FAIRNESS IN THE LAB

THE EFFECTS OF NORM ENFORCEMENT IN ECONOMIC DECISIONS

ISBN 90 5170 681 2
Cover design: Crasborn Graphic Designers bno, Valkenburg a.d. Geul

This book is no. **370** of the Tinbergen Institute Research Series, established through cooperation between Thela Thesis and the Tinbergen Institute. A list of books which already appeared in the series can be found in the back.

Fairness in the Lab

The Effects of Norm Enforcement in Economic Decisions

Academisch Proefschrift
ter verkrijging van de graad van doctor
aan de Universiteit van Amsterdam
op gezag van de Rector Magnificus
prof. mr. P.F. van der Heijden
ten overstaan van een door het college voor promoties ingestelde
commissie, in het openbaar te verdedigen in de Aula der Universiteit
op donderdag 16 februari 2006, te 12:00 uur

door

Ernesto Guillermo Reuben París

geboren te San José, Costa Rica

Promotiecommissie

Promotor:	prof. dr. F.A.A.M. van Winden
Overige leden:	prof. dr. C.K.W. De Dreu
	prof. dr. S. Gächter
	prof. dr. J. Potters
	prof. dr. H. Oosterbeek
	prof. dr. A.J.H.C. Schram

To my parents

Foreword

I acknowledge with gratitude the big debts I owe to many people who have contributed directly or indirectly to the completion of this dissertation. First, I would like to deeply thank my parents, William Reuben and Victoria París, for their unconditional support and loving guidance. They taught me to follow my dreams irrespective of how far away they might take me and how hard I might need work to accomplish them. They both have shown me, with their lives, the significance of perseverance, high ideals, but most importantly of kindness and love.

My family will always play an important role in my life. All of them have always, in one way or another, encouraged me to go forward. My grandparents have always been a good example of how one ought to lead one's life and strike that delicate balance between enjoyment and accomplishment that is so hard to achieve. I would also like to thank my brother for first, not becoming an economist (one too many of those in the family), and second for, in his own way, always supporting me. I should also thank him as well as my cousins, uncles, aunts, etc. for (constantly) reminding me that economists can be and often are wrong, and hence to keep an open mind to new and refreshing ways of thinking.

I would not have been able to even start writing this thesis if it had not been for the encouragement and guidance of my supervisor, Frans van Winden. He has patiently taught me the meaning of scientific research. His thoughtful comments and wise suggestions have helped me overcome the most challenging problems. Moreover, his enthusiasm and youthful curiosity have taught me to truly enjoy my work. From him I have learned that doing research is truly fascinating if one is not afraid of trekking into unexplored territories.

I also owe many thanks to the people of CREED: Aljaz, Arthur, Eva, Gijs, Jacob, Jens, Joep, Joris, Jos, Karin, Matthijs, Peter, Theo, and Vjollca. They have made every day at work interesting and truly enjoyable. Moreover, their helpful remarks have always provided me with new insights that have considerably improved the quality of this dissertation. A special mention is deserved by two CREEDers, Arno and Astrid, who not only encouraged and advised me, but also spent countless hours working with me in various research projects. They have made joint-work effortless and a pleasure to take part in. I am sincerely honored to be part of such a talented research group.

Throughout my dissertation, I have interacted with many individuals who have, sometimes inadvertently, given me valuable advice: Charles, Dirk, Enrique, Martin, Martine, Massimo, Nick, Nikos, Peter, Randolph, Rupert, Sandra, and Simon. To them I am grateful. I would like to highlight the help I received from Benedikt and Christian. My long conversations with them helped me sharpen my thoughts and polish my ideas.

Not necessarily related to this dissertation but nevertheless important has been the support I have received from everyone at the Tinbergen Institute. I would like to thank Maarten for giving me the opportunity to come here and to many others for making the Netherlands feel like a second home. The many hours of study and the nerve-wrecking exams were only bearable with the help of Carla, Emily, Maria, Marcos, Martijn, Miguel, Robert, Sandra, Sebi, Wendy, and many others. Triple tours have both helped me relax and get to know Amsterdam to a considerable detail. Equally important have been my friends in Costa Rica who, are too many to name, but have, by putting up with me at a considerable distance, demonstrated what true friendship is really all about.

I am especially grateful for having the opportunity to live with two great friends. Both Hugo and Astrid have provided me with a family in a foreign country. They, as well as Stephane, are also responsible for the improvement of my limited cooking skills. In spite of our inevitable separation, I hope we will often come together to take pleasure in future Sunday dinners.

These acknowledgements would be incomplete if I did not thank Ana. She of all people has been by my side throughout this project. She has always supported me by listening, unflinching, as I talk (and talk) about my research. In a myriad of cases, she has read and corrected my writing and in some instances my thinking. However, her most important contribution has been to be simply by my side, encouraging me with her gentle nature and loving ways. For the better part of the writing of this thesis, we have shared many laughs and enjoyed many moments. Fortunately, it is only the beginning.

Contents

Chapter 1

Motivation and Outline

The existence of fairness norms is an elusive and yet incredibly important characteristic of our societies. From the moment they are born, people are taught the merits of acting fairly. We commonly praise fair behavior and we deeply disapprove of unfair behavior. However, in spite of their importance, fairness norms are little understood. We still have to learn how fairness norms are formed, how their content is determined, and in what situations they play an important role. This thesis takes a closer look at the enforcement of fairness norms in order to prove an answer to this type of questions.

In many situations of interest to economists, fairness norms affect behavior in significant ways. For instance, the unwillingness of individuals to cheat or take advantage of others promotes cooperation in social dilemmas (Dawes, 1980; Elster, 1989), reduces shirking in workplaces (Fehr et al., 1993; Fehr et al., 1998), and motivates people to pay taxes (Andreoni et al., 1998; Alm et al., 1995). Moreover, the willingness of individuals to enforce fair behavior, even at a cost to themselves, can have significant effects on bargaining outcomes (Güth et al., 1982), the settling of disputes (Ellickson, 1994), and the provision of public goods (Fehr and Gächter, 2000b). Failure to properly understand the effects of fairness norms will prevent us from accurately predicting behavior and from implementing beneficial policies in these and many other cases.

