actually have a large detrimental role.

Different participant pools reacted very differently to the punishment received. In only 11 out 16 societies, those punished in one round for contributing less than the group average increased their contribution in the next round and the extent of the mean estimated increase per punishment point received varies considerably. Thus, punishment did not have an equally strong disciplinary effect on free riders in all participant pools in increasing their cooperation and in some societies punishments did not increase cooperation at all. Figure 4 summarizes the nature of pro-social and anti-social punishments across these diverse societies.

The authors suggest that "anti-social" punishments are more prevalent in societies which are characterized by (1) a lack of strong social norms of civic cooperation as expressed in people's attitudes towards tax evasion, abuse of the welfare system or dodging fare on public transport and (2) weak law enforcement.[25,26]

[25] The rule of law indicator is based on a number of variables that measure "*the extent to which agents have confidence in and abide by the rules of society, and in particular the quality of contract enforcement, the police, and the courts, as well as the likelihood of crime and violence*". (Hermann et al. 2008, pp. 1366.)

[26] Gächter and Hermann (2010) demonstrate the existence of anti-social punishments using participants from both urban and rural areas of Russia as well as participants who are "young" (less than 30 years of age) and "mature" (older than 30). The urban participants come from Kursk and Zheleznogorsk, located in the so-called "Central Black Earth Zone" about 400 miles south of Moscow. The rural participants are located in the rural areas around Kursk and in Ust-Kinel. The urban young participants are mostly university students. Participants play two *one-shot* public goods games, one with punishments and one without, in groups of three. Each punishment point meted out costs the punisher one token in earnings while it

3.4 Counter-punishments: revenge or cooperative norm enforcement?

The above suggests that counter-punishments can either be “pro-social” or “anti-social”. “Pro-social” counter-punishments can take two forms. They could be aimed at (1) those who engage in “anti-social” punishments in the first stage and punish high contributors or (2) those who fail to punish free-riders. As such “pro-social” counter-punishments should enhance cooperation by deterring free-riding or “anti-social” first stage punishments and can play a powerful cooperative *norm enforcement* role. “Anti-social” counter-punishments, on the other hand, are more like acts of revenge. They are usually carried out by free riders and aimed at cooperators.

An innovative study by Denant-Boemont et al. (2007) addresses these two aspects of counter-punishment; its role (1) in exacting revenge on those who have punished one before or (2) in enforcing cooperative norms via punishment of free riders as well as of those who engage in “anti-social” punishments.

This study uses a between subjects design with groups of four participants and a “partners” protocol. The principal innovation here is implemented at the third, counter-punishment stage. In the *full information* (FI) treatment, players are not only informed about the contributions of others but also of how each individual sanctioned each other individual. In the *revenge only* (RO) treatment, each individual is informed of the source and quantity of the sanctions directed towards him in a treatment that is analogous to Nikiforakis (2008) and allows the authors to focus on the “*anti-social*” aspect of counter-punishments. In the *no revenge* (NR) treatment, no individual is informed about who sanctioned him personally and by how much. Rather, all individuals are informed of the source and the quantity of the sanctions directed toward each player other than one’s own self. This treatment is similar to that implemented by Cinyabuguma et al. (2005) and allows for isolating the “*norm enforcement*” effect of counter-punishments. There is also a *six stage full information* treatment but with five stages of unrestricted sanctioning following the initial contribution decision in the first stage.

The availability of sanction enforcing “pro-social” punishments coupled with a restriction on revenge seeking is welfare improving in that earnings in the *no revenge* treatment, at 81% of the social optimum, are significantly higher than either the *full information* treatment (53% of the optimum) or the *revenge only* treatment (36% of the optimum). In the *full information* treatment, which allows for both “anti-social” punishments as well as “norm enforcing” punishments, increments in contributions

costs the receiver three tokens in earnings. Gächter and Hermann find that there is significant punishment of cooperators across all subject pools, with people punishing both those who contributed the same amount as the punisher as well as those who contributed significantly more than the punisher. Rural residents contribute more than urban ones and mature participants contribute more than young ones. Gächter and Hermann (2009) compare anti-social punishments among university students located in Belgorod and Yekaterinburg (both in Russia) and those located in Zurich and St. Gallen (in Switzerland). Like Hermann et al. (2008), Gächter and Hermann (2009) show that there may be a cultural component to the desire to impose anti-social punishments in that such punishments are far more pronounced among the Russian participants than among the Swiss participants. Gächter and Hermann (2009) also provide a brief overview of various mechanisms that enhance cooperation.

caused by the norm enforcing sanctions are not enough to fully offset the negative effect exerted on contributions by anti-social counter-punishments.[27]

3.5 Concluding remarks

The evidence on costly punishments suggests that providing participants with the opportunity to engage in such punishment of group members can usually help sustain high levels of contributions. But this finding is subject to at least three caveats.

First, there is the problem that the punishment itself creates a second-level public good; those who are willing to mete out costly punishment must not only punish free-riders but also those non-punishers who might contribute but do not punish free riders and hence free ride on others' punishment and so on. This requires the creation of "meta-norms" of punishment in the words of Axelrod (1986). However, the existence of punishment opportunities and of conditional cooperators who are willing to use such punishment does seem to reduce free riding. And at times, given a long enough time horizon, as in Gächter et al. (2008), the threat of punishment might be enough to sustain cooperation without the punishment actually having to be carried out.[28]

The second problem has to do with the issue of "anti-social" punishments as discussed in Sect. 3.3. The presence of such "anti-social" punishments may result not only in no increase in cooperation but also seriously reduce efficiency compared to control treatments with no punishment. The cooperation enhancing effect of punishments seems more prominent in particular participant pools, especially in western industrialized societies. Also, the exact nature of counter-punishment matters. If participants are only allowed to engage in "pro-social" punishments but not in targeted revenge, then this could be welfare improving. However, if we allow for both "anti-social" and "norm enforcing" punishments then the net effect is detrimental to cooperation because the increase in cooperation caused by norm enforcing sanctions does not fully offset the contribution reducing effect of anti-social punishments.

The third problem is that the efficiency implications of costly punishments are not clear cut. By and large, across the majority of studies cited above, efficiency is actually lower in treatments with punishment compared to control treatments without punishment. The ability of punishments to enhance *efficiency* seems to depend

[27] The authors also find that adding more stages of unrestricted punishment as in the *six stage full information treatment* is unambiguously welfare reducing. This reduction arises from two sources: a reduction in contributions and much higher levels of sanctioning. In fact the average punishment points awarded in this treatment is 2.7 times those in the next highest treatment.

