Strong Reciprocity

Norms and Preferences Governing Cooperation and Punishment Behaviour

Till Olaf Weber

Thesis submitted to the University of Nottingham for the degree of Doctor of Philosophy

September 2017

Abstract

Many problems that societies face have the character of social dilemmas, in which cooperation benefits the whole society but is costly to the individual. The recent literature in experimental economics has focused on uncovering driving factors of cooperative success in social dilemmas. This thesis contributes to this literature and includes three research studies that investigate the influence of individual cooperative dispositions, societal and cultural differences, as well as institutional differences on human cooperative behaviour. Chapter 1 introduces the research questions, discusses the research methods used, and outlines the substantive contributions of the thesis. Chapter 2 presents an experimental test of a common implicit assumption in the literature, which suggests that only people with a cooperative disposition engage in the punishment of defectors in social dilemmas. The experimental test rejects this assumption and shows that individual cooperativeness is independent of one's propensity to punish. Chapter 3 investigates the channels through which culture and societal differences affect cooperative behaviour. The experimental results show that societal differences in behaviour are mainly driven through differences in beliefs about other people's behaviour. Chapter 4 reports on an experimental comparison of informal and formal sanctioning institutions. These experiments show that informal sanctions like peer pressure are necessary to foster high and stable cooperation levels in the long run. Chapter 5 concludes.

Acknowledgements

First and foremost, I would like to thank my supervisors Professor Simon Gächter and Dr Benedikt Herrmann for their excellent guidance throughout my time of study, their enthusiasm for this research project and invaluable comments on my drafts.

Additionally, I am very grateful to Dr Ori Weisel, Dr Jonathan Schulz and all members of the ERC Research Group for their criticism and suggestions on various stages of the projects, which helped to shape and improve this thesis tremendously. Similarly, I would like to thank the members of CeDEx for their feedback and helpful comments during several seminars.

I also would like to acknowledge generous financial support from Professor Gächter's European Research Council grant ERC-AdG 295707 COOPERATION which provided a full PhD studentship as well as the funding to conduct the experimental studies included in this thesis.

I would like to thank all my friends and colleagues who have supported me in countless occasions and helped to turn the PhD into a very enjoyable experience. Last but not least, I would like to thank Eleni and my family. Without their love and encouragement, this would have not been possible.

Contents

List of Figures

List of Tables

CHAPTER 1
Introduction

Societies face many challenges like curbing pollution, restricting the exploitation of common resources or the funding of national parks. Hardin (1968) argued that all of these challenges have common characteristics. Cooperative behaviour benefits societal welfare but is costly for the individual. Hardin has termed the collapse of cooperation in these situations the 'tragedy of the commons'. Nearly 50 years later, we continue to see all too frequent examples of the tragedy of the commons. For instance, the success of the Paris agreement on climate change mitigation, which ensures a collaborative effort in curbing carbon dioxide emissions to the benefit of all societies, is threatened by a unilateral withdrawal of the US (Shear 2017, 1 June). This is a timely example which demonstrates the enduring relevance of the tragedy of the commons.

In some situations, humans display a remarkable degree of cooperativeness and bear the cost of cooperation to the benefit of others. This is even true for cooperation in large groups with strangers that extend beyond genealogical kinship (Bowles and Gintis 2011). Examples of this behaviour include periodic democratic movements, workers' strikes and even the shunning of non-cooperators in these actions. Additionally, a multitude of experimental studies have been devoted to uncovering the driving factors of human cooperative behaviour and the factors which determine the success or conversely the breakdown of human cooperation (Ledyard 1995, Balliet et al. 2011, Chaudhuri 2011).

However, the investigation of potential influences and the channels through which they shape behaviour is far from complete. This thesis contributes to the literature by investigating three important factors affecting human cooperative behaviour. These are (*a*) individual differences in cooperativeness, (*b*) the

consequences of the societal and cultural background on cooperative behaviour and (*c*) the influence of the design of sanctioning institutions. The remainder of this chapter introduces the three research topics in more detail and explains the research questions which will be subsequently explored.

The theory of strong reciprocity explains the emergence and maintenance of cooperation by the presence of individuals who are willing to incur costs to help other people who helped them (strong positive reciprocity) and to punish people who wronged them (strong negative reciprocity), even in the absence of strategic incentives to do so. In contrast to such 'strong reciprocators', self-regarding people cooperate and punish only if there are sufficiently strong future benefits from doing so. The conjecture that only a share of the population—the strong reciprocators—cooperate and punish defectors if it is costly to do so appears intuitive and is a common implicit assumption in the literature (e.g., Camerer and Fehr 2006). Yet, an unanswered question is whether this assumption holds up to an experimental test.

Societal and cultural differences have previously been identified as a key driver of individual cooperative behaviour when trading off between personal benefits and societal welfare. Cooperative behaviour has been shown to vary across small-scale societies (Henrich et al. 2005) as well as across large-scale industrialised nations (Gächter et al. 2010). Since cooperative dispositions and beliefs jointly affect cooperative efforts (Fischbacher and Gächter 2010), the question arises which of the driving factors are affected by societal and cultural differences. Are the differences in behaviour due to differences in cooperative dispositions or differences in beliefs? Furthermore, the impact of punishing free riders in repeated interactions has been shown to greatly differ across societies (Herrmann et al. 2008). This raises the question whether this result is best explained by differences in preferences for punishment or is rooted in strategic considerations unique to repeated interactions.

Apart from an intrinsic motivation to cooperate, prosocial behaviour might be induced by informal punishment (e.g., peer-pressure) or formal sanctioning institutions (i.e., police and courts). All of these factors influence cooperative behaviour simultaneously. It is therefore not possible to disentangle the relative effect of one single factor on behaviour merely by observing behaviour in the real world. This raises the questions of what the relative effects of informal and formal sanctioning institutions

are in supporting cooperative behaviour and whether cooperative dispositions explain the engagement in the respective sanctioning institutions.

This thesis uses laboratory experiments to explore the research questions introduced above. This methodology has several advantages that are key to investigate the driving factors of cooperative behaviour. First, using laboratory experiments allows for controlling the factors that might influence cooperative behaviour, such as the degree of anonymity or communication between participants. By conducting anonymous one-shot experiments with strangers, we can exclude any strategic incentive for cooperative behaviour and thus conduct a clean investigation of cooperative dispositions. Second, conducting laboratory experiments with students allows us to measure the differences in behaviour across societies. To do that, we keep the experimental set-up constant and draw on student subjects that have similar socio-economic backgrounds across societies. Third, laboratory experiments are useful to test differences between sanctioning institutions, because they allow an exogenous allocation of institutions. Thus, they exclude the possibility of self-selection into different institutional regimes and any connected spill-over effects.

This thesis includes three self-contained research studies which constitute Chapters 2–4. Each of these chapters is followed by an appendix containing supplementary analyses and the instructions used during the experimental investigations.

Chapter 2 is devoted to testing a common implicit assumption in the literature on cooperative behaviour which we call the 'Strong Reciprocators Assumption'. In a one-shot public good experiment, we measure participants' disposition towards strong positive reciprocity and classify participants as either Dispositional Conditional Cooperators (DCC) or Dispositional Free Riders (DFR). Participants then play a one-shot direct-response public goods game either with or without punishment. We find that DFR only contribute to the public good when punishment is possible, whereas DCC also cooperate without threat of punishment. Surprisingly and in contrast to the 'Strong Reciprocators Assumption', the disposition towards strong positive reciprocity is unrelated to the disposition towards strong negative reciprocity; that is, the punishment behaviour of DCC and DFR is practically identical. The 'burden of cooperation' is thus

carried by a larger set of individuals than previously assumed, which can help explain the high levels of cooperation observed when punishment opportunities are available.

Chapter 3 explores the influence of societal differences in culture and institutions on cooperative behaviour. We measure the cooperative dispositions, beliefs and the propensity to punish free riding, using variants of public goods games with student participants in four countries (Morocco, Turkey, the UK and the US). We chose these countries because Morocco and Turkey are culturally similar and so are the UK and the US. We place the culturally similar countries into cultural clusters and note that cultural differences are quite large when comparing across the two cultural clusters. We find that differences in the cooperation rates across societies cannot be explained by the variation in the distribution of cooperative dispositions alone. Beliefs about other people's cooperative efforts and propensities to punish help to explain cross-societal variation in behaviour. Furthermore, costly altruistic punishment in our one-shot game is remarkably similar across different societies. This contrasts previous studies that use repeated games to investigate cross-societal differences in punishment. Thus, societal differences in punishment are likely to be driven by strategic play or retaliation emerging in repeated interactions. Interestingly, we find that emotional responses to free riding are similar across societies meaning that negative emotions are a likely driver of costly altruistic punishment.

Chapter 4 reports on a laboratory experiment designed to disentangle and quantify the relative impact of informal and formal sanctioning institutions. We first elicit cooperative dispositions and then conduct four variations of public goods games: without punishment, with informal peer-punishment, with a formal sanctioning institution and lastly a combination of formal and informal punishment. Informal peer-punishment induced high and stable cooperation levels. The formal sanctioning institution is the least efficient in terms of social welfare. A best-reply analysis reveals that formal sanctions crowd out voluntary contributions. The combination of informal and formal sanctions leads to the highest levels of cooperation and efficiency. Yet these levels are not significantly different from the levels induced by informal punishment alone. We conclude that informal peer-punishment is crucial to stabilising cooperation levels in the long run. In contrast, the formal sanctioning institution encourages best-reply reasoning and only induces cooperation when the monetary incentives are large enough. Additionally, we find that a cooperative disposition supports the funding of

pool-punishment, showing that the costly formation of formal sanctioning institutions relies on a smaller share of supporters.

Chapter 5 summarises the three studies and discusses their main contributions to economic research as well as their implications for policy making.

CHAPTER 2

Strong Reciprocity and Selfishness Revisited[1]

2.1 Introduction

The neoclassical *Homo economicus* model of humans as rational beings in narrow pursuit of their own self-interest does not sit well with the varied range of uniquely human phenomena that depend on self-sacrificial cooperative behaviour. Risk taking in large-scale conflict, caring for the sick and disabled and periodic collective efforts, like democratic movements combating authoritarian regimes or workers engaging in collective action as well as the shunning of non-cooperators in these actions, illustrate how humans stand out among other (cooperative) species by cooperating in large groups that extend genealogical kinship (Bowles and Gintis 2011), even when cooperation entails the forgoing of private welfare in the interest of the greater good (Bowles and Gintis 1987, Gintis 2000).

A central result in the literature is that human cooperation, in situations similar to the ones described above, can be sustained through *strong reciprocity*. In the words of Herbert Gintis, who coined the term, 'a strong reciprocator is predisposed to cooperate with others and punish non-cooperators, even when this behaviour cannot be justified in terms of self-interest, extended kinship, or reciprocal altruism' (see Gintis 2000, p. 169). Strong reciprocity entails that even in the absence of material incentives, for example in one-shot interactions with strangers, there is a willingness to pay a cost for both cooperating with cooperative others (strong *positive* reciprocity) and for punishing non-cooperative others (strong *negative* reciprocity). Proximate explanations for this behaviour might be an intrinsic desire to equalise payoffs, react to the intention

[1] This chapter draws on joint work in progress with Simon Gächter and Ori Weisel.

of others or avoid guilt and relieve anger. This concept is distinct from *weak reciprocity*, which refers to settings where cooperation and punishment can be rationalised by selfish and strategic incentives (Gintis 2000).

A common assumption in the literature is that strong reciprocity is manifested in the form of s*trong reciprocators*, a set of individuals who are willing to pay a cost for both cooperating with other cooperators and punishing non-cooperators (Gintis 2000, Fehr et al. 2002, Boyd et al. 2003, Fehr and Fischbacher 2003, Bowles and Gintis 2004, Gintis et al. 2005, Camerer and Fehr 2006, West et al. 2007, Gintis et al. 2008, Bowles and Gintis 2011, Yamagishi et al. 2012). In other words, the assumption—which we refer to as the *Strong Reciprocators Assumption*—is that the combination of positive and negative reciprocity is present (and might even have a genetic component) within each strong reciprocator. It is this set of individuals—strong reciprocators—that enable the emergence and maintenance of cooperation. It follows from the Strong Reciprocators Assumption that *only* individuals who are willing to incur a personal cost to cooperate with cooperative others (strong positive reciprocators) may be willing to engage in the costly punishment of non-cooperators (i.e., engage in strong negative reciprocity). Those who are unwilling to pay a personal cost to cooperate with cooperative others will *never* pay for punishing non-cooperative others, and thus play no role in the emergence and maintenance of cooperation. The Strong Reciprocators Assumption is indeed a compelling one, especially when considering that costly punishment can reasonably be considered a form of second-order cooperation in itself (Heckathorn 1989, Panchanathan and Boyd 2004, Balafoutas et al. 2016); after all, why would individuals who are never willing to engage in self-sacrifing first-order cooperation be willing to spend resources for engaging in second-order cooperation in the form of punishment?

Strengthening the ubiquity of the Strong Reciprocators Assumption, formal reciprocity models are typically unidimensional, using a single parameter at the individual level to represent both the positive (i.e., strong positive reciprocity) and negative (strong negative reciprocity) sides of reciprocal preferences (Falk and Fischbacher 2006, West et al. 2007, Dufwenberg et al. 2011). Although compelling and often implicitly adopted, the Strong Reciprocators Assumption has yet to be put to a convincing test, and existing evidence on the links between strong negative and positive reciprocity is scarce, with mixed results (Yamagishi et al. 2012, Egloff et al. 2013,

Eriksson et al. 2014, Peysakhovich et al. 2014). A particular piece of evidence that puts the assumption into question is that while the proportion of strong positive reciprocators in the population has been estimated, based on multiple studies, at around 50% (Chaudhuri 2011), work on negative reciprocity shows that 84.3% of the people engage in costly altruistic punishment (Fehr and Gächter 2002). These figures suggest a sizeable proportion of people who are not positive reciprocators, but do engage in negative reciprocity.

The current work tests the Strong Reciprocators Assumption by making a distinction—both conceptual and operational—between agents' intrinsic disposition towards (first-order) cooperation and their actual cooperative behaviour in the presence of punishment. The distinction addresses a conceptual confound in much of the existing literature, which arises from the fact that dispositions towards cooperation (i.e., whether individuals are positive reciprocators) are not measured independently, but are typically inferred from behaviour in experimental games in which players' moves are simultaneous, their payoffs are interdependent, and there is a threat of being punished. A typical result in such settings is that individuals who cooperate are also those who tend to punish free riders (e.g., Falk et al. 2005). While this pattern is not in contrast with the Strong Reciprocators Assumption, it is also not convincing support of it and certainly not proof. The reason is that cooperation in such games is not influenced only by one's initial disposition towards cooperation, but also by *beliefs and expectations* about the cooperation and punishment behaviour of others. To illustrate, consider an individual who is in principle willing to positively reciprocate others' cooperative behaviour, but happens to be pessimistic about the chance that others will actually cooperate. Such a person may refrain from cooperation, despite her positive reciprocal tendency, and her behaviour will be identical to that of an individual who is not willing to cooperate regardless of the behaviour of others (e.g., even when others do cooperate). Similarly, cooperation in the presence of punishment opportunities is not necessarily an indication of a disposition to cooperate, but might also be a result of the changed incentives (i.e., fear of suffering from punishment if one does not cooperate).

We employ a two-phase design to overcome this problem. In the first phase, we classify participants according to their *disposition* towards strong positive reciprocity as either *Dispositional Conditional Cooperators* (strong positive reciprocators; DCC) or *Dispositional Free Riders* (not willing to positively reciprocate others' cooperation;

DFR). Then, in the second phase, we examine their cooperation and punishment *behaviour*. Importantly, the classification procedure in the first phase controls for players' beliefs about the behaviour of others and is thus a clean measure of players' disposition towards strong positive reciprocity (first-order cooperation; see below). The second phase, as it is a one-shot interaction in which there is no material incentive to use punishment, is a clean measure of players' disposition towards strong negative reciprocity (second-order cooperation in the form of costly punishment; see below).

Negative emotions—anger in particular—have been shown to be a proximate mechanism behind strong negative reciprocity (Fehr and Gächter 2002, Sanfey et al. 2003, Hopfensitz and Reuben 2009, Cubitt et al. 2011). The punishment of non-cooperators serves as an outlet for negative emotions, and the psychological reward associated with punishment can outweigh the material cost (de Quervain et al. 2004). We utilise the link between negative emotions and costly punishment to explore potential individual differences in the proximate explanation of punishment. If it is indeed the case that DFR never punish, the question arises: Do DFR experience less anger than DCC in the face of others' defection? Or do DFR experience similar levels of anger, but still refrain from punishing? To answer these questions, and in the case that DFR do engage in punishment, to understand whether they do so for similar reasons to DCC, we elicit the intensity of a range of emotions experienced by participants.

2.2 Study 1

Study 1 included two phases, which were both based on the same four-person one-shot public goods game. Each group member received an endowment of 20 money units (MU) and decided how many of these MU to keep for herself and how many to contribute to a group project. Contributions to the group project were multiplied by 1.6, and then redistributed equally among all four group members. Contribution in this game is self-sacrificial because for each MU that an individual player contributes, the return to that particular player is only 0.4 MU (1.6 divided by 4), regardless of the contributions of other group members. From an individual perspective, it is thus optimal to refrain from contribution and keep all 20 MU for oneself, regardless of the behaviour of the other players in the group. From the group perspective, however, it is best if each group member contributes all 20 MU.

The first phase was used to determine participants' disposition towards strong positive reciprocity in an incentive compatible way, using a variant of the strategy method (Selten 1967). Each player made a series of conditional contribution choices, indicating how many MU she wishes to contribute for each possible mean integer contribution of the other three group members, from 0 to 20 MU (Fischbacher et al. 2001). We refer to the resulting set of choices as the player's *contribution schedule*. DCC are characterised by a positively sloped contribution schedule; their willingness to contribute is increasing in the contributions of their group members. DFR are never willing to contribute, regardless of the contributions of their group members. To avoid participants from updating their beliefs about the behaviour of others, no feedback was provided between the first and second phase.

The second phase of Study 1 was introduced as a surprise to preclude any considerations of generalised reciprocity in the first phase. In the second phase, participants were randomly rematched with new group members, and played a one-shot public goods game this time in the direct-response method (i.e., without conditioning contributions on those of others). After all four group members decided—unconditional on the decisions of others—how many MU to contribute, they were asked for their beliefs regarding the contributions of others, and then received feedback about their group members' actual contributions. In the *With Punishment* treatment, each player had the option to assign up to 5 MU for the punishment of each of her group members. The punishment ratio was 2:1; for each MU invested in punishing, the punished group member lost 2 MU and the punisher paid 1 MU. After making the punishment decision and before learning the results of the punishment stage, participants completed an emotions questionnaire, which elicited the intensity of their emotional response to the contribution behaviour of their group members (13 positive and negative emotions; see Methods section for details). In the *Without Punishment* treatment, punishment was not available. Following feedback, participants proceeded directly to the emotions questionnaire. 184 participants took part in Study 1, 92 in the *With Punishment* treatment and 92 in the *Without Punishment* treatment.

2.2.1 Results

2.2.1.1 Dispositions towards strong positive reciprocity

Following the first phase, 48.9% of the players were classified as *Dispositional Conditional Cooperators* (DCC), whose contribution schedule is increasing in the

contributions of their group members, and 26.6% of the players as *Dispositional Free Riders* (DFR), who are not willing to contribute anything regardless of the contribution of the other group members. The remaining 24.5% are unclassified and excluded from analysis (see Appendix; Figure 2.5 for illustration, and Methods section for the classification criteria and additional details).

2.2.1.2 Beliefs and contribution behaviour

Without punishment both DCC and DFR behave in a way that is true to their disposition towards cooperation (Figure 2.1a). A full 78% of DFR are perfectly consistent with their elicited disposition and contribute nothing, whereas 76% of DCC make a positive contribution ($M_{DFR} = 1.85$, $M_{DCC} = 7.39$, Mann-Whitney $Z = 4.23$, $p < .001$). Furthermore, DCC contribute slightly less than what they believe others contribute ($M_{belief} = 8.24$; $M_{contribution} = 7.39$, Wilcoxon signed-rank test $Z = -1.08$, $p = .282$). In contrast, the contributions of DFR are significantly lower than beliefs about the others' contributions ($M_{contribution} = 1.85$; $M_{belief} = 3.30$; Wilcoxon signed-rank test $Z = -2.22$, $p = .027$). Interestingly, beliefs of DFR are significantly lower than beliefs of DCC ($M_{DFR} = 3.30$, $M_{DCC} = 8.24$, Mann-Whitney $Z = 3.93$, $p < .001$).

The presence of punishment opportunities (Figure 2.1b) has little effect on DCC, with neither beliefs nor contributions significantly differing as compared to *Without Punishment* (Mann-Whitney $Z_{belief} = -0.83$, $p_{belief} = .409$; $Z_{contribution} = -1.34$, $p_{contribution} = .180$). Beliefs and behaviour of DFR, however, are dramatically affected by punishment, increasing to DCC-like levels (Mann-Whitney $Z_{belief} = -2.00$, $p_{belief} = .046$; $Z_{contribution} = -2.00$, $p_{contribution} = .045$). Note that this shift in beliefs and contributions does not necessarily reflect a change in DFR's disposition towards cooperation; rather, it is likely to reflect the expectation that the opportunity to punish will raise overall contributions and a desire to avoid being punished by contributing (almost) as much as the others.

Overall, the results with respect to beliefs and contributions are consistent with the 'positive' side of the Strong Reciprocators Assumption. Without punishment, DCC positively reciprocate what they expect others to contribute, conditioning their own contribution on their beliefs about others' contributions, while DFR mostly refrain from contribution.

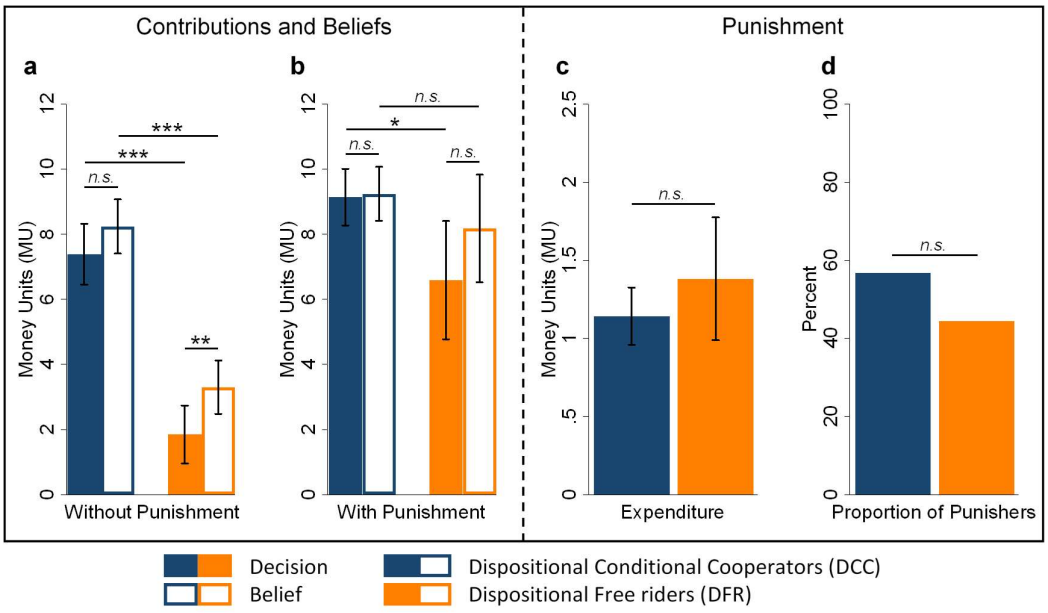


FIGURE 2.1. Contribution and prosocial punishment behaviour of Dispositional Conditional Cooperators (DCC) and Dispositional Free Riders (DFR). a In the *Without Punishment* treatment, DCC and DFR are true to their disposition. The contributions and beliefs of DCC are significantly higher than that of DFR. Mean contributions and beliefs of DCC are similar (Wilcoxon signed-rank $Z = -1.08$, $p = .282$), whereas that of DFR differ significantly (Wilcoxon signed-rank $Z = -2.22$, $p = .027$). **b** The introduction of punishment increases the contributions and beliefs of DFR only. In the *With Punishment* treatment, contributions are similar to the beliefs for both DCC (Wilcoxon signed-rank $Z = 0.60$, $p = .549$) and DFR (Wilcoxon signed-rank $Z = -0.82$, $p = .412$). **c** The mean expenditure on prosocial punishment is similar for DCC and DFR (Mann-Whitney $Z = 0.32$, $p = .749$). **d** Similar proportions of DCC and DFR engage in prosocial punishment, $\chi^2(1, N = 46) = 0.44$, $p = .506$. The error bars indicate ±1 *SEM*. Each individual as independent observation for all Mann-Whitney tests.

2.2.1.3 Punishment behaviour

The 'negative' side of the Strong Reciprocators Assumption postulates that only DCC will bear the cost of punishing defectors. In the absence of material incentives, as is the case in the one-shot game that we consider, DFR are expected to never invest resources to punish others. In sharp contrast, the mean investment in punishment that is directed at defectors is higher for DFR than for DCC (albeit not significantly different; $M_{DFR} = 1.38$, $M_{DCC} = 1.14$; Mann-Whitney $Z = 0.32$, $p = .749$; each individual as an independent observation for all tests in this section; Figure 2.1c), and DFR are not significantly different from DCC in the share of participants who use punishment when

encountering a group member who contributed less than themselves (57% vs. 44%; $\chi^2(1, N = 46) = 0.44$, $p = .506$; Figure 2.1d). Additional to the punishment of defectors shown in Figure 2.1, sanctioning cooperators (i.e., antisocial punishment) was also permitted. However, engagement in antisocial punishment is generally low and differences between DCC and DFR are not significant (Appendix; Figure 2.6).

We use a regression analysis to control for factors that can potentially influence the punishment decision (Appendix; Table 2.1; Table 2.2). The results confirm our findings reported above. Both the frequency and severity of punishment are not significantly different between DCC and DFR. Our results thus do not support the Strong Reciprocators Assumption. The only difference in the punishment behaviour of DCC and DFR is an opposing peer effect: As other group members contribute more, DCC are more likely to punish a defector, but DFR are less likely to punish a defector. A possible explanation is that DCC perceive others' high contributions as a signal for a high contribution norm, which makes them more likely to enforce this norm (Reuben and Riedl 2013). DFR's behaviour is consistent with an inclination to free ride on others' (second-order public good of) punishment.

2.2.1.4 Emotions

In line with previous studies (Bosman and van Winden 2002, Ben-Shakhar et al. 2007), we find a strong link between punishment expenditure and the intensity of negative emotions one experiences (Appendix; Table 2.3). All five negative emotions included in the questionnaire (anger, contempt, envy, irritation and jealousy) are positively correlated with punishment expenditure. In the following we focus on anger, the central moral emotion connected with norm transgressions (Nelissen and Zeelenberg 2009). Figure 2.2 illustrates the self-reported anger of DCC and DFR, depending on the degree to which others' contribution deviates from their own contribution. In both, the *Without Punishment* and *With Punishment* treatment, DCC and DFR exhibit higher anger levels as the negative deviation of others' increases, and in both treatments DCC and DFR mostly feel no anger towards group members who contributed at least as much as they did. A regression analysis shows no significant level differences between DCC and DFR in the *With Punishment* treatment (Appendix; Table 2.4). The only difference between DCC and DFR is that DFR are significantly angrier than DCC when a group member contributed more than them. Overall, DCC and DFR are highly similar in the way anger is related to the behaviour of others.

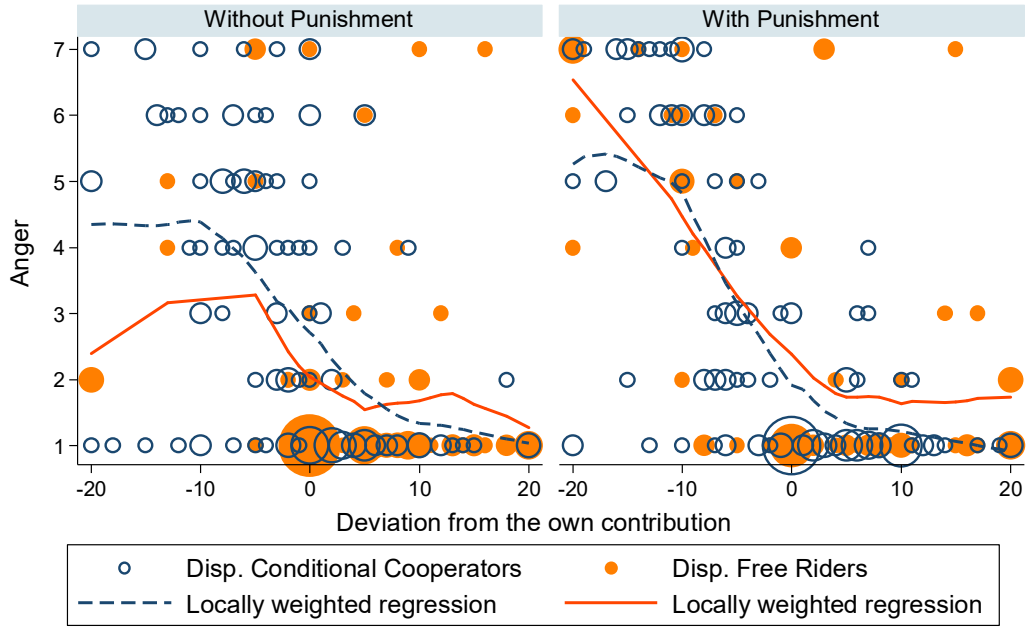


FIGURE 2.2. Self-reported anger levels depend on the deviation from own contribution. Participants indicated the intensity of anger on a scale from 1 (*not at all*) to 7 (*very much*). The size of the bubbles corresponds to the number of observations at this location. The lines indicate the locally weighted regression functions of Dispositional Conditional Cooperators (DCC; dark blue) and Dispositional Free Riders (DFR; orange), and are very similar for both types. In both treatments, DCC and DFR feel angrier when group member's negative deviations from the own contribution are higher.

2.2.2 Discussion of Study 1

The aim of Study 1 was to provide a direct test of the often-invoked assumption that strong positive reciprocity (conditional cooperation) and strong negative reciprocity (punishment) are linked: Strong reciprocators are necessarily DCC and may punish if others contribute less than them; DFR never punish and cooperate only if there is a threat of punishment. The main result of Study 1 is that—in contrast to the wide-spread belief that emerges from the literature—the individual tendency to punish defectors is independent of dispositions towards strong positive reciprocity: The punishment behaviour of DCC and DFR is virtually indistinguishable; both punish low contributors to a similar degree. This result is reminiscent of 'selfish punisher' types who do not contribute to the public good, but punish other non-contributors (Heckathorn 1989, Dufwenberg et al. 2011, Rand and Nowak 2011). The difference is that in our experiment DFR do make positive contributions to the public good when there is a

threat of punishment and then, like DCC, also punish those who contributed less than them. In fact, DFR are practically indistinguishable from DCC not only in their punishment behaviour, but also in their contributions to the public good when facing the threat of punishment.

Even if their punishment *behaviour* is highly similar, DCC and DFR might have different *motives* to punish in Study 1. Given the 2:1 punishment ratio in Study 1, (prosocial) punishment harms the punished person, but also reduces the absolute payoff differences between the punisher and the punished person. Study 2 is designed to separate these two motivations by changing the punishment ratio such that punishment no longer reduces payoff differences, while keeping contribution levels similar to those in Study 1.

2.3 Study 2

Study 2 closely follows the two-phase design of the *With Punishment* treatment in Study 1. The first phase was identical to that of Study 1. The second phase was different; it included two punishment conditions that differ in the punishment ratio (similar to Falk et al. 2005). In one condition, the ratio was 3:1, which allowed the punisher to reduce the absolute payoff difference vis-à-vis a punished free rider. Crucially, in the other condition the punishment ratio was 1:1, which did not allow the punisher to change the absolute difference in payoffs between herself and the punished person, thus excluding one of the motives for prosocial punishment present in Study 1.

A novel feature of Study 2 is that the actual punishment ratio applicable to each participant was determined—by an individual random draw—only *after* making the contribution decisions in the second phase. Each participant had a 50% chance of drawing the 3:1 punishment ratio (for each MU spent on punishment, the punished group member lost 3 MU), and a 50% chance of drawing the 1:1 punishment ratio (for each MU spent on punishment, the punished group member lost 1 MU). This procedure has two important advantages: (*a*) Contribution levels and beliefs about others' contributions were kept constant across the two conditions, allowing for a clean comparison of punishing behaviour, as everything preceding the punishment decision is identical; (*b*) the expected payoff reduction of each MU spent on punishment was 2 MU, as in Study 1, allowing for a comparison between the studies. We recruited 272

participants to take part in Study 2. 135 were randomly assigned to the 3:1 punishment ratio condition, and 137 to the 1:1 punishment ratio condition.

2.3.1 Results

2.3.1.1 Dispositions towards strong positive reciprocity

Following the first phase, 55.5% of participants were classified as DCC, 27.9% as DFR, and 16.5% remained unclassified (and were excluded from the analysis). The distribution of types in Study 2 is very similar to that of Study 1, $\chi^2(1, N = 366) = 0.12$, $p = .729$.

2.3.1.2 Beliefs and contribution behaviour

Contributions in Study 2 ($M = 9.62$, $SD = 7.12$), as well as beliefs about others' contributions ($M = 9.72$, $SD = 6.01$), were similar to those in Study 1 (Mann-Whitney $Z = -1.31$, $p = .191$; $Z = -0.97$, $p = .333$; respectively), confirming that our novel punishment procedure indeed created similar incentives to contribute as compared to the *With Punishment* treatment in Study 1 (see Appendix; Figure 2.7).

2.3.1.3 Punishment behaviour

Pooling the 3:1 and 1:1 punishment ratio conditions in Study 2, we find that the novel punishment procedure had little effect on overall prosocial punishment expenditures, with both DCC and DFR spending similarly in Study 1 and 2 (Mann-Whitney $Z_{DCC} = 1.11$, $p_{DCC} = .267$; $Z_{DFR} = 0.57$, $p_{DFR} = .572$; each individual as an independent observation for all tests in this section). See Appendix, Figure 2.8 for an illustration.

We now examine each punishment condition separately (Figure 2.3). When the punishment ratio was 3:1—which allows for the reduction of absolute payoff differences—the expenditures of DCC and DFR on prosocial punishment are nearly identical ($M_{DCC} = 1.19$, $SD_{DCC} = 1.52$; $M_{DFR} = 1.19$, $SD_{DFR} = 1.60$; Mann-Whitney $Z = -0.38$, $p = .707$; Figure 2.3a).

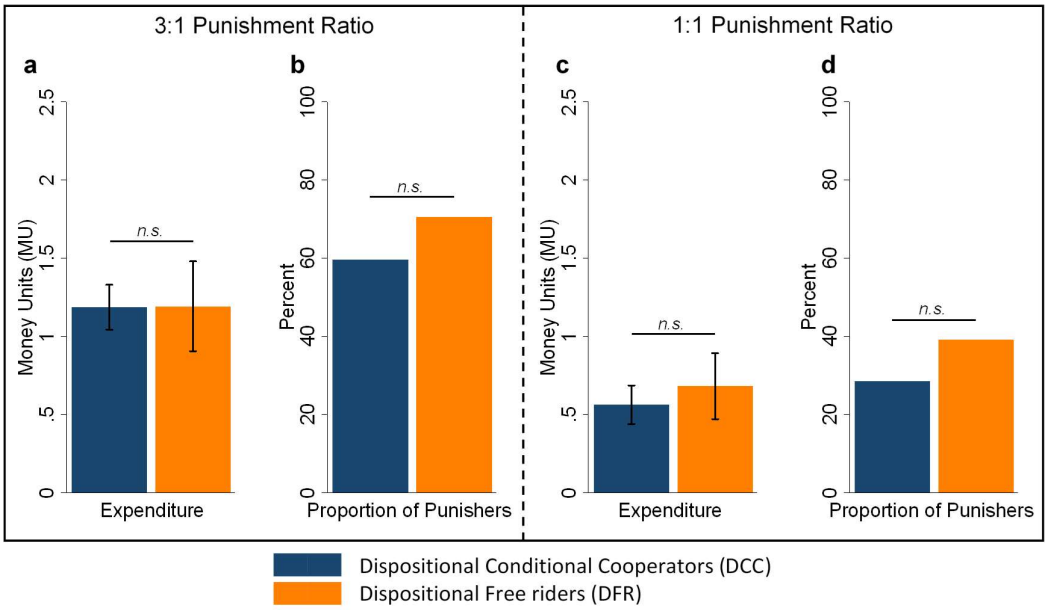


FIGURE 2.3. Prosocial punishment behaviour of Dispositional Conditional Cooperators (DCC) and Dispositional Free Riders (DFR) in the 3:1 punishment ratio condition and the 1:1 punishment ratio condition. a The mean expenditure of DCC and DFR on prosocial punishment in the 3:1 punishment ratio condition is indistinguishable. (Mann-Whitney $Z = -0.38$, $p = .707$). **b** The share of DCC and DFR engaging in prosocial punishment is similar in the 3:1 punishment ratio condition ($\chi^2(1, N = 79) = 0.67$, $p = .412$). **c** The Mean expenditure of DCC and DFR on prosocial punishment is similar in the 1:1 punishment ratio condition (Mann-Whitney $Z = -0.65$, $p = .513$). **d** The shares of DCC and DFR engaging in prosocial punishment are similar in the 1:1 punishment ratio condition ($\chi^2(1, N = 79) = 0.84$, $p = .359$). The error bars indicate bootstrapped ±1 *SEM*. Each individual as independent observation for all Mann-Whitney tests.

The main point of interest in Study 2 is the 1:1 punishment ratio condition, which excludes payoff-based motives for punishment, for example, inequity aversion or spite. If the punishment of DFR is motivated primarily by such motivations, they should not punish, or punish less than DCC, when the punishment ratio is 1:1. As in the 3:1 condition, DCC and DFR expenditures on prosocial punishment are very similar ($M_{DCC} = 0.56$, $SD_{DCC} = 1.34$; $M_{DFR} = 0.68$, $SD_{DFR} = 1.35$; Mann-Whitney $Z = -0.65$, $p = .513$; Figure 2.3c). Additionally, we do not find significant differences in the antisocial punishment behaviour of DCC and DFR (Appendix; Figure 2.9). A series of regression models support these results and do not show evidence of differences in the frequency or severity of punishment between DCC and DFR (Appendix; Table 2.5;

Table 2.6). Despite the stark difference in their disposition towards strong positive reciprocity, DCC and DFR seem to be strong negative reciprocators to the same degree, and, moreover, for a similar set of motives.

As discussed above, the 3:1 punishment ratio permits a larger set of motives for prosocial punishment. Accordingly, and in line with previous findings (Carpenter 2007, Egas and Riedl 2008, Nikiforakis and Normann 2008), both DCC and DFR spent more on prosocial punishment when the punishment ratio was 3:1 than when it was 1:1 (Mann-Whitney $Z = 3.19$, $p = .001$; $Z = 2.02$, $p = .043$; respectively). Overall, independent of the disposition towards strong positive reciprocity, the 3:1 condition seems to tap into a larger range of motives to punish and induces a higher punishment expenditure.

2.3.1.4 Emotions

Similar to Study 1, anger and punishment are positively correlated in both the 3:1 and 1:1 punishment ratio conditions (Appendix; Table 2.7). Figure 2.4 shows the anger levels reported by DCC and DFR separately for each punishment ratio. For both types and across the two ratios, anger is associated in a similar manner with negative deviations of other group members' contributions. A regression analysis does not reveal significant level differences in anger across types for neither the 3:1 nor the 1:1 punishment ratios (Appendix; Table 2.8).

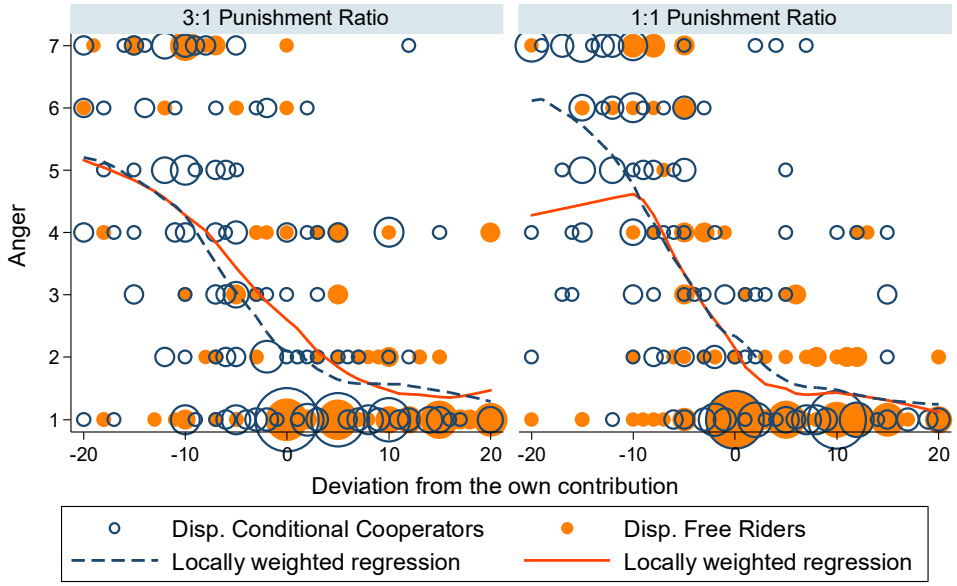


FIGURE 2.4. Self-reported anger increases for a negative deviation of others, independent of the punishment ratio or the individual cooperative disposition. The emotional reaction of Dispositional Conditional Cooperators (DCC) is similar to that of Dispositional Free Riders (DFR) in the 3:1 and 1:1 punishment ratio condition. The size of the bubbles corresponds to the number of observations at this location. The lines indicate the locally weighted regression functions for DCC (dark blue) and DFR (orange).

2.3.2 Discussion of Study 2

Study 2 shows that, independent of the punishment ratio and the set of relevant motives for punishment that it dictates, DCC and DFR use very similar levels of prosocial punishment, the type of punishment that drives and enables cooperation. Along with the replication of Study 1's finding that the relation between anger and punishment is similar for DCC and DFR, Study 2 shows that DCC and DFR are not only very similar in their punishment behaviour, but are also guided by similar motives and emotional responses.

2.4 General discussion

The findings of the present study do not support the Strong Reciprocators Assumption, which postulates that the emergence and maintenance of cooperation depends on a set of individuals—Strong Reciprocators—who are predisposed towards both positive and negative reciprocity. Rather, the intrinsic disposition towards strong positive

reciprocity, that is, whether one is willing to positively reciprocate, even at a personal cost, others' kind actions, is found to be unrelated to the willingness to pay a cost in order to reciprocate others' unkind actions by using punishment. In fact, Dispositional Conditional Cooperators are nearly indistinguishable from Dispositional Free Riders in their cooperation levels, punishment levels and motives, and the way punishment is related to anger.

The 'burden of cooperation' is thus carried by a larger set of individuals than previously assumed, which can help explain the high levels of cooperation observed when punishment opportunities are available (Fehr and Gächter 2002). The notion of strong reciprocity should be refined such that it does not necessitate the presence of a subset of strong reciprocators in a group, who are disposed towards both positive and negative reciprocity, but more generally refers to the presence of both types of reciprocal dispositions in the population at large.

The distinction between people based on their disposition towards strong positive reciprocity, that is, whether they are predisposed to conditionally cooperate with others or not, has proven to be crucial in understanding the dynamics of cooperation in the absence of punishment opportunities. In combination with expectations regarding the future contributions of others, the distinction allows for accounting of the way cooperation evolves, and typically declines over time when the punishment of non-cooperators is not possible (Fischbacher and Gächter 2010). The current results suggest that the distinction between DCC and DFR is not crucial in explaining the *maintenance* of cooperation, which requires that there are sufficiently many strong negative reciprocators in the population. Since the punishment behaviour of DCC and DFR is virtually identical, the ongoing cooperative success of groups depends more on the presence of a sufficient number of strong negative reciprocators than on its composition in terms of DCC and DFR.