A key insight into norm-guided behavior is that individuals do not blindly follow fairness norms. Instead, they estimate what the costs and benefits of their actions are and behave fairly only when they gain from doing so (Fehr and Schmidt, 2000). For example, although people might experience guilt if they call in sick to work in order to leave on vacation, they will nevertheless do so if the joy of vacationing outweighs the feelings of remorse (and there is no chance of being caught). Hence, although most individuals care about fairness, in some cases they behave selfishly.

An effective way of limiting unfair behavior is to allow individuals to sanction each other. Numerous experiments as well as casual observation demonstrate that people are willing to spend their own resources in order to punish those who act unfairly (Camerer, 2003).[1] Consequently, the availability of punishment opportunities provides individuals with a strong incentive to behave in accordance to fairness norms. In cases where selfish behavior

[1] Given that in many situations, punishment does not bring benefits to the punisher, it is still hotly debated what are the ultimate causes of this type of behavior (e.g. Fehr and Henrich, 2004). However, in this thesis, we concentrate on the proximate causes of norm enforcement.

leads to undesirable outcomes, such as the depletion of common-pool resources, the increased adherence to fairness norms induced by punishment has the added value that it can increase welfare (Ostrom et al., 1992; Fehr and Gächter, 2000b).

Punishment of unfair behavior, however, does not always make individuals better off. It only does so when benefits from additional compliance to the fairness norm outweigh the losses produced by punishment. For example, punishment has been shown to produce lower average earnings when it is very costly (Nikiforakis and Normann, 2005) or in cases where there is no repeated interaction (Egas and Riedl, 2005). Other possibly undesirable consequences of punishment are the destruction of resources due to conflicting fairness norms and the crowding out of positive reciprocity (see Knez and Camerer, 1995, and Fehr and Rockenbach, 2003). So far, our understanding of how fairness norms are enforced is too limited, and therefore, we cannot systematically predict whether punishment will have desirable or undesirable effects.

In contrast to the empirical evidence, standard economic theory (assuming self-regarding preferences) does not predict that individuals punish norm violations. Consequently, it does not give us insights into the effectiveness or desirability of punishment in different instances. For this reason, new theories are being developed to explain the willingness of individuals to enforce fairness norms. An important branch of this literature consists of theories that assume individuals possess other-regarding preferences. In other words, they assume individuals care about the earnings of others and how those earnings are achieved (for an overview, see Fehr and Schmidt, 2000). However, even though these theories successfully explain behavior in many experiments, there are many instances in which they fail to predict the way individuals punish.

In order to improve our knowledge of norm enforcement, we need a better understanding of the motivations and reactions of both the punishers and the punished. A promising line of research in this respect is the study of how emotions affect decision-making (Fehr et al., 2005). Although there is theoretical work on the effects of emotions in various situations,[2] there is some discussion on whether these models adequately capture experienced emotions (Elster, 1998). What is needed is empirical work that measures emotions and their relationship with behavior. Recent studies have revealed that emotions, particularly anger, play a crucial role by motivating individuals to punish unfair behavior in two-player games (Bosman and van Winden, 2002; Quervain et al., 2004). In this thesis, we extend this line of research in two ways. First by investigating the role of emotions as suppressors of unfair behavior, and second, by exploring whether emotions help individuals overcome the second-order public-good nature of norm enforcement.

Further insights on norm enforcement can be obtained by studying punishment in situations where different fairness notions can be applied. The majority of the experimental

[2] To name a few examples, economists have explored the value of emotions as reliable signals (Frank, 1987), their use for harnessing political support (Glaeser, 2005), their role in limiting tax evasion (Erard and Feinstein, 1994), and even their effect on the willingness to reject low offers in bargaining games (Kirchsteiger, 1994).

research on punishment focuses on games where an outcome satisfies various fairness norms. For example, in both ultimatum-bargaining games and symmetric public good games with punishment there is an outcome that maximizes efficiency, gives everyone the same earnings, and splits equally any surplus generated in the game. This gives subjects a clear reference point from which to judge unfairness. However, it is also important to study punishment behavior in cases where no such outcome exists. First, people commonly interact in situations where different interpretations of fairness imply different behavior. Second, these types of situations allow us to observe how much importance people attach to different fairness notions, and they permit us to compare competing theoretical models (see Charness and Rabin, 2002; Engelmann and Strobel, 2004). In this thesis, we study punishment in public good games where there is within-group heterogeneity. We introduce heterogeneity along two dimensions, first with respect to endowments and second with respect to the benefits from the public good. This allows us to observe the cooperation levels that are enforced in cases where efficiency, earnings equality, and the equal distribution of the surplus, cannot be simultaneously achieved.

As our research method, we use laboratory experiments. This allows us to study behavior and test various theoretical models in a controlled environment, which is especially useful when studying norm enforcement. In everyday life, people have numerous ways of punishing others in order to enforce a fairness norm. In many instances, such as public displays of disapproval, punishment is hard to observe and even if it is detectable, the costs to both the punisher and the punished are difficult to determine. In the laboratory, we are able to observe all instances of punishment, the precise behavior that motivated the punishment, and the (monetary) cost to the punisher as well as the punished. This allows us to isolate a specific aspect of norm enforcement and to analyze it in detail. As was mentioned, in addition to studying what individuals do, we also investigate the emotional motivations behind their behavior. We measure emotions through self-reports, which is an often used technique in social psychology (Robinson and Clore, 2002). Self-report measures have been shown to be correlated with physiological measures of arousal (e.g. Ben-Shakhar et al., 2004). Furthermore, they allow us to easily distinguish different emotions. We also use self-reports to measure fairness perceptions and expectations.

In the thesis, we present five different experiments (one per chapter) analyzing different aspects of norm enforcement. In chapters 2 through 4, we focus on how emotions motivate individuals to enforce and comply with fairness norms. In chapters 5 and 6, we analyze the effects of different types of heterogeneity on cooperation and punishment in public good games. In each chapter, we motivate the research question and discuss the related literature. After that, we describe the experimental design and the corresponding predictions of various theoretical models.[3] We proceed with the analysis of the data, after which, we conclude with a discussion of the main results and of how they can help improve the

[3] We concentrate on theories that successfully explain punishment behavior in similar situations.

modeling of norm enforcement. Appendixes are provided with the experiments' instructions and some descriptive statistics. A discussion of the conclusions is provided in the final chapter. Next, we give a brief overview of the different chapters and their main conclusions.