[28] Henrich and Boyd (2001) develop a theoretical model to demonstrate that "conformist transmission" (where agents use the popularity of a choice as an indirect measure of its worth) can address this problem of higher order punishments and stabilize cooperation. Suppose being punished is sufficiently costly so that co-operators have higher payoffs than defectors. Then a second-order free rider who cooperates but does not punish free-riders will achieve a higher payoff because they avoid the punishment cost. But if defections do not pay, then such defections will occur rarely and by mistake; so over time such defections become less frequent and as we ascend higher orders of punishing the difference in payoffs between punishers and second-order free-riders will start to approach zero. Thus conformist transmission, no matter how weak, will at some stage (t) be able to overcome payoff biased imitation in the form of free-riding and stabilize punishment. Once this happens, by backward induction, payoffs will favor strategies that punish at the $(t-1)^{\text{th}}$ stage, which, in turn, favors punishment in the $(t-2)^{\text{th}}$ stage and so on till cooperation is stabilized in the very first stage.

crucially on two things: (1) the cost-effectiveness of the punishment; as shown by Nikiforakis and Normann (2008) and Egas and Riedl (2008) it is only when the punishment is *low cost* and *high impact* that it also leads to an increase in efficiency over and above any increase in contributions. (2) The other factor that makes a difference is the time horizon. As Gächter et al. (2008) show, if the time horizon is sufficiently long, then even low impact punishments can generate efficiency gains that are not present over a shorter time frame.

These caveats highlight the practical difficulties of implementing such costly decentralized peer punishment. Furthermore, Guala (2010) points out that ethnographic evidence from tribal societies or the historical evidence on common pool resource usage does not provide a lot of support for either the use or the efficacy of such costly monetary punishments. Therefore, mechanisms that rely less on monetary punishments and more on other factors might be easier to adopt. This is what I explore next.

4 Sustaining cooperation without monetary punishments

Axelrod (1986) suggests that a social norm is essentially an implicit rule that members of society feel compelled to adhere to. One way of creating and sustaining such a norm is via internalization, where a norm becomes so entrenched in a society that violating it causes psychological discomfort. Below, I explore different ways in which such internalization may be achieved. I begin by exploring punitive measures that are non-monetary in nature such as social ostracism or exclusion and then go on to discuss other non-punitive mechanisms that lead to successful cooperation.

4.1 Even non-monetary punishments can sustain cooperation

Masclet et al. (2003) is a pioneering study demonstrating that non-monetary punishments such as expressions of disapproval can enhance cooperation. This study will be familiar to most readers and hence I will avoid a lengthy discussion. The authors compare the efficacy of *non-monetary punishments* with that of *monetary punishments*. The latter is similar to those in Fehr and Gächter (2000). In the *non-monetary punishment treatment*, participants are given the opportunity of expressing approval or disapproval of the actions of other group members but these do not affect monetary payoffs. As with the punishments each participant can assign between zero and ten points to another participant where zero indicates no disapproval and ten indicates maximum disapproval.

The authors find that both monetary and non-monetary sanctions initially increase contributions by a similar amount but that over time, monetary sanctions are more effective and lead to higher contributions than non-monetary sanctions. Furthermore, and not surprisingly, they find that non-monetary sanctions are more effective in the "partners" treatment as opposed to the "strangers" treatment. The authors also find that the average earnings of participants are higher with either monetary or non-monetary punishments compared to the baseline situation where no sanctions are available.[29]

Cinyabuguma et al. (2006) look at punishment from a different perspective by allowing group members the opportunity to expel free-riders by majority vote. This is a between subjects study with each session consisting of 16 participants all belonging to the same group, in a "partners" protocol. Participants play for 15 rounds with two stages in each round. In each round after learning individual contributions, each participant is given an opportunity to vote to remove an individual from the group. A group member will be removed from the group if half or more voted to expel that person and such exclusion is irreversible. Group members who are expelled play a similar public goods game except with half the endowment in each round compared to the original group and therefore will potentially earn less.

The authors find that while there are few actual expulsions, a majority of participants voted to expel another participant at least once and typically the ones getting expulsion votes or actually being expelled are either the lowest contributor or the second lowest contributor. When a participant receives an expulsion vote in a given round, even if he is not actually expelled, he responds by increasing his contribution in the next period, as in Masclet et al. (2003). Both the average contribution to the public good and the overall earnings are higher in the expulsion treatment than in the control treatment.

4.2 Sustaining cooperation via non-punitive mechanisms

The studies that look at non-punitive measures can be broadly classified into (1) those that attempt to foster cooperation among participants without making any attempt to sort them and (2) those that try to form sorted groups on the basis of similarity of behavior or preferences.

4.2.1 Sustaining cooperation in non-sorted groups

One obvious mechanism to promote cooperation is to allow for communication among participants. Ever since Dawes et al. (1977) and Isaac and Walker (1988a) we have known that communication can improve cooperation. Bochet et al. (2006) extend this literature by directly comparing the relative efficacy of communication vis-à-vis punishment in sustaining cooperation using a between subjects design.

The authors look at three types of communication; (1) face-to-face, (2) chat-room and (3) numerical cheap talk. In the treatments with face-to-face communication, each participant has the opportunity to talk to the other three group members for five minutes before the game starts. In the treatments with chat room communication, each participant can exchange verbal messages with group members via a computer chat room. In the numerical cheap talk treatment, each participant has the option of

[29] Noussair and Tucker (2005) extend the work done by Masclet et al. (2003) by looking at whether the availability of *both* monetary and non-monetary sanctions can generate higher welfare than either type on its own. They find that over time contributions in the non-monetary punishment treatment falls consistently as the non-monetary sanctions appear to lose effectiveness when not backed up by a monetary sanction.

typing in a number to indicate his/her potential contribution to the public account. No other form of communication is allowed.

The authors also look at three more treatments where each particular communication strategy is combined with the opportunity to punish one's group members. It costs 0.25 token to punish another participant by 1 token. There is also a punishment-only treatment which only allows for punishment of group members but no opportunity to communicate.

The main insight of this study is that once face-to-face communication is allowed contributions jump up to 96% of the maximum which is significantly higher than those in the control and the punishment-only treatments. Given the already high contributions in the face-to-face condition, allowing punishments on top of that leads to only a small increase in contributions to 97%. The chat room communication with punishment does almost as well as the face-to-face treatment and gets average contributions of around 96% while the chat room communication without punishment does not do as well with contributions averaging 81%. However, both versions of the chat room treatment do better than either the control or the punishment-only treatment. Unlike the other communication treatments, the numerical cheap talk treatment with or without punishments does not exhibit either higher contributions or higher earnings as compared to either the control or the punishment-only treatments.[30]

Chaudhuri et al. (2006) investigate a different type of communication scheme by allowing participants to pass advice. The focus here is on the evolution of cooperative norms using an inter-generational approach. Participants in one generation leave advice for the succeeding generation via free form messages. Such advice can be *private knowledge* (advice left by one player in generation t is given only to her immediate successor in generation $t+1$), *public knowledge* (advice left by players of generation t is made available to all members of generation $t+1$) or *common knowledge* (where the advice is not only public but also read aloud by the experimenter). Contributions in these advice treatments are compared to those in a control treatment with no opportunity to leave advice. Participants play in groups of 5 for 10 rounds. However each participant in generation t is connected to another participant in the immediately succeeding generation $t+1$ and each participant in generation t earns a second payment which is equal to 50% of the earnings of her generation $t+1$ successor.