By and large, the present results deviate from, and challenge, the current understanding of strong reciprocity. However, a careful examination of the literature reveals a number of papers that are suggestive of the current results. For example, Fehr and Schmidt's model of social preferences (Fehr and Schmidt 1999) makes a distinction between attitudes towards disadvantageous and advantageous inequality, and models each with a separate parameter (α and β, respectively). An agent with $\alpha > 0$ and $\beta = 0$

is expected to minimise disadvantageous inequality by punishing group members who contributed less than herself, but not to contribute to the public good even when expecting others to do so, because she does not mind the advantageous inequality. Such *hypocritical cooperation* (Carpenter 2007) is exhibited by the DFR in our experiment who are willing to invest in punishing others, but is in contrast with the Strong Reciprocators Assumption. Note, however, that this reasoning holds only when punishment can indeed reduce disadvantageous inequality, as is the case in Study 1 and when the punishment ratio was 3:1 in Study 2. The willingness to punish when the punishment ratio is 1:1 is not readily explained by a desire to reduce inequality.

Another example is recent work showing that there are two types of non-cooperators, namely *Homo economicus* and '*quasi-Homo economicus*', who differ only slightly in the degree to which they pursue their own self-interest without regarding the welfare of others—or, in other words, in the degree to which they are non-reciprocal—but differ significantly in their psychological composition (Yamagishi et al. 2014). This result shows that the non-reciprocal side of human behaviour is not unidimensional, but involves both choice patterns and psychological traits. Similarly, our data points to the conclusion that the reciprocal side of human behaviour is also not unidimensional; rather, there is a discontinuity when shifting from positive to negative reciprocity.

The logic of the strong reciprocity argument is that strong positive reciprocity is required for cooperation to emerge, and strong negative reciprocity is necessary for it to be sustained. At its core, the argument does not require that the same agents are predisposed towards both the positive and negative sides of reciprocity. Still, numerous scholars have adopted, with varying degrees of explicitness, the admittedly commonsensical Strong Reciprocators Assumption, which postulates that strong reciprocity is manifested through the presence of strong reciprocators, a set of individuals that hold both positive and negative reciprocal dispositions. The fact that our results challenge the Strong Reciprocators Assumption does not undermine the core idea underlying strong reciprocity, that both types of reciprocity are needed for sustained cooperation. Rather, we show that DCC and DFR, who are fundamentally different in their disposition towards positive reciprocity, hardly differ in their disposition towards negative reciprocity, and that it is not necessarily the case that there is a set of strong reciprocators that enable long-term cooperation. A likely explanation, which our data supports, is that once an individual chooses to cooperate, be it due to a

willingness to reciprocate the cooperation of others or due to fear of being punished for non-cooperation, the negative emotions associated with the free riding behaviour of other group members, and the desire to relieve oneself from these emotions, take over strategic considerations (Guala 2012). Even otherwise self-regarding people are suddenly willing to bear the cost of disciplining wrongdoers to the benefit of the group.

2.5 Methods

2.5.1 Study 1

2.5.1.1 Participants and procedures

We recruited 184 students at the University of Nottingham without prior experience in public goods experiments (101 females, mean age = 19.84, *SD* = 2.14) using ORSEE (Greiner 2015). The experiment was approved by the University of Nottingham School of Economics Ethics Committee and informed consent was obtained from all participants. The majority of participants were undergraduates from various fields of study (28% Humanities, 26% Economics and Business studies, 22% Natural Sciences and Engineering, 17% Law, Social and Political Sciences and 7% Medical Sciences). The experiment was computerised with z-Tree (Fischbacher 2007). The experimental sessions lasted for about 90 minutes, and participants' earnings were paid in private at the end of each session (*M* = £10.25, *SD* = £2.00). Each session consisted of reading the instructions, computerised control questions, two experimental games and a questionnaire. The control questions were designed to check participants' understanding of the games' payoff functions. Participants had to correctly answer all control questions before the start of the experimental games. We did not provide any feedback after the first game in order to prevent subjects from updating their beliefs, as well as to exclude potential income effects and strategic play. See the Appendix for the instructions.

2.5.1.2 The public goods game

The core of our experimental design is a one-shot public goods game played in groups of four. Each group member received an endowment of 20 tokens each, and decided how many tokens to keep for herself and how many to contribute to a common group project. Each token that a person kept for herself yielded one money unit (MU) to that person. Contributions to the project were multiplied by 1.6 and divided equally among the four group members. The social optimum is characterised by full contributions,

whereas the individually money-maximising strategy is to contribute nothing, regardless of the choices of other group members.

The experiment included two phases, each with a different variation of the one-shot public goods game. Participants were randomly assigned to four-person groups at the beginning of each phase. The first phase was used to assess each participant's individual disposition towards cooperation. In the second phase, they played a standard one-shot public goods game either with or without punishment (in different treatments).

2.5.1.3 First phase – Measuring individual dispositions towards cooperation

Individual dispositions towards cooperation were measured using the one-shot public goods game described above, played in a variant of the strategy method (Fischbacher et al. 2001). Each participant first decided on an *unconditional* contribution to the public good, and then on a series of *conditional* contributions, indicating her preferred contribution for each mean (integer) contribution of the other group members (see Appendix for the instructions and decision screens). We refer to the resulting set of choices as the *contribution schedule*. In each group, one member was randomly drawn and her contribution to the public good was determined by her contribution schedule, according to the average unconditional contribution of the other group members. We also elicited participants' beliefs about their group members' average contribution; Participants earned three MU for guessing the average contribution of others correctly, two MU for a deviation of one point, one MU for a deviation of two points, and zero MU for a higher deviation.

The contribution schedules were used to classify participants according to their disposition towards cooperation. The criteria were based on past work using this method (Fischbacher et al. 2012): Dispositional Conditional Cooperators (DCC) increase their contributions in the average contributions of others; they either have a positive Spearman's rank correlation coefficient, significant at the 1% level, between their own contribution and the others' average contribution, or have a monotonically increasing schedule, with at least one increase. Dispositional Free Riders (DFR) contribute exactly zero for each and every possible average contribution of others. Participants who were not classified as either DCC or DFR were excluded from the analysis, as the Strong Reciprocators Assumption does not make unambiguous behavioural predictions for them.

2.5.1.4 Second phase – Cooperation and punishment behaviour

In this phase, participants were randomly rematched with new group members and played the one-shot public goods game described above. Each participant decided, unconditional on the choices of others, how many tokens (out of 20) to contribute to the public good. Beliefs about the group members' average contribution were elicited and incentivised as in the first phase. Participants then learned the individual contribution of each of their group members, in order to keep the information structure constant across treatments. In the *With Punishment* treatment, after receiving feedback, participants had the option to assign up to five punishment points to each of their group members. Every punishment point cost the punisher one MU and destroyed two MU of the punished group member's income. To avoid negative payoffs, each participant received a fixed payment of 10 MU. After allocating punishment points to their group members, subjects stated their belief about how many punishment points they received from each of their group members and reported emotions as described below. In the *Without Punishment* treatment, participants proceeded to the emotion elicitation questionnaire directly after viewing feedback about others' contributions, that is, without having the opportunity to punish.

2.5.1.5 Emotions elicitation

After making the punishment decision and before receiving feedback on the outcome of the punishment stage, participants reported their emotional response to their group members' behaviour. The questionnaire, adapted from Bosman and van Winden (2002), included thirteen emotions (anger, contempt, envy, fear, guilt, happiness, irritation, jealousy, joy, sadness, shame, surprise and warmth). Participants rated the intensity with which they felt each emotion on a seven-point scale ranging from 1 (*not at all*) to 7 (*very much*).

2.5.2 Study 2

2.5.2.1 Participants and procedures

The experimental procedures in Study 2 were identical to those of Study 1. 272 participants took part in Study 2 (169 females, mean age = 21.14, $SD = 2.43$). The majority of participants were undergraduates from various fields of study (12% Humanities, 13% Economics and Business studies, 29% Natural Sciences and Engineering, 15% Law, Social and Political Sciences and 31% Medical Sciences).

Participants' earnings were paid in private at the end of the session (M = £10.85, SD = £1.98).

2.5.2.2 Random draw of the punishment ratio

The first phase of Study 2 was identical to that of Study 1. The second phase closely follows the *With Punishment* treatment in Study 1's second phase. The difference was that in Study 2 the punishment ratio was determined randomly. At the start of the game, participants were informed that each participant's individual punishment ratio will be either 3:1 or 1:1, with a 50% chance for each option. A 3:1 (1:1) punishment ratio means that for each MU, a particular participant assigns for punishment, the punished person loses three (one) points. The random draw was independent for each participant, such that individuals within each group could have different punishment ratios. This novel design allows for observing the effect of different punishment ratios on punishment behaviour while holding the contribution level constant.

2.6 Appendix

2.6.1 Eliciting cooperative dispositions

We use an anonymous one-shot public goods game, played with a variant of the strategy method to measure cooperative dispositions, following the methodology of Fischbacher et al. (2001). The details are described in the Methods Section of Chapter 2. This method allows to disentangle the disposition to cooperate from the belief about other group members' cooperative efforts, which is necessary to distinguish between types of players who would be indistinguishable from behavioural data alone. For example, in a one-shot public goods game played under strict anonymity, both a selfish player and a pessimistic strong positive reciprocator would contribute noting.

The strategy method of classifying participants has several additional advantages. It closely resembles the tasks in the second phase of the experiment and was shown to successfully predict behaviour in similar games (Fischbacher and Gächter 2010, Fischbacher et al. 2012). It has been replicated in different countries, showing that conditional cooperation and free riding are widespread cooperative dispositions across different societies (Kocher et al. 2008, Herrmann and Thöni 2009, Martinsson et al. 2013). Furthermore, there is evidence that the elicited contribution strategies are stable over time (Volk et al. 2012).

The method has been criticised (Burton-Chellew et al. 2016), arguing that the variation in cooperative types may result from confusion regarding the game's payoff-maximising strategy rather than from underlying differences in preferences or dispositions to cooperate. However, conditional cooperation and free riding (DCC and DFR in the terminology we use) are also found by studies using other elicitation methods (Burlando and Guala 2005, Kurzban and Houser 2005). Additionally, a recent

large-scale study (with over 2000 participants from the general population) found heterogeneous cooperative dispositions even after controlling for misperception of the payoff-maximising strategy (Fosgaard et al. 2017).

2.6.2 Supporting analysis for Study 1

2.6.2.1 Consistency of cooperative dispositions

Figure 2.5 shows the average contribution schedule for DCC and DFR as well as their individual contributions in the *Without Punishment* and *With Punishment* treatments. In both treatments, DCC contribute more if they hold higher beliefs and their individual contributions are close to the average contribution schedule. The quadratic fitted regression lines are very close, showing that contribution behaviour is similar in both treatments. Furthermore, contributions in both treatments are consistent with the predicted behaviour.

We calculate the predicted contribution using the individual contribution schedules from the first phase and beliefs from the second phase. 41% of DCC in the *Without Punishment* treatment and 31% of DCC in the *With Punishment* treatment are perfectly consistent with their cooperation schedule. The consistency of DCC's contributions with their individual cooperative dispositions is similar across treatments (Kolmogorov-Smirnov equality-of-distributions test, $p = .400$). 78% of DFR in the *Without Punishment* treatment and 55% in the *With Punishment* treatment contribute exactly according to their indicated cooperative disposition. This difference is weakly significant (Kolmogorov-Smirnov equality-of-distributions test, $p = .099$), indicating that punishment has a larger effect on DFR than on DCC. DFR deviate more strongly from the contribution levels predicted by their individual cooperative dispositions. The quadratic fitted regression lines of Figure 2.5 illustrate that DFR increase their contribution depending on their beliefs about their group members' contribution.

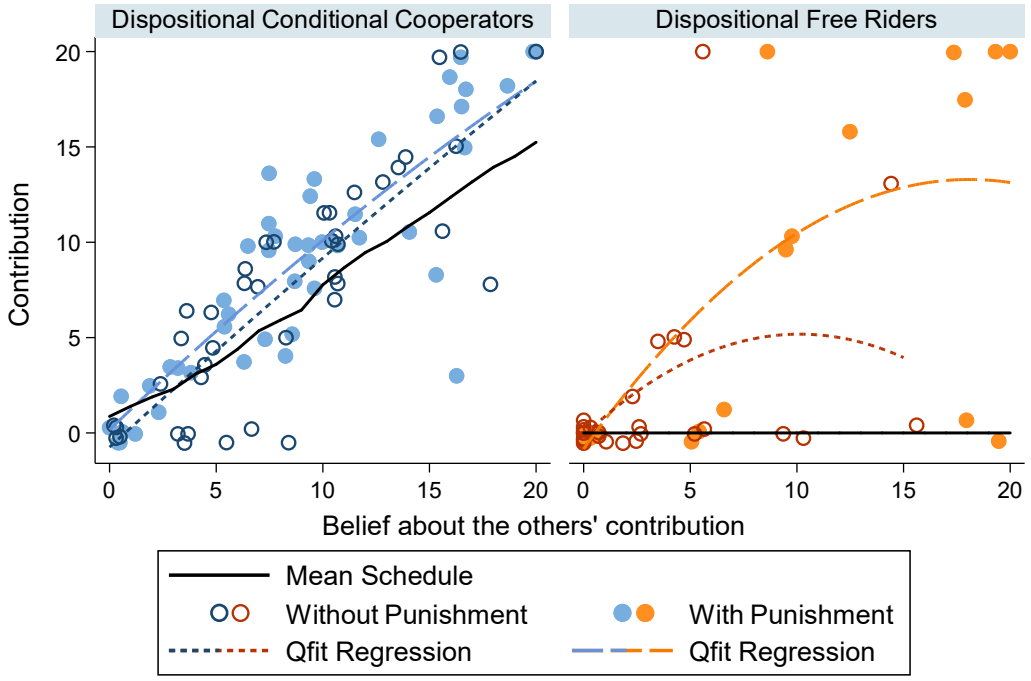


FIGURE 2.5. Consistency of behaviour with the disposition in Study 1. DCC (left panel) make higher contributions for a higher belief about other group members' mean contribution. In both treatments, DCC act in line with their mean schedule, elicited in the first part of the experiment. DFR (right panel) show a bigger deviation from their mean schedule under the threat of punishment.

2.6.2.2 Antisocial punishment by DCC and DFR

We compare the antisocial punishment in Study 1 for DCC and DFR. Antisocial punishment refers to punishment that targets group members who contributed at least as much as the punisher. Expenditure on antisocial punishment is slightly higher for DCC than DFR, but the difference is not significant ($M_{DCC} = 0.04$, $SD_{DCC} = 0.25$; $M_{DFR} = 0.00$, $SD_{DFR} = 0.00$; Mann-Whitney $Z = 0.98$, $p = .326$; Figure 2.6a). We find similar proportions of antisocial punishers (5% DCC vs. 0% DFR; $\chi^2(1, N = 56) = 0.98$, $p = .322$; Figure 2.6b). Comparing the above figures with those of prosocial punishment, we conclude that engagement in antisocial punishment is very low in our sample.

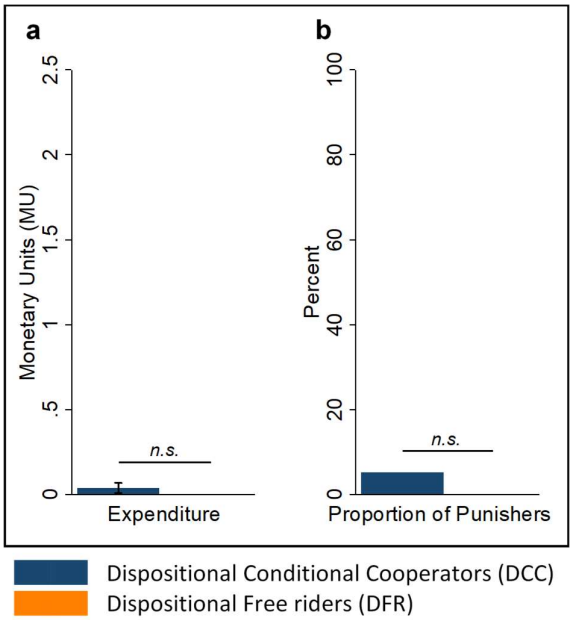


FIGURE 2.6. Antisocial punishment behaviour of Dispositional Conditional Cooperators (DCC) and Dispositional Free Riders (DFR). Mann-Whitney test for differences in the punishment expenditure with the subject as independent observation: *n.s.* $p \geq .10$. χ^2 test for differences in the proportion of punishers: *n.s.* $p \geq .10$.

2.6.2.3 Regression analysis of punishment

We use a regression analysis to control for the magnitude of the defector's deviations from the punisher's contributions and the behaviour of the other group members (Nikiforakis and Engelmann 2011). Both factors potentially influence the punishment decision. We select a two-stage regression model to disentangle the subjects' likelihood to punish, using a Probit model, from the severity of punishment, using a truncated linear regression model (Nikiforakis and Engelmann 2011). Table 2.1 reports the results. The regression models control for the absolute negative deviation of the punished group member and the average contribution deviation of the other group members. A dummy variable for DFR captures differences in the punishment behaviour of types. Two interaction terms control for potential different reactions of DFR towards negative deviations in the contributions of the punished and the others'.

TABLE 2.1.
Two-stage regression model of punishment controlling for the deviation in player *i*'s (the punisher) contribution, and player *j*'s (the recipient of punishment) contribution.

Dependent variable: Punishment dummy (Col. 1); Punishment points (Col. 3)	(1) Punishment decision	(2) Avg. marg. effects	(3) Punishment severity
Player *j*'s absolute negative deviation from player *i*	0.136*** (0.029)	0.028*** (0.006)	0.195*** (0.059)
The other players' average contribution deviation from player *i*	0.052** (0.022)	0.011** (0.004)	0.069** (0.033)
Dispositional Free Rider (DFR)	−0.417 (0.457)	−0.085 (0.094)	−0.571 (0.790)
DFR × Player *j*'s absolute negative deviation from player *i*	−0.045 (0.036)	−0.009 (0.007)	0.073 (0.094)
DFR × The other players' average contribution deviation from player *i*	−0.116*** (0.035)	−0.024*** (0.007)	0.032 (0.076)
Constant	−1.439*** (0.201)		0.970** (0.469)
N (Clusters)	213 (23)		43 (16)

Note. Only DCC and DFR are included in the analysis. Col. 1: Probit coefficients; Col. 2: Average marginal effects of the Probit model; Col. 3: Truncated linear regression. *SE* clustered on groups are given in parentheses. * $p < .10$; ** $p < .05$; *** $p < .01$.

The first column shows that the likelihood of punishment increases with a higher negative deviation of the other group member from the punisher's contribution and for a higher average contribution of other group members from the punisher's contribution. The dummy variable for DFR is not significant, indicating that there is no level difference between the punishment chosen by DCC and DFR. The interaction term between DFR and the other player's average contribution deviation is negative and highly significant. Interestingly, the average marginal effect of this coefficient outweighs the positive coefficient for the others' behaviour. We interpret this as opposing peer effects for DCC and DFR: Cooperators are more likely to punish a defector if the other group members contribute more on average to the public good. An explanation might be that a high average contribution of the others signals a high

contribution norm to DCC, which they are more likely to enforce. Conversely, DFR are less likely to punish a defector if the other group members contributed more on average. This is consistent with free riding on the other group members' punishment which resembles a second order public good.

The regression model for punishment severity is reported in the third column. Punishment is more sever for higher negative deviations of the punished group member and for a higher average contribution of the other group members. Neither the dummy variable for DFR nor the interaction terms are statistically significant. This implies that, once the decision to punish is made, there is no difference in the number of punishment points chosen between DCC and DFR.

Additionally, we check for differences between DCC's and DFR's frequency and severity of punishment depending on the contribution deviation of player j, the recipient of punishment, from the beliefs of player i, the punisher. The results are consistent with the findings on deviations from player i's contribution level and are reported in Table 2.2. We find that the likelihood of punishment and the severity significantly increase with the deviation from beliefs. The likelihood and severity also rise with the deviation of player j from the average contribution of the other group members. However, the coefficient for DFR is not significantly different from zero, showing that there are no level differences in the likelihood and severity of punishment when controlling for the deviation from beliefs.

TABLE 2.2.
Two-stage regression model of punishment controlling for the deviation in player *i*'s (the punisher) beliefs about the average contribution of others and player *j*'s (the recipient of punishment) contribution.

Dependent variable: Punishment dummy (Col. 1); Punishment points (Col. 3)	(1) Punishment decision	(2) Avg. marg. effects	(3) Punishment severity
Player *j*'s absolute negative deviation from player *i*'s belief	0.147*** (0.030)	0.030*** (0.006)	0.157*** (0.051)
The other players' average contribution deviation from player *i*'s belief	0.053** (0.024)	0.011** 0.005	0.067** (0.029)
Dispositional Free Rider (DFR)	−0.470 (0.407)	−0.094 0.081	−1.148 (0.992)
DFR × Player *j*'s absolute negative deviation from player *i*	−0.034 (0.046)	−0.007 0.009	0.144 (0.103)
DFR × The other players' average contribution deviation from player *i*	−0.103** (0.042)	−0.021** 0.008	0.039 (0.076)
Constant	−1.495*** (0.174)		1.268** (0.530)
N (Clusters)	213 (23)		43 (16)

Note. Only DCC and DFR are included in the analysis. Col. 1: Probit coefficients; Col. 2: Average marginal effects of the Probit model; Col. 3: Truncated linear regression. *SE* clustered on groups are given in parentheses. * $p < .10$; ** $p < .05$; *** $p < .01$.

2.6.2.4 Analysis of emotions

Table 2.3 shows that all five negative emotions included in the questionnaire are positively correlated with punishment expenditure, and their correlation coefficients are highly significant. Comparing the results for DCC and DFR separately reveals some differences: The number of punishment points that DFR choose is positively and significantly correlated with all five negative emotions. DCC exhibit a positive and significant correlation between punishment expenditure and negative emotions only for anger and irritation, but not for envy, jealousy and contempt. As envy and jealousy are payoff-oriented emotions, these results suggest that DFR do not only care about contribution norms but use punishment to reduce relative payoff differences.

TABLE 2.3.
Spearman's rank correlation coefficient of negative emotions and punishment expenditure.

Emotion	Sample	r_s	p	N
	All subjects	.45	< .001	276
Anger	DCC	.48	< .001	147
	DFR	.38	.002	66
	All subjects	.17	.005	276
Contempt	DCC	.13	.125	147
	DFR	.25	.041	66
	All subjects	.19	.001	276
Envy	DCC	.13	.114	147
	DFR	.25	.040	66
	All subjects	.37	< .001	276
Irritation	DCC	.33	< .001	147
	DFR	.34	.005	66
	All subjects	.16	.010	276
Jealousy	DCC	.10	.218	147
	DFR	.34	.006	66

Similar to Cubitt et al. (2011), we define an 'emotions function' describing the self-reported intensity of a particular emotion by subject i with regard to a group member j. The intensity of the emotion depends on the difference in the contribution of i and j. For example, a subject is thought to report a higher anger level regarding one of her group members for a larger negative deviation of this group member's contribution compare to the own contribution.

Table 2.4 provides the results of an ordered Probit regression model for the emotions functions. We test differences in the self-reported negative emotions between cooperative dispositions. The dependent variable—the intensity of the self-reported emotion—is censored and restricted to integers from one to seven. The regression model controls for a positive and negative contribution deviation of player j as well as for the other group members' behaviour. The model also includes a dummy variable for DFR. Additionally, the regression model includes two interaction terms to allow for the possibility of differences in the slopes of the emotions functions of DCC and DFR.

We estimate the regression separately for the *Without Punishment* treatment and the *With Punishment* treatment.

In the *Without Punishment* treatment, we find significant level differences between DCC and DFR for all negative emotions except for jealousy. Generally, DFR report a lower intensity of negative emotions compared to DFR when controlling for the negative deviation, positive deviation and the behaviour of other group members (except for jealousy).

In the *With Punishment* treatment, no significant level differences between DCC and DFR can be detected, except for jealousy. Looking at the regression model for anger, the only significant difference between types comes from DFR's reaction to positive deviations in the punishment game. DFR are significantly angrier than DCC if a group member contributed more than themselves. A reason might be the anticipation of punishment.

TABLE 2.4.
Regression analysis of self-reported negative emotions.

	Without Punishment					*With Punishment*				
Dependent variable	Anger	Contempt	Envy	Irritation	Jealousy	Anger	Contempt	Envy	Irritation	Jealousy
Player *j*'s absolute negative deviation from player *i*	0.080** (0.035)	−0.016 (0.027)	0.054* (0.029)	0.054* (0.031)	0.048 (0.033)	0.168*** (0.043)	0.062** (0.025)	0.058** (0.029)	0.167*** (0.030)	0.061** (0.029)
Player *j*'s positive deviation from player *i*	−0.121** (0.049)	−0.068** (0.032)	−0.103** (0.050)	−0.126*** (0.046)	−0.082* (0.043)	−0.066** (0.031)	0.025 (0.021)	−0.089*** (0.032)	−0.033 (0.024)	−0.096*** (0.031)
Deviation of the others' mean contribution from *i*'s input	0.028* (0.015)	0.001 (0.015)	0.015 (0.014)	0.020 (0.017)	−0.006 (0.013)	0.019 (0.015)	0.009 (0.010)	0.013 (0.015)	0.027** (0.013)	0.019 (0.016)
Dispositional Free Rider (DFR)	−0.859* (0.444)	−0.583* (0.336)	−0.794** (0.334)	−0.721* (0.399)	−0.593 (0.386)	−0.111 (0.556)	−0.347 (0.386)	−0.788 (0.490)	−0.110 (0.421)	−0.885** (0.441)
DFR × Player *j*'s absolute negative deviation	0.022 (0.045)	0.073* (0.040)	0.031 (0.038)	0.021 (0.046)	0.034 (0.048)	0.008 (0.063)	0.012 (0.052)	0.069 (0.043)	−0.012 (0.047)	0.058 (0.042)
DFR × Player *j*'s positive deviation	0.105 (0.069)	0.097** (0.043)	0.099 (0.061)	0.064 (0.068)	0.065 (0.060)	0.072** (0.036)	−0.013 (0.022)	0.082 (0.059)	0.022 (0.037)	0.078 (0.056)
Pseudo R^2	0.11	0.02	0.08	0.10	0.09	0.21	0.03	0.11	0.17	0.11
N (Clusters)	204 (23)	204 (23)	204 (23)	204 (23)	204 (23)	213 (23)	213 (23)	213 (23)	213 (23)	213 (23)

Note. Only DCC and DFR included. Ordered Probit coefficients with robust standard errors clustered on groups. * $p < .10$; ** $p < .05$; *** $p < .01$.

2.6.3 Supporting analysis for Study 2

2.6.3.1 Consistency of cooperative dispositions

Figure 2.7 shows the mean contribution schedule from the first phase, as well as contributions and beliefs from the second phase of Study 2. The contributions of DCC are largely scattered around the diagonal and rising with beliefs. The quadratic fitted regression line is above the black line, illustrating the mean cooperative dispositions of DCC.

We find that 33% of DCC and 39% of DFR make a contribution in the second phase, which is exactly consistent with the predicted contribution using their contribution schedule from the first phase and their belief from the second phase.

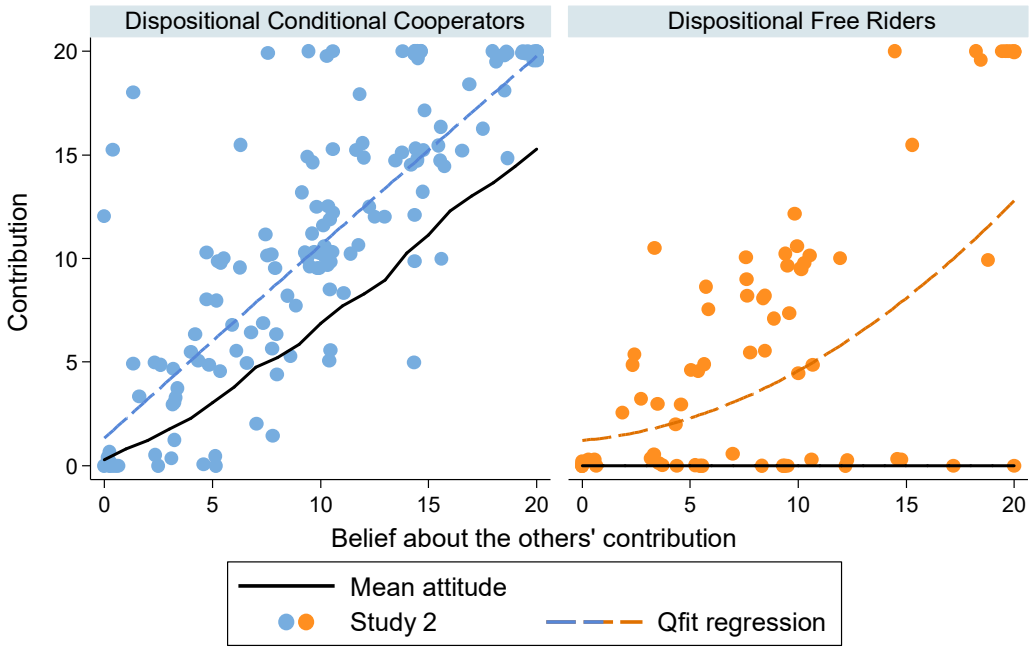


FIGURE 2.7. Consistency of behaviour with the disposition in Study 2. DCC (left panel) make higher contributions for higher beliefs about the other group members' average contribution. This is in line with their cooperative disposition indicated by the mean schedule elicited in the first phase of Study 1. A large share of DFR (right panel) also increases the own contribution for a higher belief.

2.6.3.2 Similarity of punishment by DCC and DFR

We designed the punishment stage of Study 2 in a way that keeps the incentive to contribute constant across the two punishment ratios 3:1 and 1:1. We chose the parameters so that the expected punishment ratio equals that of Study 1 (2:1). First, we compare the levels of punishment across Study 1 and 2. We find similar levels of prosocial punishment for both, DCC and DFR (Mann-Whitney $Z_{DCC} = 1.41$, $p_{DCC} = .159$; $Z_{DFR} = 0.43$, $p_{DFR} = .669$; each subject as independent observation for all tests in this section). Additionally, the expenditure on antisocial punishment was similar using the novel procedure (Mann-Whitney $Z_{DCC} = -0.72$, $p_{DCC} = .474$; $Z_{DFR} = -1.40$, $p_{DFR} = .162$).

In the following, we pool the two conditions of Study 2 (Figure 2.8). We find similar punishment expenditures on prosocial punishment ($M_{DCC} = 0.87$, $SD_{DCC} = 1.46$; $M_{DFR} = 0.90$, $SD_{DFR} = 1.47$; Mann-Whitney $Z = -0.34$, $p = .733$; subject as independent observation for all tests in this section) and antisocial punishment DFR ($M_{DCC} = 0.16$, $SD_{DCC} = 0.66$; $M_{DFR} = 0.29$, $SD_{DFR} = 1.02$; Mann-Whitney $Z = -0.39$, $p = .699$). We find similar frequencies of prosocial punishment (45% DCC vs. 53% DFR; $\chi^2(1, N = 158) = 0.69$, $p = .406$) and antisocial punishment (9% DCC vs. 10% DFR; $\chi^2(1, N = 184) = 0.11$, $p = .742$) comparing DCC and DFR.

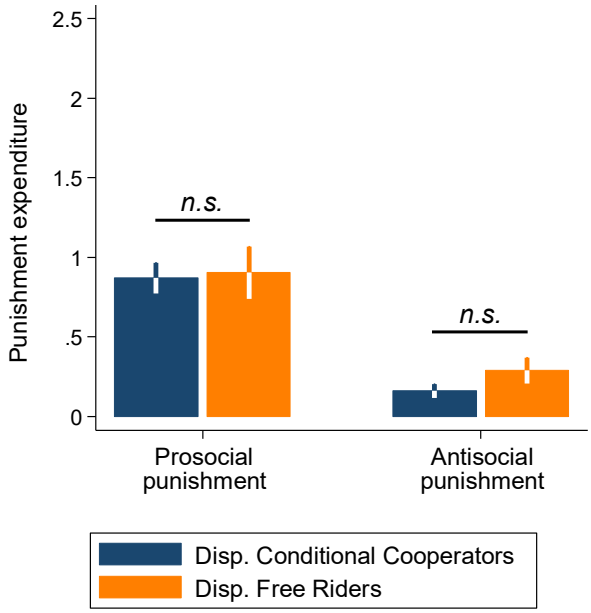


FIGURE 2.8. Average punishment in Study 2 is very similar for DCC and DFR (pooling the 3:1 punishment ratio and the 1:1 punishment ratio conditions). This holds for both, prosocial and antisocial punishment. The error bars indicate bootstrapped ±1 *SEM*. Mann-Whitney test: *n.s.* $p \geq .10$; subject as independent observation.

2.6.3.3 Antisocial punishment by DCC and DFR

Now we compare the antisocial punishment in Study 2 for the 3:1 and 1:1 punishment ratio separately. In the 3:1 punishment ratio condition, antisocial punishment is slightly higher for DFR than DCC, but the difference is not significant ($M_{DCC} = 0.23$, $SD_{DCC} = 0.86$; $M_{DFR} = 0.41$, $SD_{DFR} = 1.08$; Mann-Whitney $Z = -1.24$, $p = .215$; Figure 2.9a). Additionally, differences in the proportions of antisocial punishers are not significant (10% DCC vs. 19% DFR; $\chi^2(1, N = 94) = 1.56$, $p = .212$; Figure 2.9b). Similarly, in the 1:1 punishment ratio condition the expenditures on antisocial punishment of DCC and DFR do not significantly differ ($M_{DCC} = 0.09$, $SD_{DCC} = 0.31$; $M_{SR} = 0.18$, $SD_{DFR} = 0.94$; Mann-Whitney $Z = 0.94$, $p = .347$; Figure 2.9c). The proportion of antisocial punishers is similar for DCC and DFR (8% DCC vs. 3% DFR; $\chi^2(1, N = 90) = 0.97$, $p = .324$; Figure 2.9d).

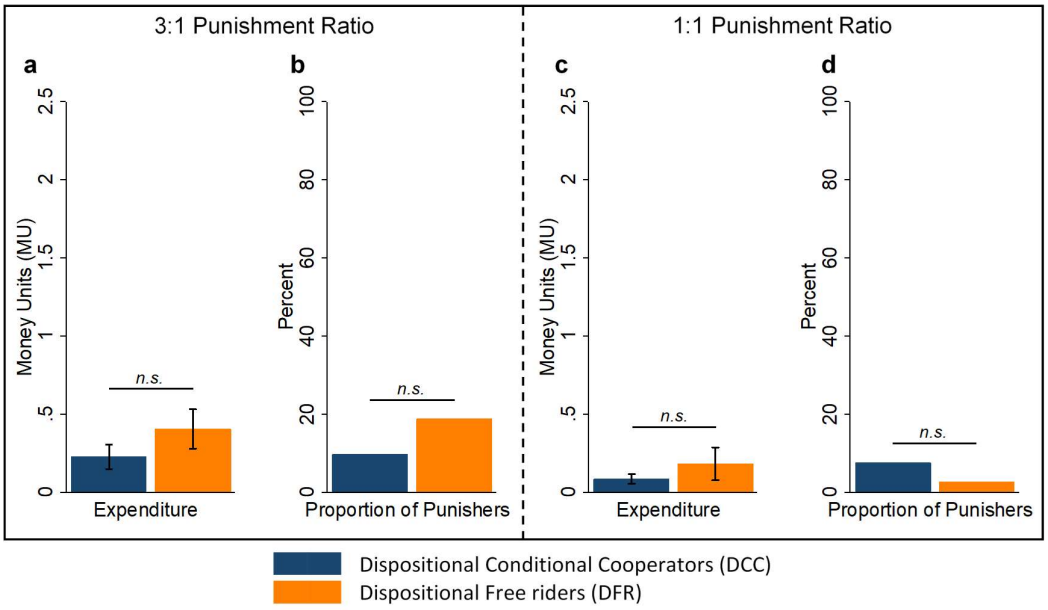


FIGURE 2.9. Antisocial punishment behaviour of Dispositional Conditional Cooperators (DCC) and Dispositional Free Riders (DFR) in the 3:1 Punishment Ratio condition and the 1:1 punishment ratio condition. Mann-Whitney test for differences in the punishment expenditure with the subject as independent observation: *n.s.* $p \geq .10$. χ^2 test for differences in the proportion of punishers: *n.s.* $p \geq .10$.

2.6.3.4 Regression analysis of punishment

Similar to Study 1, we use a two-stage regression model to disentangle potential differences between DCC and DFR in the likelihood and severity of punishment. Table 2.5 shows the results for regression models based on the assumption that punishment depends on the deviation of the punished person's contribution from that of the punisher. In both the 3:1 and 1:1 condition, the likelihood and severity of punishment are positively associated with a larger negative deviation from the punisher's contribution as well as the other player's average contribution deviation. The only exception is the punishment severity in the 1:1 condition, for which the other players' average contribution deviation from the punisher's contribution has no significant effect. The insignificant dummy variable for DFR shows that there are no level differences in the likelihood or severity of punishment between DCC and DFR. This is true for both, the 3:1 and the 1:1 conditions. In 1:1, the severity of punishment inflicted by DFR is significantly higher for a larger deviation of the other players, as compared to DCC.

TABLE 2.5.
Two-stage regression model of punishment depending on deviation from the punisher's own contributions.

	3:1		*1:1*	
	(1) Punishment decision	(2) Punishment severity	(3) Punishment decision	(4) Punishment severity
Player *j*'s absolute negative deviation	0.119*** (0.023)	0.203*** (0.076)	0.090*** (0.024)	0.134** (0.066)
The other players' average contribution deviation	0.029* (0.017)	0.123*** (0.031)	0.043** (0.017)	−0.049 (0.073)
Dispositional Free Rider (DFR)	−0.061 (0.350)	1.309 (2.152)	−0.455 (0.371)	0.518 (1.382)
DFR × Player *j*'s absolute negative deviation	−0.007 (0.053)	−0.127 (0.200)	0.080* (0.048)	0.082 (0.143)
DFR × The other players' avg. contribution deviation	0.022 (0.028)	−0.186 (0.137)	0.022 (0.039)	0.229** (0.116)
Constant	−1.065*** (0.162)	0.325 (0.759)	−1.574*** (0.265)	0.556 (0.676)
N (Clusters)	336 (63)	101 (45)	345 (61)	46 (24)

Note. Includes only DCC and DFR. Col. 1, 3: Probit coefficients with punishment dummy as dependent variable; Col. 2, 4: Truncated linear regression with punishment points as dependent variable. Robust standard errors clustered on groups are given in parentheses. * $p < .10$; ** $p < .05$; *** $p < .01$.

We find similar results when basing our regression model on the deviation of the punished person's contribution from the punisher's beliefs, rather than her actual contributions. Table 2.6 reports the results. We do not find significant level differences between DCC and DFR for the likelihood or severity of punishment in neither the 3:1 nor 1:1 conditions.

TABLE 2.6.
Two-stage regression model of punishment depending on deviation from the punisher's belief about other group members' average contributions.

	3:1		*1:1*	
	(1) Punishment decision	(2) Punishment severity	(3) Punishment decision	(4) Punishment severity
Player *j*'s absolute negative deviation from *i*'s belief	0.134*** (0.025)	0.185** (0.078)	0.103*** (0.025)	0.112 (0.089)
The others' avg. contribution deviation from *i*'s belief	0.024 (0.019)	0.116*** (0.032)	0.046** (0.019)	−0.015 (0.099)
Dispositional Free Rider (DFR)	−0.078 (0.371)	0.156 (2.223)	−0.130 (0.417)	2.271 (1.790)
DFR × Player *j*'s absolute negative deviation	−0.012 (0.050)	−0.031 (0.209)	0.024 (0.054)	−0.157 (0.204)
DFR × The other players' avg. contribution deviation	0.045 (0.032)	−0.067 (0.120)	−0.030 (0.031)	−0.049 (0.135)
Constant	−1.073*** (0.164)	0.464 (0.750)	−1.607*** (0.260)	0.768 (0.860)
N (Clusters)	336 (63)	101 (45)	345 (61)	46 (24)

Note. Includes only DCC and DFR. Col. 1, 3: Probit coefficients with punishment dummy as dependent variable; Col. 2, 4: Truncated linear regression with punishment points as dependent variable. Robust standard errors clustered on groups are given in parentheses. * $p < .10$; ** $p < .05$; *** $p < .01$.

2.6.3.5 Analysis of emotions

Next, we explore the link between negative emotions and punishment expenditure. Table 2.7 reports Spearman's rank correlation coefficient for all five negative emotions and punishment expenditure. Pooling all subjects shows that negative emotions are generally positively and significantly associated with punishment expenditure. Similar to Study 1, we find that anger and irritation have the largest correlation coefficients which are highly significant.

TABLE 2.7.
Spearman's rank correlation coefficient of negative emotions and punishment expenditure in Study 2.

		3:1			*1:1*		
Emotion	Sample	r_s	*p*	*N*	r_s	*p*	*N*
	All subjects	.46	< .001	405	.34	< .001	411
Anger	DCC	.47	< .001	231	.22	< .001	222
	DFR	.40	< .001	105	.53	< .001	123
	All subjects	.19	< .001	405	.21	< .001	411
Contempt	DCC	.23	< .001	231	.07	.269	222
	DFR	.06	.535	105	.30	< .001	123
	All subjects	.19	< .001	405	.26	< .001	411
Envy	DCC	.18	.007	231	.17	.010	222
	DFR	.11	.273	105	.40	< .001	123
	All subjects	.43	< .001	405	.29	< .001	411
Irritation	DCC	.46	< .001	231	.21	.002	222
	DFR	.29	.003	105	.46	< .001	123
	All subjects	.20	< .001	405	.27	< .001	411
Jealousy	DCC	.21	.001	231	.13	.051	222
	DFR	.01	.933	105	.45	< .001	123

We use the same approach as for Study 1 in order to investigate the driving factors of the intensity of negative emotions. Table 2.8 reports the results of regression analyses to investigate the emotions functions in the 3:1 and the 1:1 conditions. Controlling for the absolute positive and negative contribution deviations as well as the average contribution deviation of the other group members, we generally find no level differences in the intensity of self-reported negative emotions comparing DCC and DFR. A notable exception is that DFR report lower jealousy compared to DCC in the 1:1 condition, and this difference is weakly significant.

TABLE 2.8.
Regression analysis of self-reported negative emotions.

	3:1					*1:1*				
Dependent variable	Anger	Contempt	Envy	Irritation	Jealousy	Anger	Contempt	Envy	Irritation	Jealousy
Player *j*'s absolute negative deviation from player *i*	0.126*** (0.018)	0.039** (0.019)	0.072*** (0.019)	0.134*** (0.019)	0.060*** (0.018)	0.173*** (0.020)	0.049*** (0.018)	0.073*** (0.021)	0.153*** (0.020)	0.074*** (0.021)
Player *j*'s positive deviation from player *i*	−0.051* (0.028)	−0.016 (0.022)	−0.019 (0.022)	−0.055** (0.027)	−0.008 (0.024)	−0.051* (0.027)	−0.005 (0.022)	−0.059** (0.028)	−0.060** (0.028)	−0.049** (0.023)
Deviation of the others' mean contribution from *i*'s input	0.019* (0.012)	−0.007 (0.011)	−0.008 (0.011)	0.020* (0.011)	−0.018 (0.012)	0.020* (0.010)	0.016 (0.011)	0.020 (0.013)	0.014 (0.011)	0.022* (0.013)
Dispositional Free Rider (DFR)	0.207 (0.307)	0.038 (0.256)	0.165 (0.299)	−0.099 (0.331)	0.045 (0.323)	−0.106 (0.287)	−0.096 (0.312)	−0.230 (0.285)	0.129 (0.284)	−0.576* (0.346)
DFR × Player *j*'s absolute negative deviation	−0.010 (0.049)	−0.039 (0.046)	−0.007 (0.038)	0.044 (0.046)	−0.004 (0.040)	−0.001 (0.066)	0.003 (0.050)	0.072 (0.058)	−0.002 (0.065)	0.095 (0.061)
DFR × Player *j*'s positive deviation	−0.010 (0.042)	0.012 (0.030)	−0.015 (0.039)	0.015 (0.037)	0.003 (0.040)	0.021 (0.034)	0.041 (0.030)	0.055 (0.036)	0.016 (0.037)	0.076** (0.033)
Pseudo R^2	0.14	0.02	0.07	0.17	0.05	0.21	0.02	0.08	0.19	0.08
N (Clusters)	336 (63)	336 (63)	336 (63)	336 (63)	336 (63)	345 (61)	345 (61)	345 (61)	345 (61)	345 (61)

Note. Only DCC and DFR included. Ordered Probit coefficients with robust standard errors clustered on groups. * $p < .10$; ** $p < .05$; *** $p < .01$.