In Chapter 2 we study of the links between emotions and punishment behavior in a power-to-take game where there is more than one punisher.[4] Furthermore, we investigate whether social ties between punishers affects their emotional and behavioral response. Introducing more than one punisher allows us to determine the effects of emotions on punishment behavior when there are possibilities to free ride on the punishment of others. Our results indicate that anger-like emotions are the motivating force behind the decision to punish. Moreover, we find that the existence of a social tie between punishers produces different emotional reactions. The emotional reaction of individuals with a social tie facilitates coordination on punishment. In contrast, the emotional reaction of individuals who do not share a social tie hinders punishment This leads to a higher amount of overall punishment that makes opportunist behavior unprofitable.

In Chapter 3 we investigate the motivations and behavior of individuals who *receive* punishment. In particular, we concentrate on how their willingness to behave in a fair manner depends on the amount of punishment they receive, their fairness perceptions, and experienced emotions. In this chapter, we also test whether individuals who are willing to enforce fair behavior are also more likely to behave fairly. We find that punishment combined with deviations from a perceived fairness norm trigger feelings of shame and guilt. This induces individuals to behave more fairly in the future. However, we also find that fairness perceptions vary considerably between individuals. We do not observe that willingness to punish others translates into willingness to treat others fairly.

The behavior of individuals who receive punishment is studied further in Chapter 4. In this chapter, we concentrate on the willingness of punished individuals to retaliate against those who punish them. The goal of this chapter is to understand the type of motivations that must be present for punishment to be an effective institution for the support of cooperative behavior. We find that punishment sustains cooperation at high levels even though retaliation is commonly observed and it often results in an extreme reduction of payoffs. We also find that, as with people who punish unfair behavior, individuals who retaliate are motivated by anger. However, only those who feel anger and do not feel shame retaliate. Hence, by restraining punished individuals from fighting back, shame plays a crucial role in the enforcement of fairness norms.

In Chapter 5 we depart from the study of the relationship between emotions and punishment. In this chapter, we extend the study of punishment in public good settings to groups with heterogeneous endowments. In particular, we focus on how individuals punish depending on their endowment and the endowment of others and on how this affects

[4] In this way, we extend the line of research that focuses on emotions and punishment, and which has, so far, focused only on two-person games (Bosman and van Winden, 2002; Sanfey et al., 2003; Quervain et al., 2004).

cooperation. Our results indicate that endowment heterogeneity does not affect the ability of punishment to sustain cooperative behavior but it does affect its ability to do so efficiently. Our findings are explained by differences in punishment behavior between rich and poor individuals, which can be interpreted as the enforcement of different cooperation norms. Surprisingly, the cooperation norm that is enforced seems to depend on whether there is inequality in contribution possibilities and not on whether there is inequality in endowments.

In Chapter 6 we continue our study of public good provision by investigating behavior in so-called privileged groups (i.e. groups where some individuals lack an incentive to free ride). Privileged groups have the advantage that they have less free riding incentives. However, they suffer from the fact that high cooperation levels are no longer supported by fairness norms such as payoff equality or the rewarding of intentionally kind actions. Indeed, we find that punishment increases cooperation more in normal groups than in privileged groups. Surprisingly, the difference is not due to individuals giving less punishment or to reacting less when they receive punishment. In fact, it is due to individuals being unwilling to reciprocate high contributions. In normal groups, an individual who contributes a high amount induces others to contribute more (even if they are not punished). In privileged groups, high contributions do not produce the same response.

Finally, in Chapter 7 we briefly discuss how the findings of previous chapters shed light on the way people enforce and comply with fairness norms. Furthermore, we provide suggestions for future work.

In all, the goal of this thesis is to advance our understanding of how and how well fairness norms are enforced in economically relevant situations. Our results suggest that different emotions motivate individuals to enforce and comply with fairness norms. Furthermore, we find that in asymmetric situations, significantly different norms can be enforced depending on small differences in the game. The insights gained by the work presented in this thesis can hopefully improve the modeling of norm-guided behavior and allow us to better predict the consequences of fairness norms.

Chapter 2

Punishing with Friends

Social Ties, Emotions, and the Coordination of Punishment[*]

In this chapter, we study the links between emotions and punishment behavior in situations where there is more than one punisher. Specifically, we study the effect of social ties between individuals on their decision to punish a third party for behaving unkindly. Furthermore, we investigate whether this effect is explained by differences between the emotional reaction, beliefs, or fairness norms of individuals that share a tie and individuals who do not.

2.1 Introduction

Recently, a number of experimental studies have begun to explicitly investigate the links between emotions and punishment behavior (e.g. Bosman and van Winden, 2002; Sanfey et al., 2003; Quervain et al., 2004; Bosman et al., 2005b). They find substantial support for the economic significance of emotions, using various methods of measurement, and explain an individual's decision to punish as a tradeoff between the emotional satisfaction of punishing an opportunistic act and the (more cognitive) reward of a monetary gain.

The main purpose of this chapter is to extend this kind of analysis to cases where there are two instead of one punisher and where the punishers either do or do not know each other. In order to do this in a tractable setting we use a three-person version of the power-to-take game (Bosman and van Winden, 2002). In this game, a proposer (or take authority) can make a claim on the resources of one responder. Subsequently, the responder can destroy any part (including nothing and everything) of her own resources. In this chapter, we alter the game so that the proposer can make a claim on the resources of two instead of one responder. Having multiple responders makes this game more realistic for social environments characterized by appropriation of which the power-to-take game captures important aspects, such as taxation, common agency or monopolistic selling (see Bosman and van Winden, 2002). For illustration, one might think of a tax authority selecting an income tax rate, while taxpayers can destroy the income tax base (at a cost to themselves). Furthermore, we investigate whether knowing each other as responders affects the emotional and behavioral response.