The authors find that average contributions in the common knowledge of advice treatment are significantly higher than the other treatments including the control treatment. Common knowledge of advice generates a process of social learning that leads to high contributions and less free riding. In later generations of the common knowledge treatment contributions and efficiency are greater than 90% of the social optimum and the modal contribution to the public account is the entire token endowment. This behavior is sustained by advice that is generally exhortative, suggesting high contributions, which in turn creates optimistic beliefs among participants about others' contributions.

The authors also collect data on beliefs using an incentive-compatible mechanism. Average post advice beliefs are the highest in the common knowledge treatment.

[30] For two other studies that also explore the role of communication in fostering cooperation in public goods game see Brosig et al. (2003) and Bochet and Putterman (2009).

Given the strong positive correlation between beliefs and contributions, it is not surprising that this treatment generated high contributions.

Rege and Telle (2004) examine the impact of social norms via indirect social approval and framing on cooperation. The authors look at the effect of two factors: (1) *social approval* and (2) *associative framing*.

The first *no-approval* treatment is run using a *double-blind* protocol, thereby making social approval or disapproval impossible. In the *social approval* treatment this anonymity is removed where each participant's identity and contribution to the public good are revealed to the group members. In an *associative* treatment the participants are referred to as a "community" with the intention of creating social and internalized norms for cooperation, while in a *non-associative* treatment the instructions are written in more abstract language. This generates four different conditions—(1) no approval/non-associative; (2) no approval/ associative; (3) approval/non-associative and finally (4) approval/associative.

The authors find that the treatments have a significant impact on contributions. Average contributions increase from 34% in the no-approval/non-associative treatment to 55% in the no-approval/associative treatment, to 68% in the approval/non-associative treatment to 77% in the approval/associative treatment.[31]

4.2.2 Cooperation in sorted groups

Chaudhuri (2009) points out, in many of the things we do in life we actually *choose* the people we wish to interact with as when we join religious or social groups, ostensibly because these people have preferences similar to ours. Sustaining cooperation in such sorted groups might prove to be less of a challenge. Below I discuss papers where such sorting is either (1) *exogenous* (undertaken by the experimenter on the basis of a pre-determined rule which may or may not be known to the participants) or (2) *endogenous* (allowing participants to form groups or leave groups on their own accord).

4.2.2.1 Exogenously sorted groups Gunnthorsdottir et al. (2007) investigate this issue by sorting cooperative contributors in the same group. Each session includes 12

[31] Seely et al. (2005), look at the role of a "credible assignment"—essentially a non-binding public announcement—in a linear public goods game. Except they look at a game which will be terminated after a certain number of periods with a given probability. In essence Seely et al.'s endeavor amounts to a test of the folk theorem in an infinitely repeated prisoner's dilemma. Participants are assigned to different treatments where they are asked to adopt various strategies vis-à-vis their contributions to the public good. The strategy that succeeds in fostering the most cooperation involves using a "grim trigger". Here participants are asked to make the socially optimal contribution to the public account as long as the other group members do so but contribute nothing in all subsequent periods following a deviation from this norm. But given that this is not a finitely repeated game and hence the strategic considerations are quite different, I will avoid a more elaborate discussion of this study. Chaudhuri and Paichayontvijit (2010a) compare the efficacy of such public announcements with that of costly punishments in a finitely repeated linear public goods game. They find that contributions in the initial rounds are higher in the treatment with an announcement compared to the control and punishment treatments. However contributions decay much faster in the treatment with an announcement whereas contributions increase in the treatment with punishments over time. Payoffs are higher in the announcement treatment in the initial rounds but decrease over time whereas payoffs increase in the punishment treatment over time.

participants. Participants are grouped into four where they play 10 rounds of a public goods experiment with three possible values of the MPCR: 0.3, 0.5 and 0.75. The authors look at two different grouping rules: (1) participants are *randomly re-matched* into different groups at the end of each round (*Random* treatment) and (2) participants are *sorted* into groups depending on their contribution at the end of each round (*Sorted* treatment). The four participants who contribute the most to the public account are placed into one group; the fifth to eighth highest contributors are placed into another group; and the four lowest contributors are placed in the third group. Hence the grouping is dependent on the contribution in the current round. To avoid strategic behavior participants are not informed about how the groups are formed.

For a given value of the MPCR, contributions among the *sorted* groups are always greater than among *randomly* formed groups. Also the decay in contributions is much slower among *sorted* groups compared to the *randomly* formed groups with little or no decay in the two sorted treatments with $\text{MPCR} = 0.5$ and $\text{MPCR} = 0.75$.

The authors define "free-riders" as those who contribute 30% or less of their endowment to the public account. The rest are defined as "cooperators". Within each MPCR, by round 4 at the latest, contributions by cooperators in the *sorted* treatment exceed contributions by cooperators in the *random* treatment. Since the *sorted* treatment reduces the number of interactions between cooperators and free-riders, the authors conclude that higher contributions by the cooperators in this treatment are due primarily to the more efficacious nature of their prior interactions and the exposure to a history of cooperative interactions. On the other hand, the decay in contribution in the random re-matching treatment is due almost entirely to the reduction in contribution by the cooperators who experience much greater interaction with free riders.

In Gächter and Thöni (2005) participants first take part in a "ranking experiment" which consists of playing a one-shot linear public goods game with an MPCR of 0.6 in randomly formed groups of three. Participants did not receive any information about the contribution of other group members or their earnings at this point. Following this, participants take part in the main experiment which consists of playing a ten-period repeated linear public goods game.

For the main experiment, the three highest contributors in the ranking experiment are put together in one group, the next three in the second group and so on till the three lowest contributors who form the last group. Participants get to know how these groups are formed and are also informed how much their *new group members* contributed in the ranking experiment.

There is also a control treatment, where the groups are formed *randomly* and has nothing to do with what the participants contributed in the ranking experiment. The authors also combine the two grouping protocols with the opportunity to punish group members. This then gives rise to four separate conditions: (1) *Sorted no punishment*; (2) *Random no punishment*; (3) *Sorted punishment* and (4) *Random punishment*.

Sorting people into groups based on their performance in the ranking exercise led to a substantial increase in contributions. Even without any punishment opportunities, the top third of contributors in the *sorted* groups contribute significantly more than the most cooperative third in the *randomly* formed groups with average contributions of 70% of the social optimum among *sorted* groups and only 48% of the social optimum among *random* groups. Not only that, the three highest contributors in the *sorted* groups achieved the same level of contribution as the most cooperative third of *randomly* formed groups even when the latter had a punishment option at their disposal. The availability of a punishment opportunity does not make a difference in sorted groups since the three highest contributors in the treatment with *sorting but no punishment* manage to sustain the same level of cooperation as those in the treatment with *sorting and punishment*.

de Oleveira et al. (2009) also engage in exogenous sorting but in their study some participants are explicitly informed about the type of the other group members while others are not. So the focus is on the role of information regarding the type of group members. Participants first play a one-shot public goods game where they are categorized either as "Conditional Cooperators" or as "Selfish" using the same approach as in FGF. Then, on a different day they take part in a linear public goods game repeated for 15 rounds. Participants are placed into groups of three, where the groups can be (1) homogeneous, consisting of either all conditional cooperators or all selfish players or (2) heterogeneous, consisting of two players of one type and one of the other. In the *Known Distribution* treatment, participants are explicitly told about the composition of the group prior to starting the experiment, while in the *Unknown Distribution* treatment they are not given this information. In both treatments, the participant knows his own type. The composition of the groups remains unchanged for the duration of the session.