2.6.4 Instructions

2.6.4.1 Study 1: Without Punishment treatment – Part 1

You are now taking part in an economic experiment. Depending on the decisions made by you and other participants, you can earn a considerable amount of money. It is therefore very important that you read these instructions with care.

These instructions are solely for your private use. **It is prohibited to communicate with other participants during the experiment.** If you have any questions, please raise your hand. A member of the experiment team will come and answer them in private. If you violate this rule, you will be dismissed from the experiment and you will forfeit all payments.

During the experiment, we will not speak in terms of Pounds, but in Guilders. At the end your entire earnings will be calculated in Guilders. The total amount of Guilders you have earned will be converted to Pounds at the following rate:

1 Guilder = 0.20 Pounds

After this experimental session, your entire earnings from the experiment will be paid to you privately in cash.

At the end of the session, you will be asked to fill in a questionnaire. The answers you provide in this questionnaire are completely anonymous. They will not be revealed to anyone either during the experiment or after it. Furthermore, your responses to the questionnaires will not affect your earnings during the experiment.

The groups

At the beginning of the experiment, all participants will be randomly divided into groups of four. Apart from you, there will be three other members in your group. **You will not learn who the other people in your group are at any point.**

The decision situation

Each participant receives an endowment of **20 tokens**. You have to decide how many of these 20 tokens you will contribute to a group project, and how many you will keep for yourself. The three other members of your group have to make the same decision. They can also either contribute tokens to the project or keep tokens for themselves. You

and the other members of the group can each choose any amount between 0 and 20 tokens to contribute (including 0 and 20).

The payoffs

The income of every member of the group is calculated in the same way. Your income consists of two components:

(1) The first component is the amount of tokens that you keep for yourself. Every token that you do not contribute to the project automatically belongs to you and earns you one Guilder.
(2) The second component is your personal return from the group project. For all of the tokens contributed to the project the following happens: the project's value will be multiplied by 1.6 and this amount will be divided equally among all four members of the group.

For example, if 1 token is contributed to the project, the project's value increases to 1.6 Guilders. This amount is divided equally among all four members of the group. Thus every group member receives 0.4 Guilders.

The following function illustrates your income in Guilders:

$$\textit{Your Total Income} = 20 - \textit{Your Contribution} + 0.4 \times (\textit{Group Project})$$

In order to explain the income calculation we will give some examples. Please read them carefully. At the end of the introductory information, you will be asked to answer several computerised control questions which are designed to check that you have understood the decision situation.

Example 1

If each of the four members of the group contributes 0 tokens to the project, all four will receive an income from their private account of 20. Nobody receives anything from the project, because no one contributed anything. Therefore the total income of every member of the group is 20 Guilders.

Calculation of the total income of every participant: $(20 - 0) + 0.4 \times (0) = 20$

Example 2

If each of the four members of the group contributes 20 tokens, there will be a total of 80 tokens contributed to the project. The income from the private account is 0 for everyone, but each member receives an income from the project of 0.4 × 80 = 32 Guilders.

Calculation of the total income of every participant: (20 − 20) + 0.4 × (80) = 32

Example 3

If you contribute 20 tokens, the second member 10 tokens, the third member 5 and the fourth 0 tokens, the following incomes are calculated:

Because the total contribution to the project is 35 tokens, everyone will receive 0.4 × 35 = 14 Guilders from the project.

You contributed all your 20 tokens to the project. You will therefore receive 14 Guilders in total at the end of the experiment.

The second member of the group also receives 14 Guilders from the project. In addition, she receives 10 Guilders from her private account, because she contributed 10 tokens to the project. Thus, her total income is 24 Guilders altogether.

The third group member receives 14 Guilders from the project as well. Additionally, this group member will receive 15 Guilders from her private account. The total income therefore adds up to 29 Guilders.

The fourth member of the group, who did not contribute anything, also receives the 14 Guilders from the project and additionally the 20 Guilders from the private account, which means her total income is 34 Guilders.

Calculation of your total income: (20 − 20) + 0.4 × (35) = 14
Calculation of the 2nd group member's total income: (20 − 10) + 0.4 × (35) = 24
Calculation of the 3rd group member's total income: (20 − 5) + 0.4 × (35) = 29
Calculation of the 4th group member's total income: (20 − 0) + 0.4 × (35) = 34

Example 4

The three other members of your group contribute 20 tokens each to the project. You do not contribute anything. In this case the incomes will be calculated as follows:

Calculation of your total income: $(20 - 0) + 0.4 \times (60) = 44$

Calculation of the total income of each other group member: $(20 - 20) + 0.4 \times (60) = 24$

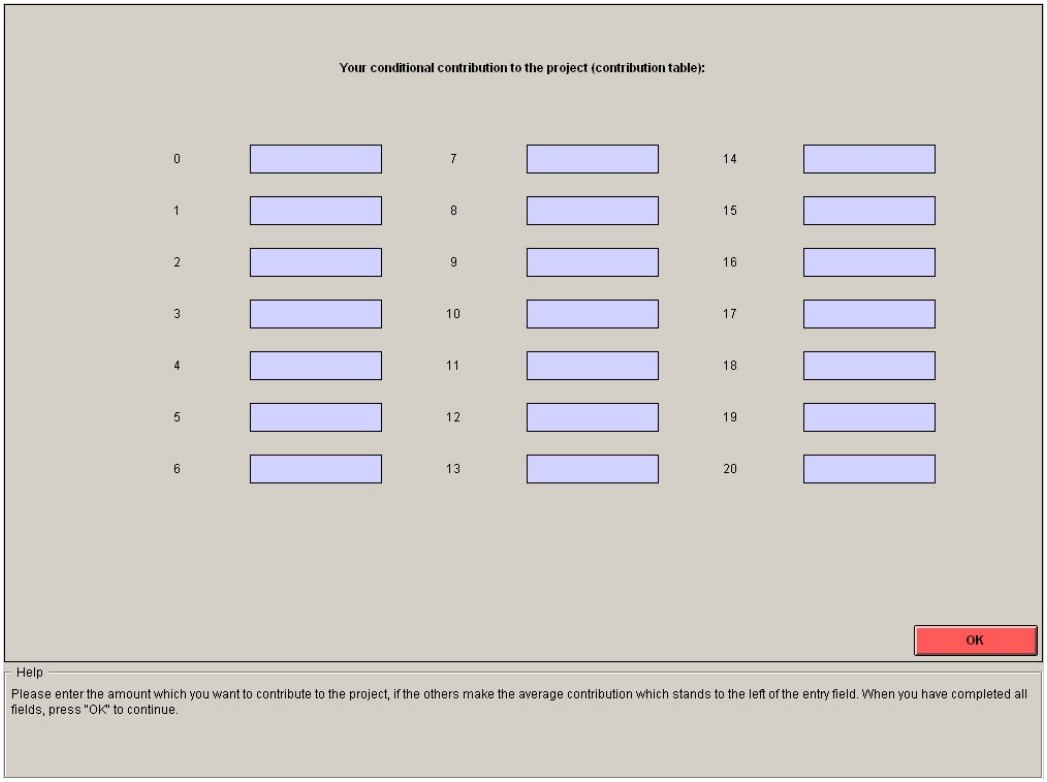
The experiment

The experiment is based on the decision situation just described to you, conducted **only once**. In this experiment you will make two types of decisions: an **unconditional contribution** and filling in a **contribution table**.

When making your **unconditional contribution**, the following screen will appear:

As mentioned above, your endowment in the experiment is 20 tokens. You have to decide how many tokens you contribute to the project by typing a number between 0 and 20 (including 0 and 20) in the box. This box can be reached by clicking on it with the mouse. By deciding how many tokens to contribute to the project, you automatically decide how many tokens you keep for yourself. After entering the amount of tokens you want to contribute you must click on the "*OK*" button. Once you have done this, your decision can no longer be revised.

Your second task is to fill in a **contribution table** on the following screen:



The contribution table indicates **how many tokens you want to contribute to the project for each possible average contribution of the other group members** (rounded to the nearest integer). The table allows for conditioning your contribution on that of the other group members.

The numbers to the left of the input fields are the possible average contributions of the **other** group members (rounded to the nearest integer). You have to enter how many tokens you want to contribute to the project, conditional on the indicated average contribution of the other group members. **You must enter a number between 0 and 20 (including 0 and 20) into each box.**

For example, in the first box you enter the amount of tokens you want to contribute to the project in case the average contribution to the project of the other three group members is 0 tokens. In the next boxes you enter how much you contribute for an average contribution of 1, 2, 3, … tokens. After entering your decisions, you must click on the "*OK*" button.

After all participants of the experiment have made an unconditional contribution and have filled their contribution table, a random mechanism will select one member from

every group. For **this group member, the contribution table** will be used to determine the contribution to the project. Whereas for **the other three group members, their unconditional contributions** will define the amount of tokens they add to the project.

You will not know whom the random mechanism will select before you make your unconditional contribution and fill in the contribution table. Therefore you must think carefully about both decisions. Either of them could determine your actual contribution to the project.

Example 5

Suppose that the **random mechanism selects you**; and that the other three group members made unconditional contributions of 0, 2, and 4 tokens, respectively. The average contribution of these three group members is, therefore, 2 tokens. If you indicated in your contribution table that you will contribute 1 token if the others contribute 2 tokens on average, then the total contribution to the project is given by 0 + 2 + 4 + 1 = 7 tokens. Each group member would, therefore, earn 0.4 × 7 = 2.8 Guilders from the project plus their respective income from their own private account. If, instead, you indicated in your contribution table that you would contribute 19 tokens if the others contribute 2 tokens on average, then the total contribution of the group to the project would be given by 0 + 2 + 4 + 19 = 25 tokens. Each group member would earn 0.4 × 25 = 10 Guilders from the project plus their respective income from their own private account.

Example 6

Suppose that the **random mechanism does not select you**; and that your unconditional contribution is 16 tokens, while those of the other two group members not selected by the random mechanism are 18 and 20 tokens respectively. Your average unconditional contribution and that of these two other group members is, therefore, 18 tokens. If the group member whom the random mechanism did select indicates in her contribution table that she will contribute 1 token if the other three group members contribute on average 18 tokens, then the total contribution of the group to the project is given by 16 + 18 + 20 + 1 = 55 tokens. Each group member will therefore earn 0.4 × 55 = 22 Guilders from the project plus their respective income from their own private account. If, instead, the randomly selected group member indicates in her contribution table that she contributes 19 if the others contribute on average 18 tokens, then the total

contribution of the group to the project is 16 + 18 + 20 + 19 = 73 tokens. Each group member would therefore earn 0.4 × 73 = 29.2 Guilders from the project plus their respective income from their own private account.

The random mechanism

Each group member is assigned a Group Member ID between 1 and 4, which denotes this participant's number inside her group. Moreover, participant number 2 was randomly selected at the very beginning of the experiment. This participant will draw a ball from an urn after all participants have made their unconditional contribution and have filled out their contribution table. Each ball in the urn has a different colour and each colour corresponds to a Group Member ID: orange = 1, blue = 2, yellow = 3, green = 4. The resulting number will be entered into the computer. If your Group Member ID is drawn, then your contribution table will determine your contribution to the project. For all other members of your group, the unconditional contributions will be relevant. Otherwise, your unconditional contribution determines your contribution.

If you have any questions, please raise your hand and a member of the experiment team will come and answer them in private.

2.6.4.2 Study 1: Without Punishment treatment – Part 2

You are now taking part in a second experiment. Your payoff from this experiment is completely unrelated to the decisions you have made in the previous one. The money you earn in this experiment will be added to what you earned in the first experiment. As before, the Guilders you have earned will be converted to Pounds at the following rate:

1 Guilder = 0.20 Pounds

As in the previous experiment, all participants will be randomly divided into groups of four. However, the composition of the group is entirely new. **You will not learn who the other people in your group are at any point.**

The decision situation

The decision situation is the same as the one described on the first instruction sheet: Each participant receives an endowment of **20 tokens**. You have to decide how many of these 20 tokens you contribute to a group project and how many you keep for yourself. The three other members of your group have to make the same decision.

However, this time you will make only an unconditional contribution to the project. There will be no contribution table.

Like in the first experiment, you will make the contribution decision in the second experiment **only once**.

You will see the following screen when making your contribution decision:

After the contribution decision, you will see how many tokens each of the other three group members has contributed to the project and their corresponding income from this contribution decision. Nonetheless, the identities of your group members will not be revealed at any time.

The payoffs

The calculation of payoffs is identical to the previous experiment. The income of every member of the group is calculated in the same way. Your income consists of two components:

(1) The first component is the amount of tokens that you keep for yourself. Every token that you do not contribute to the project automatically belongs to you and earns you one Guilder.
(2) The second component is your personal return from the group project. For all of the tokens contributed to the project the following happens: the project's value will be multiplied by 1.6 and this amount will be divided equally among all four members of the group.

For example, if 1 token is contributed to the project, the project's value increases to 1.6 Guilders. This amount is divided equally among all four members of the group. Thus every group member receives 0.4 Guilders.

The following function is exactly the same as in the previous experiment and illustrates your income in Guilders:

$$\textit{Your Total Income} = 20 - \textit{Your Contribution} + 0.4 \times (\textit{Group Project})$$

If you have any questions, please raise your hand and a member of the experiment team will come and answer them in private.

2.6.4.3 Study 1: With Punishment treatment – Part 1

Same as in Part 1 of the Without Punishment treatment.

2.6.4.4 Study 1: With Punishment treatment – Part 2

You are now taking part in a second experiment. Your payoff from this experiment is completely unrelated to the decisions you have made in the previous one. The money you earn in this experiment will be added to what you earned in the first experiment. As before the Guilders you have earned will be converted to Pounds at the following rate:

1 Guilder = 0.20 Pounds

As in the previous experiment, all participants will be randomly divided into groups of four. However, the composition of the group is entirely new. **You will not learn who the other people in your group are at any point.**

The decision situation

The decision situation is the same as the one described on the first instruction sheet: Each participant receives an endowment of **20 tokens**. You have to decide how many of these 20 tokens you contribute to a group project and how many you keep for yourself. The three other members of your group have to make the same decision. However, this time you will make only an unconditional contribution to the project. There will be no contribution table.

After the contribution decision, there will be a **second stage**. At this stage, you will see how many tokens each of the other three group members has contributed to the project and their corresponding income from this contribution decision. Nonetheless, the identities of your group members will not be revealed at any stage. You can either **decrease** or **leave unchanged** the income of each other group member by assigning **deduction points** to them. The other group members can also decrease your income, by allocating deduction points to you, if they wish to do so.

Deduction points

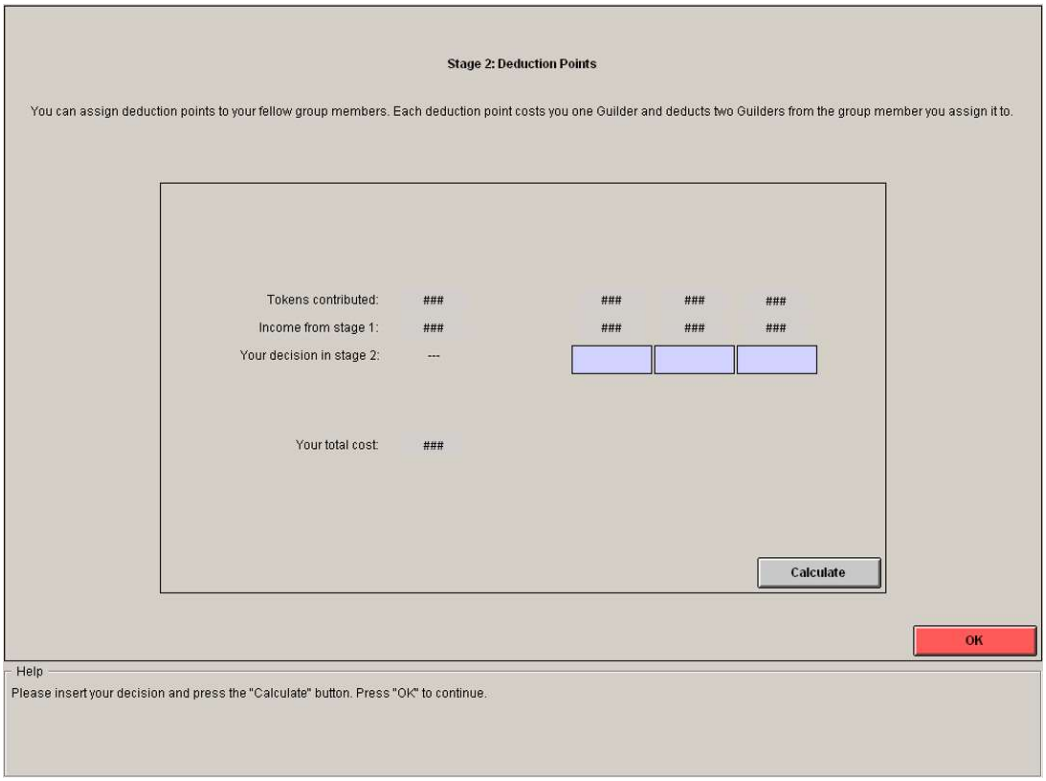
In stage 2, you can assign **between 0 and 5 deduction points to each other group member.** The maximum number of deduction points, you can allocate to the other group members together is therefore 15 deduction points.

For each deduction point that you assign, there is a cost to you of one Guilder. Thus, the total cost to you in Guilders of assigning deduction points to other group members is given by the total number of deduction points that you assign.

For each deduction point that you assign to a particular group member, you will decrease their income by 2 Guilders unless their income is already exhausted. For example, if you give a group member 2 deduction points, you will decrease this group member's income by 4 Guilders.

Your **own income will be reduced by 2 Guilders for each deduction point that is assigned to you** by the other three group members. If all of your income from the first stage of this experiment is exhausted, it cannot be reduced any further by other group members.

You will see the following screen at stage 2:



The column on the left shows your contribution and your income from the first stage. The other three columns indicate the contribution of your group members and their income from the first stage.

If you do not wish to change the income of the other group members, type "*0*" into the fields next to *"Your decision in stage 2."* In case you want to assign deduction points, enter the number of deduction points you want to assign into this field. You must enter a decision into every field and press the "*Calculate"* button. This will display the cost of your decision. Until you press the "*OK*" button, you can still change your decision. To recalculate the costs after making a change, simply press the "*Calculate*" button again.

The payoffs

Your total income in Guilders from the two stages will be calculated as follows:

Your Income From Stage 1 = 20 − *Your Contribution* + 0.4 × (*Group Project*)

Total Income After Stage 2 = Income From Stage 1 **(1)**

− 2 × (*Sum Of Deduction Points Assigned To You*) **(2)**

− (*Deduction Points Assigned By You*)

if (1) + (2) is greater or equal to 0.

Total Income After Stage 2 = 0 − (Deduction Points Assigned By You)

if (1) + (2) is less than 0.

Please note that your income in Guilders after stage 2 can be negative, if the cost of deduction points assigned by you exceeds your income from stage 1 less any reduction in your income caused by other group members.

However, at the end of the experiment and in addition to the calculation just given, you and the other members of your group will each receive a lump sum payment of **10 Guilders.** This payment is to cover losses that you could incur.

If you have any questions, please raise your hand and a member of the experiment team will come and answer them in private.

2.6.4.5 Study 2: Part 1

Same as in Study 1.

2.6.4.6 Study 2: Part 2

You are now taking part in a second experiment. Your payoff from this experiment is completely unrelated to the decisions you have made in the previous one. The money you earn in this experiment will be added to what you earned in the first experiment. As before the Guilders you have earned will be converted to Pounds at the following rate:

1 Guilder = 0.20 Pounds

As in the previous experiment, all participants will be randomly divided into groups of four. However, the composition of the group is entirely new. **You will not learn who the other people in your group are at any point.**

The decision situation

The decision situation is the same as the one described on the first instruction sheet: Each participant receives an endowment of **20 tokens**. You have to decide how many of these 20 tokens you contribute to a group project and how many you keep for yourself. The three other members of your group have to make the same decision. However, this time you will make only an unconditional contribution to the project. There will be no contribution table.

After the contribution decision, there will be a **second stage**. At this stage, you will see how many tokens each of the other three group members has contributed to the project and their corresponding income from this contribution decision. Nonetheless, the identities of your group members will not be revealed at any stage. You can either **decrease** or **leave unchanged** the income of each other group member by assigning **deduction points** to them. The other group members can also decrease your income, by allocating deduction points to you, if they wish to do so.

Deduction points

In stage 2, you can assign **between 0 and 5 deduction points to each other group member.** The maximum number of deduction points, you can allocate to the other group members together is therefore 15 deduction points.

For each deduction point that you assign, there is a cost to you of one Guilder. Thus, the total cost to you in Guilders of assigning deduction points to other group members is given by the total number of deduction points that you assign.

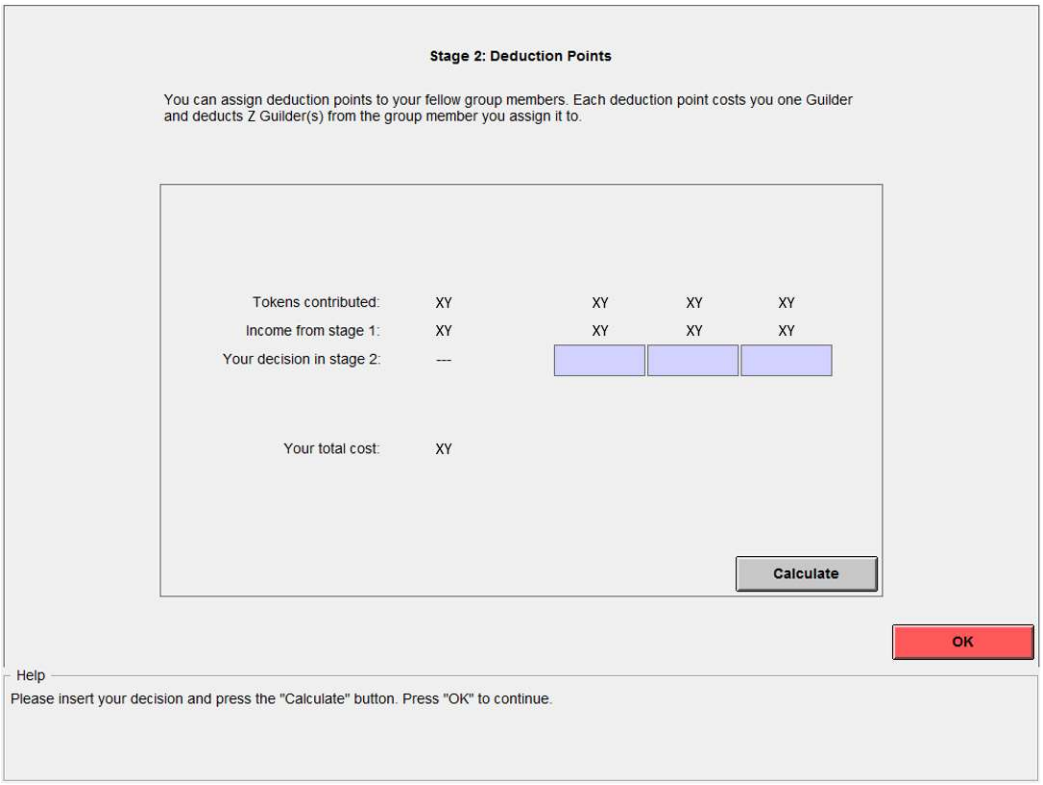
For each deduction point that you assign to a particular group member, you will decrease their income.

Before allocating deduction points, the computer will randomly draw your deduction factor, which will be either 1 or 3. The magnitude of **income reduction from each deduction point you allocate is determined by this deduction factor.** If the deduction factor is 1, then each deduction point you allocate to a group member will reduce this group member's income by 1 Guilder. If the deduction factor is 3, then each deduction

point you allocate to a group member will reduce this group member's income by 3 Guilders. There is a 50% chance of drawing 1 and a 50% chance of drawing 3. The computer will randomly draw the deduction factors of your group members, in exactly the same way. The draws are independent for each participant.

Your **own income can be reduced by the other three group members**. The income reduction depends on the number of deduction points you receive and the respective deduction factor of the other group members. If all of your income from the first stage of this experiment is exhausted, it cannot be reduced any further by other group members.

At stage 2, you will learn your deduction factor and then see the following screen:



The column on the left shows your contribution and your income from the first stage. The other three columns indicate the contribution of your group members and their income from the first stage.

If you do not wish to change the income of the other group members, type "*0*" into the fields next to *"Your decision in stage 2."* In case you want to assign deduction points, enter the number of deduction points you want to assign into this field. You must enter

a decision into every field and press the "*Calculate"* button. This will display the cost of your decision. Until you press the "*OK*" button, you can still change your decision. To recalculate the costs after making a change, simply press the "*Calculate*" button again.

The payoffs

Your total income in Guilders from the two stages will be calculated as follows (we will refer to your three group members as A, B and C):

Your Income From Stage 1 = 20 − *Your Contribution* + 0.4 × (*Group Project*)

Total Income After Stage 2 = Income From Stage 1 **(1)**

− (*Deduction Points Assigned To You By A*) × *A's deduction factor*
− (*Deduction Points Assigned To You By B*) × *B's deduction factor* **(2)**
− (*Deduction Points Assigned To You By C*) × *C's deduction factor*
− (*Deduction Points Assigned By You*)

if (1) + (2) is greater or equal to 0.

Total Income After Stage 2 = 0 − (Deduction Points Assigned By You)

if (1) + (2) is less than 0.

Please note that your income in Guilders after stage 2 can be negative, if the cost of deduction points assigned by you exceeds your income from stage 1 less any reduction in your income caused by other group members.

However, at the end of the experiment and in addition to the calculation just given, you and the other members of your group will each receive a lump sum payment of **10 Guilders.** This payment is to cover losses that you could incur.

If you have any questions, please raise your hand and a member of the experiment team will come and answer them in private.

CHAPTER 3

A Cross-Societal Comparison of Cooperative Dispositions and Norm Enforcement[2]

3.1 Introduction

All societies face challenges like air pollution, over-exploitation of common resources or tax evasion. In each of these social dilemmas, the highest social welfare is achieved by cooperation. Yet individually, it is beneficial to free ride on others' costly efforts, for example, to free ride on those who curb pollution, restrain their use of common resources, and pay taxes when audits are rare. Hardin (1968) described this problem as the 'tragedy of the commons'. Although the structure of social dilemmas is universal, the solution to the conflict between self-serving and welfare-maximising behaviour appears highly sensitive to cultural cues, institutions and social norms, which differ across societies. This phenomenon was shown in studies using behavioural economics experiments which found large variation in cooperation levels across small-scale societies (Henrich et al. 2005) as well as complex industrialised nations (Gächter et al. 2010).

Possible explanations for the large differences in voluntary cooperation across societies include the varying prevalence of social norms of cooperation and perception of norm transgressions. Experiments conducted in a country with high cooperative norms (Switzerland) found that participants in anonymous public goods games, who were given the opportunity to punish others at a cost, use this peer-punishment mechanism to enforce high cooperation norms (Fehr and Gächter 2002). This peer-punishment mechanism led to higher cooperation rates compared to games in which punishment was not possible. Yet, the use and effect of peer-punishment was found to

[2] This chapter draws on joint work in progress with Benjamin Beranek, Simon Gächter, Fatima Lambarraa and Jonathan Schulz.

vary greatly across societies (Herrmann et al. 2008, Gächter and Herrmann 2009). Although punishment of free riding emerged as a universal pattern, Herrmann et al. (2008) reported a large cross-societal variation in antisocial punishment. In some societies punishment did not foster high cooperation rates, because a similar extent of punishment was directed towards cooperators and defectors.

A better understanding of the factors driving these observed cross-societal variations in cooperation and norm enforcement is vital to fostering higher levels of cooperation in societies around the world. In the present study, we conduct variants of public goods experiments in Morocco, Turkey, the UK and the US, in order to investigate the driving factors of cross-societal variation in cooperative behaviour. Using indices of culture and institutions (Inglehart and Baker 2000), these four countries can be grouped into two distinct cultural clusters: Islamic countries (Morocco, Turkey) and English-speaking countries (UK, US).

Our study contributes to the literature on cross-societal differences in cooperation in two ways. In the first part of the study, we focus on the role of beliefs and cooperative dispositions in explaining differences in behaviour across societies. Several articles have documented that cooperative behaviour varies across societies. However, it is not clear to what degree this is driven by differences in cooperative dispositions or by differences in beliefs. Our results show similar cooperative dispositions on the aggregate level across societies. Yet, we find a strong influence of culture on beliefs and aggregate cooperation behaviour. We therefore conclude that differences in behaviour across cultural clusters do not stem from differences in dispositions alone, but rather that differences in beliefs help to explain this variation.

The second part of this study explores cross-societal differences in costly altruistic punishment. As opposed to previous studies that found a large cross-societal variation in antisocial punishment in repeated games, we implement a one-shot game that excludes any strategic incentives to punish and spill-over effects over rounds. Our results show that antisocial punishment is remarkably similar across societies. This suggests that previously reported differences in punishment are likely to be driven by strategic play or retaliation that emerge in repeated interactions. Additionally, we consider participants' emotional response as a proximate explanation for punishment.

In all four countries, encountering a defector is associated with high anger levels, but the reported strength of emotions varies across societies.

3.1.1 Determinants of cooperative behaviour across societies

The present study tests different channels through which cross-societal variation in culture and institutions can shape cooperative behaviour. This is necessary because many people have a tendency to reciprocate (Bowles and Gintis 2011), which in social dilemmas manifests itself in conditional cooperation. Therefore, a subject with a disposition to conditionally cooperate only does so if she expects others to cooperate as well. If she expects others to defect, she will not exert any cooperative effort. Thus, focusing only on the behavioural outcome of this decision process makes it impossible to distinguish between a subject with a disposition to conditionally cooperate, but who is pessimistic about the likelihood others will also cooperate and a subject with a disposition to free ride.

Societal differences in behaviour can root in differing cooperative dispositions, varying beliefs about other people's behaviour, or both. To disentangle the effects of individual dispositions and beliefs on behaviour, we draw on the 'ABC of cooperation' framework described by Gächter et al. (2017) and inspired by Fischbacher and Gächter (2010). Gächter et al. (2017) use an incentive compatible way of eliciting individual cooperative dispositions (which they call *attitudes*) in a one-shot public goods game (Fischbacher et al. 2001). Furthermore, they elicit the expected average contributions of others to the public good (*beliefs*). The authors then use the combination of both to predict participants' behaviour (*contributions*) in public goods games with a give or a take frame. Thus, cooperative behaviour (c_i) is a function of the individual cooperative disposition/attitude (a_i) and belief (b_i) about other people's cooperative behaviour, such that $a_i(b_i) \rightarrow c_i$.

This framework allows the channels through which societal differences affect behaviour to be pinpointed. First, cooperative dispositions can be directly influenced by culture and institutions. For example, the upbringing and socialisation process might shape one's disposition to cooperate. Second, the belief about how other individuals will behave in a given situation is shaped by the culture and institutions of the society. The expected cooperative behaviour of strangers might increase if the cultural background includes social norms of cooperation and if institutions exist that follow

the rule of law. These beliefs are crucial in the determination of one's own behaviour for those with conditionally cooperative dispositions. Indeed, it has been shown that the majority of participants in laboratory experiments have a disposition to condition the own cooperative effort on that of others (Chaudhuri 2011).

We follow the method proposed by Fischbacher et al. (2001) to independently measure cooperative dispositions in a public goods game. The authors use the strategy method (Selten 1967) as an incentive compatible way to measure individual cooperative dispositions independent of beliefs about other people's behaviour. Fischbacher et al. (2001) report that about 50% of participants condition their cooperative efforts on that of others and classify this group as Conditional Cooperators (CC). Approximately 30% of participants were classified as Free Riders (FR), who contribute nothing irrespective of the other participants' cooperative efforts. The rest of the participants displayed more complicated contribution patterns (14% hump shaped and 6% other patterns). The finding that Conditional Cooperators and Free Riders are the most frequent cooperative dispositions has been confirmed by studies using a variety of different experimental measures of cooperative dispositions (e.g., Burlando and Guala 2005, Kurzban and Houser 2005).

Previous studies tested for differences in cooperative dispositions across societies using the method introduced by Fischbacher et al. (2001). Herrmann and Thöni (2009) find a similar distribution of types across student samples from four universities in rural and urban Russia. Kocher et al. (2008) report significant differences in the distribution of types across student samples in Austria, Japan and the US. Similarly, Martinsson et al. (2013) show that the distribution of types differs across student samples from Colombia and Vietnam. In all societies, conditional cooperation emerged as the most common individual cooperative disposition. However, these studies do not report independent measures of beliefs about other peoples' cooperative behaviour.

Furthermore, the studies discussed above report imperfect conditional cooperation on the aggregate level across all societies included in the investigation. Although on average participants have the inclination to increase the own contribution to the public good if others do so as well, they do not perfectly match the contribution of others but undercut it. This suggests that beliefs about the behaviour of others impact

own cooperative efforts in all societies included in previous research. To our knowledge, the present study is the first to investigate the role of beliefs in combination with cooperative dispositions to explain cross-societal differences in cooperative behaviour.

3.1.2 Norm enforcement across societies

One way of increasing cooperation rates in social dilemma situations is to allow the punishment of defectors. A large share of participants in public goods games with peer-punishment are willing to engage in punishment even if it is costly and altruistic (Fehr and Gächter 2002, Balliet et al. 2011). Social preference models of payoff- or intentions-based fairness can explain such behaviour (e.g., Fehr and Schmidt 1999, Bolton and Ockenfels 2000, Falk and Fischbacher 2006). Yet, these models do not account for the influence of culture and institutions on human behaviour. This might be problematic since it has been shown that societal differences in trust and norms of civic cooperation directly link to cooperation-based economic performance (Knack and Keefer 1997, Herrmann et al. 2008).

Furthermore, the repeated games used by Herrmann et al. (2008) include a larger set of possible motives for punishment. In addition to other-regarding motives such as fairness considerations, repeated games allow for strategic and selfish intentions to drive punishment. For example, it might be optimal in repeated interactions to punish defectors in order to induce higher levels of cooperation in the following rounds and reap the benefits associated with increased cooperation. Punishment in repeated interactions might lead to feuds or counter-punishment (Nikiforakis and Engelmann 2011). Prosocial and antisocial punishment might be more frequent or severe in repeated interactions for these reasons.

The present study investigates societal differences in punishment using one-shot experiments. This experimental set-up precludes any strategic incentive for punishment as well as the possibility of retaliation. Prosocial and antisocial punishment observed in our setting should be driven by preferences alone.

Additionally, we test for differences across societies in the range of emotional response to defection. Here we focus on the ‘moral emotions’ anger and guilt, which are connected to the punishment decision. The relief of anger can serve as a proximate explanation for punishment (Sanfey et al. 2003, Hopfensitz and Reuben 2009, Nelissen

and Zeelenberg 2009, Cubitt et al. 2011). Differences in the anger levels induced by defection can therefore explain variation in the use of punishment. Guilt, on the other hand, is thought to prevent a defector from counter-punishing a cooperator and therefore strengthens the cooperation-enhancing effect of punishment (Nelissen and Zeelenberg 2009). Differences in guilt across societies can thus account for variation in the cooperation-enhancing effect of punishment.

3.2 Research methods of our cross-societal study

To investigate the impact of societal differences on behaviour, we adopt a methodology similar to Gächter and Herrmann (2009). The authors conducted experiments in four different subject pools: two Russian and two Swiss cities. Noting that some variance is expected even across subject pools that belong to the same society, Gächter and Herrmann (2009) argue that variance in behaviour can be attributed to societal and cultural differences if the variance in behaviour across culturally similar subject pools is smaller than the variance in behaviour across culturally distant subject pools. In their study, the authors claimed an effect driven by societal and cultural differences if the variance between the two Russian (Swiss) subject pools was small and the variance across countries was large.

We adapt this methodology to our cross-societal investigation of cooperation and punishment. Thus, our goal was to run experiments in four countries that belong to two different cultural clusters, maximising the differences between societies across cultural clusters and minimising the differences between societies within a cultural cluster. If culture and institutions affect cooperative dispositions, beliefs or emotions, then we expect the variation between the two countries within a cultural cluster to be smaller than that across the two cultural clusters.

We ran experiments in the UK (where the authors are based) as well as in Morocco (MO), Turkey (TR) and the US, choosing universities with an experimental economics laboratory and ongoing research collaborations with the University of Nottingham School of Economics. We rely on student samples to investigate cross-societal differences because we expect students to have a similar socio-economic status and level of education across all four countries. This procedure minimises differences in subjects' socio-economic background and provides additional support for our argument to attribute potential differences in behaviour to societal differences.

According to Inglehart and Baker (2000)'s identification of cultural clusters based on common values and beliefs, Morocco and Turkey belong to the group of *Islamic countries*. Both share similar religious values and traditions since Islam is the predominant religion in Morocco and Turkey. The UK and the US are grouped together with other *English-speaking countries*. Both societies share cultural characteristics through a common language, former colonial ties and a strong Protestant tradition.

We chose these four societies based on several country-level measures of institutions and cultural dimensions that might impact cooperative norms and norm enforcement (Figure 3.1; Appendix, Table 3.5). Differences between Morocco and Turkey and between the UK and the US are relatively small on almost all dimensions, whereas differences between the Islamic countries cultural cluster (Morocco and Turkey) and the English-speaking countries cultural cluster (the UK and US) are large. Next, we discuss several indices commonly used to measure the distance of societies in their cultural backgrounds, values, social capital, quality of institutions and prosperity.

Hofstede's cultural dimensions (Hofstede and Hofstede 2001) are a long-established set of dimensions to quantify cultural differences: *Individualism* measures the importance of the collective versus the individual. The degree of individualism or collectivism is a fundamental characteristic of any society (Greif 1994). *Power distance* measures how unequal or egalitarian a society is. *Masculinity* indicates a society's valuation of stereotypically male attributes. *Uncertainty avoidance* measures the tolerance for uncertainty or ambiguity. The two Islamic countries' score for power distance is above the sample mean and for individualism Turkey is just below and Morocco slightly above the sample mean. For the two English-speaking countries the individualism scores are amongst the highest of the sample and they score below the sample mean for power distance. The Islamic countries score above the sample mean for uncertainty avoidance. The masculinity score of Morocco is just above and that of Turkey is just below the sample mean. The two English-speaking countries score below the sample mean for uncertainty avoidance and above the sample mean for masculinity.

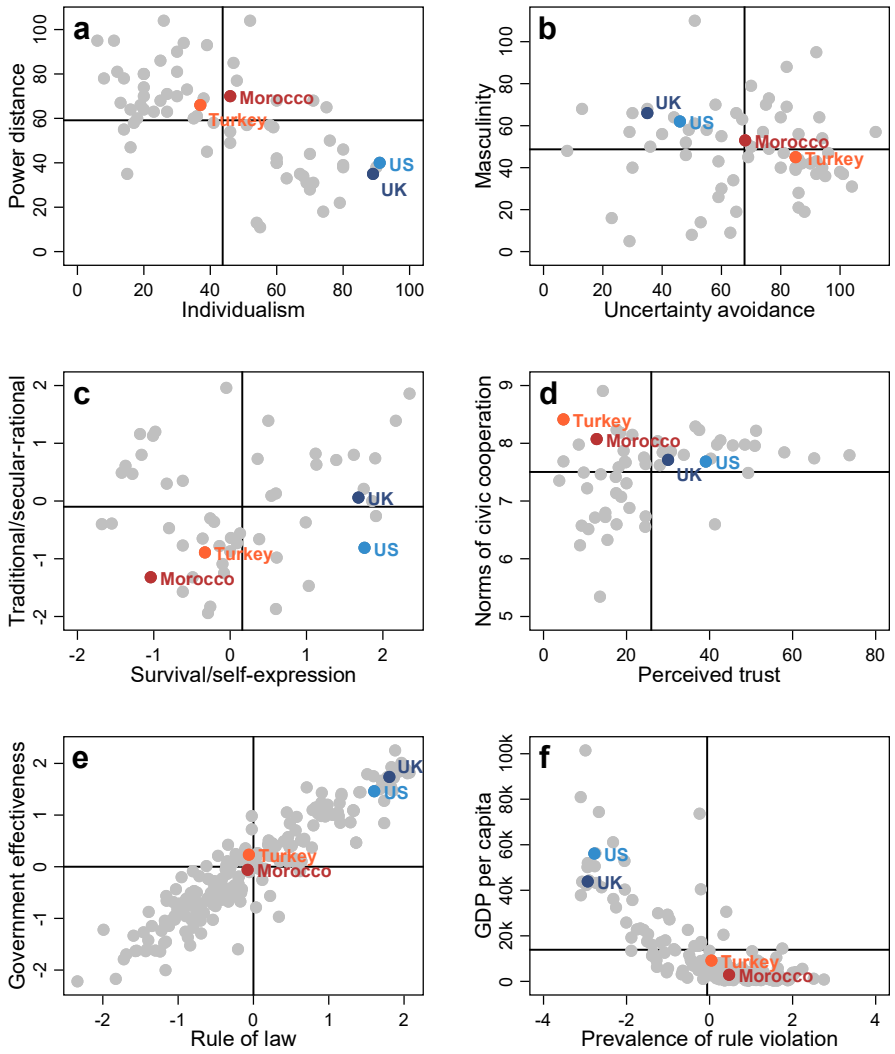


FIGURE 3.1. Measures of cultural and institutional indicators. **a**, **b** Cultural dimensions (Hofstede and Hofstede 2001). Power distance (11 = lowest; 104 = highest), Individualism (6 = least individualist; 91 = most individualist), Masculinity (5 = least masculine; 110 = most masculine) and Uncertainty avoidance (8 = most tolerant to uncertainty; 112 = most uncertainty avoidant) **c** World Values (Inglehart and Welzel 2005). Traditional vs. Secular-rational values (−1.94 = strongest emphasis on traditional values; 1.96 = strongest emphasis on secular-rational values) and Survival vs. Self-expression values (−1.68 = strongest emphasis on survival values; 2.35 = strongest emphasis on self-expression values). **d** Norms of civic cooperation (1 = very weak norm of civic cooperation; 10 = very strong norm of civic cooperation) from Knack and Keefer (1997). Perceived trust: share of people who agree with the statement 'most people can be trusted' (World Values Survey Association 2014). **e** Government effectiveness (−2.22 = lowest government effectiveness; 2.25 = highest government effectiveness) and rule of law (−2.34 = lowest rule of law; 2.07 = highest rule of law) from the Worldwide Governance Indicators (Kaufmann and Kraay 2016). **f** World Bank GDP per capita in current USD (PPP) for 2015 and Prevalence of Rule Violations (−3.10 = lowest; 2.84 = highest) from Gächter and Schulz (2016). The grey dots represent all other countries for which data was available. The black lines indicate the sample mean values.