The step from one to two responders is not at all trivial. On one hand, in the relationship between the proposer and a responder, two new factors are introduced that may affect the outcome of the game. First, if the benefit from punishing the proposer is

[*] This chapter is based on Reuben and van Winden (2004) and Reuben and van Winden (2005b).

independent of who does the punishing, an externality is introduced which opens up the opportunity to free ride with less punishment. Previous work on emotions does not tell us whether angry individuals will satisfy their desire to harm the proposers if someone else does the punishing. Furthermore, the possibility of taking money from two instead of one responder makes the situation between the proposer and each of the responders considerably more unequal. If this triggers a more intense emotional response then more punishment could result.

On the other hand, the presence of a second responder introduces another set of considerations, namely, the relationship between the two responders. Responders may care for their relative payoffs, and in addition, they may care for how they will feel about each other's behavioral response. In both cases, behavior may be affected by the belief of what the other responder is going to do. If a responder cares for relative payoffs,[5] then her belief regarding the destruction decision of the other responder will determine whether her destruction increases or decreases the expected payoff difference between the two of them. This could lead to either more or less destruction. Alternatively, if a responder is concerned with how her action will be viewed by the other responder, then how much the other responder will destroy becomes, also from this perspective, a reference point for evaluating her decision. For example, if a responder destroys much less than the other, she might be seen as not standing up to unfair behavior, and conversely, if she destroys much more, she might be considered foolish for overreacting. *A priori* it is not clear what the behavioral consequences will be.

Beliefs and emotional responses are likely to be affected by the kind of relationship that exists between the responders. For example, people put different weights on the opinions of others depending on the type and strength of the relationship between them. We care more for what our close friends think of us than what a complete stranger might think. Furthermore, it is to be expected that individuals also differ in the importance assigned to payoff differences depending on who the other person is. It would be natural to expect, for instance, that a responder would mind more if a friend gets a lower payoff than if a stranger does. Although there are a few experiments that study the effects of social distance by having subjects interact across different countries (e.g. Charness et al., 2003), there is practically no work on the effect of real social ties on the behavior of individuals in economic experiments.[6] Hence, it is hard to predict the behavioral consequences of responders being friends instead of strangers. Since ties seem to play an important role in collective action (see Chong, 1991) and are potentially relevant in many economic situations (e.g. work environments), the issue whether and how they affect behavior is in fact of much wider interest. We have therefore decided to give it a prominent place in our experimental design.

The chapter is organized as follows. In Section 2.2 we present the experimental design and link it to related studies. Section 2.3 describes the experimental procedures. Results are

[5] As in the models of Levine (1998), Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Charness and Rabin (2002), and Falk and Fischbacher (2005).

[6] An exception being Abbink et al. (2002)

presented in Section 2.4. Section 2.5 discusses the main results in relation to the existing literature. Section 2.6 concludes.

2.2 Design and Related Literature

For our study we use a three-person version of the power-to-take game (Bosman and van Winden, 2002). In this one-shot game one subject, who can be considered as the proposer (with endowment E^{prop}), is matched with a pair of other subjects, the responders (each with an endowment E_i^{resp} where $i \in \{1,2\}$ indicates the responder). The game consists of two stages. In the first stage, the proposer decides on the 'take rate' $t_i \in [0,1]$, which is the part of responder i's endowment after the second stage that will be transferred to the proposer. In the second stage, both responders decide simultaneously to destroy a part $d_i \in [0,1]$ of their own endowment. For the proposer the payoff of the game equals her endowment plus the transfer from each of the responders, i.e. $E^{prop} + t_1(1 - d_1)E_1^{resp} + t_2(1 - d_2)E_2^{resp}$, while responder i's payoff equals the part of her endowment that she does not destroy minus the amount transferred to the proposer, i.e. $(1 - t_i)(1 - d_i)E_i^{resp}$. In order not to introduce too many behavioral issues at a time, in our experiment proposers can only select a uniform take rate (that is, $t_1 = t_2 = t$) and all the endowments are equal ($E_1^{resp} = E_2^{resp} = E^{prop}$).[7]

To study the impact of responders knowing each other, the experiment consists of two treatments, one where responders are anonymous to each other ('strangers' treatment), and one where responders know each other ('friends' treatment). By comparing the results from the strangers treatment with earlier experiments involving only one responder, we can observe whether the presence of another responder appears to make a difference. By comparing the results from the strangers and friends treatments we can establish whether the existence of a tie between the responders makes a (further) difference. Furthermore, by using self-reports as research method for measuring emotions (Clore and Robinson, 2002) we can determine which emotions are important in these settings and analyze their explanatory value for observed behavior.

The simplicity of our design facilitates the study of the influence of emotions on behavior. First, each responder makes only one decision. This is useful since emotions can impact various decisions and it might be hard to disentangle which emotion influenced which decision. Second, responders cannot influence each other's monetary payoffs. Therefore, we are able to observe how a responder feels about the decision of the other responder without interference of any effect the other responder might have had on the first responder's income.

Our work is related, on the one hand, to studies exploring the economic significance of emotions and, on the other hand, to studies investigating how the presence of others affects decision-making. Although still small in number, there are some studies explicitly dealing

[7] The power-to-take game differs in three ways from the well-known ultimatum game. First, in the power-to-take game each participant has an endowment. Second, in this game only the endowment of the responder(s) is at stake. And third, the responders can destroy any amount of their endowment.

with emotions to explain responder behavior in the kind of game investigated in this chapter. However, they are all restricted to the one-proposer-one-responder case. A relatively early paper exploring this issue is Pillutla and Murnighan (1996). Responders in an ultimatum game experiment were asked, after each of a series of offers they had to accept or reject, to answer the open-ended question "How do you feel?" All offers were predetermined and afterwards a lottery selected one to determine actual payoffs. Answers to the feeling question were rated for expressions of the emotion of anger, and the rejection of offers was found to be related to this measure of anger. Their analysis also suggests that emotional reactions provide the critical link that determines when fairness perceptions tend to affect behavior. Bosman and van Winden (2002) introduced the power-to-take game with the specific purpose of explicitly investigating the importance of emotions for negative reciprocity in a situation of appropriation. In several experiments they had responders self-report on their feelings, but now concerning a whole list of different emotions (positive as well as negative) and with 7-point scales for them to indicate the felt intensity of the respective emotion (see also van Winden, 2001; Bosman et al., 2005b). In addition, they asked for responders' expectations (regarding the take rate). Their results show that the destruction of own resources by responders is related to the intensity of experienced negative emotions (particularly, contempt, irritation, and anger), which in turn is positively related to the actual take rate and negatively to the expected take rate.