There are two important insights coming out of this study. First, not surprisingly contributions in groups with three conditional cooperators are significantly higher than in those with two conditional cooperators or one. But more importantly, contributions in groups with three conditional cooperators are higher when the distribution is *known* as opposed to when it is *unknown*. This suggests that the mere presence of conditional cooperators (which can conceivably be inferred from the contribution patterns) is not enough, conditional cooperators need to know that there are no selfish types in their group for them to sustain cooperation. This latter finding echoes the results reported by Chaudhuri and Paichayontvijit (2006) that conditional cooperators cooperate more when they are made aware of the presence of other conditional cooperators.[32]

4.2.2.2 Endogenous sorting of participants Page et al.'s (2005) approach is similar to Gächter and Thöni (2005) except that in Page et al. participants can choose who they want in a group with them. There are 16 participants in each one of four sessions. At periodic intervals during a session, each participant is shown a list, without other identifying information and in a random order, of each of the other 15 participants' average contribution to the public account till that point. Participants are then given the opportunity to express a preference among possible future partners by ranking them. The four individuals with the lowest rank are then put together in the same

[32] Burlando and Guala (2005) also form sorted groups of four based on their classification of participants as described in Sect. 2 above. These sorted groups then play a linear public goods game for 10 rounds a week later. The striking finding of this study is that in the second session with sorted groups—the groups consisting of all unconditional cooperators or all conditional cooperators achieved almost full cooperation for the entire duration of the session.

group. Here the group size is always equal to four, except participants get to choose which group they wish to belong to.

After new groups are formed, participants resume play without information about whom they have been grouped with, a matter on which only indirect inference can be made by observing one's three partners' contributions. The authors also look separately at a punishment treatment and a combined treatment with regrouping and punishment.

Regrouping leads to significant increases in contribution to the public good compared to the control treatment. Moreover, average contributions in the regrouping treatment are about the same as in the punishment treatment (about 70% of the social optimum on average). Thus the participants' abilities to influence with whom they are grouped has a demonstrable positive effect on cooperation and efficiency in this study.[33]

The last two studies discussed in this section adopt a more complex mechanism where both the size and the composition of the group are determined endogenously. In Ehrhart and Keser (1999) there are five sessions with 18 participants in each playing for 30 rounds. Participants in a session are divided into two populations of 9 each. No participant knows who the other 8 members of the population are. In the first round of a session the 9 members of a population are randomly assigned to a three 3-person groups. From the second round onwards, each round consists of two stages. In the first stage, participants make contribution decisions as usual. Participants here get to see the contributions of all members of the population. In the second stage, each participant gets to decide whether to continue with the same group or whether to leave the group at a cost. It is possible for a participant to leave and form a one member group during the particular round.

The returns to the public good are set in a way that the socially efficient outcome would be to form a "grand coalition" with all 9 members of the population belonging to the same group. However, Ehrhart and Keser find that this grand coalition is seldom achieved. Because the returns are higher with increasing group size, groups with higher contributions tend to grow in size. But as they do, the extent of free riding within the group increases as well preventing the groups from reaching the optimal size. Participants who make high contributions are more likely to leave a larger group and form a smaller one (or even a "single") while free-riders are more likely to join larger groups in order to reap the economies of scale.

Charness and Yang (2007) undertake a more elaborate investigation where participants are not only free to leave their current groups as in Ehrhart and Keser (1999) but they can also vote to expel group members as in Cinyabuguma et al. (2006).

[33] Ones and Putterman (2007) also study the impact of costly punishments in a situation where groups are sorted according to their levels of cooperation and find that groups consisting of participants with similar cooperative tendencies outperform randomly composed groups. They extend the findings of Page et al. (2005) and Gächter and Thöni (2005) by testing whether cooperative preferences are stable over time and whether differences in group outcomes can be predicted by knowing the types of participants who compose those groups. Like Gunnthorsdottir et al. (2007), Ones and Putterman (2007) find that early contributions can serve as a significant predictor of contributions in later periods. Moreover the combination of own type measures coupled with measures of experiences of interacting with other groups members can explain substantial parts of the variation in later contributions by participants.

However, here the expulsion vote is less punitive because expelled members are free to join other groups or remain as "singles" with the same endowment. Thus expulsion need not imply a reduction in payoff. Beyond this, there are also opportunities for mergers among groups as a whole. There are two experimental treatments and a control treatment, each consisting of two blocks of 15 rounds each.

The main focus of interest in this study is their treatment 2, where 9 people in a "society" are placed into three groups of 3 participants each and play a public goods game for 3 rounds. The social value of an allocation to the public account depends on the group size and as in Ehrhart and Keser's study, the greatest group returns are achieved by forming a "grand coalition" where all 9 members of society belong to the same group. After the first three periods, participants learn about the average contribution of each other individual in their society (by identification number only) for those three periods. At that point participants can choose to either exit the group or vote to expel other group members. Groups are allowed to merge as well. After the end of the first segment of 15 rounds, groups are re-formed and play a second set of 15 rounds which proceed along similar lines.

Clearly endogenous group formation enhances contributions to the public good in comparison to exogenously formed groups as in the control treatment. While the contribution rate in exogenously formed groups in a control treatment steadily decline to around 25% of the social optimum, in endogenously formed groups the rate increases to above 95% in the later periods.

Contrary to Ehrhart and Keser (1999), the most commonly occurring group composition in this treatment is in fact the grand coalition of with all 9 members of society belonging to the same group followed by 8-1 and 7-2 splits respectively and these larger groups tend to be quite stable over time. The authors also find that participants are less likely to exclude another group member the higher that member's contribution vis-à-vis the group average and individuals/groups are more likely to merge with another group, when that latter group is larger and achieves higher average contributions vis-à-vis the contributions in the former group. Given the ability to sort cooperators in this treatment, profit-maximizing participants find that it pays to cooperate given that they manage to belong to groups where others also contribute.[34]

4.3 Concluding remarks

The evidence presented in this section suggests that in the presence of conditional cooperators, contributions to the public good can be sustained by means other than

[34] See Ahn et al. (2009) for a study looking at endogenous group formation in a congestible, rather than pure, public goods game. In this game, the payoff function is such that the contributions become increasingly more expensive the higher the contribution level. Payoff is also decreasing in increasing group size. Kosfeld et al. (2009) also examine the issue of endogenous formation of institutions using theory and experiments. Prior to taking part in a public goods game, participants get to decide whether to join a group that will allow for punishment of free-riders or a group where no such punishment is possible. They find that subjects frequently implement an organization with punishment and like Charness and Yang (2007) the majority of these consist of the "grand coalition" of all four group members. However, the experimental implementation of their theoretical model raises questions because in the actual experiment those who join the groups with sanction are constrained to contribute their endowment to the public account while the possible punishment of non-contributors is a major factor behind the formation of the sanctioning groups in the first place. This renders their conclusions somewhat difficult to interpret.

monetary punishments. These may include non-monetary punitive measures such as expressions of disapproval or social exclusion. They could also include other interventions such as different types of communication including advice giving from one generation to next and assortative matching of like-minded participants. While in some cases such assortative matching is achieved exogenously with the experimenter sorting participants into groups based on similarities in their behavior or preferences, in some cases, participants left to themselves can form cooperative groups endogenously and can sustain cooperation via either expulsion of free-riders or via leaving less cooperative groups for more cooperative ones.