Inglehart and Welzel (2005) introduce a measure of cultural values based on the World Value Survey (WVS). They argue that societal differences can be measured using two dimensions: *traditional values versus secular-rational values* and *survival values versus self-expression values*. The first dimension measures the importance of authority, traditional family values and religion in a society. The second dimension indicates the valuation of self-expression, individual wellbeing and quality of life. The two dimensions are extracted from the WVS survey responses using factor analysis and together account for 71% of the cross-national variation (Inglehart and Welzel 2005, p. 50). Morocco and Turkey score below the sample mean in both dimensions showing a high emphasis on traditional values and survival values. The two English-speaking countries score high on self-expression values, but the importance of traditional and secular-rational values differs. The UK scores higher on secular-rational values, whereas the US scores higher on traditional values.

In addition to cultural indicators, we draw on indicators of institutional quality to measure the distance between societies. *Norms of civic cooperation* and *perceived trust* relate to the strength of social norms in a society and are taken from the World Value Survey (World Values Survey Association 2014). The score for norms of civic cooperation measures the acceptability of claiming government benefits one is not entitled to, fare-dodging on public transport and cheating on taxes (rescaled average value of World Value Survey Questions V198–V200; Knack and Keefer 1997). All four countries' scores for norms of civic cooperation are higher than the sample mean for norms of civic cooperation. Morocco and Turkey score lower than the world mean on perceived trust, whereas the scores for the UK and the US lie above the sample mean.

Government effectiveness and *Rule of law* are drawn from the Worldwide Governance Indicators 2015 (Kaufmann et al. 2011, Kaufmann and Kraay 2016) and measure institutional quality. The two Islamic countries score below the sample mean for rule of law and are scattered around the mean for government effectiveness. The two English-speaking countries score higher than the sample mean on both indicators. These measures are directly linked to the quality of formal institutions such as the government and the judicial system.

Further indicators of prosperity and institutional quality are *GDP per capita* and *Prevalence of Rule Violations* (PRV, Gächter and Schulz 2016). The GDP per capita of the two Islamic countries is below the sample mean, whereas the English-speaking countries have a GDP per capita far above the sample mean. The PRV measures how common rule violations like corruption, tax evasion or political fraud are in a society. The PRV is above the sample mean in the two Islamic countries and below the sample mean for the two English-speaking countries.

Additionally, we calculate the Euclidian distances between each pair of countries for the measures of culture and institutions discussed above (Table 3.1). This confirms that Morocco and Turkey (UK and US) are relatively close according to our cultural and institutional indicators. Conversely, the Euclidian distances between countries from different cultural clusters are larger. We thus conclude that the four countries included in the present study fulfil the necessary conditions for our analysis of variance approach.

TABLE 3.1.
Euclidean distances of the economic and cultural dimensions.

	Morocco	*Turkey*	*UK*	*US*
Morocco	0			
Turkey	22.13	0		
UK	77.72	88.53	0	
US	79.65	87.34	17.77	0

3.3 Experimental methods

3.3.1 Participants and procedures

We recruited 80 students at the École Nationale d'Agriculture de Meknès in Morocco, 88 students at Istanbul Bilgi University in Turkey, 92 students at the University of Nottingham in the UK and 128 students at Stony Brook University in the US.[3] Table 3.2 summarises the characteristics of the four student samples. The experiments were computerised and conducted with z-Tree (Fischbacher 2007). We used ORSEE

[3] We excluded all participants who indicated in the post-experimental questionnaire that they were not citizens of the respective countries (two participants from the Turkish sample, four participants from the UK sample and 22 participants from the US sample).

(Greiner 2015) to recruit participants at Istanbul Bilgi University, the University of Nottingham and Stony Brook University. At the École Nationale d'Agriculture de Meknès, the experimenters recruited from the local student population ensuring that subjects would only participate once in the experiment.

TABLE 3.2.
Characteristics of the four student samples.

	Morocco	Turkey	UK	US
Age	20.84 (1.44)	22.02 (1.93)	19.63 (1.85)	19.56 (2.75)
Female	65%	40%	56%	59%
Business and economics	1%	22%	23%	9%
Urban background	48%	67%	63%	46%
Number of siblings	3.03 (2.18)	1.79 (2.01)	1.69 (1.37)	1.69 (1.35)
Middle class	89%	42%	84%	58%
N	80	86	88	106
Average payoff	MAD 100.46 (21.64)	TRY 40.93 (7.95)	GBP 11.50 (1.53)	USD 13.87 (3.32)

Note. SD in parentheses. Business and Economics: participants self-reported studying business or economics. Urban background: percent of participants who lived most of their life in a town with at least 10,000 inhabitants. Middle class: participant self-reported their family's income to be equal or greater than the average.

We follow the established rules for conducting cross-cultural economic experiments (Roth et al. 1991, Herrmann et al. 2008). In all four laboratories, the experiments were run by local research assistants. The instructions and software were presented in the local language. Participants made their decisions in private with visual separations between workstations. The experimental sessions lasted for about 90 minutes and were conducted according to a strict protocol in order to minimise the differences in the way sessions were run across countries. Participants' earnings were paid in private at the end of the session. The stake sizes were chosen for each country to reflect local purchasing power and student wages. Each session consisted two experimental phases and a socio-economic questionnaire. Participants were randomly re-matched after the first phase. We did not provide any feedback after the first phase in order to prevent participants from updating their beliefs and to exclude potential income effects and strategic play.

3.3.2 Experimental games

The present study is based on an anonymous one-shot public goods game played in groups of four. This game allows us to study an incentivised social dilemma situation in the laboratory. The game was described to participants using neutral language to avoid framing effects. Each participant received an endowment of 20 tokens and decided how much to contribute to a common project. Each token kept yielded one monetary unit (MU). All contributions to the project were multiplied by 1.6 and divided equally among the four group members. In this game, the social optimum is characterised by full contributions, whereas the money-maximising strategy is to contribute nothing. During the experimental session, participants played two different versions of the game. The first phase consisted of the cooperative disposition elicitation game (D-Game) and the second phase consisted of the punishment game (P-Game); both described in detail below.

3.3.2.1 Elicitation of cooperative dispositions (D-Game)

We used the game described above played with the strategy method to measure cooperative dispositions (Fischbacher et al. 2001). First, each subject made an unconditional contribution to the project. Afterwards, participants filled in a contribution table that allowed for conditioning their own contribution on the average contribution of the group members. We ensured incentive compatibility by randomly choosing one participant per group for whom the contribution table determined the actual contribution to the common project based on the average contribution of her three co-players. After making their two contribution decisions, we elicited the participants' beliefs about the other group members' average contribution (similar to Gächter and Renner 2010). Participants earned three MU for guessing correctly, two MU for a deviation of one point, one MU for a deviation of two points and zero MU for a higher deviation.[4]

We classified the subjects according to their responses on the contribution table for each of the 21 possible average contributions of their group members. The criteria are taken from Fischbacher et al. (2012). *Conditional Cooperators* (CC) showed a positive Spearman's rank correlation coefficient, significant at the 1% level, between own contributions and the average contributions of others or at least one increase in

[4] Belief hedging might be a potential concern in this set up. However, we believe this to be an unlikely: Blanco et al. (2010) find no evidence for belief hedging in a similar social dilemma experiment.

their contribution schedule. *Free Riders* (FR) contributed nothing for every possible average contribution of their group members. Participants who did not fit the criteria of the two categories above were deemed *Unclassified Others* (OT).

3.3.2.2 Punishment game (P-Game)

This game was a variation of the public goods game and consisted of two stages. First, participants chose their contribution to the common project and indicated their belief about the average contribution of the other group members. In the P-Game, beliefs were not incentivised in order to avoid punishment motivated by disappointment due to wrong beliefs or income effects (Cubitt et al. 2011). In the second stage of the game, participants learned their fellow group members' individual contributions and could assign up to five punishment points to each. Every punishment point cost the punisher one MU and destroyed two MU of the targeted person's income. To cover potential losses, each participant received a lump sum payment of 10 MU. After allocating punishment points to their group members, participants were asked to state how many punishment points they expected to receive from each of their fellow group members. Additionally, participants reported their emotional responses to each of their group members' contribution behaviour in the first stage. They rated the intensity of thirteen emotions (anger, contempt, envy, fear, guilt, happiness, irritation, jealousy, joy, sadness, shame, surprise and warmth) on a seven-point scale ranging from 1 (*not at all*) to 7 (*very much*).

3.4 Results

3.4.1 Part A: Driving factors of cooperation

3.4.1.1 Cooperative dispositions

To test for societal and cultural differences in the distribution of types of cooperative dispositions, we adopt the following strategy using non-parametric tests. First, we check for cross-societal variation by testing for differences across all four societies. Then, we test for differences across the two cultural clusters pooling participants of the culturally similar countries. Additionally, we check for within-cluster variation by testing for differences between the culturally similar countries. This strategy allows us to attribute cross-societal variation to the influence of culture and institutions when differences across cultural clusters are large and the variation between culturally similar countries is small.

After classifying each participant according to the individual cooperative disposition (see Methods for details) as either Conditional Cooperator (CC), Free Rider (FR) or Unclassified Others (OT), we compare the distribution of dispositions across societies as a possible driver of behavioural differences.

We find significant differences in the distributions of cooperative dispositions across the four societies, $\chi^2(6) = 14.33$, $p = .026$. The type distributions differ weakly significantly across the two cultural clusters, pooled $\chi^2(2) = 4.71$, $p = .095$ (Table 3.3). Within cultural clusters, variation appears small with only weakly significant differences for the comparison of Morocco and Turkey, $\chi^2(3) = 5.44$, $p = .066$, and no significant differences when comparing the UK and the US, $\chi^2(3) = 4.35$, $p = .113$.

Across all four societies we find that CC account for the largest share of participants with 49% in Morocco, 45% in Turkey, 55% in the UK and 57% in the US. The differences in the share of CC across the four societies are not significant, $\chi^2(3) = 2.97$; $p = .396$. However, the shares of FR vary significantly across societies from 8% in Morocco to 22% in the UK, $\chi^2(3) = 9.19$; $p = .027$.

TABLE 3.3.
Distribution of cooperative dispositions in the four countries.

Disposition	Morocco	(*a*)	Turkey	(*b*)	UK	(*c*)	US
Conditional Cooperators (CC)	49%		45%		55%		57%
Free Riders (FR)	8%	*	20%	*	22%	*n.s.*	11%
Unclassified Others (OT)	44%		35%		24%		32%
N	80		86		88		106

Note. Column (*a*) shows the result of a χ^2 test comparing Morocco and Turkey. Column (*c*) indicates the result of a test comparing the UK and the US. Column (*b*) shows the result of a pooled χ^2 test comparing Islamic and English-speaking cultural clusters. CC include perfect conditional cooperators, who equal the others' contribution: 3 subjects in Morocco, 3 in Turkey, 9 in the UK and 13 in the US. OT include unconditional cooperators: 9 subjects in Turkey, 2 in the UK, 3 in the US. *n.s.* $p \geq .10$; * $p < .10$.

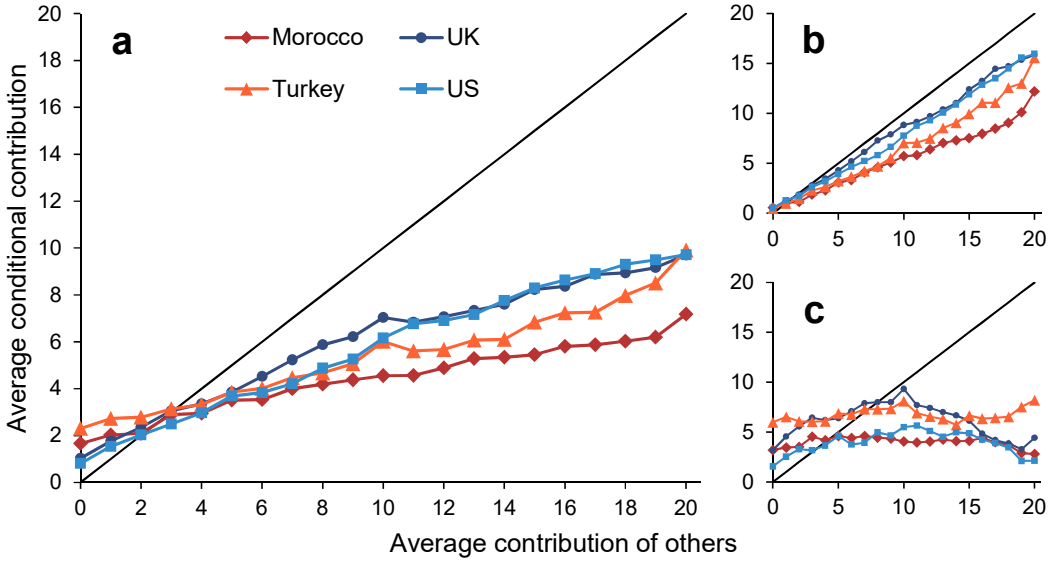


FIGURE 3.2. Average conditional contributions by societies. a The average contribution schedule across all subjects. **b** The average contribution schedule of Conditional Cooperators only. **c** The average conditional contributions of Unclassified Others. By definition, the average contribution schedule of Free Riders is along the x-axis. The 45º-line indicates perfect conditional cooperation.

Since the design of the D-Game precludes any strategic incentive to cooperate, we interpret conditional contributions, elicited with the strategy method, as an indicator of the cooperative dispositions. We find increasing average conditional contributions for a higher average contribution of others across all four societies (Figure 3.2a). A steeper increase in the average conditional contributions indicates a stronger societal inclination to match others' cooperative efforts. Perfect conditional cooperation, that is, exactly matching the average contribution of others, would result in values on the 45°-line (Figure 3.2a; black line). However, the average conditional contributions for all societies fall short of perfect conditional cooperation. This indicates that on average participants undercut the contributions of others. Thus, in all four societies conditional cooperation is imperfect.

We use a regression analysis to test for societal differences in the conditional contributions (Table 3.4, Col. 1–4). First, we regress the conditional contribution on the average contribution of other group members while controlling for the socio-economic background of participants. We estimate this regression model separately for each of the four societies. For all four societies, the coefficient for the average contributions of

others is positive and highly significant. This coefficient would be one if contributions were perfectly conditional. Imperfect conditional cooperation is implied by these coefficients being less than one. The Moroccan subjects on aggregate increase their conditional contributions by 0.240 tokens for each average additional token contributed by their fellow group members. The coefficients for the average contributions of others are larger in Turkey (b = 0.328), the UK (b = 0.416) and the US (b = 0.456) suggesting that subjects in these countries are on aggregate more conditionally cooperative.

TABLE 3.4.
A regression analysis of conditional contributions.

Dependent variable: conditional contribution	(1) Morocco	(2) Turkey	(3) UK	(4) US	(5) Pooled
Average contributions of others	0.240*** (0.041)	0.328*** (0.045)	0.416*** (0.052)	0.456*** (0.044)	0.456*** (0.044)
Islamic countries					1.171 (0.755)
Morocco					−0.150 (0.838)
UK					0.851 (0.588)
Average contribution of others × Islamic countries					−0.127** (0.063)
Average contribution of others × Morocco					−0.089 (0.060)
Average contribution of others × UK					−0.040 (0.068)
Socio-economic controls	Yes	Yes	Yes	Yes	Yes
Constant	4.806 (7.224)	−0.086 (6.338)	8.329 (5.533)	2.580 (3.566)	3.062 (2.710)
R^2	0.08	0.12	0.23	0.22	0.15
N	1680	1806	1848	2226	7560

Note. OLS estimation with robust *SE* clustered on individuals in parentheses. The control variables (age, female, urban background, middle class, single child, economics/business student) are insignificant, apart from a positive coefficient of female in Col. 3, significant at the 5% level. * $p < .10$; ** $p < .05$; *** $p < .01$.

Additionally, we estimate a pooled regression model to test for differences between the Islamic countries cultural cluster and the English-speaking countries cultural cluster, as well as for differences within these cultural clusters (Table 3.4, Col. 5). We include dummy variables for the Islamic countries cultural cluster, Morocco, the UK, as well as interaction terms between the dummy variables and the average contribution of others.

The insignificant dummy variable 'Islamic countries' suggests no level differences in conditional contributions across cultural clusters. However, the negative and significant interaction 'Average contribution of others × Islamic countries' indicates a lower inclination to match the average contributions of others in the Islamic countries cultural cluster compared to the English-speaking countries cultural cluster, ceteris paribus. We find no evidence of within-cluster differences for the Islamic countries. The insignificant dummy variable 'Morocco' and the insignificant interaction 'Average contribution of others × Morocco' indicate similar conditional contributions comparing the two Islamic countries. This also holds true when testing for joint significance of the two coefficients, $F(2, 359) = 1.36$, $p = .259$. Similarly, we do not find differences within the cluster of English-speaking countries, comparing the UK and the US. This follows from the insignificant dummy variable for the UK and its interaction term. The two coefficients are jointly insignificant, $F(2, 359) = 1.06$, $p = .348$.

Since the most common cooperative disposition in all four societies is CC, we repeat the analysis above for only those subjects classified as CC. The conditional contributions of these subjects show a steeper increase, closer to perfect conditional cooperation (Figure 3.2b). The regression analysis reveals no significant differences in the conditional contributions across cultural clusters (Appendix; Table 3.6). However, the conditional contributions vary significantly within the Islamic countries cultural cluster, but not within the English-speaking countries cultural cluster.

The conditional contribution of FR is zero for every possible level of average contributions of others by definition. Thus, there is no variation in conditional contributions across societies. However, there might be some variation in the conditional contributions of subjects classified as neither CC nor FR. We find that these

subjects show a relatively stable pattern of conditional contributions with societal differences in the level of conditional contributions (Figure 3.2c).

In summary, we find that all societies on aggregate are characterised by imperfect conditional cooperation, that is, the desire to partially match the average contribution of other group members. However, there are differences in conditional cooperation across cultural clusters. Subjects in the English-speaking countries cultural cluster are on average more willing to contribute than their counterparts in the Islamic countries cultural cluster for a higher average contribution of others. This does not hold when comparing only CC. The shares of CC are comparable across cultural clusters and their conditional contributions are similar. This highlights the role of beliefs in determining the outcome of the decision-making process.

3.4.1.2 Beliefs and cooperative behaviour

Figure 3.3 compares unconditional beliefs and unconditional contributions in the D-Game. We find significant variation in unconditional beliefs across societies, Kruskal-Wallis $\chi^2(3) = 18.56$, $p < .001$. First, we compare the differences in unconditional beliefs across cultural clusters by pooling the data for the two English-speaking countries and the two Islamic countries. Unconditional beliefs are significantly different when comparing the English-speaking countries cultural cluster with the Islamic countries cultural cluster (pooled Mann-Whitney $Z = 3.19$ $p = .001$). Additionally, we look at within-cluster variation. We find highly significant differences when comparing Morocco and Turkey ($M_{MO} = 6.22$, $SD_{MO} = 5.21$; $M_{TR} = 8.38$, $SD_{TR} = 5.20$; Mann-Whitney $Z = -2.86$, $p = .004$), but no differences in the average unconditional beliefs between the UK and the US ($M_{UK} = 8.72$, $SD_{UK} = 5.69$; $M_{US} = 8.88$, $SD_{US} = 3.84$; Mann-Whitney $Z = -0.12$, $p = .908$). We also test for differences in the accuracy of beliefs, defined as the deviation from the actual contribution behaviour of others, and find a similar belief accuracy across societies (Appendix; Table 3.7).

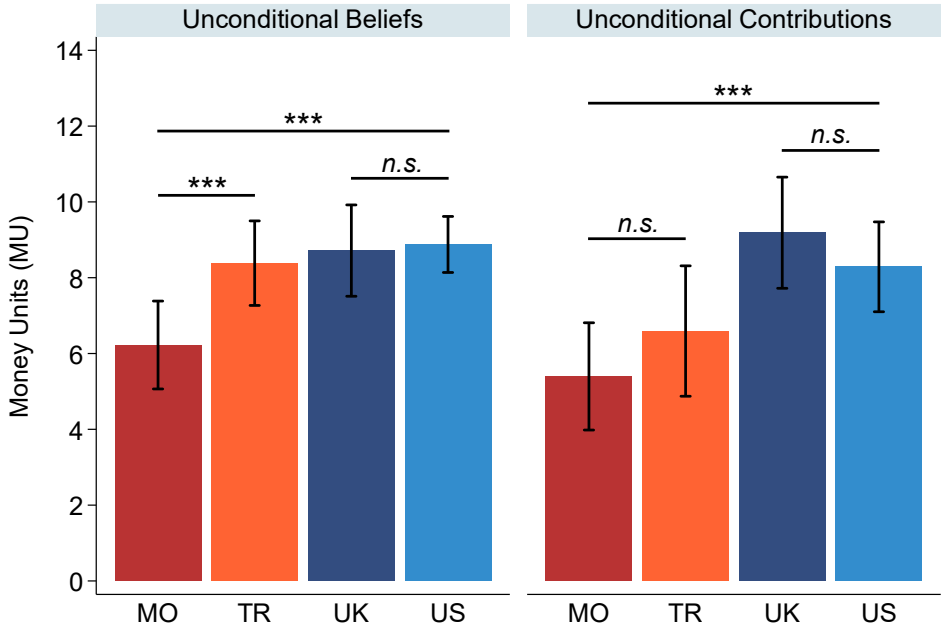


FIGURE 3.3. Mean unconditional beliefs and unconditional contributions in the D-Game by country. The error bars indicate the 95% CI. (Pooled) Mann-Whitney test: *n.s.* $p \geq .10$; *** $p < .01$.

Similarly, unconditional contributions vary significantly across societies, Kruskal-Wallis $\chi^2(3) = 19.62$, $p < .001$. Comparing unconditional contributions across cultural clusters shows significant differences between the English-speaking and the Islamic countries cultural clusters (pooled Mann-Whitney $Z = 4.36$, $p < .001$). Our within cultural cluster analysis of unconditional contributions reveals no differences when comparing the UK and the US ($M_{UK} = 7.88$, $SD_{UK} = 5.93$; $M_{US} = 7.10$, $SD_{US} = 5.28$; Mann-Whitney $Z = 1.05$, $p = .293$) or Morocco and Turkey ($M_{MO} = 4.63$, $SD_{MO} = 5.45$; $M_{TR} = 5.65$, $SD_{TR} = 6.88$; Mann-Whitney $Z = 0.18$, $p = .858$).

Furthermore, we use the individual cooperative disposition and unconditional beliefs to calculate point predictions for the unconditional contributions in the D-Game. Comparing predicted with actual contributions in the D-Game yields accuracy of more than 61% for all four societies and no significant differences in the accuracy of predicted contributions (Appendix; Figure 3.8).

The results presented above show that cultural and institutional conditions affect beliefs and contributions in the absence of a possibility to punish others. For both,

beliefs and contributions, we observe little or no variation *within* cultural clusters, while observing large and significant variations in both *across* cultural clusters.

3.4.2 Part B: Societal differences in punishment

3.4.2.1 Beliefs and cooperative behaviour in the P-Game

Figure 3.4 illustrates average beliefs and contributions in the P-Game for all four societies. First, we investigate the differences in *beliefs* across cultural clusters. This shows significant differences in beliefs across the two cultural clusters (pooled Mann-Whitney $Z = 3.67$, $p < .001$). Next, we explore the variation in beliefs within cultural clusters. Beliefs are similar for the two Islamic countries ($M_{MO} = 6.95$, $SD_{MO} = 5.76$; $M_{TR} = 8.22$, $SD_{TR} = 5.91$; Mann-Whitney $Z = -1.43$, $p = .153$) and weakly significantly different for two English-speaking countries ($M_{UK} = 10.08$, $SD_{UK} = 5.10$; $M_{US} = 8.89$, $SD_{US} = 4.41$; Mann-Whitney $Z = 1.77$, $p = .076$). We find no differences in the accuracy in beliefs across societies, which is defined as the deviation from the actual contribution of others (Appendix; Table 3.7).

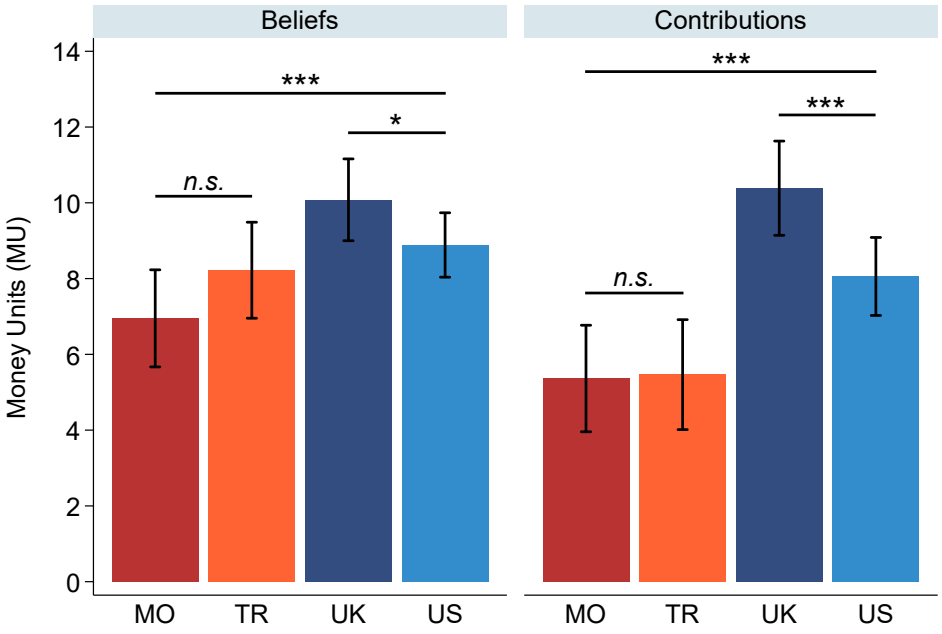


FIGURE 3.4. Mean beliefs and contributions in the P-Game by country. The error bars indicate the 95% CI. (Pooled) Mann-Whitney test: *n.s.* $p \geq .10$; * $p < .10$; *** $p < .01$.

Now we turn to societal differences in *contributions*. Contributions are significantly different across the two cultural clusters (pooled Mann-Whitney $Z = 6.48$, $p < .001$). A comparison within cultural clusters shows similar contribution levels for the two Islamic countries ($M_{MO} = 5.36$, $SD_{MO} = 6.32$; $M_{TR} = 5.47$, $SD_{TR} = 6.77$; Mann-Whitney $Z = 0.61$, $p = .540$) and significant differences for the two English-speaking countries ($M_{UK} = 10.39$, $SD_{UK} = 5.87$; $M_{US} = 8.06$, $SD_{US} = 5.34$; Mann-Whitney $Z = 2.79$, $p = .005$).

Although we find some variation in beliefs and contributions within cultural clusters, the variation across cultural clusters is highly significant. Thus, we conclude that there is a clear effect of cultural and institutional differences on beliefs and contributions in the P-Game. The difference in contributions between subjects in the Islamic and English-speaking countries cultural clusters appear especially noteworthy and warrants further study.

Next, we investigate potential societal differences in the effect of punishment on beliefs and cooperative behaviour. For our investigation, we compare the levels of beliefs and behaviour across the D-Game without punishment and the P-Game with punishment. If participants believe that defectors will be punished, then contribution levels and beliefs might increase in the P-Game compared to the D-Game. Thus, comparing beliefs and contributions across the D-Game and the P-Game gives some insights into a society's norms of cooperation and expected punishment of norm transgressors.

Mean beliefs about the other group members' contributions in the two games are similar for Morocco, Turkey and the US (Wilcoxon signed-ranks tests, all $p > .339$). Only participants in the UK hold weakly significantly higher beliefs in the P-Game compared to the D-Game (Wilcoxon signed-ranks $Z = -1.72$, $p = .086$). In contrast, the effect of introducing punishment on cooperative behaviour differs across cultural clusters. Mean contributions with punishment are significantly higher in the UK and the US compared to the game without punishment (Wilcoxon signed-ranks $Z = -3.65$, $p < .001$; $Z = -2.42$, $p = .016$; respectively). However, introducing punishment does not affect average contribution levels for Morocco and Turkey (Wilcoxon signed-ranks $Z = -0.78$ $p = .435$; $Z = -0.45$ $p = .656$; respectively). This behaviour is in line with the findings of Herrmann et al. (2008) who show that the effect of introducing punishment

on the level of cooperation in public good games differs starkly across societies. In some societies, contributors and defectors are punished to similar extents and thus punishment does not foster cooperative behaviour.

In conclusion, we find that cooperation behaviour in the presence of punishment varies across cultural clusters with lower contribution levels in Morocco and Turkey. A reason for this might be that subjects in these countries expect an arbitrary use of punishment which targets high contributors as well as defectors. Consequently, punishment is not expected to have cooperation enhancing effects. The following section investigates societal differences in anticipated punishment behaviour as a potential explanation for the societal variation in contributions.

3.4.2.2 Expected punishment

A credible threat of punishment can change the incentives in our one-shot experiment substantially, so that rational and selfish agents might increase contributions in order to avoid being punished. However, the direction of expected punishment is crucial in order for it to be an effective deterrent of defection. For example, we only expect punishment of defection (prosocial punishment) to foster high cooperation levels. In contrast, cooperation levels are likely to decline if participants expect high levels of punishment for cooperative behaviour (antisocial punishment).

In order to measure expected punishment in the P-Game, we asked participants to state the number of punishment points they expect to receive from each of their fellow group members. Using an OLS regression model, we estimate expected punishment across societies (Appendix; Table 3.8). The regression model explains expected number of punishment points as a function of the positive or negative contribution deviation from the other group member (the potential punisher). Figure 3.5 illustrates the results separately for negative deviations from the other group member's contribution level (i.e., prosocial punishment) and positive deviations (i.e., antisocial punishment).

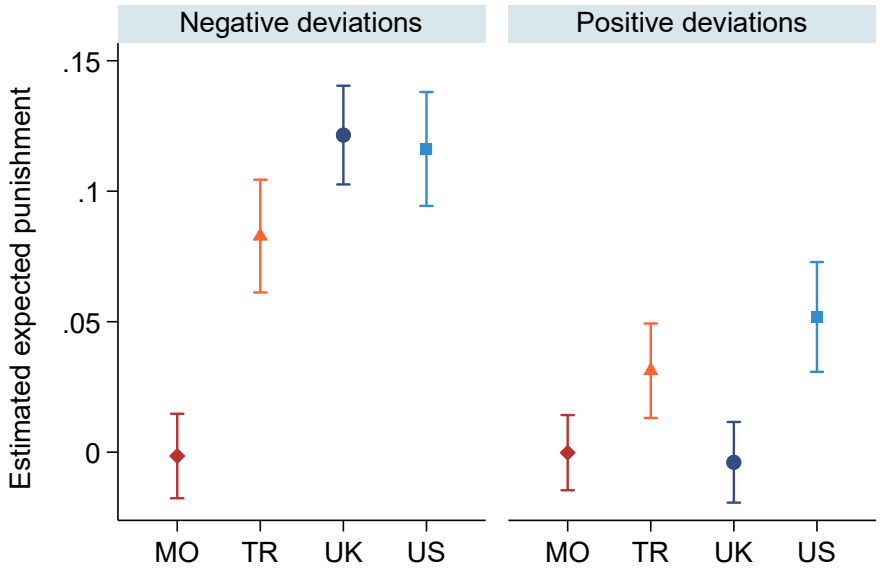


FIGURE 3.5. Estimated expected punishment (±1 *SEM*) by country for a negative or positive one-unit deviation from the punisher's contribution. The estimated expected punishment is obtained from a pooled OLS regression. The estimates are the linear combination of the relevant coefficients (Appendix; Table 3.8, Col. 5).

For Morocco, a negative one-MU deviation from the other group member's contribution level is associated with an expected punishment of approximately zero points (Figure 3.5; left panel). For the other three societies, negative deviations by one MU are linked with an expected punishment of 0.083 points in Turkey, 0.122 points in the UK and 0.116 points in the US. Although the results above suggest considerable variation in the expected punishment across societies, our regression analysis (Appendix; Table 3.8, Col. 5) shows no significant differences in the punishment of *negative deviations* across cultural clusters. However, we find substantial within-cluster variation in expected punishment for the Islamic countries cultural cluster, with expected prosocial punishment being higher in Turkey compared to Morocco. Expected punishment of negative deviation is similar for the UK and the US.

We draw on the same regression analysis to compare differences in expected punishment of *positive deviations* across societies (Figure 3.5; right panel). For all four societies, expected punishment for a positive one-MU deviation from another subject's contribution level is close to zero points. We find no significant differences in the expected number of punishment points across cultural clusters. However, we find significant within-cluster variation for the Islamic countries cultural cluster with higher

expected antisocial punishment in Turkey compared to Morocco. Additionally, we find higher expected antisocial punishment in the US compared to the UK (Appendix; Table 3.8, Col. 5).

A further question is whether the accuracy of expected punishment differs across societies. We find a significantly higher accuracy of expected punishment for the Islamic countries cultural cluster compared to the English-speaking countries cultural cluster (Appendix; Figure 3.9).

In conclusion, expected punishment can help to explain contribution behaviour. Within the English-speaking countries cultural cluster, the expectation is that defection will be punished almost exclusively leading to high contribution levels in the P-Game. Conversely, for Morocco no punishment is expected, which explains the low contribution levels. The findings for Turkey are however more complex: We find that the expected prosocial punishment lies in the medium range and some antisocial punishment is expected. This relatively undirected punishment is likely to be driving down cooperative behaviour.

3.4.2.3 Realised punishment

Societal differences in realised punishment are an indicator for the prevalence of norm enforcement in a society (Figure 3.6; Appendix; Table 3.9). We use an OLS regression model, similar to the one described above, to estimate punishment expenditure for each society. Again, we look at the punishment of negative deviations of a subject's contribution from the punisher's contribution (prosocial punishment) and the punishment of positive contribution deviations (antisocial punishment) separately.

Although Figure 3.6 (left panel) suggests a considerable variation in the realised punishment of negative deviations across societies, we find no significant differences across the two cultural clusters (Appendix; Table 3.9, Col. 5). However, the variation within the Islamic countries cultural cluster is large with a significantly higher estimated punishment of negative deviations in Turkey compared to Morocco. The estimated prosocial punishment is similar for the UK and the US.

The estimated antisocial punishment is low in all four societies (Figure 3.6; right panel) and significantly lower in the Islamic countries cultural cluster compared to the English-speaking countries cultural cluster (Appendix; Table 3.9, Col. 5). We

find no significant differences in the estimated antisocial punishment within cultural clusters.

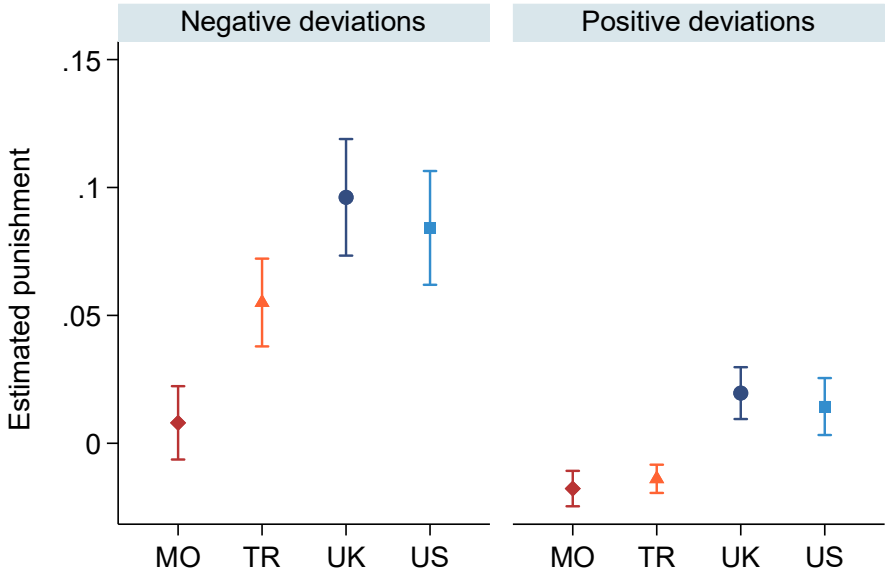


FIGURE 3.6. Estimated realised punishment (±1 *SEM*) by country for a negative or positive one-unit deviation from the punisher's contribution. The estimated expected punishment is obtained from a pooled OLS regression. The estimates are the linear combination of the relevant coefficients (Appendix; Table 3.9, Col. 5).

Our finding of very little antisocial punishment across all four societies (approximately zero for the two Islamic countries) is particularly interesting because it is at odds with Herrmann et al. (2008). There the authors report substantial levels of antisocial punishment for Turkey and low levels of antisocial punishment for the UK and the US. Morocco was not included in their sample. In the ten-times repeated game used by Herrmann et al. (2008), participants, who played in fixed group of four members, received feedback on the cooperation behaviour of others and the punishment received from other group members in every period. Thus, antisocial punishment could root in a strategic desire to increase or maintain the cooperation of others or even the desire for revenge. The one-shot design of our study excludes any selfish incentive to punish and does neither allow for within-period retaliatory punishment (since there is only one punishment stage) nor for retaliation across periods. Therefore, we interpret our results as an elicitation of the preference for the enforcement of cooperation in a society. Therefore, the low levels of antisocial punishment across all four societies indicate no *preferences* for antisocial punishment. We conclude that antisocial

punishment is rooted in strategic considerations or retaliation only possible in repeated games.

Additionally, we use this dataset to test the *Strong Reciprocators Assumption* (see Chapter 2) across societies. This common implicit assumption in the literature postulates that individuals with a cooperative disposition engage in costly altruistic punishment. Individuals with a disposition for free riding would never engage in costly altruistic punishment. Like in Chapter 2, we test this assumption by comparing the punishment frequency and expenditure of CC and FR. We find some variation of the use of punishment with cooperative dispositions across societies, but no evidence supporting the Strong Reciprocators Assumption (Appendix; Figure 3.10; Figure 3.11).

3.4.2.4 Emotions

Finally, we focus on societal differences in subjects' self-reported moral emotions of anger and guilt (Nelissen and Zeelenberg 2009). In all four societies, subjects report higher anger levels when a group member contributed less than they did themselves (Figure 3.7). Thus, anger increases for negative contribution deviations.

We use an ordered Probit regression to test for differences in self-reported anger across societies when controlling for the magnitude of the negative or positive contribution deviation. Although all societies report an increase in anger for larger negative contribution deviations (Appendix; Table 3.11, Col. 1–4), we find significant level differences across cultural clusters with higher anger levels, but a significantly lower emotional reaction to negative contribution deviations in the Islamic countries cultural cluster (Appendix; Table 3.11, Col. 5). Additionally, the regression analysis yields significant variations within cultural clusters. The reaction to negative contribution deviations is significantly lower in Morocco compared to Turkey. We also find variations within the cluster of English-speaking countries with higher anger levels reported in the UK and a lower reaction to positive deviations in the UK compared to the US.

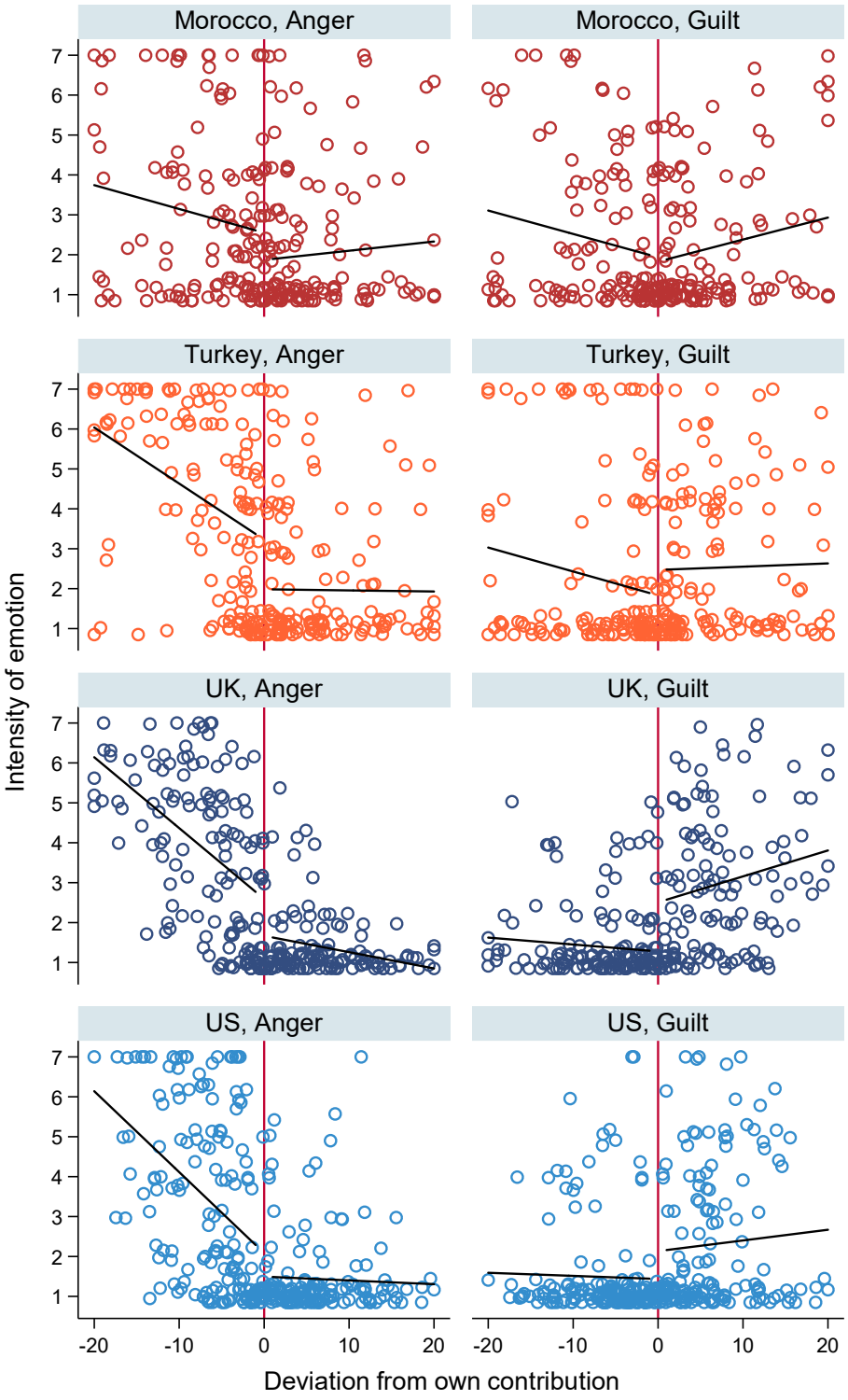


FIGURE 3.7. Self-reported anger and guilt depending on the contribution deviation between the reporting subject and another group member. The seven-point scale measures the extent with which the emotion is felt, ranging from 1 (*not at all*) to 7 (*very much*). The black lines indicate linear regression lines estimated separately for negative and positive deviations.

The link between guilt and contribution deviation appears weaker especially for Turkey. An ordered Probit regression analysis yields significantly higher levels of guilt in the Islamic countries cultural cluster compared to the English-speaking countries cultural cluster and higher reported levels of guilt for negative deviations only (Appendix; Table 3.12, Col. 5). Additionally, we find no differences within the Islamic countries cultural cluster, but some variation within the English-speaking countries cultural cluster.

In conclusion, we find that anger seems to be a universal reaction to defection, although the intensity of reported anger differs across societies. Since anger can serve as a proximate explanation for punishment (Fehr and Gächter 2002, Sanfey et al. 2003, Cubitt et al. 2011), differences in anger levels might help to explain societal differences in punishment. We expect this to be an important explanation of the observed behaviour in our experiment since the one-shot nature of the game does not allow for the self-serving motivations of punishment. The association with guilt is weak in some societies. The cooperation-enhancing effect of guilt, as inhibitor of counter punishment (Hopfensitz and Reuben 2009), does not seem to play a role in our one-shot game.

3.5 Discussion

The present study investigates differences in cooperative dispositions and beliefs as a driving factor of societal variation in cooperative behaviour. Cooperative dispositions across all societies are characterised by conditional cooperation, but variation in the distribution of individual cooperative dispositions across societies is prevalent. In both, the one-shot games with and without punishment, we found differences in beliefs and cooperative behaviour across cultural clusters. Additionally, we found that the introduction of punishment had a societal component: Only participants in the UK and the US significantly increased contributions in the game with punishment compared to the game without punishment. Expected punishment and realised punishment were relatively similar across cultural clusters. Exploring the subjects' emotional reaction to the social dilemma game, we find anger increases in all societies when a group member contributed more than her co-player.