Recently, for both games evidence has been found of a biological substrate for the negative reciprocity exhibited by responders. Sanfey et al. (2003), using fMRI of ultimatum game players, find that 'unfair' offers elicited activity in brain areas related to both emotion and cognition, and significantly heightened activity in an area related to emotions in case of rejection.[8] Regarding the power-to-take game, Ben-Shakhar et al. (2004), using skin conductance as physiological measure of emotional arousal as well as self-reports, find that both self-reported anger and physiological arousal are related to destruction, with frustrated expectations playing an important and consistent role. Moreover, the self-reported measures of emotions appeared to be correlated with the physiological measures, which is reassuring for the use of self-reports in the study of reciprocity.

In this study we take the important next step in this line of research by studying what happens when a third person is introduced. In this respect, our work is related to papers on three-person ultimatum games. For instance, various authors have conducted experiments using ultimatum games that involve an inactive dummy player (Güth and van Damme, 1998; Kagel and Wolfe, 2001; Bereby-Meyer and Niederle, 2005). They find that responders seem to concentrate on their own as well as the proposers' payoffs and mostly ignore the welfare of the dummy players. Knez and Camerer (1995) use the strategy method to observe if a pair of responders playing with the same proposer condition their acceptance on the amount offered

[8] In a similar study, Quervain et al. (2004) show that the effective punishment of norm violators produces activity in areas of the brain associated with the processing of rewards.

to the other responder. They find that about half of the responders will condition their response on the income the other responder would get. Riedl and Vyrastekova (2003) ran a three-person ultimatum game experiment in which they varied the effect the rejection of one responder has on the payoffs of another responder. They find that responders are more likely to reject proposals if this does not negatively affect their standing with respect to the other responder. However, these experiments were not designed for an analysis of emotions and their explanatory value. Hence, important variables from that perspective, such as expectations, were not measured. Our experimental design is a first shot at exploring head-on the affective side of reciprocity in case of multiple potential reciprocators.

Psychological studies suggest that people may react quite differently emotionally to the same situation when others are present, and the more so if the other person is a friend rather than a stranger (see e.g. Jakobs et al., 1999). We want to explore the economic relevance of this literature by investigating whether the presence of another responder in the power-to-take game and the nature of the relationship between the responders has an effect not only on their emotional responses but also on their behavior and, furthermore, whether the emotional response is linked to behavior.

2.3 Experimental Procedures

The computerized experiment was run in November 2003 and May 2005 in the CREED laboratory of the University of Amsterdam. In total 261 subjects, almost all undergraduate students from the University of Amsterdam, participated in the experiment. About 44% of the subjects were students of economics. The other 56% were students from various fields such as biology, political science, law, and psychology. About 43% of the subjects were female. Subjects received a show-up fee of 5 euros, independent of their earnings in the experiment, and 10 euros as endowment. On average, subjects were paid out 13.59 euros. The whole experiment took about one hour.

The experiment consisted of two treatments: a ‘strangers’ treatment, where the two responders in the game did not know each other, and a ‘friends’ treatment, where the responders knew each other. Recruitment for the friends treatment was done in the following way. Subjects were allowed to sign up only if they did so as a pair, that is, they had to provide the name of someone they knew and with whom they would take part in the experiment. If a subject signed up with someone else but nevertheless showed up alone to the experiment, he or she was not allowed to participate. In this way we hoped to recruit subjects with social ties. This approach, which is similar to the one used by Abbink et al. (2002), gives the opportunity to employ individuals with stronger bonds than one can establish in the laboratory. In an attempt to measure the strength of each pair’s social tie, we asked each individual to describe the type of relationship they had with their partner and how frequently they saw each other. For the strangers treatment, recruitment was done in two different ways. In half of the sessions, we recruited subjects in exactly the same way as in the friends treatment. We will refer to these sessions as ‘paired-strangers’. In the other half of the sessions we recruited

subjects independently as is done in most experiments. We will refer to these sessions as 'single-strangers'. We did this in order to check whether forcing people to attend the experiment in pairs attracts different subjects than the normal recruiting procedure. Knowing this is important when comparing our results to other experiments. However, we found no significant differences between the behavior, beliefs, or emotional reaction of subjects who were recruited in pairs and subjects who were recruited independently. Hence, for most of the analysis presented in this chapter we pool the data from all sessions of the strangers treatment.[9]

After arrival in the lab's reception room, each pair of subjects drew a card to be randomly assigned to two seats in the laboratory. Once everyone was seated the instructions for the (one-shot) power-to-take game were read, followed by a few exercises to check the subjects' understanding of the procedures (a translation of the instructions is provided in Appendix 2A). After these exercises the subjects were informed, by opening an envelope on their desk, to which role (that of proposer or responder) they had been randomly assigned. The game was framed as neutral as possible, avoiding any suggestive terms. Subsequently, the subjects were randomly assigned into groups of three. In the friends treatment, each group included a proposer and a pair of responders who signed up together for the experiment. Consequently, in this treatment anonymity was ensured between proposers and responders but not between the responders themselves. In other words, proposers knew that the responders in their group were people that came together to the experiment, but they did not know which responders they were. Similarly, responders knew that the other responder in their group was the individual with whom they came to the experiment whereas they did not know the identity of the proposer. In the strangers treatment, subjects that came independently to the experiment were randomly assigned to groups, whereas subjects that came in pairs were assigned to *different* groups. In either case, complete anonymity was ensured since none of the members of a group knew who the other group members were. The group assignment was clearly explained in the instructions.

Subjects then played the three-person power-to-take game via the computer.[10] During the game, subjects were asked to fill out a few forms indicating not only their decisions but also how they felt, which take rate they expected, and which take rate they considered to be fair. Since in this chapter we will concentrate on the responders' behavior, Figure 2.1 shows the precise order in which the responders' decisions, emotions and expectations were measured. Note that we asked subjects to report what they expected others to do before they observed their actual behavior. As in Bosman and van Winden (2002) subjects' emotions towards other players were measured through self-reports after the subject observed what the others did. We asked for the fair take rate, at the end, in the debriefing questionnaire.