5 Conclusion: where to from here?

We now have a fairly clear picture about the preference heterogeneity among participants and the preponderance of conditional cooperators. We also have a good idea of how we can go about creating institutions—particularly those relying on costly punishments—that exploit such conditional aspects of behavior to sustain cooperation. I think that the value added by yet another paper exploring these issues is going to be limited.

What then would be fruitful avenues of exploration? One obvious way forward is to apply the lessons learned to "field" settings in designing institutions that deal with social dilemmas. There is already a fairly robust literature in the area. Nevertheless, the evidence cited by Frey and Meier (2004), Gächter (2007), Ostrom (1990) and Ostrom et al. (1994) and Ostrom's being awarded the 2009 Nobel memorial prize in economics suggests a renewed and wider interest in policy implications of these findings.[35]

I think that the phenomenon of contributions decay might lead to further work in the area in terms of both experiments and theory. Existing theoretical models in this area all assume complete information regarding player types. It is possible that one potential area of advance would involve assuming asymmetric information regarding types. There are two ways to think about such incomplete information; one is to assume heterogeneity in types (such as conditional cooperators and free riders) and the other is to assume heterogeneity in prior beliefs among conditional cooperators and then look for sequential equilibria in such games. Of course, it is hard to predict how tractable these models might be and how much additional light they might shed on the controversies at hand.

It is also clear that there will continue to be substantial contributions to this literature from a neuroeconomic perspective as in de Quervain et al. (2004), Fehr and Camerer (2007), Knoch et al. (2010) and Spitzer et al. (2007) especially in terms of understanding the motives behind altruistic punishments and norm compliance.

[35] The possible applications are numerous including charitable contributions (Andreoni 2006; Andreoni and Petrie 2004; List and Lucking-Reiley 2002; Martin and Randal 2008; Vesterlund 2003), tax compliance (Andreoni et al. 1998; Frey and Torgler 2007), managing natural resources (Cardenas et al. 2002; Carpenter and Seki 2009; Ostrom 1990; Ostrom et al. 1994), labor relations (Bewley 1999, 2005), legal enforcement (Bohnet et al. 2001; Kahan 2005) and many others, possibly unexplored as of yet.

Finally, this line of work is expanding upon traditional socio-biological theories of human cooperation with their emphasis on individual selection such as kin selection (Hamilton 1964), reciprocal altruism (Trivers 1971) or costly signaling (Zahavi and Zahavi 1997). In fact, the emerging literature on "strong reciprocity" (see Gintis et al. 2005 for a broad overview) argues that the presence of *homo reciprocans*—conditional cooperators who are willing to punish free riders even if such punishment is costly to the punishers—may be the primary driving force behind sustaining cooperation in a variety of social settings. These insights seem to provide new evidence in favor of group (or multi-level) selection as well as gene-culture co-evolution (Boyd and Richerson 1985; Cavalli-Sforza and Feldman 1981; Sober and Wilson 1998) as opposed to selection at the level of the individual. The findings of this line of work, especially those carried out among small-scale tribal societies, as in Henrich et al. (2004) or Efferson et al. (2010), will likely provide valuable clues to human evolutionary processes.

Acknowledgements I am deeply grateful to Jordi Brandts for his advice over multiple revisions. I am also indebted to Indira Basu, Jeff Carpenter, Gary Charness, Ernst Fehr, Simon Gächter, Martin Kocher, Nikos Nikiforakis, Louis Putterman, Andrew Schotter, Barry Sopher, Steve Tucker and four anonymous referees for extensive feedback. I thank Meg Paichayontvijit, Geoff Brooke, Parapin Prak and Lisa Meehan for research assistance. I gratefully acknowledge the financial assistance provided by the University of Auckland Research Council. I thank the University of Auckland Business School for granting me research and study leave which allowed me to work on the paper as well as Marie-Claire Villeval, Izabela Jelovac and Phillippe Polome of GATE (CNRS) at the University of Lyon and the Department of Economics at Rutgers University for their hospitality during my visits to those places while working on this survey. I dedicate this article to my parents-in-law Nandita and Samir Basu for looking after the kids and thereby giving me time to carry out my research for this article. I am responsible for all errors in the paper, though Ishannita (5) and Ananrita (2) should share some of the blame.

References

Ahn, T., Isaac, R. M., & Salmon, T. (2009). Coming and going: experiments on endogenous group sizes for excludable public goods. *Journal of Public Economics*, *93*(1–2), 336–351.

Ambrus, A., & Pathak, P. (2009). *Cooperation over finite horizons: a theory and experiments*. Working Paper, Department of Economics, Harvard University.

Anderson, C., & Putterman, L. (2006). Do non-strategic sanctions obey the law of demand? The demand for punishment in the voluntary contribution mechanism. *Games and Economic Behavior*, *51*(1), 1–24.

Anderson, S., Goeree, J., & Holt, C. (1998). A theoretical analysis of altruism and decision error in public goods games. *Journal of Public Economics*, *70*, 297–323.

Andreoni, J. (1988). Why free ride? Strategies and learning in public goods experiments. *Journal of Public Economics*, *37*, 291–304.

Andreoni, J. (1990). Impure altruism and donations to public goods: a theory of warm-glow giving. *Economic Journal*, *100*, 464–477.

Andreoni, J. (1995). Cooperation in public goods experiments: kindness or confusion? *American Economic Review*, *85*(4), 891–904.

Andreoni, J. (2006). Philanthropy. In: S.-C. Kolm & J. Mercier Ythier (Eds.), *Handbook of giving, reciprocity and altruism* (pp. 1201–1269). Amsterdam: North Holland.

Andreoni, J., & Croson, R. (2008). Partners versus strangers: random rematching in public goods experiments. In C. Plott & V. L. Smith (Eds.), *Handbook of experimental economics results* (pp. 776–783). Amsterdam: North-Holland.

Andreoni, J., & Petrie, R. (2004). Public goods experiments without confidentiality: a glimpse into fundraising. *Journal of Public Economics*, *88*, 1605–1623.