The prevalence of conditional cooperation across all four societies highlights the importance of beliefs about the behaviour of other group members in making one's own contribution decision. Thus, the 'ABC of cooperation' framework (Gächter et al.

2017) of dispositions and beliefs driving cooperative behaviour is useful to explain societal and cultural variations. This framework helps to demonstrate that differences in cooperation levels across societies are not solely rooted in differing preferences for cooperation but are influenced by differing beliefs of the likelihood of others' cooperation as well. This holds for experimental games both with and without costly peer-punishment.

If peer-punishment is available, expected punishment is likely to influence one's own contribution decision as well. This link became apparent through the relatively low contribution levels in Morocco and Turkey in the game with punishment. The fact that the contribution levels are below the expected contribution can have two causes: This behaviour is optimal if little punishment of defection is expected as is the case for Morocco, or if some antisocial punishment is expected as is the case for Turkey.

Thus, we conclude that the effects of societal differences on cooperative behaviour are far from trivial. On the one hand, culture and institutions are likely to shape norms of behaviour and thus influence peoples' behaviour directly. On the other hand, people's everyday experiences about the cooperativeness of others influence beliefs about other peoples' behaviour in a given situation, that is, the expected contributions of others to the public good in our experiments.

We found no differences across cultural clusters in prosocial punishment. The levels of antisocial punishment are low across all four societies, but significantly lower in the Islamic countries cultural cluster compared to the English-speaking countries cultural cluster. This is at odds with the results reported by Herrmann et al. (2008). One reason for the difference might be that Herrmann et al. (2008) used repeated games with partner matching to investigate differences in punishment, whereas this study relied on one-shot experiments that exclude any strategic or self-beneficial incentive to punish. Our results show that societies are very similar in their preferences for the enforcement of cooperation. Thus, antisocial punishment seems likely to be a feature of repeated interactions with the same group of people, driven by strategic motives or counter punishment (Nikiforakis and Engelmann 2011).

Our findings on emotions are in line with the interpretation that costly peer-punishment provides an outlet for anger in all four societies. Yet, we find significantly higher levels of self-reported anger and in the Islamic countries cultural cluster

compared to the English-speaking countries cultural cluster. This might be due to societal differences in the acceptability of showing emotions in encounters with strangers.

However, our findings are at odds with Gächter and Herrmann (2009) who conducted one-shot experiment with punishment in Russia (attributed to the cultural cluster of Orthodox countries) and Switzerland (belonging to the cultural cluster of Protestant-European countries). The authors reported significant differences in antisocial punishment across the two countries in their one-shot setting. It is worth mentioning that, although absolute levels differ across countries, the pattern of antisocial punishment is similar: Most punishment is directed at defectors.

An additional benefit of our study is extending the experimental research beyond western, educated, industrialised, rich and democratic countries (Henrich et al. 2010). Like most experimental research, we draw on participants from the UK and the US, but now are able to compare the behaviour of subjects from more traditional experimental economics subject pool with participants from Morocco and Turkey. These four countries included in this study belong to two different cultural clusters and differ on many cultural and institutional dimensions. Thus, we can make a strong case for our two main findings. First, cultural differences in beliefs and dispositions drive cooperation levels across societies. Second, the comparable punishment behaviour across societies hints at similar preferences for norm enforcement across cultural clusters.

3.6 Appendix

3.6.1 Supporting analysis

3.6.1.1 Cultural distance

Cross-cultural differences in cooperative norms and norm enforcement are likely to be captured by various macro-level indicators (Herrmann et al. 2008). We compared important measures of socio-economic and cultural differences for the four countries we study (Table 3.5).

TABLE 3.5.
Cultural and institutional indicators for Morocco, Turkey, the UK and the US.

Dimension	Indicator	Morocco	Turkey	UK	US
World values	Traditional vs. secular-rational	−1.32	−0.89	0.06	−0.81
	Survival vs. self-expression	−1.04	−0.33	1.68	1.76
Cultural dimensions	Power distance	70	66	35	40
	Individualism	46	37	89	91
	Masculinity	53	45	66	62
	Uncertainty avoidance	68	85	35	46
Prosperity	GDP per capita	2.9	9.1	43.9	56.1
Quality of institutions	Prevalence of Rule Violations	0.47	0.04	−2.93	−2.78
	Government effectiveness	−0.06	0.23	1.74	1.46
	Rule of law	−0.08	−0.06	1.81	1.60
Social capital	Norms of civic cooperation	8.07	8.41	7.71	7.68
	Perceived trust	0.13	0.05	0.30	0.39

Note. Cultural dimensions (Hofstede and Hofstede 2001). World Values (Inglehart and Welzel 2005). World Bank GDP per capita (PPP) in current USD for 2015 and Prevalence of Rule Violations (Gächter and Schulz 2016). Government effectiveness and rule of law (Worldwide Governance Indicators 2015). Norms of civic cooperation (Knack and Keefer 1997). Perceived trust (World Values Survey Association 2014).

3.6.1.2 Conditional contributions across societies

The regression analysis of conditional contributions of CC follows Table 3.4 in the main text. The only difference is that Table 3.6 only includes subjects classified as CC. Col. 1–4 show large and significant coefficients for the average contribution of others. This indicates that for an increase in the average contribution of other by one MU, CC in Morocco increase their conditional contribution by 0.510 MU (in Turkey by 0.695 MU, in the UK by 0.788 MU, in the US by 0.789 MU).

TABLE 3.6.
A regression model of conditional contributions of *Conditional Cooperators*.

Dependent variable: conditional contribution	(1) Morocco	(2) Turkey	(3) UK	(4) US	(5) Pooled
Average contributions of others	0.510*** (0.049)	0.695*** (0.056)	0.788*** (0.045)	0.789*** (0.038)	0.789*** (0.037)
Islamic countries					0.228 (0.714)
Morocco					0.566 (0.670)
UK					0.630 (0.586)
Average contribution of others × Islamic countries					−0.094 (0.067)
Average contribution of others × Morocco					−0.185** (0.074)
Average contribution of others × UK					−0.002 (0.058)
Socio-economic controls	Yes	Yes	Yes	Yes	Yes
Constant	21.770*** (7.658)	8.759 (8.800)	6.863** (3.357)	−0.666 (3.802)	4.573* (2.602)
R^2	0.43	0.45	0.60	0.63	0.53
N	819	819	1008	1260	3906

Note. Includes only Conditional Cooperators. OLS estimation with robust *SE* clustered on individuals in parentheses. Control variables: age, female, urban background, middle class, single child, economics/business student. The following effects are significant at the 5% level: negative for age in Col. 1 and Col. 3, as well as single child in Col. 2, all significant at the 5% level. * $p < .10$; ** $p < .05$; *** $p < .01$.

The insignificant dummy variable 'Islamic', its insignificant interaction effect with the average contribution of others, and joint insignificance of both coefficients, $F(2, 185) = 0.99$, $p = .374$, suggest no differences in the conditional contributions of CC across cultural clusters (Col. 5). Similarly, within the cluster of English-speaking countries, we find no evidence for level differences, no significant interaction effect and joint insignificance of both, $F(2, 185) = 0.72$, $p = .490$. Yet, comparing the Islamic countries yields significantly lower conditional contributions in Morocco for a higher contribution of others. The insignificant dummy variable for Islamic countries suggests no level differences in conditional contributions when comparing Morocco and Turkey.

3.6.1.3 Accuracy of beliefs

In all four societies included in the present study, participants are on average imperfect conditional cooperators who condition their own kindness on their expected behaviour of others. This conditional cooperation strategy works best if people are able to predict the behaviour of others correctly. Furthermore, the accuracy of beliefs might be an important indicator for the saliency of contribution norms within a society. We define belief accuracy as the absolute deviation of beliefs about other group members' behaviour from their actual unconditional contribution (Table 3.7).

TABLE 3.7.
Absolute deviation of beliefs from actual contributions (i.e., belief accuracy).

	D-Game	P-Game
Morocco ($N = 80$)	5.27 (4.02)	5.09 (4.56)
Turkey ($N = 86$)	5.47 (4.23)	5.68 (4.46)
UK ($N = 88$)	5.39 (3.38)	4.97 (3.60)
US ($N = 106$)	4.30 (3.14)	4.25 (3.42)
Kruskal-Wallis test p	.134	.177

Note. SD in parentheses.

In the D-Game, the accuracy of beliefs is similar across cultural clusters (pooled Mann-Whitney $Z = -0.731$, $p = .465$). Looking at the variation in the accuracy of beliefs within a cultural cluster, we do not find a significant difference comparing Morocco and Turkey (Mann-Whitney $Z = -0.15$, $p = .878$), but significant differences across the UK and the US (Mann-Whitney $Z = 2.38$, $p = .017$). In the P-Game, there are no significant differences in accuracy of beliefs comparing the two cultural clusters

(pooled Mann-Whitney $Z = -1.17$, $p = .244$). Likewise, the accuracy of beliefs within cultural clusters is similar for the Islamic countries (Mann-Whitney $Z = -1.19$, $p = .233$) as well as the English-speaking countries (Mann-Whitney $Z = 1.57$, $p = .116$). These results show that there is some uncertainty about other people's behaviour, but this uncertainty does not vary with culture.

3.6.1.4 Accuracy of predicted contributions

To test differences in the determinants of cooperative behaviour across countries, we investigate the predictive power of individual cooperative dispositions. Following Fischbacher et al. (2012), we calculate the predicted unconditional contribution for each participant using their schedule and unconditional belief to then compare the predicted with the actual contribution ±2 tokens (Figure 3.8). 61.3% of participants in Morocco, 62.8% in Turkey, 61.4% in the UK and 61.3% in the US are consistent with predicted contributions. The average deviation from predicted contributions are not significantly different across countries, Kruskal-Wallis $\chi^2(3) = 1.14$, $p = .768$.

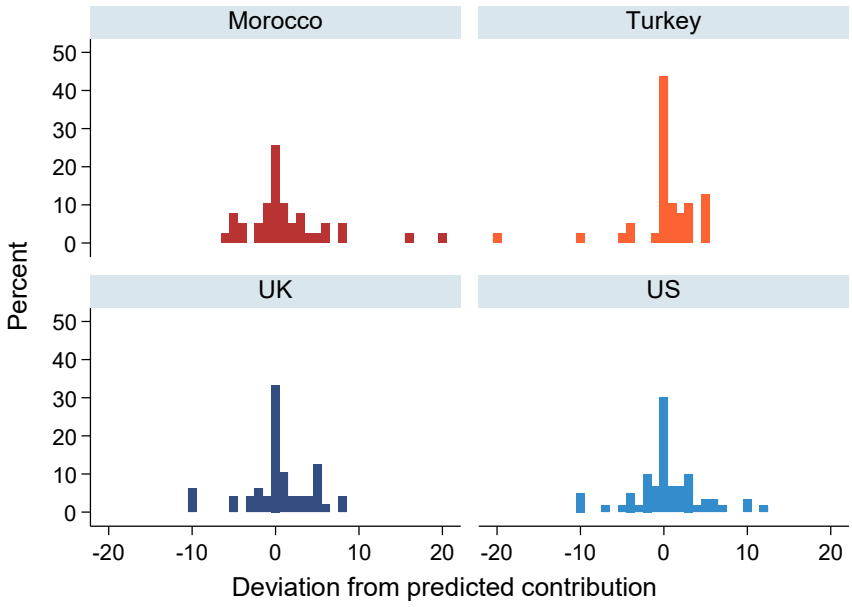


FIGURE 3.8. Deviations from predicted contributions in the four countries.

3.6.1.5 Accuracy of expected punishment

We investigate whether there are differences in the accuracy of expected punishment across the three countries (Figure 3.9). In Morocco, 70% of punishment actions are correctly predicted. Participants in Morocco overestimate the number of punishment points in 22% of instances. In Turkey, 51% of punishment actions are correctly predicted and 38% are overestimated. In the UK, 47% of instances are correctly predicted and 44% are overestimated. In the US, 48% of beliefs about punishment are correct and 40% overpredict punishment. The accuracy of beliefs (absolute deviation of expected punishment from actual punishment) is highly significantly different across cultural clusters (pooled Mann-Whitney $Z = 3.67$, $p < .001$). We find a highly significant difference in accuracy of beliefs across the two Islamic countries (Mann-Whitney $Z = -4.17$, $p < .001$) and a similar accuracy of beliefs across the two English-speaking countries (Mann-Whitney $Z = 0.30$, $p = .766$). These results hint at differences in the salience of punishment norms across countries. In Morocco, punishment is less prevalent, and a large share of people correctly predict the number of punishment points that they actually receive. Therefore, the norm to punish defectors seems to be relatively weak. In the UK, punishment of defectors is more severe, but people overestimate the use of punishment.

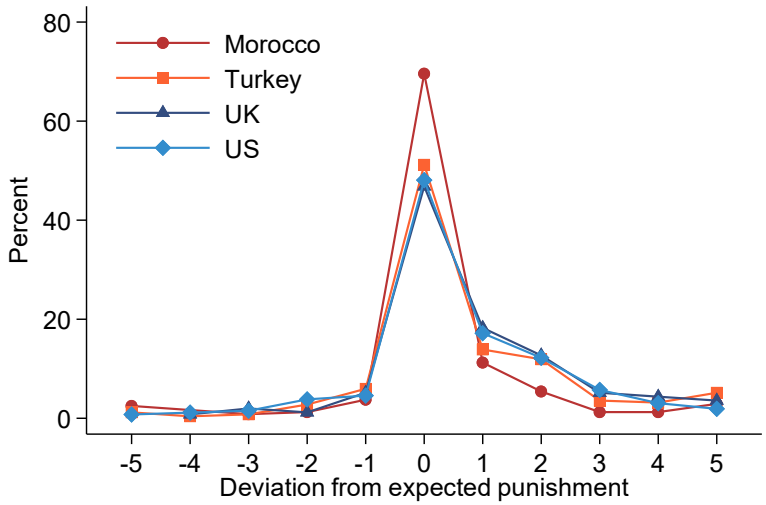


FIGURE 3.9. Accuracy of beliefs about punishment. If the deviation of expected punishment from the actual punishment equals zero, then beliefs are correctly predicted. A positive deviation shows an overestimation of punishment and a negative deviation indicates an underestimation.

3.6.1.6 Regression analysis of punishment

We use an OLS regression model to test for differences in expected punishment across cultural clusters (Table 3.8). Although some authors suggest a two-stage regression model in order to separate between the likelihood of punishment and the punishment severity (Nikiforakis and Engelmann 2011, Chapter 2), splitting the regression analysis in two steps is not feasible due to a small sample size and relatively few punishment incidences in our one-shot game.

As independent variables, we include the absolute negative deviation in the contribution levels between the punisher and the person receiving the punishment. Additionally, we include the positive contribution deviation between the punisher and the person receiving the punishment. We also include socio-economic controls (age, gender, urban background, middle class, single child, economics or business student). We first estimate the models separately for the four societies (Table 3.8, Col. 1–4). In all societies but Morocco, subjects expect a significant increase in punishment for higher negative deviation from the punisher's contribution. In Turkey and the US, subjects' expected punishment also increases in the positive deviation from the punisher's contribution.

We also estimate a pooled model with further explanatory variables to test for differences between and within cultural clusters (Table 3.8, Col. 5). We include a dummy variable for Islamic countries, Morocco, the UK, as well as interaction terms between the dummy variables and the contribution deviation.

We find no level differences in expected punishment between cultural clusters. The interaction terms 'Absolute negative deviation × Islamic countries' and 'Positive negative deviation × Islamic countries' are insignificant as well, suggesting no differences in the reaction to negative or positive contribution deviations across cultural clusters. Additionally, we find no joint significance of the dummy variable for Islamic countries with its interaction terms for negative deviations, $F(2, 1062) = 0.59$, $p = .553$, or positive deviation, $F(2, 1062) = 0.31$, $p = .735$.

TABLE 3.8.
Regression analysis of expected punishment.

Dependent variable: expected punishment	(1) Morocco	(2) Turkey	(3) UK	(4) US	(5) Pooled
Absolute negative deviation from punisher's contribution	0.007 (0.016)	0.076*** (0.021)	0.119*** (0.019)	0.120*** (0.022)	0.116*** (0.022)
Positive deviation from punisher's contribution	0.003 (0.016)	0.030* (0.018)	−0.000 (0.016)	0.058*** (0.022)	0.052** (0.021)
Islamic countries					0.113 (0.178)
Morocco					−0.100 (0.192)
UK					0.308* (0.160)
Absolute negative deviation × Islamic countries					−0.033 (0.031)
Absolute negative deviation × Morocco					−0.084*** (0.027)
Absolute negative deviation × UK					0.005 (0.029)
Positive deviation × Islamic countries					−0.021 (0.028)
Positive deviation × Morocco					−0.031 (0.023)
Positive deviation × UK					−0.056** (0.026)
Socio-economic control variables	Yes	Yes	Yes	Yes	Yes
Constant	−2.845** (1.106)	2.117** (1.058)	1.938** (0.912)	0.154 (0.592)	0.428 (0.414)
R^2	0.18	0.13	0.20	0.18	0.15
N	240	258	264	318	1080

Note. OLS estimates. Robust standard errors in parentheses. Control variables: age, female, urban background, middle class, single child, economics/business student. The following effects are significant at the 5% level. Col. 1: positive for age, single child; negative for economics/business student. Col. 2: positive for urban; negative for economics/business student. Col. 4: positive for female; negative for urban. * $p < .10$. ** $p < .05$. *** $p < .01$.

We find no level differences in expected punishment within the cultural cluster of Islamic countries. However, the negative and highly significant interaction term 'Absolute negative deviation × Morocco' indicates a lower increase in expected punishment for higher negative contribution deviations in Morocco compared to Turkey. The interaction term for positive deviations is not significant and the interaction term and the dummy variable for Morocco are jointly insignificant, $F(2, 1062) = 2.18$, $p = .114$.

Comparing within-cluster differences for the English-speaking countries reveals weakly significantly higher levels of expected punishment in the UK compared to the US. The insignificant interaction term 'Absolute negative deviation × UK' suggests differences in the reaction to negative contribution deviations comparing the UK and the US. However, the interaction term 'Absolute positive deviation × UK' is negative and significant, indicating a lower expected punishment of positive contribution deviations in the UK compared to the US.

To investigate societal differences in realised punishment, we estimate an OLS model with a similar specification as discussed above (Table 3.9). Now, the dependent variable is punishment expenditure. The absolute negative deviation now refers to the deviation of a group member from the punisher and shows prosocial punishment. A positive deviation indicates higher contributions compared to the punisher which implies antisocial punishment.

First, we estimate regression analyses for each society separately (Table 3.9, Col. 1–4). For all countries but Morocco, the punishment expenditure increases significantly for a higher negative contribution deviation. For Morocco and Turkey, punishment decreases significantly if the other group member contributes more than the punisher.

TABLE 3.9.
Regression analysis of realised punishment.

Dependent variable: punishment expenditure	(1) Morocco	(2) Turkey	(3) UK	(4) US	(5) Pooled
Absolute negative deviation from punisher's contribution	0.008 (0.015)	0.054*** (0.017)	0.097*** (0.022)	0.084*** (0.022)	0.084*** (0.022)
Positive deviation from punisher's contribution	−0.013** (0.006)	−0.018*** (0.006)	0.023** (0.010)	0.015 (0.011)	0.014 (0.011)
Islamic countries					−0.052 (0.113)
Morocco					0.157 (0.128)
UK					−0.163 (0.113)
Absolute negative deviation × Islamic countries					−0.029 (0.028)
Absolute negative deviation × Morocco					−0.047** (0.022)
Absolute negative deviation × UK					0.012 (0.032)
Positive deviation × Islamic countries					−0.028** (0.012)
Positive deviation × Morocco					−0.004 (0.009)
Positive deviation × UK					0.005 (0.015)
Socio-economic control variables	Yes	Yes	Yes	Yes	Yes
Constant	−2.906*** (1.080)	0.040 (0.521)	−0.865 (0.601)	−0.456 (0.532)	−0.678** (0.310)
R^2	0.11	0.21	0.28	0.15	0.15
N	240	258	264	318	1080

Note. OLS estimates. Robust standard errors in parentheses. Control variables: age, female, urban background, middle class, single child, economics/business student. The following effects are significant at the 5% level. Col. 1: positive for age; negative for urban background, economics/business student. Col. 2: negative for female, single child; positive for urban background, middle class. Col. 4: negative for economics/ business student. * $p < .10$. ** $p < .05$. *** $p < .01$.

Comparing differences across cultural clusters yields no significant level differences in punishment expenditure (Table 3.9, Col. 5). Additionally, the punishment of negative deviations is similar, with an insignificant interaction term 'Absolute negative deviation × Islamic countries' and joint insignificance of the dummy variable for Islamic countries with the interaction term, $F(2, 1062) = 1.22$, $p = .297$. This shows that punishment of free riding is similar when comparing the two English-speaking and the two Islamic countries. However, there are differences in antisocial punishment across cultural clusters. The negative and significant interaction term 'Positive deviation × Islamic countries' indicates lower punishment of contributions that are higher than the contribution level of the punisher.

Within the cultural cluster of Islamic countries, we find considerable variance. Although the insignificant dummy Morocco suggests no level differences in punishment expenditure between Morocco and Turkey, the reaction to negative deviations is significantly lower in Morocco, which is indicated by the negative and significant coefficient for 'Absolut negative deviation × Morocco'. However, the insignificant interaction term 'Positive deviation × Morocco' and the test for joint significance between this interaction term and the dummy variable for Morocco suggest a similar reaction to positive deviations comparing Morocco and Turkey, $F(2, 1062) = 1.32$, $p = .268$.

Investigating differences in punishment expenditures within the cluster of English-speaking countries reveals no level differences, which is implied by the insignificant dummy variable UK. Additionally, the interaction term 'Absolute negative deviation × UK' is insignificant, implying no differences in the punishment response to free riding. A test for joint significance of the two coefficients corroborates this finding, $F(2, 1062) = 1.46$, $p = .234$. Similarly, we find no differences in the reaction in punishment to positive deviations (or antisocial punishment). This is evident from the insignificant interaction term 'Positive deviation × UK' as well as the test for joint significance with the dummy variable UK, $F(2, 1062) = 1.72$, $p = .179$.

3.6.1.7 Cooperative dispositions and punishment

With our dataset we can test whether the engagement in punishment and the severity of punishment depend on the individual cooperative dispositions. We can thus conduct a replication of the analysis conducted in Chapter 2 across societies.

We find that a similar share of CC and FR engage in prosocial punishment as well as antisocial punishment (Figure 3.10). This finding holds for all four societies. In Morocco, no FR punishes. However, the share of CC who punish is not significantly different from zero.

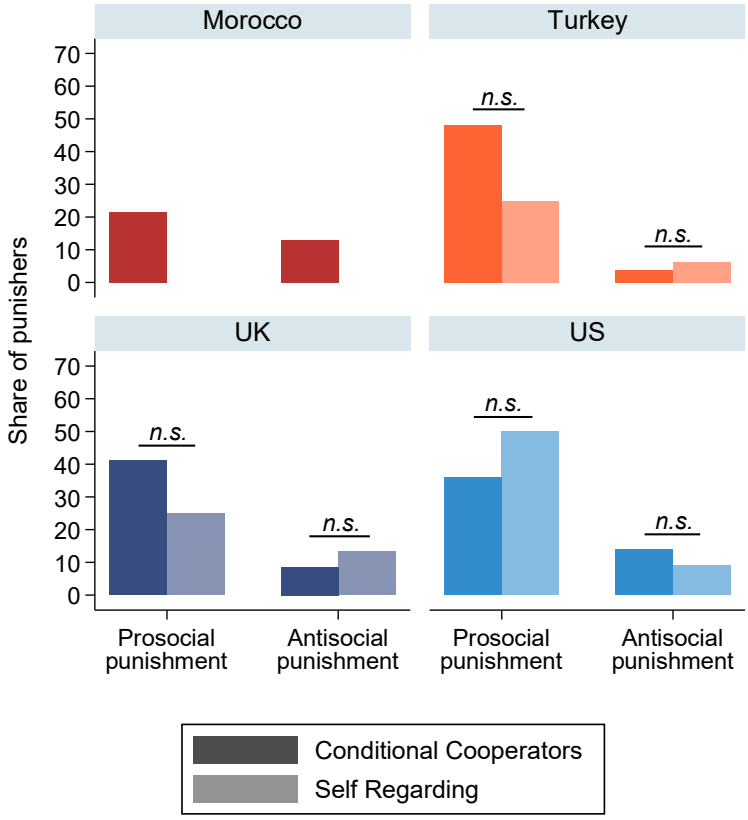


FIGURE 3.10. The share of punishers by cooperative disposition across the four societies. Due to the low number of observations, we do not report the test results for Morocco. χ^2 test: *n.s.* $p \geq .10$.

Additionally, we compare the average punishment expenditure for CC and FR across societies (Figure 3.11). We find a similar expenditure on prosocial and antisocial punishment across the four societies. One exception is the US, with FR making an even higher expenditure on prosocial punishment compared to CC. The punishment expenditure of FR in Morocco is zero because no FR chose to punish. The punishment expenditure of CC is however not significantly different from zero.

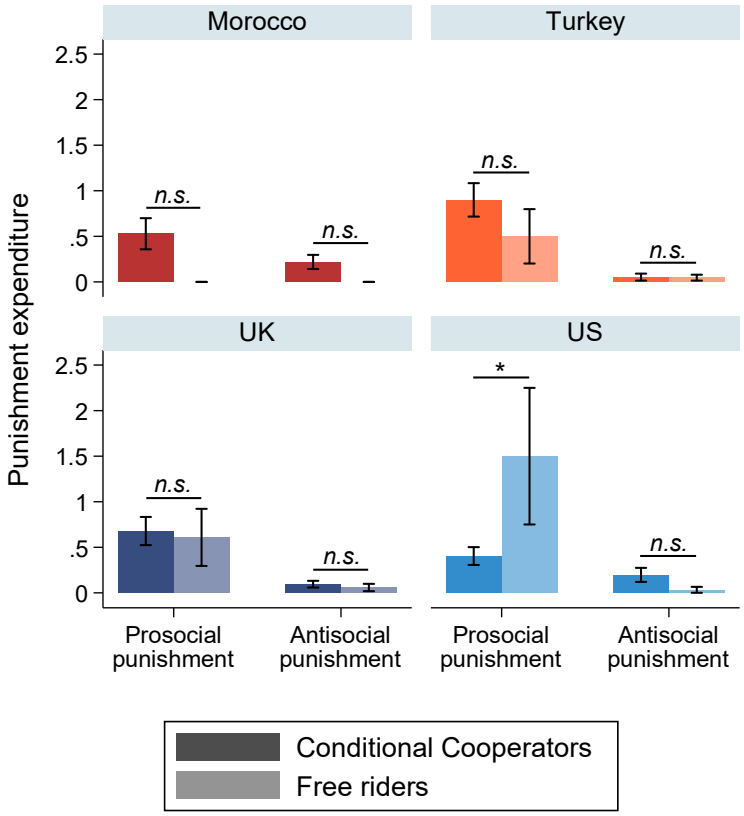


FIGURE 3.11. The average punishment expenditure by cooperative disposition across the four societies. The error bars indicate the bootstrapped 95% CI. Mann-Whitney test: *n.s.* $p \geq .10$; * $p < .10$.

We use a two-stage regression analysis of punishment (see Chapter 2) to test for differences in the punishment decision and the severity of punishment separately (Table 3.10). We run the regressions for Turkey, the UK and the US. We cannot run the regression analysis for Morocco because in Morocco no FR chose to punish. The regression models show some variation across societies but overall confirm the findings of Chapter 2: Engagement in punishment is prevalent in FR and thus does not require a cooperative disposition. For Turkey, the engagement in punishment and the severity are even higher for FR. In the UK, the punishment decision is independent of cooperativeness, but the severity is not. We find a lower level of punishment in FR, but a higher increase of severity for larger negative deviations and deviation of others. In the US, both the punishment decision and the severity are independent of the cooperative disposition.

TABLE 3.10.
Cooperative dispositions, punishment decision and severity of punishment by country.

	Turkey			UK			US		
Dependent variable	(1) Punishment decision	(2) Avg. marg. effect	(3) Punishment severity	(4) Punishment decision	(5) Avg. marg. effect	(6) Punishment severity	(7) Punishment decision	(8) Avg. marg. effect	(9) Punishment severity
Absolute negative deviation from the punisher	0.027	0.005	2.689***	0.149**	0.030**	0.207***	0.494*	0.099*	0.306***
	(0.053)	(0.011)	(0.436)	(0.069)	(0.014)	(0.029)	(0.273)	(0.053)	(0.037)
Deviation of the others' mean contribution from the punisher	0.055	0.011	0.026	0.073**	0.015**	−0.025***	0.061	0.012	−0.003
	(0.045)	(0.010)	(0.091)	(0.031)	(0.006)	(0.007)	(0.040)	(0.008)	(0.042)
Dispositional Free Rider (DFR)	0.101	0.020	3.740***	0.744	0.152	−1.133**	1.142	0.229	2.211
	(0.451)	(0.090)	(0.887)	(0.528)	(0.107)	(0.570)	(0.755)	(0.150)	(1.370)
DFR × Absolute negative deviation	0.154*	0.030*	2.748***	−0.003	−0.001	0.160**	−0.434	−0.087	−0.200
	(0.089)	(0.016)	(0.428)	(0.067)	(0.014)	(0.069)	(0.275)	(0.054)	(0.148)
DFR × Deviation of the others' mean contribution	−0.002	−0.000		0.019	0.004	0.231***	0.017	0.003	0.036
	(0.063)	(0.012)		(0.038)	(0.008)	(0.086)	(0.047)	(0.009)	(0.118)
Constant	−1.685***		−1.409*	−2.205***		0.195	−2.319***		−1.177
	(0.339)		(0.768)	(0.492)		(0.445)	(0.737)		(1.105)
N (Clusters)	168 (21)		30 (11)	201 (23)		36 (16)	216 (30)		33 (21)

Note. Only CC and FR included. Col. 1, 4, 7: Probit coefficients; Col. 2, 5, 8: Average marginal effects of the Probit model; Col. 3, 6, 9: Truncated linear regression. *SE* clustered on groups are given in parentheses. Col. 3 one coefficient omitted due to the lack of variation. * $p < .10$; ** $p < .05$; *** $p < .01$.

3.6.1.8 Regression analysis of anger and guilt

We test for differences in the emotional response to the social dilemma by estimating several ordered Probit regression models. Self-reported anger (Table 3.11) and guilt (Table 3.12) serve as the dependent variables. The independent variables in both regression models contain the absolute negative deviation of a group member's contribution from the subject's contribution who is reporting the emotional response. We also include the absolute positive contribution deviation of a group member from the subject who is self-reporting the emotions and dummy variables for Islamic countries, Morocco and the UK. Moreover, we add interaction terms between the dummy variables and the positive/negative contribution deviation. Finally. We include socio-economic control variables to control for potential differences across subject pools.

We find that the driving factors of anger are similar across countries. In all societies, participants feel angrier for a higher negative deviation of a group member from the own contribution. This link is strongest for the UK and weakest for Morocco (Table 3.11, Col. 1–4).

Additionally, we estimate a pooled model to test for differences between and within cultural clusters (Table 3.11, Col. 5). We find significantly higher anger levels in the Islamic countries compared to the English-speaking countries. Although for all societies anger increases for larger negative deviations, the negative and weakly significant interaction effect 'Absolute negative deviation × Islamic countries' shows that the intensity of the reaction to negative deviations is lower for Islamic countries. We also find significant within-cluster variation for Islamic countries. The negative and highly significant interaction effect 'Absolute negative deviation × Morocco' indicates a smaller increase in anger for a higher negative contribution deviation in Morocco compared to Turkey. The reaction to positive differences is similar across Morocco and Turkey, indicated by an insignificant interaction term and the dummy variable and its interaction term being jointly insignificant, $\chi^2(2) = 0.35$, $p = .839$. Additionally, we find level differences in anger comparing the two English-speaking countries with significantly higher anger levels in the UK compared to the US.

TABLE 3.11.
Explaining self-reported anger through deviation from the punisher's own contribution.

Dependent variable: anger	(1) Morocco	(2) Turkey	(3) UK	(4) US	(5) Pooled
Absolute negative deviation	0.039***	0.103***	0.160***	0.151***	0.148***
	(0.015)	(0.017)	(0.015)	(0.017)	(0.017)
Positive deviation	−0.025	−0.030**	−0.107***	−0.036*	−0.042**
	(0.018)	(0.015)	(0.021)	(0.019)	(0.020)
Islamic countries					0.549***
					(0.162)
Morocco					0.045
					(0.164)
UK					0.267*
					(0.153)
Absolute negative deviation × Islamic countries					−0.042*
					(0.023)
Absolute negative deviation × Morocco					−0.067***
					(0.023)
Absolute negative deviation × UK					−0.029
					(0.020)
Positive deviation × Islamic countries					0.012
					(0.025)
Positive deviation × Morocco					0.007
					(0.023)
Positive deviation × UK					−0.046*
					(0.027)
Socio-economic control variables	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.03	0.10	0.23	0.14	0.11
N	240	258	264	318	1080

Note. Ordered Probit coefficients. Robust standard errors in parentheses. Control variables: age, female, urban background, middle class, single child, economics/ business student. The following effects are significant at the 5% level. Col. 1: negative for middle class. Col. 3: negative for urban, economics/business student. Col. 4: positive for economics/business student. * $p < .10$. ** $p < .05$. *** $p < .01$.

Next, we analyse self-reported guilt using ordered Probit regression models with the same specification as described above (Table 3.12). Col. 1–4 reveal differences in the effect of contribution deviations on guilt. For Morocco, guilt increases in both, positive and negative deviations from another group member. For Turkey, guilt appears independent of the contribution deviation from the other group members. However, the emotions reaction across the two English-speaking countries seems consistent. For both, guilt increases if the other group member contributed more than the reporting subject, with positive and significant coefficients for positive deviations.

We estimate a pooled model to test for differences across cultural clusters and within cultural clusters (Table 3.12, Col. 5). The positive and highly significant dummy variable for Islamic countries reveals higher levels of guilt in the two Islamic countries compared to the two English-speaking countries. The weakly significant interaction terms 'Absolute negative deviation × Islamic countries' indicates higher guilt for negative deviations in the Islamic countries. The insignificant interaction 'Positive deviation × Islamic countries' shows that the reaction to positive deviations are similar across cultural clusters.

Morocco and Turkey are very similar in the self-reported guilt. The dummy variable for Morocco is insignificant as well as the interaction term 'Absolute negative deviation × Morocco'. These two variables are jointly insignificant, $\chi^2(2) = 1.07$, $p = .586$. Similarly, the interaction term 'Positive deviation × Morocco' is insignificant, and we find no joint significance with the dummy variable Morocco, $\chi^2(2) = 1.09$, $p = .581$.

Next, we look at within-cluster differences for the two English-speaking countries. For negative contribution deviations, we find an insignificant interaction term and no joint significance with the dummy variable, $\chi^2(2) = 1.68$, $p = .432$. Although, the dummy variable 'UK' and its interaction term with positive contribution deviations are individually insignificant, we find joint significance of the two coefficients, $\chi^2(2) = 8.33$, $p = .016$.

TABLE 3.12.
Explaining self-reported guilt through deviation from the punisher's own contribution.

Dependent variable: guilt	(1) Morocco	(2) Turkey	(3) UK	(4) US	(5) Pooled
Absolute negative deviation	0.050***	0.019	−0.021	−0.019	−0.020
	(0.016)	(0.016)	(0.019)	(0.019)	(0.019)
Positive deviation	0.031*	0.020	0.109***	0.058***	0.059***
	(0.017)	(0.013)	(0.013)	(0.017)	(0.017)
Islamic countries					0.449**
					(0.175)
Morocco					−0.166
					(0.165)
UK					0.166
					(0.158)
Absolute negative deviation × Islamic countries					0.043*
					(0.025)
Absolute negative deviation × Morocco					0.017
					(0.024)
Absolute negative deviation × UK					−0.000
					(0.025)
Positive deviation × Islamic countries					−0.034
					(0.022)
Positive deviation × Morocco					0.007
					(0.021)
Positive deviation × UK					0.025
					(0.020)
Socio-economic control variables	Yes	Yes	Yes	Yes	Yes
Pseudo R^2	0.05	0.03	0.10	0.04	0.04
N	240	258	264	318	1080

Note. Ordered Probit estimates. Robust standard errors in parentheses. Control variables: age, female, urban background, middle class, single child, economics/ business student. The following effects are significant at the 5% level. Col. 1: negative for age, middle class, economics/business student. Col. 2: negative for age; positive for female, middle class. Col. 3: positive for female. Col. 4: positive for economics/ business student. * $p < .10$. ** $p < .05$. *** $p < .01$.

3.6.2 Instructions

We used the instructions in English as shown below for the experiments run in the UK and the US. For the experiments in Morocco and Turkey, the instructions and experimental software were translated into the local languages.

3.6.2.1 D-Game

You are now taking part in an economic experiment. Depending on the decisions made by you and other participants, you can earn a considerable amount of money. It is therefore very important that you read these instructions with care.

These instructions are solely for your private use. **It is prohibited to communicate with other participants during the experiment.** If you have any questions, please raise your hand. A member of the experiment team will come and answer them in private. If you violate this rule, you will be dismissed from the experiment and you will forfeit all payments.

During the experiment, we will not speak in terms of Pounds, but in Guilders. At the end your entire earnings will be calculated in Guilders. The total amount of Guilders you have earned will be converted to Pounds at the following rate:

1 Guilder = 0.20 Pounds

After this experimental session, your entire earnings from the experiment will be paid to you privately in cash.

At the end of the session, you will be asked to fill in a questionnaire. The answers you provide in this questionnaire are completely anonymous. They will not be revealed to anyone either during the experiment or after it. Furthermore, your responses to the questionnaires will not affect your earnings during the experiment.

The groups

At the beginning of the experiment, all participants will be randomly divided into groups of four. Apart from you, there will be three other members in your group. **You will not learn who the other people in your group are at any point.**

The decision situation

Each participant receives an endowment of **20 tokens**. You have to decide how many of these 20 tokens you will contribute to a group project, and how many you will keep

for yourself. The three other members of your group have to make the same decision. They can also either contribute tokens to the project or keep tokens for themselves. You and the other members of the group can each choose any amount between 0 and 20 tokens to contribute (including 0 and 20).

The payoffs

The income of every member of the group is calculated in the same way. Your income consists of two components:

(1) The first component is the amount of tokens that you keep for yourself. Every token that you do not contribute to the project automatically belongs to you and earns you one Guilder.
(2) The second component is your personal return from the group project. For all of the tokens contributed to the project the following happens: the project's value will be multiplied by 1.6 and this amount will be divided equally among all four members of the group.

For example, if 1 token is contributed to the project, the project's value increases to 1.6 Guilders. This amount is divided equally among all four members of the group. Thus every group member receives 0.4 Guilders.

The following function illustrates your income in Guilders:

$$\textit{Your Total Income} = 20 - \textit{Your Contribution} + 0.4 \times (\textit{Group Project})$$

In order to explain the income calculation we will give some examples. Please read them carefully. At the end of the introductory information, you will be asked to answer several computerised control questions which are designed to check that you have understood the decision situation.

Example 1

If each of the four members of the group contributes 0 tokens to the project, all four will receive an income from their private account of 20. Nobody receives anything from the project, because no one contributed anything. Therefore the total income of every member of the group is 20 Guilders.

Calculation of the total income of every participant: $(20 - 0) + 0.4 \times (0) = 20$

Example 2

If each of the four members of the group contributes 20 tokens, there will be a total of 80 tokens contributed to the project. The income from the private account is 0 for everyone, but each member receives an income from the project of 0.4 × 80 = 32 Guilders.

Calculation of the total income of every participant: (20 − 20) + 0.4 × (80) = 32

Example 3

If you contribute 20 tokens, the second member 10 tokens, the third member 5 and the fourth 0 tokens, the following incomes are calculated:

Because the total contribution to the project is 35 tokens, everyone will receive 0.4 × 35 = 14 Guilders from the project.

You contributed all your 20 tokens to the project. You will therefore receive 14 Guilders in total at the end of the experiment.

The second member of the group also receives 14 Guilders from the project. In addition, she receives 10 Guilders from her private account, because she contributed 10 tokens to the project. Thus, her total income is 24 Guilders altogether.

The third group member receives 14 Guilders from the project as well. Additionally, this group member will receive 15 Guilders from her private account. The total income therefore adds up to 29 Guilders.

The fourth member of the group, who did not contribute anything, also receives the 14 Guilders from the project and additionally the 20 Guilders from the private account, which means her total income is 34 Guilders.

Calculation of your total income: (20 − 20) + 0.4 × (35) = 14
Calculation of the 2nd group member's total income: (20 − 10) + 0.4 × (35) = 24
Calculation of the 3rd group member's total income: (20 − 5) + 0.4 × (35) = 29
Calculation of the 4th group member's total income: (20 − 0) + 0.4 × (35) = 34

Example 4

The three other members of your group contribute 20 tokens each to the project. You do not contribute anything. In this case the incomes will be calculated as follows:

Calculation of your total income: $(20 - 0) + 0.4 \times (60) = 44$

Calculation of the total income of each other group member: $(20 - 20) + 0.4 \times (60) = 24$

The experiment

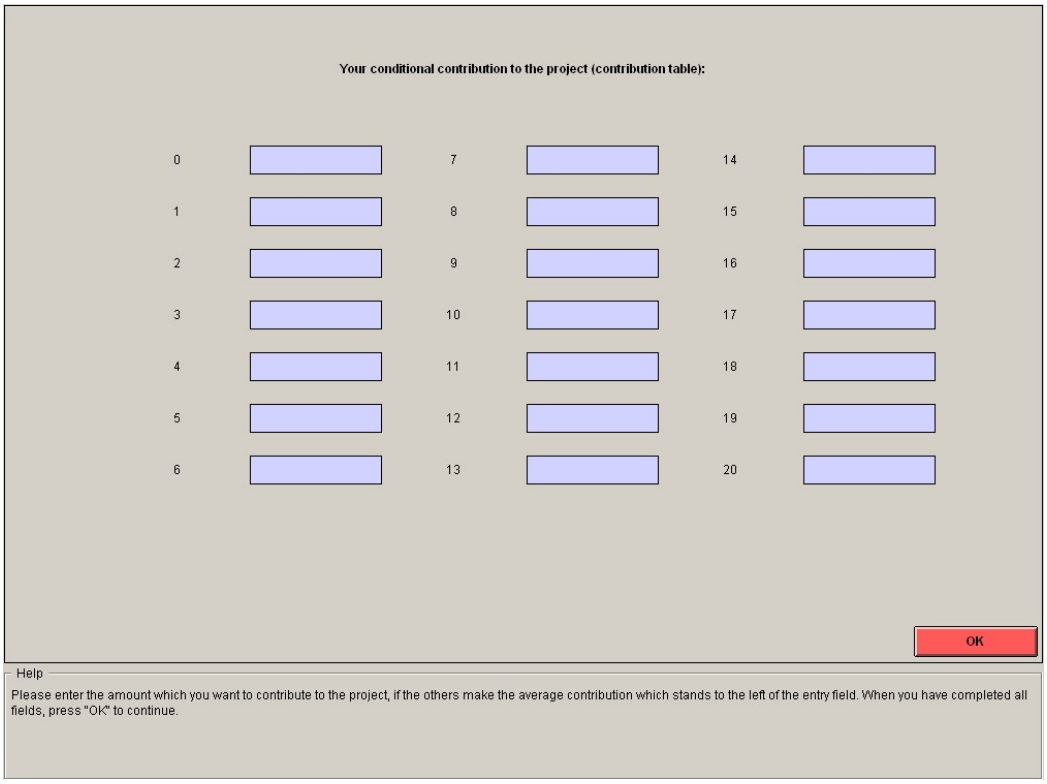
The experiment is based on the decision situation just described to you, conducted **only once**. In this experiment you will make two types of decisions: an **unconditional contribution** and filling in a **contribution table**.