---

[9] In fact, all of our results hold whether we use only the data from the paired-strangers sessions (see Reuben and van Winden, 2005b) or only the data from the single-strangers sessions (see Reuben and van Winden, 2004).

[10] The experiment was programmed and conducted with the software z-Tree (Fischbacher, 1999).

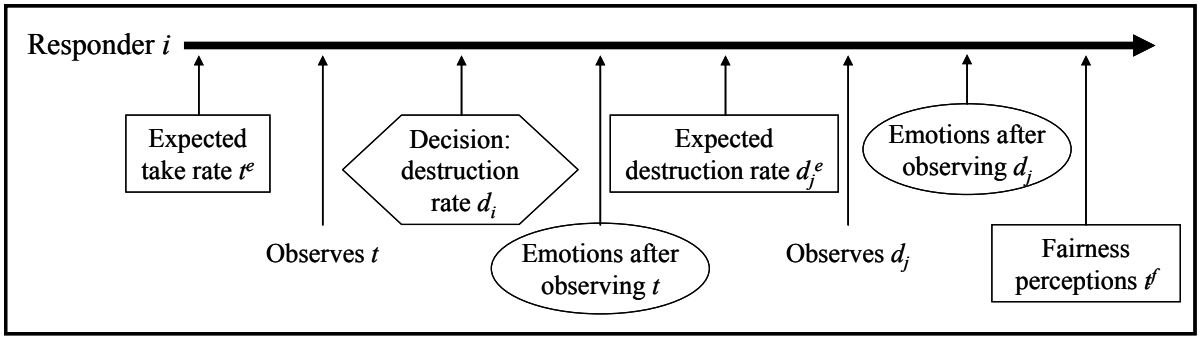

FIGURE 2.1 – SEQUENCE OF EVENTS FOR RESPONDERS

Expectations were measured by asking subjects to indicate the most likely value for t or d_j.[11] In addition to the point estimate, we asked subjects to indicate on a 7-point scale how confident they were of their expectation. Emotions were measured by providing subjects with a list of fourteen emotion names and asking them to report on a 7-point scale with what intensity they experienced each emotion. The scale ranged from "no intensity at all" (1) to "very intensely" (7). The list included the following emotions: pride, envy, anger, guilt, joy, shame, irritation, gratitude, surprise, contempt, disappointment, admiration, regret, and sadness. A variety of emotions was included to avoid pushing subjects in a particular direction.

2.4 Results

In this section we present and analyze the decisions that were taken and the emotions that subjects experienced. Furthermore, we investigate whether the reported emotions help explain the behavior of the responders in both treatments. A summary of the individual data is provided in Appendix 2B.

2.4.1 Observed behavior

On average, the take rate was 59.5% in the strangers treatment and 62.3% in the friends treatment. Interestingly, these take rates are very similar to the 60.0% mean take rate reported in the comparable one-responder power-to-take game (Bosman et al., 2005b). The similarity between the take rates in both treatments leads to our first result.

[11] We decided to measure expectations in this way since subjects might have difficulty in reporting a probability distribution of a continuous variable (over the interval [0,1]).

RESULT 2.1: *Take rates do not differ between the friends treatment and strangers treatment*.

Support: Using a Wilcoxon-Mann-Whitney (WMW) or a Kolmogorov-Smirnov test, the hypothesis that the take rates are drawn from the same distribution in both treatments cannot be rejected ($p = 0.861$ and $p = 0.722$).[12]

Turning now to the responders, in both treatments, a considerable number of responders destroyed some or all of their endowment. In the strangers (friends) treatment 26.0% (40.0%) of the responders destroyed a positive amount. On aggregate, responders in the strangers (friends) treatment destroyed 15.6% (29.4%) of their endowment. For both measures, these results are respectively below (strangers treatment) and above (friends treatment) the ones obtained for the one-responder game. In Bosman et al. (2005b) 37.5% of the responders destroyed a positive amount, while on aggregate 24.70% of the endowment was destroyed. The difference in destruction between strangers and friends is even starker if we concentrate on responders who faced high take rates, that is, take rates that are above the average take rate. This leads us to our second result.

RESULT 2.2: *Friends destroy more and more frequently than strangers. This difference is due to high destruction rates among friends when faced with high take rates*.

Support: Using WMW tests, one can reject the hypothesis that friends and strangers destroy equal quantities ($p = 0.025$) and equally often ($p = 0.052$). Of the responders who faced a take rate that was higher than the average take rate, strangers destroyed on aggregate less than friends, namely 30.9% vs. 67.3% ($p = 0.001$). Moreover, only 37.5% of the strangers destroyed some of their endowment whereas 78.6% of friends decided to do so ($p = 0.001$). There are no significant differences between friends and strangers for responders who faced below average take rates ($p > 0.531$). This result is partly driven by the fact that friends are more likely to destroy all of their endowment than strangers: 24.3% of the friends destroyed everything while only 9.6% of the strangers did so ($p = 0.025$).[13]

These two results suggest that, whereas at high take rates friends are likely to destroy more and more frequently than strangers, the behavior of proposers does not depend on whether

[12] Throughout the chapter, unless it is otherwise noted, we always use a Wilcoxon-Mann-Whitney test. Furthermore, all tests are two-sided.

[13] It would certainly be of interest to know if the strength of the social tie between responders has an effect on destruction. Unfortunately, we have very little variation in the two variables used to measure the strength of social ties, and hence, we cannot make a meaningful analysis. Roughly 60% of all pairs reported their type of relationship as a 'friendship' while the other 40% was evenly distributed among four categories. Similarly, 60% of all pairs reported their frequency of contact as 'very frequent' while the other 40% was evenly distributed among three categories.

they are facing a pair of friends or a pair of strangers. This can be further substantiated by looking at the earnings of proposers. In the strangers treatment, proposers who chose a high take rate earned on average 2.27 euros *more* than proposers who chose a low take rate. They also faced more risk, however, in the sense of a higher variance in earnings. In contrast, in the friends treatment, proposers who chose a high take rate earned on average 4.04 euros *less* than those choosing a low take rate, even though they too faced a larger variation in earnings. Hence, while in the strangers treatment it might make sense to choose a high take rate and risk some variation in income, in the friends case this is clearly an inferior choice. Nevertheless, it turns out that the proportion of proposers choosing a high take rate is the same in both treatments.