Andreoni, J., Erard, B., & Feinstein, J. (1998). Tax compliance. *Journal of Economic Literature*, *XXXVI*, 818–860.

Asch, S. (1951). Effects of group pressure upon the modification and distortion of judgment. In H. Guetzkow (Ed.), *Groups, leadership and men*. Pittsburgh: Carnegie Press.

Asch, S. (1955). Opinions and social pressure. *Scientific American*, *193*, 31–35.

Axelrod, R. (1986). An evolutionary approach to norms. *American Political Science Review*, *80*(4), 1095–1111.

Bardsley, N., & Sausgrüber, R. (2005). Conformity and reciprocity in public good provision. *Journal of Economic Psychology*, *26*, 664–681.

Berg, J., Dickhaut, J., & McCabe, K. (1995). Trust, reciprocity and social history. *Games and Economic Behaviour*, *10*, 122–142.

Bewley, T. (1999). *Why don't wages fall in a recession?* Cambridge: Harvard University Press.

Bewley, T. (2005). Fairness, reciprocity and wage rigidity. In H. Gintis, S. Bowles, R. Boyd, & E. Fehr (Eds.), *Moral sentiments and material interests: the foundations of cooperation in economic life* (pp. 303–338). Cambridge: MIT Press.

Bochet, O., & Putterman, L. (2009). Not just babble: Opening the black box of communication in a voluntary contribution experiment. *European Economic Review*, *53*(3), 309–326.

Bochet, O., Page, T., & Putterman, L. (2006). Communication and punishment in voluntary contribution experiments. *Journal of Economic Behavior & Organization*, *60*(1), 11–26.

Bohm, P. (1972). Estimating demand for public goods: an experiment. *European Economic Review*, *3*, 111–130.

Bohm, P. (1983). Revealing demand for an actual public good. *Journal of Public Economics*, *24*, 135–151.

Bohnet, I., Frey, B., & Huck, S. (2001). More order with less law: on contract enforcement, trust, and crowding. *American Political Science Review*, *95*, 131–144.

Bolton, G., & Ockenfels, A. (2000). ERC—a theory of equity, reciprocity and competition. *American Economic Review*, *90*, 166–193.

Bowles, S., & Gintis, H. (2002). Homo reciprocans. *Nature*, *415*, 125–128.

Boyd, R., & Richerson, P. (1985). *Culture and the evolutionary process*. Chicago: Chicago University Press.

Brandts, J., & Charness, G. (2000). Hot vs. cold: sequential responses and preference stability in experimental games. *Experimental Economics*, *2*(3), 227–238.

Brandts, J., & Charness, G. (2009). *The strategy versus the direct response method: a survey of experimental comparisons*. Working paper, Department of Economics, University of California—Santa Barbara and Department of Business Economics, U. Autonoma de Barcelona.

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

Brandts, J., Saijo, T., & Schram, A. (2004). How universal is behavior? A four country comparison of spite, cooperation and errors in voluntary contribution mechanisms. *Public Choice*, *119*, 381–424.

Brosig, J., Ockenfels, A., & Weimann, J. (2003). The effect of communication media on cooperation. *German Economic Review*, *4*(2), 217–241.

Bryan, J., & Test, M. (1967). Models and helping: naturalistic studies in aiding behavior. *Journal of Personality and Social Psychology*, *6*, 400–407.

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

Cardenas, J., Stranlund, J., & Willis, C. (2002). Economic inequality and burden-sharing in the provision of local environmental quality. *Ecological Economics*, *40*, 379–395.

Carpenter, J. (2004). When in Rome: conformity and the provision of public goods. *Journal of Socio-Economics*, *33*(4), 395–408.

Carpenter, J. (2007a). Punishing free-riders: how group size affects mutual monitoring and the provision of public goods. *Games and Economic Behavior*, *60*(1), 31–51.

Carpenter, J. (2007b). The demand for punishment. *Journal of Economic Behavior and Organization*, *62*(4), 522–542.

Carpenter, J., & Seki, E. (2009, forthcoming). Do social preferences increase productivity? Field experimental evidence from fishermen in Toyama Bay. *Economic Inquiry*, doi:10.1111/j.1465-7295.2009.00268.x.

Cavalli-Sforza, L., & Feldman, M. (1981). *Cultural transmission and evolution: a quantitative approach*. Princeton: Princeton University Press.

Charness, G., & Rabin, M. (2002). Understanding social preferences with simple tests. *Quarterly Journal of Economics*, *117*, 817–869.

Charness, G., & Yang, C. (2007). *Endogenous group formation and public goods provision: exclusion, exit, mergers and redemption*. Working paper, Department of Economics, University of California-Santa Barbara. Available. at SSRN: http://ssrn.com/abstract=932251.

Chaudhuri, A. (2009). *Experiments in economics: playing fair with money*. London: Routledge.

Chaudhuri, A., & Paichayontvijit, T. (2006). Conditional cooperation and voluntary contributions to a public good. *Economics Bulletin*, *3*(8), 1–14.

Chaudhuri, A., & Paichayontvijit, T. (2010a, forthcoming). Recommended play versus costly punishments in a laboratory voluntary contributions mechanism. In K. Ghosh Dastidar, H. Mukhopadhyay, & U. Sinha (Eds.), *Dimensions of economic theory and policy: essays for Anjan Mukherji*. New Delhi: Oxford University Press.

Chaudhuri, A., & Paichayonvijit, T. (2010b). *Does strategic play explain the decay in contributions in laboratory public goods games*? Working paper, Department of Economics, University of Auckland.

Chaudhuri, A., Graziano, S., & Maitra, P. (2006). Social learning and norms in a public goods experiment with intergenerational advice. *Review of Economic Studies*, *73*(2), 357–380.

Cinyabuguma, M., Page, T., & Putterman, L. (2005). Can second order punishment deter perverse punishment? *Experimental Economics*, *9*, 265–279.

Cinyabuguma, M., Page, T., & Putterman, L. (2006). Cooperation under the threat of expulsion in a public goods experiment. *Journal of Public Economics*, *89*, 1421–1435.

Cooper, D., & Kagel, J. (2009, forthcoming). Other regarding preferences: a survey of experimental results In J. Kagel & A. Roth (Eds.), *The handbook of experimental economics* (Vol. 2). Princeton: Princeton University Press.

Cox, J., Friedman, D., & Gjerstad, S. (2007). A tractable model of reciprocity and fairness. *Games and Economic Behavior*, *59*, 17–45.

Cox, J., Friedman, D., & Sadiraj, V. (2008). Revealed altruism. *Econometrica*, *76*(1), 31–69.

Croson, R. (2007). Theories of commitment, altruism and reciprocity: Evidence from linear public goods games. *Economic Inquiry*, *45*(2), 199–216.

Dawes, R. (1980). Social dilemmas. *Annual Review of Psychology*, *31*, 169–193.

Dawes, R., McTavish, J., & Shaklee, H. (1977). Behavior, communication and assumptions about other people's behavior in a common dilemma situation. *Journal of Personality and Social Psychology*, *35*, 1–11.