When making your **unconditional contribution**, the following screen will appear:

As mentioned above, your endowment in the experiment is 20 tokens. You have to decide how many tokens you contribute to the project by typing a number between 0 and 20 (including 0 and 20) in the box. This box can be reached by clicking on it with the mouse. By deciding how many tokens to contribute to the project, you automatically decide how many tokens you keep for yourself. After entering the amount of tokens

you want to contribute you must click on the "*OK*" button. Once you have done this, your decision can no longer be revised.

Your second task is to fill in a **contribution table** on the following screen:



The contribution table indicates **how many tokens you want to contribute to the project for each possible average contribution of the other group members** (rounded to the nearest integer). The table allows for conditioning your contribution on that of the other group members.

The numbers to the left of the input fields are the possible average contributions of the **other** group members (rounded to the nearest integer). You have to enter how many tokens you want to contribute to the project, conditional on the indicated average contribution of the other group members. **You must enter a number between 0 and 20 (including 0 and 20) into each box.**

For example, in the first box you enter the amount of tokens you want to contribute to the project in case the average contribution to the project of the other three group members is 0 tokens. In the next boxes you enter how much you contribute for an

average contribution of 1, 2, 3, … tokens. After entering your decisions, you must click on the "*OK*" button.

After all participants of the experiment have made an unconditional contribution and have filled their contribution table, a random mechanism will select one member from every group. For **this group member, the contribution table** will be used to determine the contribution to the project. Whereas for **the other three group members, their unconditional contributions** will define the amount of tokens they add to the project.

You will not know whom the random mechanism will select before you make your unconditional contribution and fill in the contribution table. Therefore you must think carefully about both decisions. Either of them could determine your actual contribution to the project.

Example 5

Suppose that the **random mechanism selects you**; and that the other three group members made unconditional contributions of 0, 2, and 4 tokens, respectively. The average contribution of these three group members is, therefore, 2 tokens. If you indicated in your contribution table that you will contribute 1 token if the others contribute 2 tokens on average, then the total contribution to the project is given by 0 + 2 + 4 + 1 = 7 tokens. Each group member would, therefore, earn 0.4 × 7 = 2.8 Guilders from the project plus their respective income from their own private account. If, instead, you indicated in your contribution table that you would contribute 19 tokens if the others contribute 2 tokens on average, then the total contribution of the group to the project would be given by 0 + 2 + 4 + 19 = 25 tokens. Each group member would earn 0.4 × 25 = 10 Guilders from the project plus their respective income from their own private account.

Example 6

Suppose that the **random mechanism does not select you**; and that your unconditional contribution is 16 tokens, while those of the other two group members not selected by the random mechanism are 18 and 20 tokens respectively. Your average unconditional contribution and that of these two other group members is, therefore, 18 tokens. If the group member whom the random mechanism did select indicates in her contribution table that she will contribute 1 token if the other three group members contribute on average 18 tokens, then the total contribution of the group to the project is given by 16

+ 18 + 20 + 1 = 55 tokens. Each group member will therefore earn 0.4 × 55 = 22 Guilders from the project plus their respective income from their own private account. If, instead, the randomly selected group member indicates in her contribution table that she contributes 19 if the others contribute on average 18 tokens, then the total contribution of the group to the project is 16 + 18 + 20 + 19 = 73 tokens. Each group member would therefore earn 0.4 × 73 = 29.2 Guilders from the project plus their respective income from their own private account.

The random mechanism

Each group member is assigned a Group Member ID between 1 and 4, which denotes this participant's number inside her group. Moreover, participant number 2 was randomly selected at the very beginning of the experiment. This participant will draw a ball from an urn after all participants have made their unconditional contribution and have filled out their contribution table. Each ball in the urn has a different colour and each colour corresponds to a Group Member ID: orange = 1, blue = 2, yellow = 3, green = 4. The resulting number will be entered into the computer. If your Group Member ID is drawn, then your contribution table will determine your contribution to the project. For all other members of your group, the unconditional contributions will be relevant. Otherwise, your unconditional contribution determines your contribution.

If you have any questions, please raise your hand and a member of the experiment team will come and answer them in private.

3.6.2.2 P-Game

You are now taking part in a second experiment. Your payoff from this experiment is completely unrelated to the decisions you have made in the previous one. The money you earn in this experiment will be added to what you earned in the first experiment. As before the Guilders you have earned will be converted to Pounds at the following rate:

1 Guilder = 0.20 Pounds

As in the previous experiment, all participants will be randomly divided into groups of four. However, the composition of the group is entirely new. **You will not learn who the other people in your group are at any point.**

The decision situation

The decision situation is the same as the one described on the first instruction sheet: Each participant receives an endowment of **20 tokens**. You have to decide how many of these 20 tokens you contribute to a group project and how many you keep for yourself. The three other members of your group have to make the same decision. However, this time you will make only an unconditional contribution to the project. There will be no contribution table.

After the contribution decision, there will be a **second stage**. At this stage, you will see how many tokens each of the other three group members has contributed to the project and their corresponding income from this contribution decision. Nonetheless, the identities of your group members will not be revealed at any stage. You can either **decrease** or **leave unchanged** the income of each other group member by assigning **deduction points** to them. The other group members can also decrease your income, by allocating deduction points to you, if they wish to do so.

Deduction points

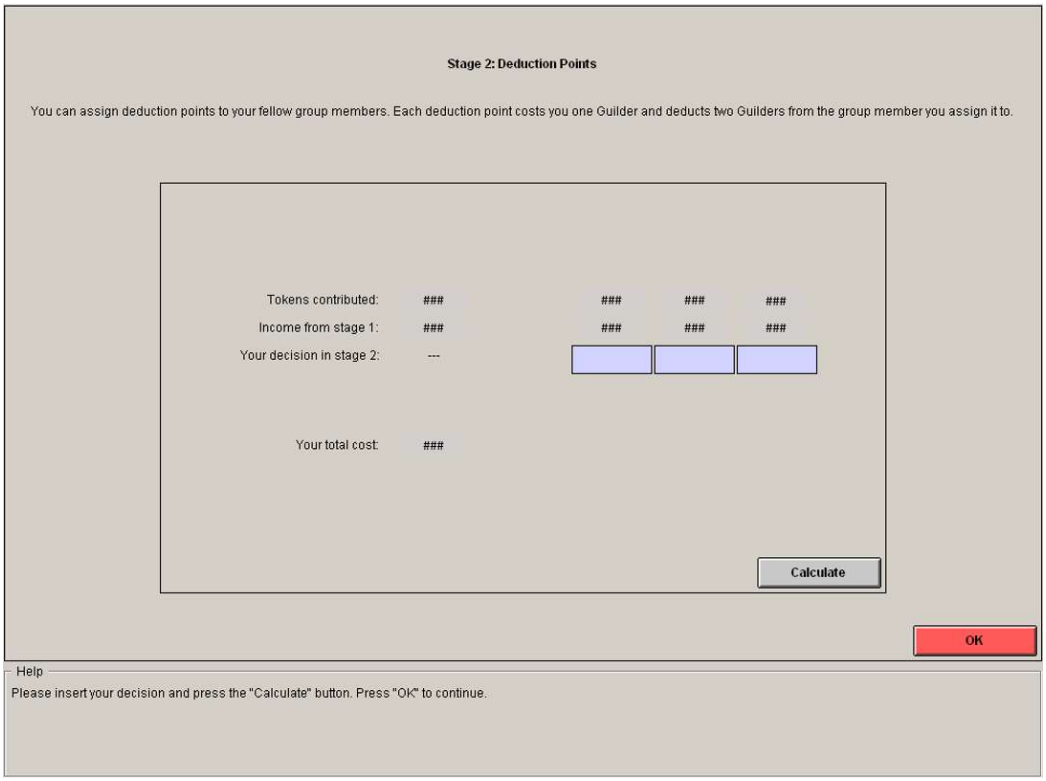
In stage 2, you can assign **between 0 and 5 deduction points to each other group member.** The maximum number of deduction points, you can allocate to the other group members together is therefore 15 deduction points.

For each deduction point that you assign, there is a cost to you of one Guilder. Thus, the total cost to you in Guilders of assigning deduction points to other group members is given by the total number of deduction points that you assign.

For each deduction point that you assign to a particular group member, you will decrease their income by 2 Guilders unless their income is already exhausted. For example, if you give a group member 2 deduction points, you will decrease this group member's income by 4 Guilders.

Your **own income will be reduced by 2 Guilders for each deduction point that is assigned to you** by the other three group members. If all of your income from the first stage of this experiment is exhausted, it cannot be reduced any further by other group members.

You will see the following screen at stage 2:



The column on the left shows your contribution and your income from the first stage. The other three columns indicate the contribution of your group members and their income from the first stage.

If you do not wish to change the income of the other group members, type "*0*" into the fields next to *"Your decision in stage 2."* In case you want to assign deduction points, enter the number of deduction points you want to assign into this field. You must enter a decision into every field and press the "*Calculate"* button. This will display the cost of your decision. Until you press the "*OK*" button, you can still change your decision. To recalculate the costs after making a change, simply press the "*Calculate*" button again.

The payoffs

Your total income in Guilders from the two stages will be calculated as follows:

Your Income From Stage 1 = 20 − *Your Contribution* + 0.4 × (*Group Project*)

Total Income After Stage 2 = Income From Stage 1 **(1)**

− 2 × (*Sum Of Deduction Points Assigned To You*) **(2)**

− (*Deduction Points Assigned By You*)

if (1) + (2) is greater or equal to 0.

Total Income After Stage 2 = 0 − (Deduction Points Assigned By You)

if (1) + (2) is less than 0.

Please note that your income in Guilders after stage 2 can be negative, if the cost of deduction points assigned by you exceeds your income from stage 1 less any reduction in your income caused by other group members.

However, at the end of the experiment and in addition to the calculation just given, you and the other members of your group will each receive a lump sum payment of **10 Guilders.** This payment is to cover losses that you could incur.

If you have any questions, please raise your hand and a member of the experiment team will come and answer them in private.

CHAPTER 4

Sustaining Cooperation: A Comparative Evaluation of Cooperative Dispositions, Peer Pressure and Formal Punishment[5]

4.1 Introduction

In 2014, the German football manager Uli Hoeneß was found guilty of evading €27.2m in taxes (Oltermann 2014, 13 March). This high-profile case of tax evasion spurred a public debate about whether the wealthy contribute their fair share to fund public goods. The *formal* punishment was a prison sentence of 3.5 years. However, *informal* punishment also played a big role, taking the form of the negative image and social stigma attached (Schöneberg 2014, 13 March). For example, even the German chancellor, Angela Merkel, publicly expressed her disappointment in Uli Hoeneß (Medick 2013, 23 April).

The decision to pay taxes—when a profitable opportunity for tax evasion exists—is an example of a social dilemma prevalent in many societies. People who are willing to pay their taxes engage in costly cooperation for the benefit of the whole society. Other examples of social dilemmas include restricting the exploitation of common resources or the costly avoidance of pollution (Hardin 1968). The example above also illustrates that different sanctioning institutions are simultaneously present in a society, all affecting contribution decisions. We therefore expect three factors to impact the levels of compliance and cooperation in a society. These are cooperative dispositions, informal sanctioning mechanism and formal sanctioning mechanisms.

A *cooperative disposition* is an intrinsic inclination to cooperate, which is associated with high contributions to public goods (Fischbacher et al. 2001, Fischbacher et al. 2012). Although subjects with an intrinsic disposition for cooperation

[5] This chapter draws on joint work in progress with Simon Gächter and Ori Weisel.

make higher contributions initially, their presence is not sufficient to prevent the decline of cooperation in repeated games (Fischbacher and Gächter 2010).

Informal sanctioning mechanisms like peer pressure, gossip or shaming can form incentives to cooperate when targeted at free riders. An important feature of informal sanctions is their decentralised structure, relying on each individual to assess rule transgressions and punish the behaviour of others. Thus, this type of sanctions is often called 'peer-punishment' in the literature. Although engagement in informal sanctions or peer-punishment is usually costly for the punisher (there is always a risk of retribution when confronting transgressors), a large share of people are willing to bear this burden and punish non-cooperators (e.g., Fehr and Gächter 2000, Fehr and Gächter 2002).

Formal sanctioning mechanisms, like the police and court system, also create costs for non-cooperators and thus change the payoff structure of the decision situation. Compared to informal sanctions, formal sanctions are costly to establish (i.e., taxes need to be paid to hire police officers or to fund courts). Additionally, formal sanctions usually follow structured and centralised mechanisms, assessing behaviour and allocating punishment according to predefined and well-known rules (e.g., Yamagishi 1986, Baldassarri and Grossman 2011, Andreoni and Gee 2012). Rather than allowing for direct punishment of others, group members make a contribution to a punishment mechanism or punishment pool, which in turn punishes defectors. Accordingly, such sanctioning mechanisms are often referred to as 'pool-punishment'.

Similar to the tax evasion example above, most cooperation problems outside the controlled laboratory environment include all three factors which simultaneously influence cooperation levels. Although one might argue that most sanctions in modern societies are carried out by formal institutions, it is important to acknowledge that this does not preclude informal punishment. Even if the police and court system are well funded and thus constitute a deterrent, the possibility of peer-punishment can be present and still influence behaviour.

The present study is motivated by this inability to separate the relative influence of these sanctioning mechanism outside the laboratory. Our aim is to analyse the effectiveness and efficiency of all three factors described above in fostering

cooperation. To this end, we employ an experimental design that allows us to separate and measure the influence of each sanctioning mechanism independently.

We explore two main research questions: First, what are the characteristic differences in the effectiveness and efficiency of formal and informal sanctioning mechanisms in fostering cooperative behaviour? Second, what is the role of cooperative dispositions in the emergence and maintenance of formal and informal sanctioning mechanisms?

Our experiment comprises four treatments, which consist of repeated public goods games with different punishment mechanisms: a baseline treatment without punishment and three treatments with peer-punishment, pool-punishment and a combination of both, respectively. We find that peer-punishment induces high and stable cooperation levels. Pool-punishment is least efficient in terms of social welfare and cooperation levels were highly volatile over time. A best-reply analysis reveals that it also crowds out voluntary contributions. The combined treatment, with both peer- and pool-punishment does not differ significantly from peer-punishment alone in terms of cooperation and efficiency levels. Additionally, subjects with a cooperative disposition have a higher likelihood of engaging in pool-punishment compared to subjects with a free-rider disposition.

4.2 Related literature and our contribution

We are not the first to study different sanctioning institutions in the laboratory as a mean to foster high cooperation levels in social dilemma games.[6] Most studies that compare informal sanctioning institutions implement a version of peer-punishment similar to that proposed by Fehr and Gächter (2000). Public goods games with peer-punishment usually comprise two stages. First, subjects decided on the contribution to the public good. Second, they are informed about the contribution levels of each group member and can then decide on punishment for each group member individually. The cost effectiveness of peer-punishment (i.e., the factor by which punishment reduces the targeted group member's income) is an important feature, which determines the

[6] Here we focus on second-order punishment only, which is the punishment decision taken by people negatively affected by a defector. For a discussion of third-party punishment, see Fehr and Fischbacher (2004).

sanctioning mechanism's effectiveness in fostering high cooperation rates (Nikiforakis and Normann 2008).

The implementation of formal sanctioning institutions in the laboratory is far less coherent with various mechanisms, costs and efficiency levels. 'Pool-punishment' was introduced by Yamagishi (1986) and allowed subjects to fund a punishment pool in addition to the decision about their contribution to the public good. The lowest contributor in the group then received a payoff deduction equal to the contributions to the punishment pool (or twice the contributions to the punishment pool in a 'high-sanctioning' condition).

A variation of the previous mechanism is 'fixed cost pool-punishment' (Traulsen et al. 2012, Zhang et al. 2014). Here, in addition to the cooperation decision, subjects can choose to pay a set contribution to the punishment pool, which in turn allocates a fixed fine to each defector. Thus, the cost effectiveness of punishment varies depending on the number of defectors. Punishment is relatively cheap for a large number of free riders and relatively expensive if only a few defect.

The challenge of second-order free riding, which is relying on others to exert costly punishment of defectors, can be addressed with 'institutional sanctions'. Markussen et al. (2014) and Kamei et al. (2015) conduct a repeated public goods game with a majority voting rule to determine the sanctioning mechanism, which can be either no sanctions, peer-punishment or institutional sanctions. For the latter, every subject pays a fixed tax (without a possibility of evasion). The fine is then applied as a proportion of the income retained and not contributed to the public good.

Andreoni and Gee (2012) compare different sanctioning mechanisms in a public goods game including the 'gun for hire' mechanism. This mechanism targets the lowest contributor in a cooperation game by imposing a fine that just makes this person worse off than the second-lowest contributor. Andreoni and Gee (2012) show experimentally that this mechanism drives up cooperation levels to full contributions. Additionally, subjects are willing to pay to establish this sanctioning mechanism which also leads to higher efficiency compared to peer-punishment. Furthermore, in a combined treatment that makes available both mechanisms—the informal and formal sanctions—the formal sanctioning mechanism crowds out peer-punishment. However, a drawback of this design is that the formal sanctioning mechanism does not hold the fine-to-fee ratio

constant and equal to that of peer-punishment, making a direct comparison to peer-punishment impossible.

A number of studies experimentally tested the emergence of different sanctioning institutions through voting (Traulsen et al. 2012, Markussen et al. 2014, Zhang et al. 2014, Kamei et al. 2015). For these studies, a comparison of the effectiveness and efficiency of sanctioning institutions is not possible because of the endogenous selection of institutions. For example, more cooperative people might self-select or vote for harsher sanctioning institutions, thus increasing average cooperation.

The contribution of the present study lies in ensuring the comparability of sanctioning mechanisms in the laboratory. We use a between-subject paradigm to allow for a comparison between treatments precluding any self-selection into punishment mechanisms. Another novelty in our experimental design is keeping the effectiveness of punishment constant across punishment mechanisms. Furthermore, the present study is the first to investigate the effect of combining informal and formal punishment mechanisms, a setting which is closest to real-world sanctioning institutions. Additionally, we investigate the link between the individual cooperative disposition and the engagement in the different punishment mechanisms.

4.3 Methods

4.3.1 Participants and procedures

We recruited a total of 325 students at the University of Nottingham to participate in our experiments (211 females, mean age = 20.91, *SD* = 2.77) using ORSEE (Greiner 2015). Students were enrolled in a variety of subjects (28% Natural Sciences and Engineering, 22% Economics and Business studies, 20% Law, Social and Political Sciences, 17% Humanities and 12% Medical Sciences). The experiment consisted of two parts, an online experiment and laboratory sessions. The online experiment was conducted using Qualtrics and participants received the access information to the online experiment two days before the laboratory session via email. Subjects were informed that they had to complete the online experiment before attending the laboratory session and were excluded from the experiment if they failed to comply with this requirement. The laboratory experiment was computerised and run with z-Tree (Fischbacher 2007). Each experimental session lasted for approximately two hours, and participants

received their total payoffs (from the online and the laboratory experiment) in private at the end of laboratory session (mean payoff = £23.88, *SD* = £4.65).

4.3.2 Experimental games and treatments

The online experiment elicited the subjects' individual cooperative disposition using an incentivised experiment introduced by Fischbacher et al. (2001). Each laboratory session consisted of two parts, Phase 1 and Phase 2. Participants only learned about Phase 2 once Phase 1 was concluded. After the experiment, subjects were asked to fill in a socio-economic background questionnaire. We conducted four treatments between subjects, varying the game played in Phase 2. Phase 1 was identical in every treatment (Table 4.1).

TABLE 4.1.
Sequence and treatments.

Treatments	Online	Laboratory		*N* (Groups)
		Phase 1: 10 Periods	Phase 2: 30 Periods	
No Punishment	Elicitation of cooperative disposition	No Punishment	No Punishment	85 (17)
Peer			Informal sanctions	80 (16)
Pool			Formal sanctions	80 (16)
Pool+Peer			Formal & Informal	80 (16)

We chose to implement a between-subjects design with four treatments to allow for an unbiased measurement of the punishment mechanisms' effectiveness and efficiency. No Punishment serves as a baseline and excludes any form of informal or formal sanctions. In this treatment, cooperation is therefore only affected by cooperative dispositions. Peer introduces informal sanctions only. Similarly, Pool only includes formal sanctions. Pool+Peer is closest to the real world in allowing both formal and informal sanctions. We can therefore test whether combining the two punishment mechanisms has an additional effect on their effectiveness and efficiency, or whether one mechanism is favoured over another.

The core of our experimental investigation in Phase 2 was a repeated binary public goods game played in groups of five. We chose groups of five to allow for within-group variation in cooperative behaviour. Group allocations were determined

randomly at the start of each laboratory session and the groups were fixed throughout Phase 1 and Phase 2 of the laboratory session.

Phase 1 consisted of a public goods game without punishment repeated for 10 periods. This allowed participants to familiarise themselves with the public good problem and the social dilemma situation. At the start of each of the ten periods, participants received an endowment of 20 tokens and decided to contribute either 0 or 20 tokens to a common project. We restricted participants' actions to binary contributions (either zero or full contributions) to make a clear distinction between free riding and cooperative behaviour. Then, participants indicated their belief about how many of the other four group members would contribute to the common project. All contributions to the project were multiplied by two and equally split amongst all group members. Thus, the social optimum is characterised by full contributions, whereas the individually money-maximising strategy is to contribute nothing. The individual payoff π_i of subject i is in a single period is given by (4.1). Subject i's contribution to the common project is indicated by $g_i \in \{0, 20\}$.

$$\pi_i = 20 - g_i + 0.4 \times \sum_{j=1}^{5} g_j. \tag{4.1}$$

In Phase 2, participants played different variants of a public goods game repeated for 30 periods. The group allocation was fixed throughout the whole laboratory session. We implemented four different treatments varying the sanctioning mechanisms available. *No Punishment* served as a baseline treatment and did not include any punishment mechanism. Thus, the game played in No Punishment was identical to Phase 1 and the individual payoff is given by (4.1).

The three other treatments (Peer, Pool, Pool+Peer) included a punishment mechanism. Subjects could therefore decide whether or not to contribute to the common project and additionally could allocate punishment points. Subjects could buy up to 12 punishment points to sanction group members. Apart from the maximum number of punishment points a group member could buy, we also kept the punishment effectiveness constant across the three treatments. One punishment point cost the punisher one money unit (MU) and reduced the income of the punished subject by two MU. Although we implemented the identical punishment effectiveness across the treatments, the mechanism for allocating the punishment points differed.

4.3.3 Informal and formal punishment

In this section, we discuss the differences in the punishment mechanisms implemented in the laboratory. We focus on the timing, mechanics of funding the mechanisms and the allocation of punishment. Based on these differences, we derive hypotheses concerning the mechanisms' effectiveness and efficiency.

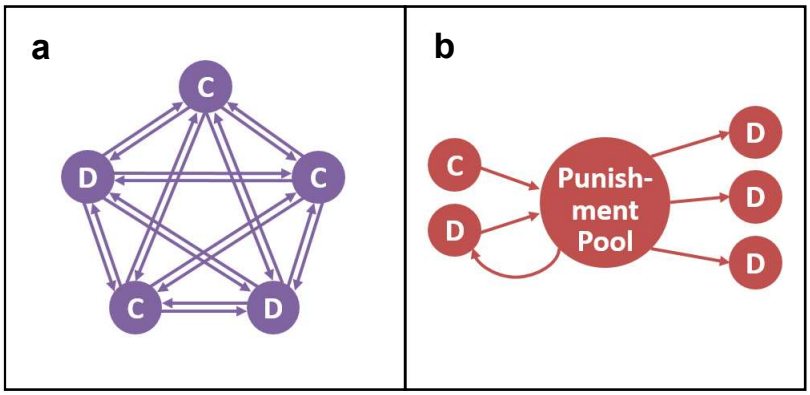


FIGURE 4.1. A schematic illustration of the peer- and pool-punishment mechanism. The five small circles indicate players who can choose to either cooperate (C) or defect (D) in a public goods game. **a** For peer-punishment, players choose whether to cooperate or defect, receive information on the behaviour of their group members and then decide whether or not to punish each of their group members. **b** For pool-punishment, players first decide how much to contribute to the punishment pool, then receive information on the total amount of punishment available and decide whether or not to contribute to the public good. Then, the mechanism allocates the available punishment equally to the defectors, and players receive information on their group members' contributions.

Peer-punishment (Figure 4.1a) captures the vigilante justice and the enforcement of social norms by members of the public and is inspired by Fehr and Gächter (2000). We use a monetary equivalent for peer-punishment that models gossip and stigmatisation to make this punishment mechanism comparable with formal sanctions and to be able to set the effectiveness of punishment. The sequence of the peer-punishment mechanisms is as follows: First, subjects decide on contributing to the public good. They are then informed about the contribution decisions of their fellow group members. Subjects decide to allocate punishment points to each of their group members separately and can buy up to 12 punishment points in total. Each MU spent

on punishment reduces the payoff of the subject it is allocated to by two MU. Subjects are then asked to indicate the total number of punishment points they expect to receive from their group members for the case they defected or contributed. This is followed by feedback on each group member's contribution to the public good as and the own earnings in the current period.

This mechanism is unstructured and allows for both prosocial and antisocial punishment (Herrmann et al. 2008). Thus, subjects can punish all of their group members independent of their contribution to the public good. This mechanism creates uncertainty about the magnitude and the probability of punishment when subjects take the decision on whether or not to contribute to the public good.

To be more precise, let $g_i \in \{0, 20\}$ be subject i's contribution to the public good and $p_{ij} \in \{0 \,..\, 12\}$ the number of punishment points which subject i allocates to another group member j. p_{ji} indicates the number of punishment points which subject i receives from j. Equation (4.2) shows subject i's individual payoff in a single period.

$$\pi_i = 20 - g_i + 0.4 \times \sum_{j=1}^{5} g_j - \sum_{j=1}^{4} p_{ij} - 2 \times \sum_{j=1}^{4} p_{ji}. \tag{4.2}$$

Drawing on previous studies allows to formulate a number of hypothesis about the performance of peer-punishment. Peer-punishment has been studied extensively and shown to induce high cooperation rates (Chaudhuri 2011) and high levels of efficiency in the long run (Gächter et al. 2008). Yet, due to the unstructured nature of peer-punishment, engagement in antisocial punishment can destroy the cooperation enhancing effect of peer-punishment and curb efficiency (Herrmann et al. 2008). Since the peer-punishment decision is made right after learning the behaviour of others, it might be an emotionally loaded or 'hot' decision (Hopfensitz and Reuben 2009). This suggests a high engagement in peer-punishment compared to other punishment mechanisms. However, once a credible threat of peer-punishment is established, it can deter free riding without further punishment expenditure (Gächter 2012). This would suggest a lower engagement in and expenditure on peer-punishment once a credible threat of punishment is established.

Pool-punishment (Figure 4.1b) resembles the decision to fund formal law enforcement institutions, which sanction ex-ante defined behaviour, and is inspired by Yamagishi (1986). The sequence of the pool-punishment mechanism is as follows:

Subjects decide how many punishment points they contribute to the punishment pool. Like in Peer, subjects can buy up to 12 punishment points, with each punishment point yielding a deduction of two MU from the punished group member's payoff. Afterwards, subjects are informed about the total punishment available through the punishment pool and then make the decision whether or not to contribute to the public good. The punishment is then equally divided amongst all defectors. Subjects then learn the contributions of their fellow group members as well as their own payoff from this period.

We implement an idealised version of this institution for which the players' actions are fully observable and only defectors are punished. Defectors receive the expected amount of punishment rather than a probability of being punished to not confound behaviour with risk preferences. A crucial feature of pool-punishment is that it is costly to establish even if there are no defectors around who ought to be punished. For example, the wages of police officer have to be paid even if no crimes are committed. Additionally, in lawful societies the level of punishment of a formal sanctioning mechanism is approximately known even before the contribution decision is made. This feature allows for a cost-benefit calculation when making the contribution decision. In this respect, our pool-punishment mechanism in the laboratory mirrors a crucial feature of law enforcement in the real world. In societies governed by the rule of law, fines and sentences for different crimes are codified and thus the punishment severity is public knowledge.

If a subject contributes to the public good, then she cannot be targeted by pool-punishment. Her income might be reduced by her own punishment expenditure. More precisely, let $p_i^P \in \{0\,..\,12\}$ be subject i's contribution to the punishment pool. If $g_i = 20$, subject i's payoff in a single period is given by (4.3).

$$\pi_i = 20 - g_i + 0.4 \times \sum_{j=1}^{5} g_j - p_i^P. \tag{4.3}$$

If a subject defects, then the payoff reduction from punishment depends on the total contributions to the punishment pool as well as the number of free riders in the group. Let n^{FR} be the total number of free riders in the group. If $g_i = 0$, subject i's payoff in a single period is given by (4.4).

$$\pi_i = 20 - g_i + 0.4 \times \sum_{j=1}^{5} g_j - p_i^P - \frac{2}{n^{FR}} \sum_{j=1}^{5} p_j^P . \quad (4.4)$$

We expect the performance of pool-punishment to be influenced by the costs of punishment. In contrast to peer-punishment, pool-punishment cannot rely on an underlying threat of punishment when people make the contribution decision to the public good. Compared to Andreoni and Gee (2012)'s formal sanctioning mechanism, which is established for a fixed cost and can conduct severe punishment, our pool-punishment uses a fixed punishment effectiveness (i.e., one punishment point costs the punisher one MU and deduces two MU of the punished group member's payoff) and thus strong sanctions are relatively costly. Consequently, the expenditure on pool-punishment is expected to be relatively high. Furthermore, pool-punishment exclusively sanctions defectors and does not allow for any antisocial punishment. This feature can greatly enhance the effectiveness of punishment in fostering high cooperation levels. The pool-punishment decision can be seen as a 'cold' decision which is not driven by affect because this decision is taken before learning the behaviour of others. Thus, punishment expenditure might be lower compared to peer-punishment and more prone to a cost-benefit calculation.

Additionally, we implement a treatment that combines formal and informal sanctioning mechanisms (Pool+Peer). This treatment is closest to the real world where—although formal sanctioning mechanisms exist—informal sanctions are always possible. This is reflected in the example of tax evasion above for which, in addition to the prison sentence, public shaming and stigmatisation are possible. The sequence follows that of Pool with an additional peer-punishment stage: After being informed of their group members' contributions to the punishment pool and common project, each subject can allocate the remaining punishment points directly to their group members.

Next, we discuss our hypothesis concerning the link between individual cooperative dispositions and engagement in costly punishment. Individual cooperativeness might be related to punishment since 'a sanctioning system is also a public good because its benefits can be enjoyed by all members regardless of their contribution to its provision' (Yamagishi 1986, p. 110). This suggests that people with a more cooperative disposition would contribute to the punishment pool, which can be rationalised by the theory of strong reciprocity (Bowles and Gintis 2011). In contrast, Chapter 2 found no difference in the peer-punishment behaviour of cooperative and

self-regarding individuals. This suggests that the link between individual cooperativeness and the engagement punishment might be dependent on the design of the punishment mechanism.

4.4 Results

4.4.1 Cooperation and punishment

First, we investigate the dynamics of cooperation and punishment over time (Figure 4.2). The game played in Phase 1 is identical across all four treatments and punishment is not available. In line with previous studies (e.g., Fischbacher and Gächter 2010), Phase 1 is characterised by a sharp decline in cooperation across all four treatments. Very different cooperation dynamics emerge in Phase 2, in which we implemented different sanctioning mechanisms. In our baseline treatment, No Punishment, we find a decline in the cooperation rate over time. In Peer and Pool+Peer, cooperation rates are high and stable over time suggesting a large effect of the punishment mechanism. In contrast, cooperation rates in Pool remain constant on a lower level.

We estimate Probit regressions to further investigate differences in the contribution behaviour in Phase 1 (Table 4.2, Col. 1, 2). The dependent variable is a contribution dummy. The independent variables are the period number and dummy variables for the Peer, the Pool and the Pool+Peer treatments. Additionally, we include interaction terms for the treatment dummies with the period number. The negative and highly significant period effect in conjunction with the insignificant interaction terms indicates a decreasing probability to contribute over time for all treatments. The positive and weakly significant dummy variables for Pool and Pool+Peer show that the initial cooperation levels are higher in these two treatments compared to the baseline treatment without punishment.

The contribution behaviour in Phase 2 is characterised by more heterogeneity across treatments (Table 4.2, Col. 3, 4). The positive and highly significant coefficient for Pool+Peer shows that the initial probability to contribute is significantly higher compared to the baseline treatment without punishment. The insignificant coefficients for Peer and Pool show that the intercept for those treatments is similar to that of No Punishment.

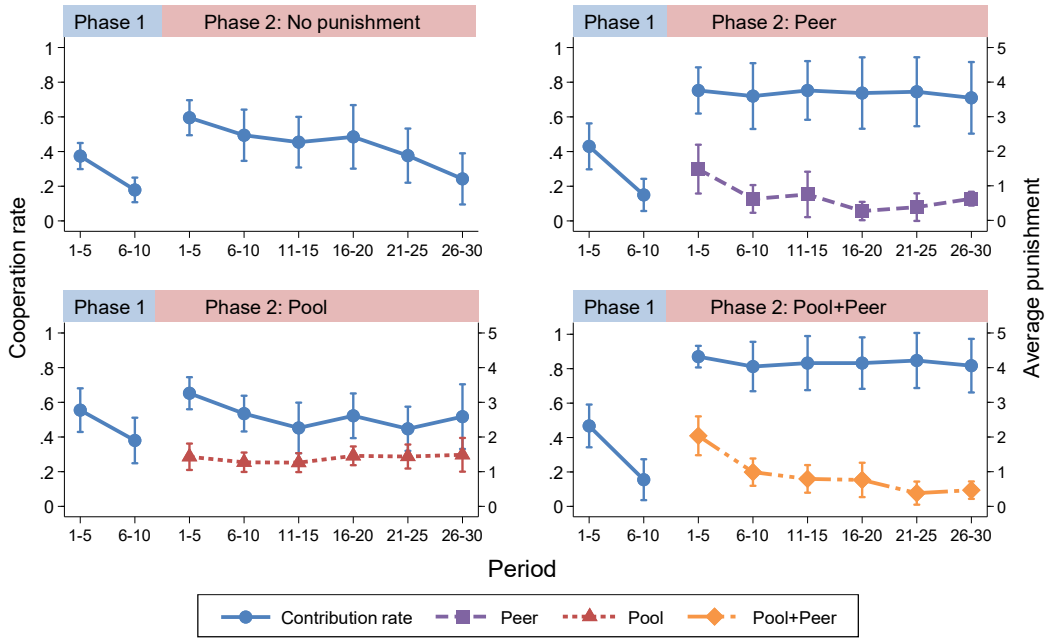


FIGURE 4.2. Cooperation rates and punishment expenditure differ across the mechanisms. Phase 1 is characterised by a decline in cooperation. In Phase 2, cooperation is high and stable in the two treatments including peer-punishment. Cooperation declines when peer pressure is excluded. Average punishment declines when peer-punishment is available, but remains at a high level for the treatment in which only pool-punishment was permitted. The markers show the average across five periods. The error bars indicate the bootstrapped 95% CI. Group averages as independent observations.

Next, we investigate differences in contribution behaviour over time. The negative and highly significant period effect shows that the probability to contribute declines in the comparison treatment without punishment. For Peer and Pool+Peer the interactions between 'Period' and the respective treatment dummies are positive and highly significant and for Pool the interaction is positive and weakly significant. This indicates a less pronounced decline in cooperation rates for all three treatments with punishment. The average marginal effect 'Peer × Period' has a similar absolute size compared to that of 'Period', $\chi^2(1) = 0.24$, $p = .625$. The same holds true for 'Pool × Period', $\chi^2(1) = 2.62$, $p = .105$, and 'Pool+Peer × Period', $\chi^2(1) = 0.21$, $p = .644$. This shows that the contribution rates remain relatively stable over time for the three treatments.

TABLE 4.2.
Regression analysis of contribution behaviour.

Dependent variable: contribution dummy	Phase 1: No punishment		Phase 2: Treatments	
	(1) Probit coefficient	(2) Avg. marg. effect	(3) Probit coefficient	(4) Avg. marg. effect
Peer	0.141	0.046	0.339	0.115
	(0.222)	(0.073)	(0.292)	(0.095)
Pool	0.360*	0.118*	−0.092	−0.031
	(0.217)	(0.070)	(0.196)	(0.067)
Pool+Peer	0.383*	0.125*	0.701***	0.237***
	(0.204)	(0.067)	(0.267)	(0.082)
Period	−0.138***	−0.045***	−0.032***	−0.011***
	(0.019)	(0.006)	(0.007)	(0.002)
Peer × Period	−0.020	−0.007	0.029***	0.010***
	(0.033)	(0.010)	(0.009)	(0.003)
Pool × Period	0.035	0.012	0.019*	0.007*
	(0.030)	(0.010)	(0.010)	(0.003)
Pool+Peer × Period	−0.057	−0.019	0.028**	0.009***
	(0.039)	(0.012)	(0.011)	(0.003)
Constant	0.120		0.338**	
	(0.113)		(0.152)	
Pseudo R^2	0.10		0.10	
N (Clusters)	3250 (65)		9750 (65)	

Note. Probit estimation. *SE* clustered on groups are given in parentheses. * $p < .10$; ** $p < .05$; *** $p < .01$.

Moreover, we compare the average expenditure on punishment. We find differing dynamics of punishment over time across treatments (Figure 4.2, dotted lines). In Peer and Pool+Peer, average punishment expenditures decline over time, showing that the latent threat of punishment is enough to stabilise cooperation even in later periods. In Pool, expenditure remains high and stable over time. Despite the constant and high expenditure on pool-punishment, contribution levels remain at a low level.

To take into account the dynamics in punishment expenditure over time, we test for differences across treatments separately for the first and the second half of Phase 2

(Figure 4.3). In the first half of Phase 2, the punishment expenditure is similar across Peer, Pool and Pool+Peer, Kruskal-Wallis $\chi^2(2) = 4.60$, $p = .100$; groups as independent observations for all non-parametric tests in this section. Significant differences in the punishment expenditure emerge in the second half of Phase 2, Kruskal-Wallis $\chi^2(2) = 16.76$, $p < .001$. Punishment expenditure is highest in the Pool treatment and significantly different from Peer (Mann-Whitney $Z = -4.58$, $p < .001$) and Pool+Peer (Mann-Whitney $Z = 3.75$, $p < .001$). The punishment expenditure in Peer and Pool+Peer is of a comparable magnitude (Mann-Whitney $Z = 0.85$, $p = .398$). These results show that punishment expenditure is high at the start of Phase 2 for all three sanctioning mechanisms. However, punishment expenditure decreases when peer-punishment is available and a credible threat of punishment is established. In contrast, for pool-punishment it is necessary to maintain the threat of punishment by funding the sanctioning mechanism even in later periods.

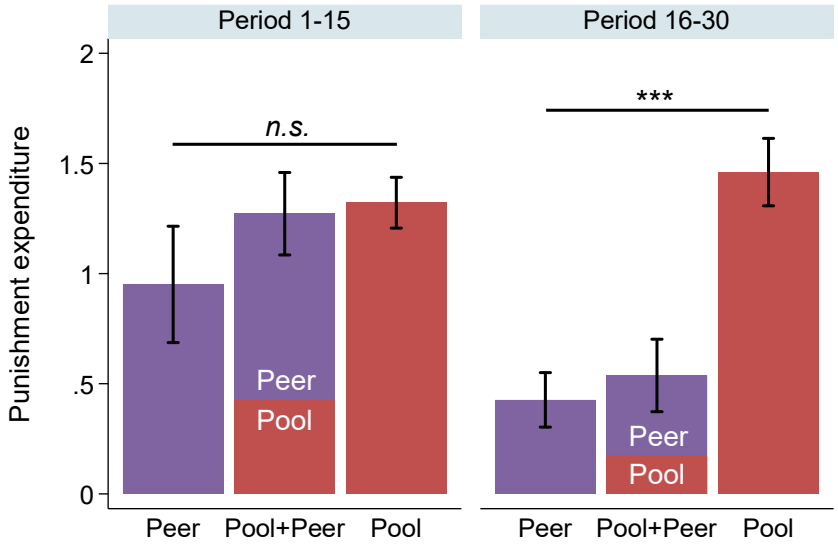


FIGURE 4.3. The average expenditure on punishment in the first and the second half of Phase 2. In the first half, the punishment expenditure is similar across the three punishment mechanisms, Kruskal-Wallis $\chi^2(2) = 4.60$, $p = .100$. The second half of Phase 2 is characterised by large differences in spending on punishment, Kruskal-Wallis $\chi^2(2) = 16.76$, $p < .001$. The error bars indicate ±1 *SEM*. Kruskal-Wallis test: *n.s.* $p \geq .10$; *** $p < .01$. Group averages as independent observations for all test statistics.

Interestingly, subjects in the Pool+Peer treatment use both sanctioning mechanisms, pool-punishment as well as peer-punishment. Although the average punishment expenditure in the second half of Phase 2 is lower compared to the first half, the share of pool-punishment used remains constant. In the first and the second half of Phase 2, pool-punishment accounts for about 33% of the punishment expenditure in this treatment.

An important consequence of the unstructured nature of peer-punishment is that it can be used to punish any group member independent of her actual contribution. Thus, peer-punishment can target defectors, but also contributors which we call antisocial punishment. This is different from pool-punishment that excludes antisocial punishment. The effectiveness of peer-punishment in fostering high cooperation therefore does not only depend on the expenditure on punishment but also on the targeted behaviour. We find a low share of antisocial punishment in the expenditure on peer-punishment. This is consistent with previous evidence from experiments conducted in the UK and is substantially lower compared to other societies (Herrmann et al. 2008). In Peer, 7% of the punishment expenditure in the first half and 9% of punishment expenditure in the second half is directed at cooperators. In Pool+Peer, 12% of the expenditure on peer-punishment in the first half and 7% in the second half is classified as antisocial punishment. The low share of antisocial punishment helps to explain the high effectiveness of peer-punishment in raising cooperation levels.

Next, we test for treatment differences in the punishment behaviour by estimating a Tobit model (Table 4.3). The dependent variable is the punishment expenditure ranging from zero to twelve MU. We include the period number, treatment dummies and interaction terms between the period number and the respective treatment dummy as explanatory variables. Peer serves as comparison treatment. The positive and weakly significant coefficient of Pool+Peer hints at a higher initial punishment expenditure in Pool+Peer compared to Peer. The initial punishment expenditure in Pool is similar to that of Peer. The negative and highly significant period effect indicates a declining punishment expenditure over time in Peer. The punishment expenditure also declines in Pool+Peer at a similar rate compared to Peer, which is shown by the insignificant coefficient for the interaction term. However, the punishment expenditure in Pool is stable over time. This follows from the positive and significant interaction

'Pool × Period', which has a similar absolute size compared to the coefficient of 'Period', $F(1, 7195) = 1.16$, $p = .281$.

TABLE 4.3.
Regression analysis of the punishment expenditure

Dependent variable: punishment expenditure	Tobit coefficients
Pool	2.339 (2.048)
Pool+Peer	3.738* (2.248)
Period	−0.260*** (0.077)
Pool × Period	0.214** (0.086)
Pool+Peer × Period	−0.118 (0.118)
Constant	−6.747*** (2.020)
Sigma	9.234*** (0.725)
N (Clusters)	7200 (48)

Note. Phase 1 and the No Punishment treatment are excluded. Tobit estimation taking into account the censoring of the dependent variable. *SE* clustered on groups in parentheses. * $p < .10$; ** $p < .05$; *** $p < .01$.

Our investigation of contribution behaviour and punishment shows a large heterogeneity across punishment mechanisms. The regressions indicate that significant differences emerge over time in Phase 2. At first, punishment is not significantly different. Peer-punishment is used to punish defectors. In later periods, very little peer-punishment is actually necessary to maintain high cooperation levels because the underlying threat of punishing defection persists. Conversely, pool-punishment is used consistently throughout the course of Phase 2. Since contributions to the punishment pool were common knowledge, expenditure on punishment is necessary in every period to maintain the deterrent effect of punishment.