2.4.2 Determinants of behavior

To investigate what is motivating a responder to destroy, we estimated a multivariate tobit model for the probability of destruction, using as explanatory variables: demographic data (gender and area of study), the take rate, the expected take rate, the perceived fair take rate, and treatment dummies. In addition, we checked for interaction between the explanatory variables and used any significant interaction term (see Table 2C.1 in Appendix 2C). The following result is obtained.

RESULT 2.3: *Destruction is positively related to the take rate and to the difference between the actual and the expected take rate.*

Support: Judged by the signs of the significant coefficients ($p < 0.050$), it appears that responders who are likely to destroy some or all of their endowment are responders who: faced high take rates and/or experienced large positive differences between the take rate and their expected take rate. Furthermore, friends have a bigger coefficient than strangers for the influence of the take rate on the amount of destruction (Wald test, $p = 0.013$). This mirrors our previous result where we found that friends destroy more frequently than strangers when faced with high take rates. Contrary to what one would expect given the emphasis on fairness in the literature, but in agreement with the results of Pillutla and Murnighan (1996), the variable measuring the differencebetween the take rate and the fair take rate has a high *p*-value ($p = 0.052$). Finally, the regression also suggests that women destroy less than men do ($p = 0.022$).

2.4.3 Experienced Emotions

The intensity scores of emotions reveal that subjects experienced a variety of emotions. Concentrating on how emotions influenced the destruction decision gives us a clear result.

RESULT 2.4: *Destruction is positively (negatively) related to the intensity of experienced negative (positive) emotions*.

Support: We estimated a tobit model for each emotion separately, with destruction as the dependent variable and censored at $d = 0$ and $d = 100$ (robust standard errors and clustering within groups). The resulting coefficients and their level of significance are presented in Table 2.1.

The table shows that, in both treatments, similar negative (as well as positive) emotions are involved in destruction behavior. WMW tests give further support for this finding: in both treatments, responders who destroyed reported significantly higher intensities of anger, contempt, irritation, and disappointment, and significantly lower intensities of joy and gratitude ($p < 0.014$). Furthermore, for none of the emotions we find a significant difference in the average intensity scores between the two treatments (for both responders who destroyed and responders who did not destroy).

TABLE 2.1 – TOBIT REGRESSIONS WITH DESTRUCTION AS THE DEPENDENT VARIABLE

	Strangers		**Friends**	
Explanatory variable	**Coef.**	**Std. Err.**	**Coef.**	**Std. Err.**
disappointment	31.823**	10.208	56.900**	20.511
joy	–34.344**	11.478	–68.441**	25.891
gratitude	–36.929**	13.024	–87.711**	27.295
anger	29.699**	10.706	54.561**	18.602
contempt	29.037**	10.648	55.870**	15.437
irritation	24.814**	9.699	55.219**	16.898
admiration	–20.028*	12.061	–95.017**	36.771
envy	28.516**	10.863	15.280	13.777
sadness	17.541	11.597	49.216**	22.316
regret	18.060	15.627	11.786	27.264
shame	13.394	14.375	18.045	27.656
surprise	2.627	8.851	23.085	20.051
guilt	–4.988	16.886	–22.905	27.824
pride	2.517	9.731	25.921	18.256

Note: ** Significant at the 5 percent level. * Significant at 10 percent level.

Having found that destruction is related to experienced emotions, the question arises what explains the different emotional responses. To answer this question we estimated a multivariate ordered probit model, with the intensity of different positive and negative emotions as the dependent variable. For the analysis, we concentrated on the emotions that turned out to be good predictors of destruction in both treatments (i.e. disappointment, joy, gratitude, anger, contempt, irritation, and admiration). Again, we used as explanatory variables: demographic data (gender and area of study), take rate, expected take rate,

perceived fair take rate, and treatment dummies. In addition, any significant interaction term was included. The following result is obtained.

RESULT 2.5: *The intensity of negative (positive) emotions is positively (negatively) related to the take rate and to the difference between actual and expected take rates*.

Support: In all the regressions for negative (positive) emotions the coefficient of the take rate has a positive (negative) and significant sign ($p < 0.023$). The same holds true for the coefficient of the difference between the take rate and the expected take rate ($p < 0.036$). For illustration, in Table 2C.2 we present the results of one such regression. In this regression we used as dependent variable the average of the three anger-like emotions: anger, irritation, and contempt. The regression also points at a possible gender effect. It appears that female subjects are more likely to report lower intensities of anger-like emotions.[14]

Summarizing, the same variables that are good predictors of destruction behavior (see Result 2.3) are also good predictors of the intensities of experienced emotions. In combination with the finding that emotions are good predictors of destruction, this suggests the following intuitive explanation for destruction: the higher the take rate and the larger the difference between the take rate and the expected take rate, the stronger the intensity of anger-like emotions experienced by a responder, which in turn makes it more likely that (s)he will destroy in order to punish the proposer. These findings are consistent with, but strengthen and extend, those obtained for the one-responder power-to-take game.