Dawes, R., Orbell, J., Simmons, R., & van de Kragt, A. (1986). Organizing groups for collective action. *American Political Science Review*, *8*, 1171–1185.

de Oleveira, A., Croson, R., & Eckel, C. (2009). *One bad apple: uncertainty and heterogeneity in public good provision*. Working paper, Department of Resource Economics, University of Massachusetts-Amherst. .

de Quervain, D., Fischbacher, U., Treyer, V., Schallhammer, M., Schnyder, U., Buck, A., & Fehr, E. (2004). The neural basis of altruistic punishments. *Science*, *305*, 1254–1258.

Denant-Boemont, L., Masclet, D., & Noussair, C. (2007). Punishment, counterpunishment and sanction enforcement in a social dilemma experiment. *Economic Theory*, *33*, 145–167.

Dufwenberg, M., & Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, *47*, 268–298.

Efferson, C. H. Bernhard, Fischbacher, U., & Fehr, E. (2010). *The ultimate origins of human prosocial behavior: an empirical test*. Working paper, Institute for Empirical Research in Economics, University of Zurich.

Egas, M., & Riedl, A. (2008). The economics of altruistic punishment and the maintenance of cooperation. *Proceedings of the Royal Society B: Biological Sciences*, *275*(1637), 871–878.

Ehrhart, K.-M., & Keser, C. (1999). *Cooperation and mobility: on the run*. Working paper, CIRANO and University of Karlsruhe.

Ertan, A., Page, T., & Putterman, L. (2009). Who to punish? Individual decisions and majority rule in mitigating the free rider problem. *European Economic Review*, *53*(5), 495–511.

Falk, A., & Fischbacher, U. (2006). A theory of reciprocity. *Games and Economic Behavior*, *54*, 293–315.

Fehr, E., & Camerer, C. (2007). Social neuroeconomics: the neural circuitry of social preferences. *Trends in Cognitive Science*, *11*(10), 419–427.

Fehr, E., & Fischbacher, U. (2004a). Third party punishment and social norms. *Evolution and Human Behavior*, *25*, 63–87.

Fehr, E., & Fischbacher, U. (2004b). Social norms and human cooperation. *Trends in Cognitive Sciences*, *8*(4), 185–190.

Fehr, E., & Fischbacher, U. (2005a). Altruists with green beards. *Analyse & Kritik*, *27*(1), 73–84.

Fehr, E., & Fischbacher, U. (2005b). Human altruism: proximate patterns and evolutionary origins. *Analyse & Kritik*, *27*(1), 6–47.

Fehr, E., & Gächter, S. (2000). Cooperation and punishment in public goods experiments. *American Economic Review*, *90*(4), 980–994.

Fehr, E., & Gächter, S. (2002). Altruistic punishment in humans. *Nature*, *415*, 137–140.

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

Fischbacher, U., & Gächter, S. (2009). *On the behavioral validity of the strategy method in public good experiments*. Discussion Paper No. 2009-25, Centre for Decision Research and Experimental Economics, University of Nottingham.

Fischbacher, U., & Gächter, S. (2010). Social preferences, beliefs, and the dynamics of free riding in public good experiments. *American Economic Review*, *100*(1), 541–556.

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

Frey, B., & Meier, S. (2004). Social comparisons and pro-social behavior. Testing 'conditional cooperation' in a field experiment. *American Economic Review*, *94*(5), 1717–1722.

Frey, B., & Torgler, B. (2007). Tax morale and conditional cooperation. *Journal of Comparative Economics*, *35*, 136–159.

Gächter, S. (2007). Conditional cooperation. Behavioral regularities from the lab and the field and their policy implications. In B. Frey & A. Stutzer (Eds.), *Economics and psychology. A promising new cross-disciplinary field. CESifo Seminar Series*. Cambridge: MIT Press.

Gächter, S., & Hermann, B. (2009). Reciprocity, culture and human cooperation: previous insights and a new cross cultural experiment. *Philosophical Transactions of the Royal Society B*, *364*, 791–806.

Gächter, S., & Herrmann, B. (2010, forthcoming). The limits of self-governance when cooperators get punished: experimental evidence from urban and rural Russia. *European Economic Review*. doi:10.1016/j.euroecorev.2010.04.003.

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

Gächter, S., & Thöni, C. (2007). Rationality and commitment in voluntary cooperation: insights from experimental economics. In P. Fabienne & H. B. Schmidt (Eds.), *Rationality and Commitment*. Oxford: Oxford University Press.

Gächter, S., Renner, E., & Sefton, M. (2008). The long run benefits of punishment. *Science*, *322*, 1510.

Geanakoplos, J., Pearce, D., & Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, *1*, 60–79.

Gintis, H., Bowles, S., Boyd, R., & Fehr, E. (Eds.) (2005). *Moral sentiments and material interests*. Cambridge: MIT Press.

Griesinger, D., & Livingston, J. (1973). Toward a model of interpersonal motivation in experimental games. *Behavioral Science*, *18*, 173–188.

Guala, F. (2010). *Reciprocity: weak or strong? what punishment experiments do (and do not) demonstrate*. IDEAS Working Paper 2010-23, University of Milan. http://ideas.repec.org/p/mil/wpdepa/2010-23.html.

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

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

Hamilton, W. (1964). The genetical evolution of social behavior. *Journal of Theoretical Biology*, *37*, 1–52.

Henrich, J., & Boyd, R. (2001). Why people punish defectors: weak conformist transmission can stabilize costly enforcement of norms in cooperative dilemmas. *Journal of Theoretical Biology*, *208*, 79–89.

Henrich, J., Heine, S., & Norenzayan, A. (2010). Most people are not WEIRD. *Nature*, *466*, 29.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., & Gintis, H. (Eds.) (2004). *Foundations of human sociality: economic experiments and ethnographic evidence in fifteen small-scale societies*. New York: Oxford University Press.

Hermann, B., & Thöni, C. (2008). Measuring conditional cooperation: a replication study in Russia. *Experimental Economics*, *12*(1), 87–92.

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

Isaac, R., & Walker, J. (1988a). Communication and free riding behavior: the voluntary contributions mechanism. *Economic Inquiry*, *26*(4), 585–608.

Isaac, R., & Walker, J. (1988b). Group size effects in public goods provision: the voluntary contributions mechanism. *Quarterly Journal of Economics*, *103*, 179–199.

Isaac, R., Walker, J., & Thomas, S. (1984). Divergent evidence on free riding: an experimental examination of possible explanations. *Public Choice*, *43*, 113–149.

Isaac, R., McCue, K., & Plott, C. (1985). Public goods provision in an experimental environment. *Journal of Public Economics*, *26*, 51–74.

Isaac, R., Walker, J., & Williams, A. (1994). Group size and the voluntary provision of public goods. *Journal of Public Economics*, *54*, 1–36.