4.4.2 Efficiency

The efficiency of the punishment mechanisms is determined by the combination of punishment expenditure and the cooperation rate. We define efficiency as the average net payoff, that is, the payoff from the public good minus the expenditure on punishment and the income reduction caused by punishment. Maximum efficiency implies full contributions and no punishment, leading to the highest possible payoff of 40 MU per person and period. Full free riding, on the other hand, yields a payoff of only 20 MU per person and period. Moreover, the punishment expenditure affects efficiency, since punishment is costly for the punisher as well as the punished person. Punishment can increase efficiency if the payoff from cooperation outweighs the cost of punishing. Conversely, punishment decreases efficiency if the cost of punishing exceeds the gain from increased cooperation.

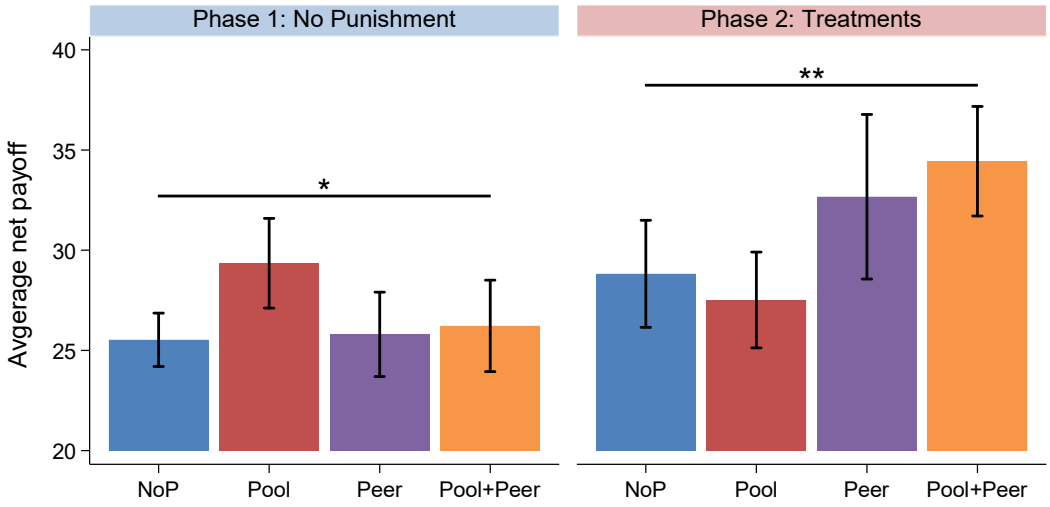


FIGURE 4.4. Efficiency is highest if peer-punishment is available. Efficiency is defined as the average net payoff. The efficiency across treatments is weakly significantly different in Phase 1 and differs significantly in Phase 2. The error bars indicate the bootstrapped 95% CI. Kruskal-Wallis test: * $p < .10$; ** $p < .05$. Group averages as independent observations.

We investigate potential differences in efficiency separately for Phase 1 and Phase 2 (Figure 4.4). Since Phase 1 was identical across treatments and Phase 2 was

only announced after Phase 1 was finished, we did not expect significant differences in efficiency. However, differences in the efficiency across treatments are weakly significant in Phase 1, Kruskal-Wallis $\chi^2(3) = 6.60$, $p = .086$; groups as independent observation for all non-parametric tests in this section. This is due to the efficiency in Pool being significantly higher than in No Punishment, Peer or Pool+Peer (Mann-Whitney $Z = -2.45$, $p = .014$; $Z = -1.81$, $p = .070$; $Z = 1.96$, $p = .050$; respectively). The efficiency across the other three treatments is of a similar magnitude, Kruskal-Wallis $\chi^2(2) = 0.04$, $p = .983$.

Comparing the efficiency across treatments in Phase 2 yields significant differences, Kruskal-Wallis $\chi^2(3) = 11.10$, $p = .011$. Here we find that the efficiency is similar comparing No Punishment and Pool (Mann-Whitney $Z = 0.19$, $p = .851$) as well as for Peer and Pool+Peer (Mann-Whitney $Z = 0.36$, $p = .719$). Yet, the efficiency in Pool is significantly lower compared to Peer (Mann-Whitney $Z = 2.22$, $p = .026$) and Pool+Peer (Mann-Whitney $Z = -3.32$, $p < .001$). Although, the efficiency in No Punishment is lower compared to Peer, this difference is not significant (Mann-Whitney $Z = -1.12$, $p = .264$). Yet, comparing No Punishment and Pool+Peer yields significant differences (Mann-Whitney $Z = -2.41$, $p = .016$). Additionally, we use regression analyses to investigate differences in the efficiency across treatments while controlling for the different periods (Appendix; Table 4.5). The results from the regression analyses corroborate the findings described above.

4.4.3 Volatility

Volatility refers to the degree with which cooperation levels fluctuate over time (Figure 4.5). The volatility of cooperation levels is an indicator of the effectiveness of punishment in fostering cooperation even when it is not used. We define volatility as the average standard deviation of cooperation rates within a group over time. Thus, we first calculate the average cooperation rates for each group and period. Then, we compute the standard deviation of average cooperation rates over time for each group separately and take the average over all groups in a given treatment. The levels of volatility in Phase 1 of the experiment appear similar, yet they are weakly significantly different across treatments, Kruskal-Wallis $\chi^2(3) = 6.35$, $p = .096$. In Phase 2, differences in volatility across treatments are more pronounced and highly significant across treatments, Kruskal-Wallis $\chi^2(3) = 15.83$, $p = .001$. The volatility in Pool is almost three times higher compared to Peer (Mann-Whitney $Z = 2.34$, $p = .018$) or

Pool+Peer (Mann-Whitney $Z = -3.58$, $p < .001$), but it is comparable to the volatility in No Punishment (Mann-Whitney $Z = -1.19$, $p = .234$). The volatility in Peer and Pool+Peer is similar (Mann-Whitney $Z = 0.25$, $p = .806$).

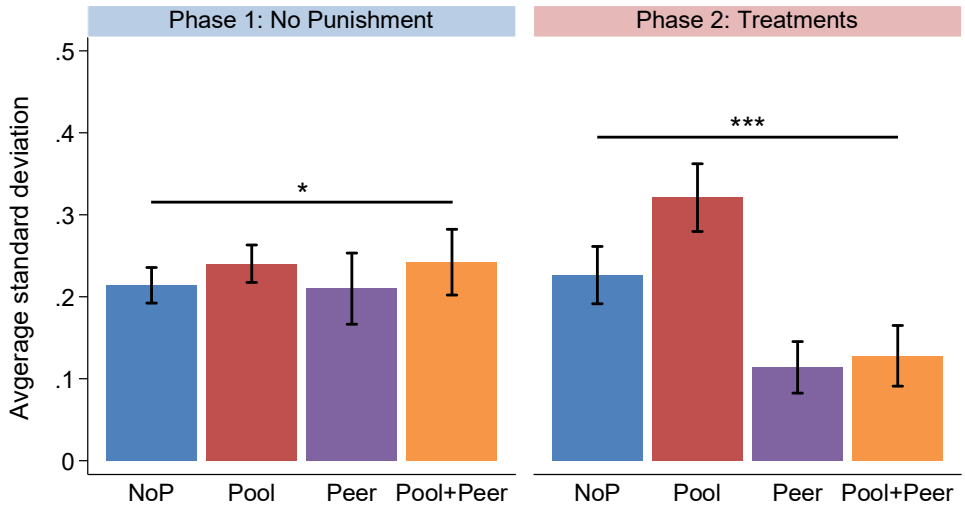


FIGURE 4.5. Volatility of cooperation levels. In Phase 1, the volatility—defined as the average standard deviation of cooperation rates over time—has a similar magnitude across treatments and the differences are weakly significant, Kruskal-Wallis $\chi^2(3) = 6.35$, $p = .096$. In Phase 2, differences across treatments in the volatility of cooperation levels are more pronounced, Kruskal-Wallis $\chi^2(3) = 15.83$, $p = .001$. The volatility in Pool is almost three times larger compared to Peer (Mann-Whitney $Z = 2.34$, $p = .018$) or Pool+Peer (Mann-Whitney $Z = -3.58$, $p < .001$), but similar to No Punishment (Mann-Whitney $Z = -1.19$, $p = .234$). The error bars indicate the bootstrapped 95% CI. Kruskal-Wallis test: * $p < .10$; *** $p < .01$. Group averages as independent observations for all non-parametric tests above.

A reason for the large differences in volatility when comparing Peer and Pool is that for Peer expected punishment drives cooperation behaviour instead of actually executed punishment. Once a credible threat of punishment for defection is established, deviations from the full contribution level are unlikely and volatility is low.

In Pool, contributions to the punishment pool depend on the belief about the number of defectors. Since the size of the punishment pool is announced ahead of the contribution decision, it defines the incentive to contribute to the public good. This allows subjects to learn about the looming punishment of defection and condition their contribution decision in each period on the size of the punishment pool. Thus, a best-

reply reasoning is induced by the pool-punishment mechanism. This can lead to a larger volatility if subjects decide to contribute only when the expected pool-punishment is large enough and defect for lower contributions to the punishment pool.

4.4.4 Best-reply analysis

We define an individual's *best reply* as the payoff-maximising action (contribute or defect) contingent on the punishment the individual expects in a given period. Due to the linear payoff function of the public goods game (see Methods section), defection always yields a payoff that is 12 MU higher than the payoff from cooperation. Therefore, the punishment mechanisms can induce a rational and self-regarding agent to contribute if the expected payoff reduction from punishment of defection exceeds 12 MU, offsetting any benefit from defection. Conversely, if the expected payoff reduction from punishment is smaller than 12 MU, defection is the payoff-maximising action. In case a subject believes the payoff reduction from punishment is exactly 12 MU, then she is indifferent between cooperation and defection. Thus, both actions constitute a best reply.

Since contribution and punishment decisions are made simultaneously, beliefs about the other group members' behaviour are crucial in determining the best reply. We elicit the expected number of defectors in all treatments as well as the expected number punishment points in Peer and Pool+Peer (see Methods section). In Pool, the number of total punishment points allocated to defectors is common knowledge. Yet, group members simultaneously decide whether or not to contribute to the common project and thus the number of defectors is unknown. Therefore, we can calculate the expected punishment depending on the total number of punishment points available and the expected number of defectors in the group. In Peer, the group members have the possibility to punish defectors after making the contribution decisions. Hence, the individual best reply depends on the expected total number of punishment points for defection and for cooperation. In Pool+Peer, the best reply depends on both factors described above. We use the expected number of defectors in the contribution stage and the expected peer-punishment to calculate the individual best reply.

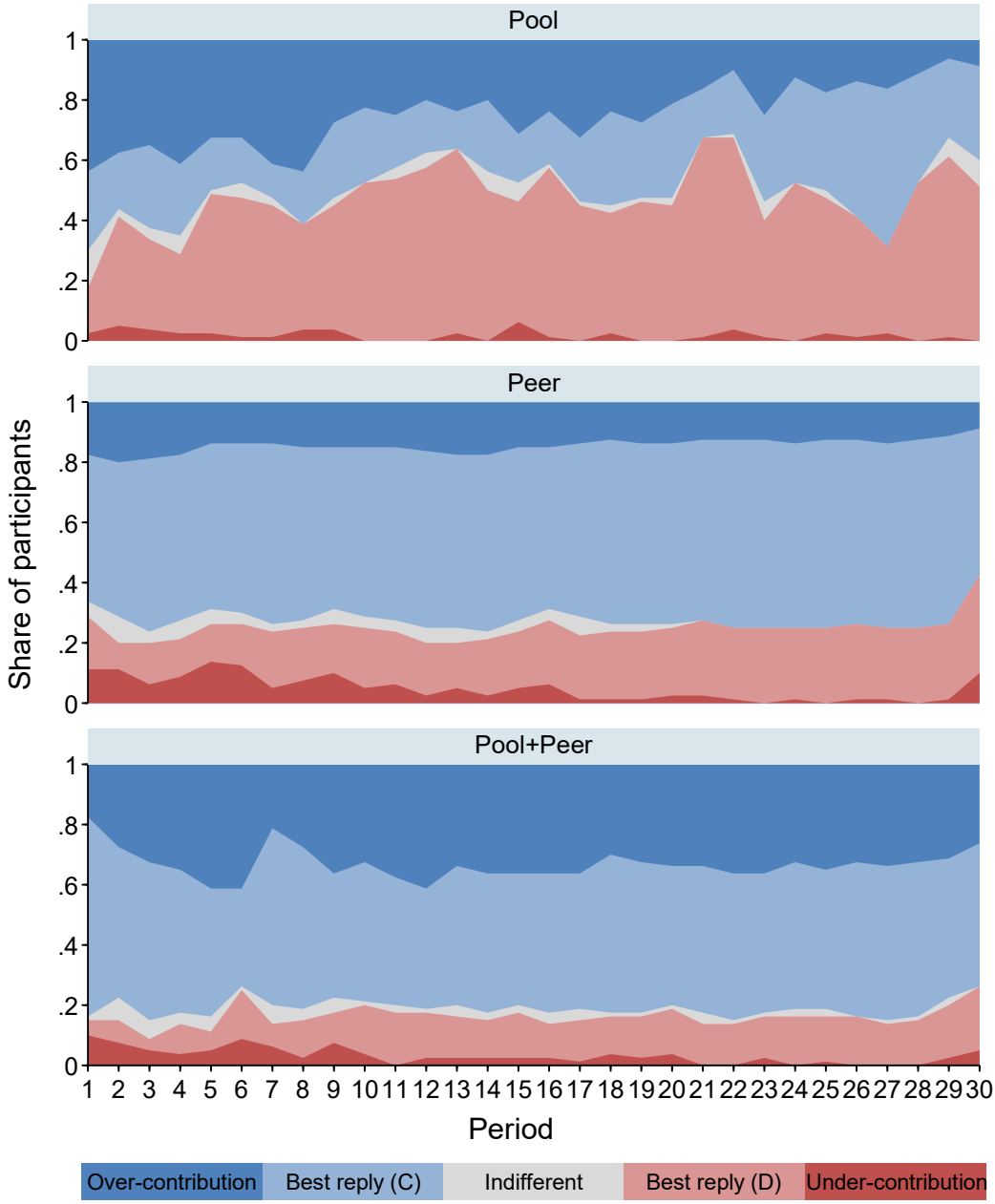


FIGURE 4.6. A best-reply analysis of the contribution decision across the three treatments. The two dark shaded areas show the share of participants whose contribution decision does not correspond with the best reply given their beliefs about the punishment behaviour of others. The dark blue area shows over-contribution and the dark red area indicates under-contribution. The three light shaded areas indicate the share of participants who play the best reply given their beliefs. Light blue corresponds to cooperation (C), light grey shows the share of participants who are indifferent between cooperation and defection, and light red shows defection (D).

Our best-reply analysis shows evidence for crowding out of over-contributions in Pool, but not for the other mechanisms (Figure 4.6). That is, the share of participants contributing more than suggested by their beliefs (over-contributions) is large in the early periods but declines over time. At the same time, the share of people who play the best reply (all light shaded areas) increases. In Peer, the decline over the course of the thirty periods in the share of participants who over-contribute is less pronounced. In Pool+Peer, we do not find evidence for a decline of over-contributions over time. Yet, the share of subjects playing the best reply increases over the thirty periods.

We use an OLS regression to test for differences in the levels and trends of over-contribution across the three punishment mechanisms (Table 4.4, Col. 1). The dependent variable is the share of participants who contribute more than suggested by their best reply. The explanatory variables include treatment dummies for Pool and Pool+Peer. Peer serves as the comparison group. Additionally, we include the variable 'Period' and interaction effects between Period and the treatment dummies. The positive and highly significant coefficients for Pool and Pool+Peer indicate a higher share of subjects who initially over-contribute in these treatments. The highly significant and negative coefficient for period shows a very small decline over time in the share of people who over-contribute in Peer. The negative and highly significant interaction 'Period × Pool' is a sign of a stronger decrease in the share of subjects who over-contribute in Pool relative to the comparison group. The weakly significant and very small positive interaction 'Period × Pool+Peer' indicates a relatively stable share of over-contribution in this treatment. We conclude that crowding out of over-contributions is strongest under the pool-punishment mechanism.

Next, we investigate treatment differences in the share of subjects who play the best reply (Table 4.4, Col. 2). The explanatory variables are the same as in the regression model described above. The negative and highly significant dummy variables for Peer and Pool+Peer indicate that—relative to our comparison group—a lower share of subjects initially play the best reply in these treatments. The highly significant positive coefficient for period indicates a slight increase over time in the share of people who play the best reply in Peer. The positive and significant interaction 'Period × Pool' shows that this increase is even higher in Pool. The negative and highly significant coefficient for 'Period × Pool+Peer' is of a similar size than the coefficient for period, implying a stable share of participants playing the best reply in Pool+Peer.

TABLE 4.4.
Regression analysis of best replies.

Dependent variable: share of participants who…	(1) over-contribute	(2) play best reply
Pool	0.234***	−0.162***
	(0.020)	(0.023)
Pool+Peer	0.147***	−0.111***
	(0.030)	(0.028)
Period	−0.002***	0.005***
	(0.000)	(0.001)
Period × Pool	−0.008***	0.006***
	(0.001)	(0.001)
Period × Pool+Peer	0.003*	−0.004***
	(0.001)	(0.001)
Constant	0.177***	0.723***
	(0.007)	(0.011)
N	90	90
R^2	0.82	0.81

Note. OLS estimates. Peer serves as the comparison group. Robust *SE* in parenthesis. * $p < .10$. ** $p < .05$. *** $p < .01$.

The best-reply analysis shows that when pool-punishment is available, people gradually rely more on best-reply reasoning when making their cooperation decision. In this treatment, we also see a significant decline in over-contributions over time. Thus, peer-punishment supports *voluntary contributions* that exceed the contribution levels suggested by the best-reply analysis. Peer-punishment, by itself, stabilises the share of over-contributions relative to pool-punishment. In combination with pool-punishment, we find an even higher and stable share of over-contributions.

4.4.5 Individual cooperative dispositions

We now focus on our second research question and explore the role of individual differences in the emergence and maintenance of different punishment mechanisms. We use the subjects' cooperative disposition, elicited during the online experiment, as a measure of individual cooperativeness. Drawing on the criteria outlined in Fischbacher et al. (2012), we classify 68% of subjects as Conditional Cooperators (CC)

and 21% as Free Riders (FR). 11% of subjects are unclassified others and excluded from the analysis. The reason for excluding unclassified subjects lies in the fact that CC and FR constitute the largest shares of subjects and that their dispositions are easy to reconcile with models of other-regarding preferences. Thus, the remainder of this section focuses on the comparison of subjects with conditional cooperative or free-rider dispositions only.

Figure 4.7 shows the characteristic patterns of punishment expenditure for the different mechanisms in Phase 2, with high and stable punishment expenditure in Pool and a declining expenditure if peer-punishment is available. We find that both, CC and FR, engage in punishment independent of the punishment mechanism. However, the expenditure on punishment of CC and FR is likely to differ across punishment mechanisms.

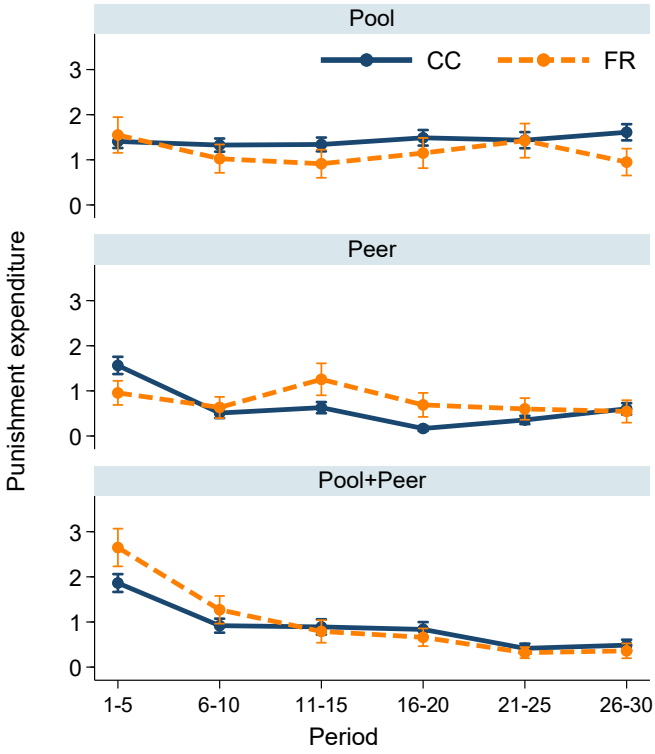


FIGURE 4.7. Average punishment expenditure by cooperative disposition. The error bars indicate ±1 *SEM*; subjects as independent observations.

We conduct regression analyses separately for each treatment to test for differences in the punishment behaviour of CC and FR (Appendix; Table 4.6; Table 4.7). We draw on a two-stage regression model to separate the likelihood of engaging in punishment from the severity of punishment (Nikiforakis and Engelmann 2011).

For Pool, we find significant differences in the CC's and FR's likelihood to engage in pool-punishment. FR are 17% less likely to make a contribution to the punishment pool compared to CC. However, if FR choose to contribute to the punishment pool, then they choose a higher contribution than CC.

For Peer, we find no significant differences in the likelihood to engage in peer-punishment when comparing CC and FR. Yet, for the case that FR engage in punishment, they choose sanctions that are even more severe than that of CC.

For Pool+Peer, we find significant differences in the likelihood of CC and FR to engage in peer-punishment, with FR being significantly more likely to punish a defector compared to CC. Additionally, we find that FR who selected to engage in punishment choose a more severe peer-punishment compared to CC. A reason for this might be the restriction on the total amount of punishment each subject can buy. If CC are more likely to engage in pool-punishment, as suggested above, then their possibility to engage in peer-punishment might be more restricted, since they would have already used parts of their total number of deduction points available.

The findings above also support the results of Chapter 2, rejecting the assumption that only subjects with a cooperative disposition engage in punishment. Similarly, we find no differences in the likelihood of peer-punishment and an even more severe peer-punishment chosen by FR. However, we find significant differences for pool-punishment, with FR being less likely to engage in pool-punishment. This suggests that the design of the sanctioning institution determines whether the cooperative disposition matters for engagement in punishment. Funding the pool-punishment mechanism might be more challenging because a smaller group of people are willing to engage in it. A reason for the difference in engagement might be that contributions to the punishment pool are made before learning the behaviour of others. This decision is therefore less likely to be driven affect compared to peer-punishment.

4.5 Discussion

The present study explores the effectiveness and efficiency of different punishment mechanisms using controlled laboratory experiments. We find that the informal punishment mechanism induces high and stable cooperation at low cost in the long run. Formal punishment leads to the lowest efficiency level and highest volatility. Combining both mechanisms achieves the highest efficiency level, although not significantly different from informal punishment. Additionally, the best-reply analysis shows that crowding out of over-contributions is highest if peer-punishment is not available. The high volatility in the pool-punishment treatment as well as the best-reply analysis indicate that, without peer pressure, formal sanctions encourage best-reply reasoning. Thus, peer pressure helps to stabilise cooperation in the long run.

We also investigate the influence of the individual cooperative disposition on the emergence and maintenance of different punishment mechanisms. Due to the nature of repeated games, there might be strategic motives to engage in punishment even for rational and self-regarding subjects.

We find no evidence for a lower likelihood to engage in peer-punishment or a lower punishment severity chosen in subjects with a free-rider disposition. In fact, we found an even higher likelihood and severity of peer-punishment in these subjects. This is in line with the findings of Chapter 2. Although the present study does not focus on the subjects' emotional response, the link between negative emotions like anger and punishment of defectors is well established (Sanfey et al. 2003, Hopfensitz and Reuben 2009, Cubitt et al. 2011, Chapter 2). Thus, peer-punishment might be an outlet for negative emotions.

The likelihood to engage in pool-punishment decreases for subjects with a free-rider disposition. However, subjects with a free-rider disposition who chose to engage in pool-punishment contribute more to the punishment pool. Since subjects contribute to the punishment pool before learning the outcome of the cooperation decision, this decision is 'cold' and less likely to be influenced by affect.

This finding might also be of importance for the design of effective large-scale sanctioning institutions that require an up-front investment to establish the institution. Pool-punishment relies more on subjects with a cooperative disposition to fund the punishment mechanism, whereas the maintenance of peer-punishment expenditure is

less affected by subjects' individual cooperativeness. Thus, for peer-punishment the second-order public good of punishment is funded by a larger group of subjects.

Our finding of peer-punishment leading to high and stable cooperation rates and efficient outcomes might be idiosyncratic to countries similar to the UK, which are characterised by pronounced norms of civic cooperation and a high rule of law. Peer-punishment typically thrives in such countries because antisocial punishment is limited (Herrmann et al. 2008). The possibility of high antisocial punishment can decrease or fully diminish the cooperation enhancing effect of peer-punishment. To test this possibility, it would be necessary to conduct a similar experiment in a low-trust country, where peer-punishment alone has been shown to be inefficient in sustaining cooperation.

4.6 Appendix

4.6.1 Supporting analysis

4.6.1.1 Regression analysis of efficiency

Table 4.5 shows the results of a regression analysis of the average net payoffs across treatments. We run regressions for Phase 1 (i.e., without punishment) and Phase 2 (i.e., the treatment period) separately. The dependent variable is the average net payoff at the group level of each period. The independent variables include dummies for the Pool, Peer and Pool+Peer treatments as well as the period number.

Although, we did not expect any differences in the average net payoff across treatments in Phase 1, we find a significantly higher efficiency in Pool. Since Phase 1 was the same for all treatments and Phase 2 was not announced until Phase 1 was finished, it might be that participants in Pool were characterised by greater cooperativeness. Yet, we see in Phase 2 that the efficiency in Pool is not significantly different from our baseline treatment without punishment. Efficiency is significantly higher in Peer and Pool+Peer. Thus, we conclude that if participants in Pool displayed a higher cooperativeness, this would actually mean we underestimate the differences in the effectiveness and efficiency of punishment mechanisms in Phase 2.

TABLE 4.5.
Differences in the average net payoff across treatments.

Dependent variable: average net payoff	(1) Phase 1	(2) Phase 2
Pool treatment dummy	5.587** (2.176)	−2.762 (3.094)
Peer treatment dummy	−0.401 (2.185)	7.338** (3.110)
Pool+Peer treatment dummy	0.557 (2.182)	8.016** (3.110)
Period dummies included	Yes	Yes
Constant	29.930*** (1.630)	28.942*** (2.319)
Sigma (Group)	5.993*** (0.591)	8.783*** (0.812)
Sigma (Residual)	5.074*** (0.201)	6.808*** (0.138)
N (Groups)	650 (65)	1950 (65)

Note. Random-effects Tobit regressions with group averages per period as independent observations. The Tobit model accounts for the censoring of the dependent variables. The No Punishment treatment is the reference group. *SE* in parentheses. * $p < .10$. ** $p < .05$. *** $p < .01$.

4.6.1.2 Cooperative dispositions and punishment

For Pool, we first estimate a Probit model with a punishment dummy taking the value one for a positive contribution to the punishment pool (Table 4.6). This model allows to investigate potential differences in the likelihood of engaging in punishment. The explanatory variables are the expected contribution of others to the punishment pool, a dummy variable taking the value one for a free-rider disposition (FR), and the interaction term 'FR × Expected avgerage pool-punishment'. Additionally, we control for the total contributions to the punishment pool in the previous round as well as the number of defectors in the previous round.

TABLE 4.6.
Explaining pool-punishment.

Dependent variable: (1) punishment dummy; (3) punishment expenditure	(1) Punishment decision	(2) Avg. marg. effects	(3) Punishment severity
Expected avg. pool-punishment	0.005 (0.029)	0.002 (0.010)	−0.619* (0.352)
FR	−0.493** (0.196)	−0.168** (0.066)	4.265** (1.710)
FR × Expected avg. pool-punishment	−0.015 (0.061)	−0.005 (0.021)	0.119 (0.393)
Total contributions to the punishment pool in $t-1$	−0.003 (0.004)	−0.001 (0.001)	0.066 (0.050)
Number of defectors in $t-1$	−0.011 (0.034)	−0.004 (0.012)	0.270 (0.424)
Period	−0.011** (0.004)	−0.004*** (0.001)	0.163*** (0.043)
Constant	−0.233 (0.144)		−1.054 (1.930)
N (Groups)	2117 (16)		638 (16)

Note. Only CC and FR are included in the analysis. (1) Probit coefficients; (2) Average marginal effects of the Probit model; (3) Truncated linear regression. *SE* clustered on groups in parentheses. * $p < .10$; ** $p < .05$; *** $p < .01$.

We find a significant level difference in the likelihood of CC and FR to engage in pool-punishment. FR are significantly less likely to make a contribution to the punishment pool, with the average marginal effect suggesting a difference in the likelihood of nearly 17%. Additionally, we find a significant decrease in the likelihood to engage in punishment over time. However, the average marginal effect suggests that this decrease is very small.

We use a truncated linear regression to investigate potential differences in the punishment severity for subjects who engage in punishment. The dependent variable is punishment expenditure and explanatory variables are the same as in the first-stage model described above.

The punishment severity is negatively and significantly associated with the expected contribution of other group members to the punishment pool. This suggests that subjects contribute less for a higher contribution of others. We find a significant level difference in the punishment severity of CC and FR, with FR choosing a more severe punishment than CC. Additionally, punishment severity increases with time.

For Peer, we also estimate a two-stage regression model to separate between the likelihood and the severity of punishment (Table 4.7, Col. 1–3). First, we estimate a Probit model with a punishment dummy as dependent variable that takes the value one if the subject engages in punishment. Since we expect deviation from the own behaviour to be punished, we include a dummy variable taking the value one for the case that the punisher cooperates and the targeted subject defects. Additionally, we include a dummy variable taking the value one for the case that the punisher defects and the other subject cooperates. We also include a dummy variable taking the value one if the punisher is classified as FR, as well as the interaction terms between FR and the other two dummy variables. Furthermore, we include the period number to control for time effects.

The likelihood of punishment significantly increases if the potential punisher cooperates and the other defects. Conversely, if the potential punisher defects and the other cooperates, the likelihood increases as well. The insignificant dummy variable for FR and its insignificant interaction term suggest no differences in the likelihood of engaging in punishment when comparing CC and FR. Additionally, the dummy FR and its interaction with CD are jointly insignificant, $\chi^2(2) = 1.68$, $p = .432$. The same holds true for the interaction with DC, $\chi^2(2) = 0.84$, $p = .658$.

The severity of punishment is significantly higher if the potential punisher cooperates and the other subject defects. If the situation is reversed, the punishment severity is not significantly impacted. There are significant level differences in the severity of punishment when comparing CC and FR, with FR punishing more severely. The interactions ‘FR × CD’ and ‘FR × DC’ are negative and significant, indicating a lower punishment by FR in these situations. Furthermore, we find that the punishment severity increases with time.

For Pool+Peer, we estimate a similar two-stage regression model that additionally includes a variable for the deviation of the other’s contribution from the

own contribution to the punishment pool as well as an interaction term with FR (Table 4.7, Col. 4–6). The Probit regression for the likelihood of punishment shows a significant increase in the likelihood if the punisher cooperates and the other defects. Similarly, the likelihood increases for the case that the punisher defects and the other contributes. There are no level differences in the likelihood of punishment comparing CC and FR. However, the positive and significant interaction 'FR × CD' indicates a higher likelihood of punishment for FR in this situation. However, the average marginal effect suggests only an increase in the punishment likelihood of about 1%. Furthermore, the punishment likelihood increases for a larger deviation of the other from the punisher's pool-punishment expenditure. This suggest that in Pool+Peer, peer-punishment is to some extent used to enforce contributions to the punishment pool.

Looking at the results for the punishment severity shows a significant increase for the case that the punisher contributed and the targeted subject defects. The reverse situation does not significantly impact the punishment severity. Additionally, we find significant level differences in the punishment severity of CC and FR, with a more severe punishment chosen by FR. However, the negative and significant interaction 'FR × CD' indicates that FR choose a lower punishment expenditure in this situation compared to CC. We also find that the punishment severity rises for a larger deviation in contributions to the punishment pool. Furthermore, the insignificant period effect suggests no change in the punishment severity over time.

TABLE 4.7.
Explaining peer-punishment.

	Peer			Pool+Peer		
	(1) Punish. decision	(2) Avg. marg. effects	(3) Punish. severity	(4) Punish. decision	(5) Avg. marg. effects	(6) Punish. severity
Punisher cooperates, other defects (CD)	2.385***	0.155***	6.010***	2.528***	0.113***	19.623***
	(0.291)	(0.020)	(1.859)	(0.163)	(0.020)	(5.908)
Punisher defects, other cooperates (DC)	0.562**	0.036**	−2.058	0.333*	0.015	−0.955
	(0.260)	(0.018)	(2.786)	(0.200)	(0.012)	(3.872)
FR	0.112	0.007	5.783***	0.101	0.005	11.679**
	(0.333)	(0.022)	(2.033)	(0.116)	(0.005)	(5.270)
FR × CD	−0.351	−0.023	−3.785**	0.309*	0.014*	−11.101*
	(0.322)	(0.020)	(1.682)	(0.184)	(0.007)	(5.665)
FR × DC	−0.421	−0.027	−9.042*	0.422	0.019	−7.536
	(0.460)	(0.033)	(5.244)	(0.511)	(0.020)	(6.906)
Difference in pool-punishments				0.037*	0.002	2.217***
				(0.021)	(0.001)	(0.696)
FR × Difference in pool-punishment				0.023	0.001	−1.548
				(0.070)	(0.003)	(0.996)
Period	−0.021***	−0.001**	0.149**	−0.033***	−0.001***	−0.084
	(0.008)	(0.001)	(0.061)	(0.007)	(0.000)	(0.132)
Constant	−1.805***		−6.487**	−1.966***		−16.839***
	(0.239)		(2.734)	(0.174)		(5.309)
N (Groups)	8760 (16)		441 (16)	8520 (16)		345 (15)

Note. Only CC and FR are included in the analysis. Col. 1, 4: Probit coefficients with the punishment dummy as dependent variable; Col. 2, 5: Average marginal effects of the Probit model; Col. 3, 6: Truncated linear regression with punishment expenditure as dependent variable. *SE* clustered on groups in parentheses. * $p < .10$; ** $p < .05$; *** $p < .01$.

4.6.2 Instructions

4.6.2.1 Phase 1: All treatments

You are now taking part in a decision making experiment. Depending on the decisions made by you and other participants, you can earn a considerable amount of money. It is therefore very important that you read these instructions with care.

It is prohibited to communicate with other participants during the experiment. If you have any questions, please raise your hand. A member of the experiment team will come and answer them in private. If you violate this rule, you will be dismissed from the experiment and you will forfeit all payments.

During the experiment, we will not speak in terms of Pounds, but in Points. At the end your entire earnings will be calculated in Points. The total amount of Points you have earned will be converted to Pounds at the following rate:

100 Points = 2 Pounds

After this experimental session, your entire earnings from the experiment will be paid to you privately in cash.

This experiment consists of two parts. You will receive a second set of instructions once the first part of the experiment is finished. After the second part of the experiment, you will be asked to fill in a questionnaire. Your responses to the questionnaire will not affect your earnings during the experiment.

The groups

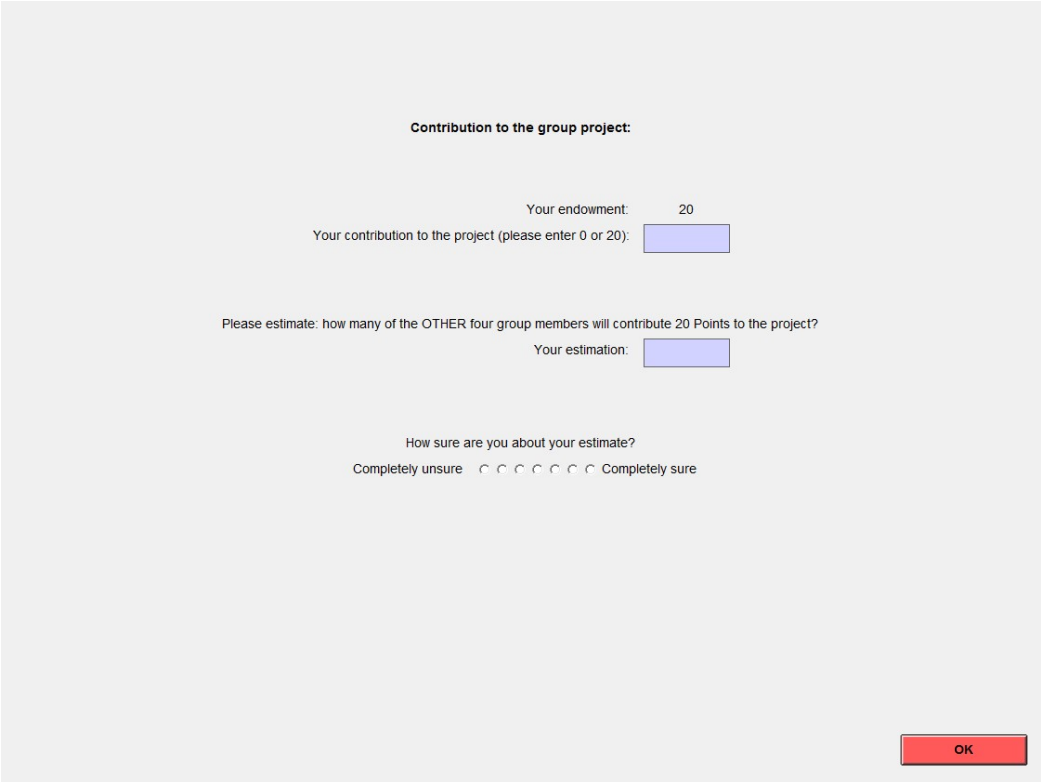
At the beginning of the experiment, all participants will be randomly divided into groups of five. Apart from you, there will be four other members in your group. **You will not learn who the other people in your group are at any point.** The groups will stay the same throughout the whole session. That means you will be paired with the same four persons throughout the first part of today's experiment as well as during the second part.

The decision situation

Each participant receives an endowment of **20 Points**. You have to decide whether to contribute or not to contribute to a group project. You can contribute either 0 Points or 20 Points to the project. You will keep the Points that you do not contribute to the group

project for yourself. The four other members of your group have to make the same decision. They can also contribute either 0 Points or 20 Points to the project.

You will see the following screen when making your contribution decision:



As mentioned above, your endowment is 20 Points. You have to decide whether to contribute 0 or 20 Points to the project by typing the number 0 or 20 in the box. By deciding how many Points to contribute to the project, you automatically decide how many Points you keep for yourself. After entering the amount of Points you want to contribute, you are asked to estimate the number of group members who contribute 20 Points to the project and indicate how confident you are that your estimation is true. After making all the entries, you must click on the "*OK*" button. Once you have done this, your decision can no longer be revised.

After the contribution decision, you will see how many Points each of the other four group members has contributed to the project and their corresponding income from this contribution decision. Nonetheless, the identities of your group members will not be revealed at any time.

The payoffs

The income of every member of the group is calculated in the same way. It consists of two components:

(1) Private account: Points that you keep for yourself. The Points that you do not contribute to the project automatically belong to you.
(2) Group project: your return from the group project. All of the Points contributed to the project will be multiplied by 2 and this amount will be divided equally among all five members of the group.

Your income in a given period can also be illustrated by the following formula:

$$\textit{Your Total Income in Points} = 20 - (\textit{Your Contribution}) + 0.4 \times (\textit{Group Project})$$

In this formula, "*20*" is your endowment, "*Your Contribution*" is either 0 or 20, and "*Group Project*" is the total number of points contributed to the group project by all group members. It is multiplied by 0.4 because that is the result of multiplying by 2 and then dividing by 5.

The periods

This decision situation will be **repeated for ten periods**. The groups are not going to change during these ten periods. You will therefore be interacting with the same persons in all periods. Your incomes from the ten periods will be summed up. This means that what you earn in each period will be added to what you earned in previous periods.

Examples

In order to explain the income calculation we will give some examples. Please read them carefully. At the end of the introductory information, you will be asked to answer several computerised control questions which are designed to check that you have understood the decision situation.

Example 1

Each of the five members of the group contributes 0 Points to the project:

- Private account: each group member will have 20 points in his/her private account.

- Group project: No one will receive anything from the group project, because no one contributed anything.

Therefore the total income of each group member is 20 Points.

Calculation: $(20 - 0) + 0.4 \times 0 = 20$

Example 2

One group member contributes 20 Points and the other four group members all contribute 0 Points to the project:

- Private account: The group member who contributed 20 Points has nothing in his/her private account. Each of the four group members who contributed 0 Points has 20 Points in their private accounts.
- Group project: A total of 20 Points are contributed to the project. Each member receives $0.4 \times 20 = 8$ Points from the project.

The group member who contributed 20 Points to the project earns 0 Points from the private account and 8 Point from the group account, for a total of 8 Points.

Calculation: $(20 - 20) + 0.4 \times 20 = 8$

Each of the four group members who contributed 0 Points earn 20 Points from their private accounts and 8 Points from the group account, for a total of 28 Points.

Calculation: $(20 - 0) + 0.4 \times 20 = 28$

If you have any questions, please raise your hand and a member of the experiment team will come and answer them in private.

4.6.2.2 Phase 2: No Punishment treatment

You are now taking part in the second part of the experiment. The money you earn in this part will be added to what you earned in the first part. As before the Points you have earned will be converted to Pounds at the following rate:

100 Points = 2 Pounds

The groups

In the second part of the experiment, groups will stay exactly the same as in the first part. That means you will be **paired with the same four persons again** in this part of the experiment.

The decision situation

The decision situation is **identical** to the one described on the first instruction sheet.

As before, each participant receives an endowment of 20 Points. You have to decide whether to contribute or not to contribute to a group project. You can contribute either 0 Points or 20 Points to the project. You will keep the Points that you do not contribute to the group project for yourself. The other four members of your group have to make the same decision. They can also contribute either 0 Points or 20 Points to the project.

The payoff

Your income in a given period can also be illustrated by the following formula:

$$\textit{Your Total Income in Points} = 20 - (\textit{Your Contribution}) + 0.4 \times (\textit{Group Project})$$

In this formula, "*20*" is your endowment, "*Your Contribution*" is either 0 or 20, and "*Group Project*" is the total number of points contributed to the group project by all group members. It is multiplied by 0.4 because that is the result of multiplying by 2 and then dividing by 5.

The periods

This decision situation will be **repeated for thirty periods**. The groups are not going to change during these thirty periods. You will therefore be interacting with the same persons in all periods. Your incomes from the thirty periods will be summed up. This means that what you earn in each period will be added to what you earned in previous periods.

If you have any questions, please raise your hand and a member of the experiment team will come and answer them in private.

4.6.2.3 Phase 2: Peer treatment

You are now taking part in the second part of the experiment. The money you earn in this part will be added to what you earned in the first part. As before the Points you have earned will be converted to Pounds at the following rate:

100 Points = 2 Pounds

The groups

In the second part of the experiment, groups will stay exactly the same as in the first part. That means you will be **paired with the same four persons again** in this part of the experiment.

The decision situation

The decision situation is similar to the one described on the first instruction sheet. But this time, there are **two stages** in each period:

The **first stage is exactly like before:** each participant receives an endowment of **20 Points**. You have to decide whether you contribute 0 or 20 Points to a group project. The four other members of your group have to make the same decision.

After the contribution decision, there will be a **second stage**. At this stage, you will see how many Points each of the other four group members has contributed to the project and their corresponding income from this contribution decision. Nonetheless, the identities of your group members will not be revealed at any stage. You can either **decrease** or **leave unchanged** the income of each other group member by assigning **Deduction Points** to them. The other group members can also decrease your income, by allocating Deduction Points to you, if they wish to do so.

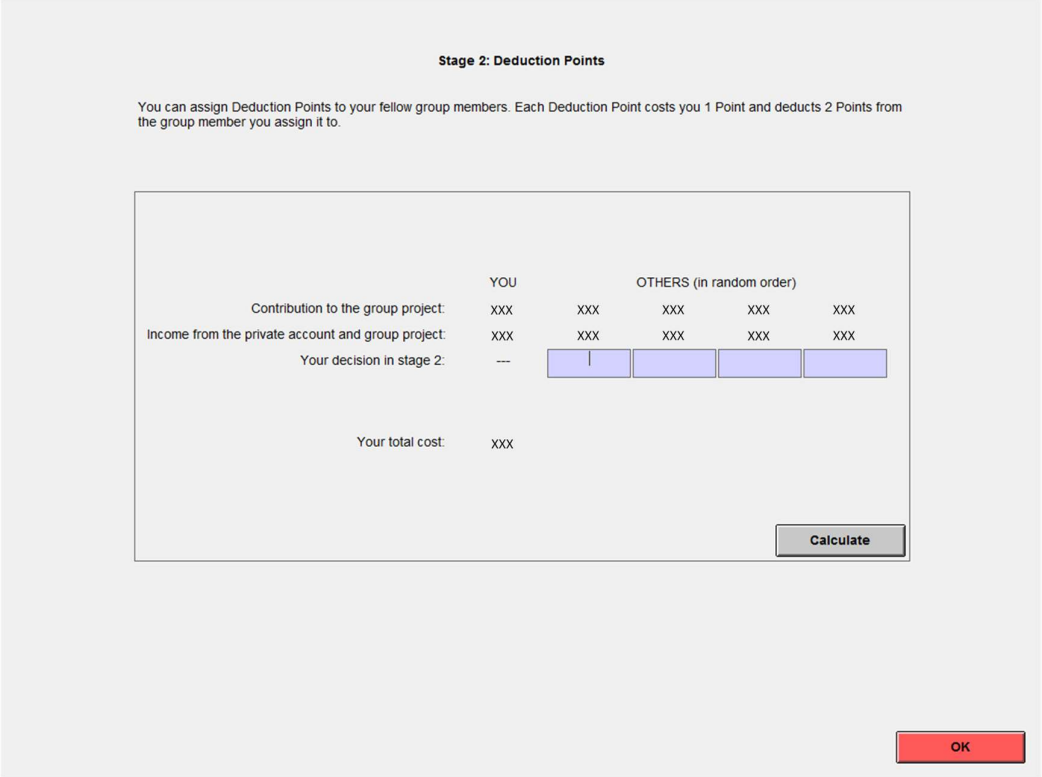
Deduction Points

In stage 2 you can assign **between 0 and 12 Deduction Points to the other group members.** Each Deduction Point that you assign costs you **one Point.** For example, if you assign 3 Deduction Points to your group members then the cost to you is 3 Points.

Each Deduction Point that you assign to a particular group member **decreases their income by 2 Points.** For example, if you assign 2 Deduction Points to a particular group member, you will decrease this group member's income by 4 Points.

Your **own income will be reduced by 2 Points for each Deduction Point that is assigned to you** by the other four group members.

You will see the following screen at stage 2:



The column on the left shows your contribution and your income from the first stage. The other four columns indicate the contributions of your group members and their incomes from the first stage.

If you do not wish to change the income of the other group members, type "0" into the fields next to "*Your decision in stage 2.*" In case you want to assign Deduction Points, enter the number of Deduction Points you want to assign into this field. You must enter a decision into every field and press the "*Calculate*" button. This will display the cost of your decision. Until you press the "*OK*" button, you can still change your decision. To recalculate the costs after making a change, simply press the "*Calculate*" button again.

The payoffs (stage 1 and 2)

Your income in a given period can be illustrated by the following formula:

$$\textit{Your Total Income in Points} = 20 - (\textit{Your Contribution}) + 0.4 \times (\textit{Group Project}) - (\textit{Deduction Points Assigned By You To Others}) - 2 \times (\textit{Sum Of Deduction Points Assigned To You By Others})$$

In a given period, you cannot lose – through Deduction Points assigned to you by other group members – more than you have earned in the period. But you will always have to incur the cost of allocating Deduction Points.

The periods

This decision situation will be **repeated for thirty periods**. The groups are not going to change during these thirty periods. You will therefore be interacting with the same persons in all periods. Your incomes from the thirty periods will be summed up. This means that what you earn in each period will be added to what you earned in previous periods.

If you have any questions, please raise your hand and a member of the experiment team will come and answer them in private.

4.6.2.4 Phase 2: Pool treatment

You are now taking part in the second part of the experiment. The money you earn in this part will be added to what you earned in the first part. As before the Points you have earned will be converted to Pounds at the following rate:

100 Points = 2 Pounds

The groups

In the second part of the experiment, groups will stay exactly the same as in the first part. That means you will be **paired with the same four persons again** in this part of the experiment.

The decision situation

The decision situation is similar to the one described on the first instruction sheet. But this time, there are **two stages** in each period:

In the **first stage**, you can assign **between 0 and 12 Deduction Points to your group's deduction pool.** The other four members of the group make the same decision.

The **second stage is exactly like before:** each participant receives an endowment of **20 Points**. You have to decide whether you contribute 0 or 20 Points to a group project. The four other members of your group have to make the same decision.

The group's deduction pool

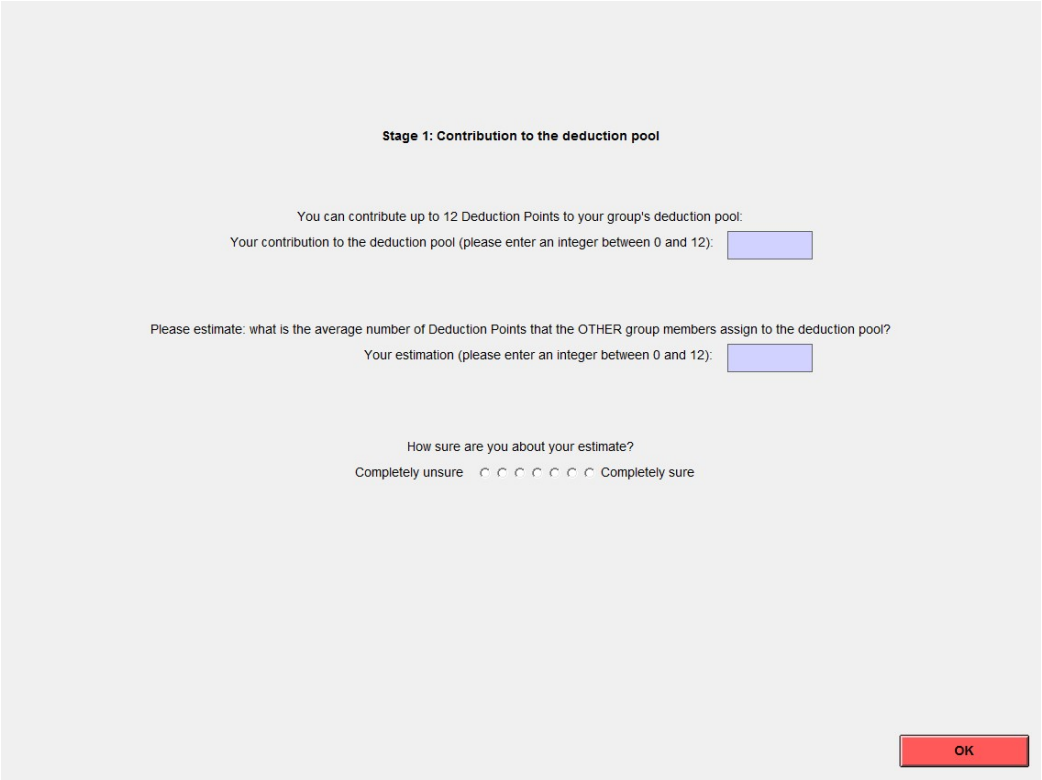
In the first stage you can assign between 0 and 12 Deduction Points to the group's deduction pool. **Each deduction point that you assign costs you one Point.** For example, if you assign 3 Deduction Points to the deduction pool then the cost to you is 3 Points.

The sum of Deduction Points in the group's deduction pool is assigned in equal shares to all group members who contribute 0 in the second stage.

Each Deduction Point that a particular group member receives from the deduction pool, decreases their income by 2 Points. For example, if there are 11 Deduction Points in the deduction pool and two group members contributed 0 in the second stage then 5.5 Deduction points will be assigned to each of them, and their income will be decreased by 11 Points each.

Your **own income will be reduced by 2 Points for each Deduction Point that is assigned to you** from the Deduction Pool.

You will see the following screen at stage 1:



As mentioned above, you can assign up to 12 Deduction Points to your group's deduction pool. You have to decide how many Deduction Points you assign to the deduction pool by typing a number between 0 and 12 (including 0 and 12) in the box. After entering the amount of Deduction Points you want to assign, you are asked to estimate the average number of Deduction Points that your group members assign to the deduction pool and indicate how confident you are that your estimation is true. After making all the entries, you must click on the "*OK*" button. Once you have done this, your decision can no longer be revised.

After everyone decided on their allocations to the deduction pool, you will learn the total amount of Deduction Points that your group assigned to the deduction pool.

In the second stage, you will make the contribution decision to the project. After the contribution decision, you will see how many Points each of the other four group members has contributed to the project and their corresponding income from this contribution decision. Nonetheless, the identities of your group members will not be revealed at any time.

The payoffs (stage 1 and 2)

Your income in a given period can be illustrated by the following formula:

Your Total Income in Points = 20 − (*Your Contribution)* + 0.4 × (*Group Project*)
− (*Deduction Points Assigned By You To The Deduction Pool*)
− 2 × (*Deduction Points Assigned To You From The Deduction Pool*)

You will only receive Deduction Points from the deduction pool if you contributed 0 at the second stage. Furthermore, in a given period you cannot lose more than you have earned through Deduction Points you receive from the deduction pool. But you will always have to incur the cost of allocating Deduction Points.

The periods

This decision situation will be **repeated for thirty periods**. The groups are not going to change during these thirty periods. You will therefore be interacting with the same persons in all periods. Your incomes from the thirty periods will be summed up. This means that what you earn in each period will be added to what you earned in previous periods.

If you have any questions, please raise your hand and a member of the experiment team will come and answer them in private.

4.6.2.5 Phase 2: Pool+Peer treatment

You are now taking part in the second part of the experiment. The money you earn in this part will be added to what you earned in the first part. As before the Points you have earned will be converted to Pounds at the following rate:

100 Points = 2 Pounds

The groups

In the second part of the experiment, groups will stay exactly the same as in the first part. That means you will be **paired with the same four persons again** in this part of the experiment.

The decision situation

The decision situation is similar to the one described on the first instruction sheet. But this time, there are **three stages** in each period:

In the **first stage**, you can assign **Deduction Points to your group's deduction pool.** The other four members of the group make the same decision.

The **second stage is exactly like before:** each participant receives an endowment of **20 Points**. You have to decide whether you contribute 0 or 20 Points to a group project. The four other members of your group have to make the same decision.

After the contribution decision, there will be a **third stage**. At this stage, you will see how many Points each of the other four group members has contributed to the deduction pool and to the project. You will also learn their corresponding income from these contribution decisions. Nonetheless, the identities of your group members will not be revealed at any stage. You can either **decrease** or **leave unchanged** the income of each other group member by assigning **Deduction Points** to them. The other group members can also decrease your income, by allocating Deduction Points to you, if they wish to do so.

You can **assign up to 12 Deduction Points** in stage 1 and stage 3 together.

The group's deduction pool

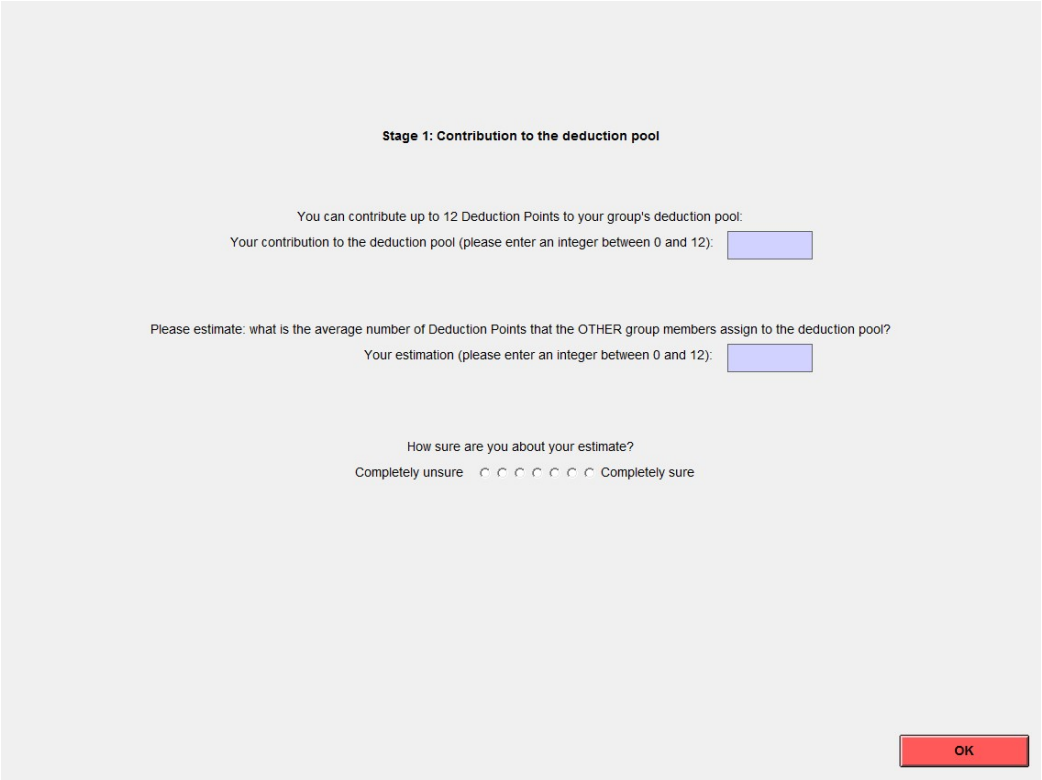
In the first stage you can assign between 0 and 12 Deduction Points to the group's deduction pool. **Each deduction point that you assign costs you one Point.** For example, if you assign 3 Deduction Points to the deduction pool then the cost to you is 3 Points.

The sum of Deduction Points in the group's deduction pool is assigned in equal shares to all group members who contribute 0 in the second stage.

Each Deduction Point that a particular group member receives from the deduction pool, decreases their income by 2 Points. For example, if there are 11 Deduction Points in the deduction pool and two group members contributed 0 in the second stage then 5.5 Deduction points will be assigned to each of them, and their income will be decreased by 11 Points each.

Your **own income will be reduced by 2 Point for each Deduction Point that is assigned to you** from the Deduction Pool.

You will see the following screen at stage 1:



As mentioned above, you can assign up to 12 Deduction Points to your group's deduction pool. You have to decide how many Deduction Points you assign to the deduction pool by typing a number between 0 and 12 (including 0 and 12) in the box. After entering the amount of Deduction Points you want to assign, you are asked to estimate the average number of Deduction Points that your group members assign to the deduction pool and indicate how confident you are that your estimation is true. After making all the entries, you must click on the "*OK*" button. Once you have done this, your decision can no longer be revised.

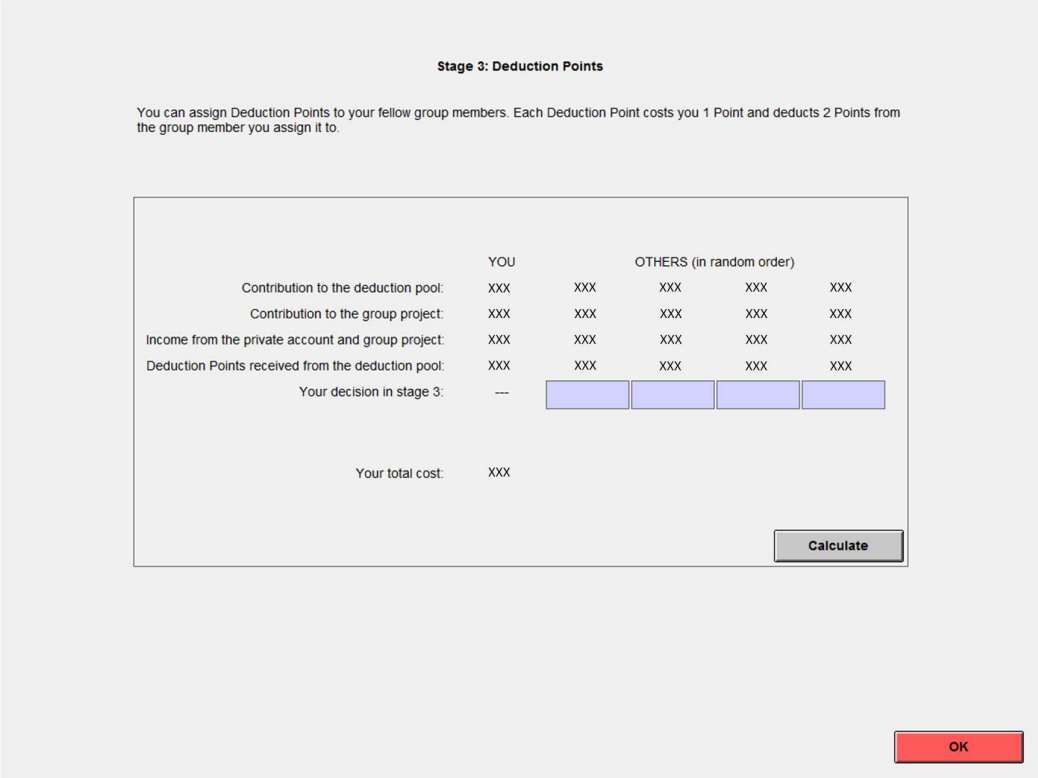
After everyone decided on their allocations to the deduction pool, you will learn the total amount of Deduction Points that your group assigned to the deduction pool.

In the second stage, you will make the contribution decision to the project.

Deduction Points

In stage 3 you can assign the rest of your 12 Deduction Points to the other group members. As before, each Deduction Point that you assign costs you one Point and decreases the income of the group member it is assigned to by 2 Points.

You will see the following screen at stage 3:



The column on the left shows your contribution to the deduction pool, your contribution to the group project, income from the private account and group project as well as Deduction Points you received from the deduction pool. The other four columns indicate the same for each of your group members.

If you do not wish to change the income of the other group members, type "0" into the fields next to "*Your decision in stage 3.*" In case you want to assign Deduction Points, enter the number of Deduction Points you want to assign into this field. You must enter a decision into every field and press the "*Calculate*" button. This will display the cost of your decision. Until you press the "*OK*" button, you can still change your decision. To recalculate the costs after making a change, simply press the "*Calculate*" button again.

The payoffs (stage 1, 2 and 3)

Your income in a given period can be illustrated by the following formula:

Your Total Income in Points = 20 − (*Your Contribution)* + 0.4 × (*Group Project*)
− (*Deduction Points Assigned By You To The Deduction Pool*)
− (*Deduction Points Assigned By You To Others*)
− 2 × (*Deduction Points Assigned To You From The Deduction Pool*)
− 2 × (*Sum Of Deduction Points Assigned To You By Others*)

You will only receive Deduction Points from the deduction pool if you contributed 0 at the second stage. Furthermore, in a given period you cannot lose more than you have earned through Deduction Points you receive from the deduction pool in stage 1 or other group members in stage 3. But you will always have to incur the cost of allocating Deduction Points.

The periods

This decision situation will be **repeated for thirty periods**. The groups are not going to change during these thirty periods. You will therefore be interacting with the same persons in all periods. Your incomes from the thirty periods will be summed up. This means that what you earn in each period will be added to what you earned in previous periods.

If you have any questions, please raise your hand and a member of the experiment team will come and answer them in private.

CHAPTER 5
Summary and Conclusion

This thesis explored individual cooperative dispositions, societal and cultural differences, as well as the design of sanctioning institutions as three important factors driving cooperative behaviour in humans. Chapter 2 investigated the first of the three factors focusing on the influence of individual cooperative dispositions on cooperative behaviour and peer-punishment. This chapter put a common implicit assumption of strong reciprocity (Gintis 2000) to the test. The assumption implies that people motivated by strong reciprocity are willing to engage in costly cooperation and punish defector at a private cost. Conversely, people with a disposition to free ride would never incur the cost of punishing defectors to the benefit of the whole group. This implicit assumption—although not explicitly discussed—is common in the economic literature of social preferences (e.g., Camerer and Fehr 2006).

The second factor—societal and cultural differences—and its influence on cooperative behaviour was the focus of Chapter 3. The investigation went beyond previous literature that looked merely at behavioural differences (e.g., Gächter et al. 2010) or differences in cooperative dispositions across countries (Kocher et al. 2008, Herrmann and Thöni 2009, Martinsson et al. 2013) by implementing and testing a comprehensive framework in which dispositions and beliefs drive cooperative behaviour. This procedure allowed us to identify the channels which influence cooperative behaviour. The chapter also measured the engagement in and the influence of altruistic peer-punishment across societies. In contrast to previous literature showing a vastly different punishment behaviour in repeated games across societies (Herrmann et al. 2008), the experimental game implemented in this chapter excludes strategic motives for punishment and rather elicits punishment preferences.

Chapter 4 explored the third factor—institutional difference—by focusing on the effectiveness and efficiency of formal and informal punishment mechanisms. Outside the laboratory both, formal and informal punishment mechanisms, are present and jointly influence behaviour, making it impossible to separate their relative effects. Chapter 4 reported on laboratory experiments designed to disentangle the relative influence of these mechanisms by implementing treatments with informal peer-punishment, formal pool-punishment and a combination of both sanctioning mechanisms. A further contribution of this chapter was to investigate the link between individual cooperative dispositions and the engagement in the different punishment mechanisms.

Drawing on the experimental evidence presented in this thesis, several key insights emerge on the influence of individual differences, societal and cultural differences, and institutions on cooperative behaviour, as well as the interaction between the three factors.

1. *Cooperative dispositions are not associated with peer-punishment*. In contrast to a common implicit assumption in the literature, cooperative dispositions are not needed to explain peer-punishment. Thus, the burden of fostering cooperation through costly peer-punishment is carried by a larger set of individuals than previously assumed. The relief of negative emotions is a likely driver of this behaviour independent of cooperative dispositions (e.g., Sanfey et al. 2003). These results were shown in Chapter 2 and also hold across different societies as is confirmed in Chapter 3. Additionally, in the repeated games of Chapter 4, participants with free-rider dispositions spend even higher amounts on punishment than participants with cooperative dispositions. Yet, this result is sensitive to the design of the sanctioning institution, with formal pool-punishment in Chapter 4 depending on cooperative dispositions.
2. *Beliefs are an important driver of societal differences in cooperative behaviour*. Societal differences in cooperative behaviour cannot be fully accounted for by differences in cooperative dispositions. All societies included in Chapter 3 are characterised by varying degrees of conditional cooperation on the aggregate level, and subjects with a disposition to conditionally cooperate are remarkably similar across societies. With conditional

cooperators making their own cooperative efforts contingent on that of others, beliefs are a key driver of societal differences in cooperative behaviour.

3. *Altruistic peer-punishment is similar across societies*. In all societies included in Chapter 3, peer-punishment was used to punish defectors. A small and similar expenditure on antisocial punishment occurred across societies in the one-shot experiment. This suggests that differences in antisocial punishment found in previous research are likely to stem from strategic considerations or feuds emerging in repeated interactions. The results from Chapter 3 imply that there are no innate differences in the preferences for antisocial punishment across societies.
4. *Peer-punishment is needed to stabilise cooperation even if formal punishment exists*. The comparison of punishment mechanisms conducted in Chapter 4 shows the importance of peer-punishment in fostering high and stable cooperation levels. The relatively high effectiveness and efficiency of this mechanism is likely to root in the credible threat of punishment of defection achieved at a low cost. Formal sanctions might promote a best-reply reasoning when making the contribution decision and the crowding out of voluntary contributions.

The experimental findings discussed above have implications for economic theory. They challenge the current understanding of strong reciprocity and call for a refinement of theoretical models of reciprocal behaviour. Strong reciprocity is not unidimensional since the two sides of strong reciprocity, strong positive reciprocity (the inclination to return kind actions) and strong negative reciprocity (the inclination to punish unkind actions), are not correlated. Although, some models do in theory separate between the two sides of reciprocity (e.g., Fehr and Schmidt 1999), others use a unidimensional representation of reciprocity (e.g., Falk and Fischbacher 2006). Thus, models of social preferences could be adapted to reflect this finding by using different parameters to represent the positive and the negative side of strong reciprocity.

The novel experimental designs used in this thesis constitute methodological contributions to experimental economics. The experimental design adopted in Chapter 2 allows for exploring the differences in the use of peer-punishment without altering the incentive to contribute to the public good in the first place. This is important because changing these incentives could in turn influence punishment behaviour. The

novel experimental design tackles this problem by holding the incentive to contribute on the first stage constant and thus allows for a clean comparison of punishment mechanisms.

Moreover, Chapter 3 uses the 'ABC of cooperation' framework (Gächter et al. 2017) in a cross-societal setting and further confirms the validity of this method. This framework is useful to compare differences in the channels that influence behaviour and thus allows us to go beyond comparing behaviour only. It has been previously used to explain the decline of behaviour in repeated public goods games (Fischbacher and Gächter 2010), as well as to investigate structural differences in the dilemmas of providing and maintaining a public good (Gächter et al. 2017). This thesis is the first attempt to apply this framework to societal and cultural differences and to explore how they affect cooperative behaviour.

The main results of this thesis can also support policy makers faced with the problem of free riding on the provision of public goods, which is prevalent in all societies. Chapter 3 highlights that—although cooperative behaviour might differ across societies—this is not due to differing cooperative dispositions. Policies aiming at shifting the beliefs about other people's behaviour are therefore a promising measure to impact cooperative behaviour. For example, advertising cooperative behaviour and leading by example might shift beliefs and influence behaviour. This is the case for all societies included in the investigation of Chapter 3.

The comparison of sanctioning institutions in Chapter 4 suggests that the threat of peer-punishment is important in fostering behaviour in line with social norms. This does not mean that formal punishment in the real world is obsolete. Especially in large-scale societies, these institutions are vital in ensuring the rule of law. However, the comparison of sanctioning institutions highlights a problem inherent to formal institutions: the possible crowding out of voluntary compliance. The challenge is therefore to optimise existing institutions and design mechanisms that minimise this factor. Additionally, Chapter 4 suggests that the institutional design interacts with individual cooperative dispositions. Therefore, policy measures should take into account that people have different cooperative dispositions. Here the peer-punishment mechanism does better, because the cost of punishment is shared independent of the

individual cooperative disposition. It would be important to engage subjects independent of their cooperative dispositions.

It has become clear that exploring the effect of individual differences, societal and cultural differences and institutions can lead to important new insights on cooperative behaviour. Yet, further research is needed to extend the findings presented in this thesis in several ways. Our research on societal differences is based on four countries only. An experimental study on a larger scale which includes more societies from different cultural clusters would be vital to generalise the findings. Additionally, further research is needed in the interaction between institutions and cooperative dispositions. This can lead to the design or more efficient policy and institutions. Lastly, interactions between culture and the design of sanctioning institutions could not be explored with the data presented in this thesis. Since the effectiveness and efficiency of sanctioning institutions might depend on the societal and cultural background, this could be vital factor in explaining a varying effect of institutions across societies and would constitute an interesting field for future research.

Bibliography

Andreoni, J. and L. K. Gee (2012): 'Gun for hire: Delegated enforcement and peer punishment in public goods provision'. *Journal of Public Economics,* Vol. 96, No. 11–12: 1036–1046.

Balafoutas, L., N. Nikiforakis and B. Rockenbach (2016): 'Altruistic punishment does not increase with the severity of norm violations in the field'. *Nature Communications,* Vol. 7, No. 13327.

Baldassarri, D. and G. Grossman (2011): 'Centralized sanctioning and legitimate authority promote cooperation in humans'. *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 108, No. 27: 11023–11027.

Balliet, D., L. B. Mulder and P. A. M. Van Lange (2011): 'Reward, punishment, and cooperation: A meta-analysis'. *Psychological Bulletin,* Vol. 137, No. 4: 594–615.

Ben-Shakhar, G., G. Bornstein, A. Hopfensitz and F. van Winden (2007): 'Reciprocity and emotions in bargaining using physiological and self-report measures'. *Journal of Economic Psychology,* Vol. 28, No. 3: 314–323.

Blanco, M., D. Engelmann, A. K. Koch and H. T. Normann (2010): 'Belief elicitation in experiments: is there a hedging problem?' *Experimental Economics,* Vol. 13, No. 4: 412–438.

Bolton, G. E. and A. Ockenfels (2000): 'ERC: A theory of equity, reciprocity, and competition'. *American Economic Review,* Vol. 90, No. 1: 166–193.

Bosman, R. and F. van Winden (2002): 'Emotional hazard in a power-to-take experiment'. *Economic Journal,* Vol. 112, No. 476: 147–169.

Bowles, S. and H. Gintis (1987). *Democracy and capitalism: Property, community, and the contradictions of modern social thought*. New York, NY: Basic Books.

Bowles, S. and H. Gintis (2004): 'The evolution of strong reciprocity: cooperation in heterogeneous populations'. *Theoretical Population Biology,* Vol. 65, No. 1: 17–28.

—— (2011). *A cooperative species: Human reciprocity and its evolution*. Princeton, NJ: Princeton University Press.

Boyd, R., H. Gintis, S. Bowles and P. J. Richerson (2003): 'The evolution of altruistic punishment'. *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 100, No. 6: 3531–3535.

Burlando, R. M. and F. Guala (2005): 'Heterogeneous agents in public goods experiments'. *Experimental Economics,* Vol. 8, No. 1: 35–54.

Burton-Chellew, M. N., C. El Mouden and S. A. West (2016): 'Conditional cooperation and confusion in public-goods experiments'. *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 113, No. 5: 1291–1296.

Camerer, C. F. and E. Fehr (2006): 'When does "economic man" dominate social behavior?' *Science,* Vol. 311, No. 5757: 47–52.

Carpenter, J. P. (2007): 'The demand for punishment'. *Journal of Economic Behavior & Organization,* Vol. 62, No. 4: 522–542.

Chaudhuri, A. (2011): 'Sustaining cooperation in laboratory public goods experiments: a selective survey of the literature'. *Experimental Economics,* Vol. 14, No. 1: 47–83.

Cubitt, R. P., M. Drouvelis and S. Gächter (2011): 'Framing and free riding: emotional responses and punishment in social dilemma games'. *Experimental Economics,* Vol. 14, No. 2: 254–272.

de Quervain, D. J. F., U. Fischbacher, V. Treyer, M. Schelthammer, U. Schnyder, A. Buck and E. Fehr (2004): 'The neural basis of altruistic punishment'. *Science,* Vol. 305, No. 5688: 1254–1258.

Dufwenberg, M., S. Gachter and H. Hennig-Schmidt (2011): 'The framing of games and the psychology of play'. *Games and Economic Behavior,* Vol. 73, No. 2: 459–478.

Egas, M. and A. Riedl (2008): 'The economics of altruistic punishment and the maintenance of cooperation'. *Proceedings of the Royal Society B-Biological Sciences,* Vol. 275, No. 1637: 871–878.

Egloff, B., D. Richter and S. C. Schmukle (2013): 'Need for conclusive evidence that positive and negative reciprocity are unrelated'. *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 110, No. 9: 786–786.

Eriksson, K., D. Cownden, M. Ehn and P. Strimling (2014): '"Altruistic" and "antisocial" punishers are one and the same'. *Review of Behavioral Economics,* Vol. 1, No. 3: 209–221.

Falk, A., E. Fehr and U. Fischbacher (2005): 'Driving forces behind informal sanctions'. *Econometrica,* Vol. 73, No. 6: 2017–2030.

Falk, A. and U. Fischbacher (2006): 'A theory of reciprocity'. *Games and Economic Behavior,* Vol. 54, No. 2: 293–315.

Fehr, E. and U. Fischbacher (2003): 'The nature of human altruism'. *Nature,* Vol. 425, No. 6960: 785–791.

—— (2004): 'Third-party punishment and social norms'. *Evolution and Human Behavior,* Vol. 25, No. 2: 63–87.

Fehr, E., U. Fischbacher and S. Gächter (2002): 'Strong reciprocity, human cooperation, and the enforcement of social norms'. *Human Nature,* Vol. 13, No. 1: 1–25.

Fehr, E. and S. Gächter (2000): 'Cooperation and punishment in public goods experiments'. *American Economic Review,* Vol. 90, No. 4: 980–994.

—— (2002): 'Altruistic punishment in humans'. *Nature,* Vol. 415, No. 6868: 137–140.

Fehr, E. and K. M. Schmidt (1999): 'A theory of fairness, competition, and cooperation'. *Quarterly Journal of Economics,* Vol. 114, No. 3: 817–868.

Fischbacher, U. (2007): 'z-Tree: Zurich toolbox for ready-made economic experiments'. *Experimental Economics,* Vol. 10, No. 2: 171–178.

Fischbacher, U. and S. Gächter (2010): 'Social preferences, beliefs, and the dynamics of free riding in Public Goods Experiments'. *American Economic Review,* Vol. 100, No. 1: 541–556.

Fischbacher, U., S. Gächter and E. Fehr (2001): 'Are people conditionally cooperative? Evidence from a public goods experiment'. *Economics Letters,* Vol. 71, No. 3: 397–404.

Fischbacher, U., S. Gächter and S. Quercia (2012): 'The behavioral validity of the strategy method in public good experiments'. *Journal of Economic Psychology,* Vol. 33, No. 4: 897–913.

Fosgaard, T. R., L. G. Hansen and E. Wengstrom (2017): 'Framing and misperception in Public Good Experiments'. *Scandinavian Journal of Economics,* Vol. 119, No. 2: 435–456.

Gächter, S. (2012): 'In the lab and the field: Punishment is rare in equilibrium'. *Behavioral and Brain Sciences,* Vol. 35, No. 1.

Gächter, S. and B. Herrmann (2009): 'Reciprocity, culture and human cooperation: previous insights and a new cross-cultural experiment'. *Philosophical Transactions of the Royal Society B: Biological Sciences,* Vol. 364, No. 1518: 791–806.

Gächter, S., B. Herrmann and C. Thöni (2010): 'Culture and cooperation'. *Philosophical Transactions of the Royal Society B: Biological Sciences,* Vol. 365, No. 1553: 2651–2661.

Gächter, S., F. Kölle and S. Quercia (2017): 'Reciprocity and the tragedies of maintaining and providing the commons'. *Nature Human Behaviour,* Vol. 1: 650–656.

Gächter, S. and E. Renner (2010): 'The effects of (incentivized) belief elicitation in public goods experiments'. *Experimental Economics,* Vol. 13, No. 3: 364–377.

Gächter, S., E. Renner and M. Sefton (2008): 'The long-run benefits of punishment'. *Science,* Vol. 322, No. 5907: 1510–1510.

Gächter, S. and J. F. Schulz (2016): Data from: 'Intrinsic honesty and the prevalence of rule violations across societies', Dryad Data Repository. DOI: 10.5061/ dryad.9k358.

—— (2016): 'Intrinsic honesty and the prevalence of rule violations across societies'. *Nature,* Vol. 531, No. 7595: 496–499.

Gintis, H. (2000): 'Strong reciprocity and human sociality'. *Journal of Theoretical Biology,* Vol. 206, No. 2: 169–179.

Gintis, H., S. Bowles, R. Boyd and E. Fehr (2005). *Moral sentiments and material interests: The foundations of cooperation in economic life*. Cambridge, MA: MIT Press.

Gintis, H., J. Henrich, S. Bowles, R. Boyd and E. Fehr (2008): 'Strong reciprocity and the roots of human morality'. *Social Justice Research,* Vol. 21, No. 2: 241–253.

Greif, A. (1994): 'Cultural beliefs and the organization of society: A historical and theoretical reflection on collectivist and individualist societies'. *Journal of Political Economy,* Vol. 102, No. 5: 912–950.

Greiner, B. (2015): 'Subject pool recruitment procedures: organizing experiments with ORSEE'. *Journal of the Economic Science Association,* Vol. 1, No. 1: 114–125.

Guala, F. (2012): 'Reciprocity: Weak or strong? What punishment experiments do (and do not) demonstrate'. *Behavioral and Brain Sciences,* Vol. 35, No. 1: 1–15.

Hardin, G. (1968): 'Tragedy of Commons'. *Science,* Vol. 162, No. 3859: 1243–1248.

Heckathorn, D. D. (1989): 'Collective action and the second-order free-rider problem'. *Rationality and Society,* Vol. 1, No. 1: 78–100.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, R. McElreath, M. Alvard, A. Barr, J. Ensminger, N. S. Henrich, K. Hill, F. Gil-White, M. Gurven, F. W. Marlowe, J. Q. Patton and D. Tracer (2005): '"Economic man" in cross-cultural perspective: Behavioral experiments in 15 small-scale societies'. *Behavioral and Brain Sciences,* Vol. 28, No. 6: 795–815.

Henrich, J., S. J. Heine and A. Norenzayan (2010): 'The weirdest people in the world?' *Behavioral and Brain Sciences,* Vol. 33, No. 2–3: 101–102.

Herrmann, B. and C. Thöni (2009): 'Measuring conditional cooperation: a replication study in Russia'. *Experimental Economics,* Vol. 12, No. 1: 87–92.

Herrmann, B., C. Thöni and S. Gächter (2008): 'Antisocial punishment across societies'. *Science,* Vol. 319, No. 5868: 1362–1367.

Hofstede, G. H. and G. Hofstede (2001). *Culture's consequences: Comparing values, behaviors, institutions and organizations across nations*. Thousand Oaks, CA: Sage Publications.

Hopfensitz, A. and E. Reuben (2009): 'The importance of emotions for the effectiveness of social punishment'. *Economic Journal,* Vol. 119, No. 540: 1534–1559.

Inglehart, R. and W. E. Baker (2000): 'Modernization, cultural change, and the persistence of traditional values'. *American Sociological Review,* Vol. 65, No. 1: 19–51.

Inglehart, R. and C. Welzel (2005). *Modernization, cultural change, and democracy: The human development sequence*. Cambridge, UK: Cambridge University Press.

Kamei, K., L. Putterman and J. R. Tyran (2015): 'State or nature? Endogenous formal versus informal sanctions in the voluntary provision of public goods'. *Experimental Economics,* Vol. 18, No. 1: 38–65.

Kaufmann, D. and A. Kraay (2016): Dataset: Worldwide Governance Indicators. Retrieved from: http://www.govindicators.org.

Kaufmann, D., A. Kraay and M. Mastruzzi (2011): 'The Worldwide Governance Indicators: Methodology and analytical issues'. *Hague Journal on the Rule of Law,* Vol. 3, No. 2: 220–246.

Knack, S. and P. Keefer (1997): 'Does social capital have an economic payoff? A cross-country investigation'. *Quarterly Journal of Economics,* Vol. 112, No. 4: 1251–1288.

Kocher, M. G., T. Cherry, S. Kroll, R. J. Netzer and M. Sutter (2008): 'Conditional cooperation on three continents'. *Economics Letters,* Vol. 101, No. 3: 175–178.

Kurzban, R. and D. Houser (2005): 'Experiments investigating cooperative types in humans: A complement to evolutionary theory and simulations'. *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 102, No. 5: 1803–1807.

Ledyard, J. (1995). Public goods: A survey of experimental research. In *Handbook of Experimental Economics*. J. Kagel and A. Roth (Eds.), pp. 111–194. Princeton, NJ: Princeton University Press.

Markussen, T., L. Putterman and J. R. Tyran (2014): 'Self-organization for collective action: An experimental study of voting on sanction regimes'. *Review of Economic Studies,* Vol. 81, No. 1: 301–324.

Martinsson, P., P. K. Nam and C. Villegas-Palacio (2013): 'Conditional cooperation and disclosure in developing countries'. *Journal of Economic Psychology,* Vol. 34: 148–155.

Medick, V. (2013, 23 April): Hoeness tax evasion case a headache for Merkel. *Der Spiegel*. Retrieved from: http://www.spiegel.de.

Nelissen, R. M. A. and M. Zeelenberg (2009): 'Moral emotions as determinants of third-party punishment: Anger, guilt, and the functions of altruistic sanctions'. *Judgment and Decision Making,* Vol. 4, No. 7: 543–553.

Nikiforakis, N. and D. Engelmann (2011): 'Altruistic punishment and the threat of feuds'. *Journal of Economic Behavior & Organization,* Vol. 78, No. 3: 319–332.

Nikiforakis, N. and H. T. Normann (2008): 'A comparative statics analysis of punishment in public-good experiments'. *Experimental Economics,* Vol. 11, No. 4: 358–369.

Oltermann, P. (2014, 13 March): Uli Hoeness sentenced to three-and-a-half years in jail for tax evasion. *The Guardian*. Retrieved from: http://www.theguardian.com.

Panchanathan, K. and R. Boyd (2004): 'Indirect reciprocity can stabilize cooperation without the second-order free rider problem'. *Nature,* Vol. 432, No. 7016: 499–502.

Peysakhovich, A., M. A. Nowak and D. G. Rand (2014): 'Humans display a "cooperative phenotype" that is domain general and temporally stable'. *Nature Communications,* Vol. 5, No. 4939.

Rand, D. G. and M. A. Nowak (2011): 'The evolution of antisocial punishment in optional public goods games'. *Nature Communications,* Vol. 2, No. 1442.

Reuben, E. and A. Riedl (2013): 'Enforcement of contribution norms in public good games with heterogeneous populations'. *Games and Economic Behavior,* Vol. 77, No. 1: 122–137.

Roth, A. E., V. Prasnikar, M. Okunofujiwara and S. Zamir (1991): 'Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An experimental study'. *American Economic Review,* Vol. 81, No. 5: 1068–1095.

Sanfey, A. G., J. K. Rilling, J. A. Aronson, L. E. Nystrom and J. D. Cohen (2003): 'The neural basis of economic decision-making in the ultimatum game'. *Science,* Vol. 300, No. 5626: 1755–1758.

Schöneberg, K. (2014, 13 March): Der Kick, das pure Adrenalin. *Die Tageszeitung*. Retrieved from: http://www.taz.de.

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

Shear, M. D. (2017, 1 June): Trump will withdraw U.S. from Paris Climate Agreement. *The New York Times*. Retrieved from: http://www.nytimes.com.

Traulsen, A., T. Rohl and M. Milinski (2012): 'An economic experiment reveals that humans prefer pool punishment to maintain the commons'. *Proceedings of the Royal Society B-Biological Sciences,* Vol. 279, No. 1743: 3716–3721.

Volk, S., C. Thöni and W. Ruigrok (2012): 'Temporal stability and psychological foundations of cooperation preferences'. *Journal of Economic Behavior & Organization,* Vol. 81, No. 2: 664–676.

West, S. A., A. S. Griffin and A. Gardner (2007): 'Social semantics: altruism, cooperation, mutualism, strong reciprocity and group selection'. *Journal of Evolutionary Biology,* Vol. 20, No. 2: 415–432.

World Values Survey Association (2014): World Values Survey Wave 5 2005–2008.

Yamagishi, T. (1986): 'The provision of a sanctioning system as a public good'. *Journal of Personality and Social Psychology,* Vol. 51, No. 1: 110–116.

Yamagishi, T., Y. Horita, N. Mifune, H. Hashimoto, Y. Li, M. Shinada, A. Miura, K. Inukai, H. Takagishi and D. Simunovic (2012): 'Rejection of unfair offers in the ultimatum game is no evidence of strong reciprocity'. *Proceedings of the National Academy of Sciences of the United States of America,* Vol. 109, No. 50: 20364–20368.

Yamagishi, T., Y. Li, H. Takagishi, Y. Matsumoto and T. Kiyonari (2014): 'In search of Homo economicus'. *Psychological Science,* Vol. 25, No. 9: 1699–1711.

Zhang, B. Y., C. Li, H. De Silva, P. Bednarik and K. Sigmund (2014): 'The evolution of sanctioning institutions: an experimental approach to the social contract'. *Experimental Economics,* Vol. 17, No. 2: 285–303.