Kahan, D. (2005). The logic of reciprocity: trust, collective action, and law. In H. Gintis, S. Bowles, R. Boyd, & E. Fehr (Eds.), *Moral sentiments and material interests: the foundations of cooperation in economic life* (pp. 339–378). Cambridge: MIT Press.

Kelley, H., & Stahelski, A. (1970). Social interaction basis of cooperators' and competitors' beliefs about others. *Journal of Personality and Social Psychology*, *16*, 66–91.

Keser, C., & van Winden, F. (2000). Conditional cooperation and voluntary contributions to public goods. *Scandinavian Journal of Economics*, *102*(1), 23–39.

Kim, O., & Walker, M. (1984). The free rider problem: experimental evidence. *Public Choice*, *43*, 3–24.

Knoch, D., Giannoti, L., Baumgartner, T., & Fehr, E. (2010). A neural marker of costly punishment behavior. *Psychological Science*, *21*(3), 337–342.

Kocher, M., Cherry, T., Kroll, S., Netzer, R., & Sutter, M. (2008). Conditional cooperation on three continents. *Economics Letters*, *101*(3), 175–178.

Kosfeld, M., Okada, A., & Riedl, A. (2009). Institution formation in public goods games. *American Economic Review*, *99*(4), 1335–1355.

Kreps, D., Milgrom, P., Roberts, J., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, *27*, 245–252.

Kurzban, R., & Houser, D. (2005). An experimental investigation of cooperative types in human groups: a complement to evolutionary theory and simulations. *Proceedings of the National Academy of Sciences*, *102*(5), 1803–1807.

Ledyard, O. (1995). Public goods: some experimental results. In J. Kagel & A. Roth (Eds.), *Handbook of experimental economics*. Princeton: Princeton University Press (Chap. 2).

Liebrand, W. (1984). The effects of social motives, communication and group size on behavior in an n-person multi stage mixed motive game. *European Journal of Social Psychology*, *14*, 239–264.

List, J., & Lucking-Reiley, D. (2002). The effects of seed money and refunds on charitable giving: experimental evidence from a university capital campaign. *Journal of Political Economy*, *110*(1), 215–233.

Martin, R., & Randal, J. (2008). How is donation behavior affected by the donations of others? *Journal of Economic Behavior and Organization*, *67*(1), 228–238.

Marwell, G., & Ames, R. (1979). Experiments on provision of public goods I: resources, interest, group size, and the free riding problem. *American Journal of Sociology*, *84*(6), 1335–1360.

Marwell, G., & Ames, R. (1980). Experiments on provision of public goods II: provision point, stake, experience, and the free riding problem. *American Journal of Sociology*, *85*(4), 926–937.

Marwell, G., & Ames, R. (1981). Economists free ride, does anyone else? *Journal of Public Economics*, *15*, 295–310.

Masclet, D., Noussair, C., Villeval, M., & Tucker, S. (2003). Monetary and nonmonetary punishment in the voluntary contributions mechanism. *American Economic Review*, *93*(1), 366–380.

Moscovici, S. (1985). Social influence and conformity. In G. Lindzey & E. Aronson (Eds.), *The Handbook of Social Psychology*. New York: Random House.

Nikiforakis, N. (2008). Punishment and counter-punishment in public good games: can we really govern ourselves? *Journal of Public Economics*, *92*, 91–112.

Nikiforakis, N. (2010). Feedback, punishment and cooperation in public goods experiments. *Games and Economic Behavior*, *68*, 689–702.

Nikiforakis, N., & Normann, H. (2008). A comparative statics analysis of punishment in public good experiments. *Experimental Economics*, *11*, 358–369.

Noussair, C., & Tucker, S. (2005). Combining monetary and social sanctions to promote cooperation. *Economic Inquiry*, *43*(3), 649–660.

Offerman, T., Sonnemans, J., & Schram, A. (1996). Value orientations, expectations and voluntary contributions in public goods. *Economic Journal*, *106*(437), 817–845.

Olson, M. (1965). *The logic of collective action: public goods and the theory of groups*. Cambridge: Harvard University Press.

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

Orbell, J., Dawes, R., & van de Kragt, A. (1990). The limits of multilateral promising. *Ethics*, *100*, 616–627.

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

Ostrom, E., Walker, J., & Gardner, R. (1992). Covenants with and without a sword: self-governance is possible. *American Political Science Review*, *86*(2), 404–417.

Ostrom, E., Gardner, R., & Walker, J. (Eds.) (1994). *Rules, games, and common pool resources*. Ann Arbor: University of Michigan Press.

Page, T., Putterman, L., & Unel, B. (2005). Voluntary association in public goods experiments: reciprocity, mimicry, and efficiency. *Economic Journal*, *115*, 1032–1053.

Palfrey, T., & Prisbrey, J. (1997). Anomalous behavior in public goods experiments: how much and why? *American Economic Review*, *87*, 829–846.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, *80*(5), 1281–1302.

Rege, M., & Telle, K. (2004). The impact of social approval and framing on cooperation in public good situations. *Journal of Public Economics*, *88*(7–8), 1625–1644.

Rockenbach, B., & Milinski, M. (2006). The efficient interaction of indirect reciprocity and costly punishment. *Nature*, *444*, 718–723.

Seely, B., Van Huyck, J., & Battalio, R. (2005). Credible assignments can improve efficiency in laboratory public goods games. *Journal of Public Economics*, *89*, 1437–1455.

Sefton, M., Shupp, R., & Walker, J. (2007). The effect of rewards and sanctions in provision of public goods. *Economic Inquiry*, *45*(4), 671–690.

Selten, R. (1967). Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments. In H. Sauermann (Ed.), *Beiträge zur experimentellen Wirtschaftsforschung* (Vol. I, pp. 136–168). Tübingen: Mohr.

Sober, E., & Wilson, D. (1998). *Unto others: the evolution and psychology of unselfish behavior*. Cambridge: Harvard University Press.

Sonnemans, J., Schram, A., & Offerman, T. (1999). Strategic behavior in public good games: when partners drift apart. *Economics Letters*, *62*(1), 35–41.

Spitzer, M., Fischbacher, U., Herrnberger, B., Grön, G., & Fehr, E. (2007). The neural signature of social norm compliance. *Neuron*, *56*, 185–196.

Trivers, R. (1971). The evolution of reciprocal altruism. *Quarterly Review of Biology*, *46*, 36–57.

Vesterlund, L. (2003). The informational value of sequential fundraising. *Journal of Public Economics*, *87*, 627–657.

Walker, J., & Halloran, M. (2004). Rewards and sanctions and the provision of public goods in one-shot settings. *Experimental Economics*, *7*(3), 235–247.

Yamagishi, T. (1986). The provision of a sanctioning system as a public good. *Journal of Personality and Social Psychology*, *51*(1), 110–116.

Yamagishi, T. (1988). The provision of a sanctioning system in the United States and Japan. *Social Psychology Quarterly*, *51*(3), 265–271.

Zahavi, A., & Zahavi, A. (1997). *The handicap principle: a missing piece of Darwin's puzzle*. New York: Oxford University of Press.