Further evidence that is easily explained with an emotion-driven account of destruction (but is hard to explain otherwise) is the time responders take to make their decision. In our experiment, responders that destroyed a positive amount not only reported higher intensities of negative emotions, they also took more time to decide (t test, $p = 0.083$). However, if we focus on responders who destroyed everything, we find that, even though they reported the highest intensities of negative emotions, they did not take more time to decide than responders who did not destroy (t test, $p = 0.432$). In other words, the slowest responders turn out to be those who reported intermediate intensities of negative emotions and destroyed intermediate amounts. Standard economic theory gives us no reason to think why making the decision to destroy requires more time than making the decision not to destroy. However, research on emotions suggests the following. At low intensities of negative emotions a decision can take little time because there is no real conflict between the (cognitive) interest to earn as much money as possible and the (emotional) urge to punish the proposer. At higher

[14] In the regressions of individual emotions we find that women are more likely to report lower intensities of anger and contempt and higher intensities of gratitude. However, we also find that women show a stronger reaction to differences between the take rate and the expected take rate in the regressions for disappointment and anger. Hence, it appears that we have a mixed outcome when it comes to gender. Women seem to report lower (higher) intensities of negative (positive) emotions, but this difference disappears if the take rate is higher than expected.

intensities, this conflict does arise and hence one would expect subjects to take more time in order to sort it out. However, if the intensity of the negative emotions becomes very high it can push subjects over a threshold beyond which they are less prone to think and simply follow the emotion's action tendency, which entails less time to reach a decision (Frijda, 1986; Frijda, 1988; Goleman, 1995).

Nevertheless, there is an important aspect of the data that is not explained by the emotional reaction of responders towards proposers. As was pointed out in Result 2.2, at high take rates friends destroy more and more frequently than strangers. However, we do not find that at high take rates the emotional reaction of responders differs between treatments. This means that friends and strangers are equally angry and unhappy at high take rates ($p > 0.350$). A more detailed look at the data reveals that the disparity between destruction and anger is caused by the fact that angry strangers appear to destroy less frequently and smaller amounts than angry friends. This can be seen in Figure 2.2.

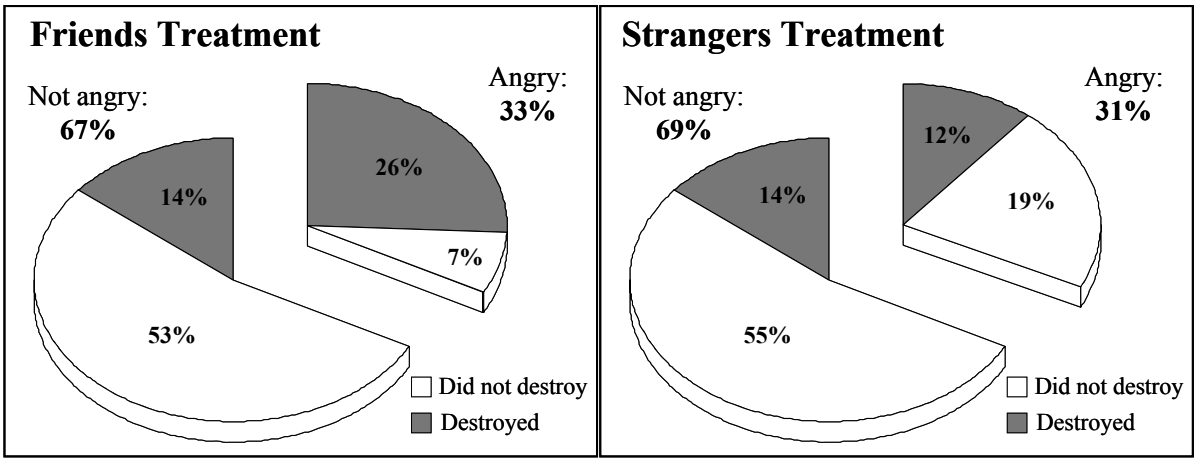

FIGURE 2.2 – DESTRUCTION BY ANGRY AND NON-ANGRY RESPONDERS

Note: Subjects are classified as angry if they reported an intensity of 6 or more (on the 7-point scales) for at least one of the following emotions: anger, irritation, or contempt. Similar results are obtained with different definitions of being angry.

As can be seen in the figure, in both treatments the ratio of non-angry and angry responders is roughly the same (2.25 to 1 for strangers and 2.04 to 1 for friends). Moreover, the frequency of destruction among non-angry responders is quite similar (20.8% for strangers vs. 21.3% for friends, $p = 0.954$). However, the frequency of destruction among angry responders is considerably different: whereas only 37.5% of the angry strangers destroyed a positive amount, 78.3% of angry friends decided to do so ($p = 0.003$). If we look at the amounts destroyed we also find a difference. On aggregate, angry strangers destroyed 28.4% of their endowment while angry friends destroyed 73.5% ($p = 0.001$).

2.4.4 Coordination

In order to explain this difference, we further analyze the behavior and emotional response of the responders. What we are interested in is to see whether pairs of friends behaved markedly different from pairs of strangers. The following results are obtained.

RESULT 2.6: *Friends are better at coordinating destruction than strangers*.

Support: Overall, in both treatments a comparable number of pairs of responders coordinate on similar destruction rates (i.e. within 10 percentage points of each other), specifically, 57.7% of the pairs of strangers, and 62.9% of the pairs of friends. However, if we concentrate on pairs in which at least one of the two responders destroyed, we find a significant difference between treatments. Among these pairs, in the strangers treatment only 11.1% coordinate on similar destruction rates whereas 42.8% do so in the friends treatment ($p = 0.009$). There is not a significant difference for pairs where at least one responder did not destroy ($p = 0.796$). This result can also be observed if we look at the correlation between the destruction rates within pairs of responders. The correlation coefficient for friends is significantly higher than the one for strangers, 0.560 vs. –0.067 (z test, $p = 0.001$, this includes pairs of responders in which there was no destruction).

A possible explanation for the better coordination of friends compared to strangers is that friends tend to be more alike (which could also explain why they are friends in the first place). However, sharing a similar preference for destruction cannot explain why we observe such a big difference between the destruction rate in the friends treatment (where pairs of friends played in the same group) and the destruction rate in the paired-strangers sessions where pairs of friends played in separate groups (destruction rates are significantly different, $p = 0.017$). A more plausible explanation for the better coordination of friends is that, first, they are better at predicting each other's behavior and, second, they have a preference for coordinating on the same action.

Overall, half of the responders accurately predict the destruction rate of the other responder in their group. Specifically, 49.0% of the strangers and 54.3% of the friends correctly predict the destruction rate of the responder they were paired with (within 10 percentage points). Although we do not see that in general friends predict better than strangers, we do find that friends are better at predicting positive destruction. In total, 39.4% of the strangers and 50.0% of the friends thought the other responder would destroy a positive amount. However, in the strangers treatment, only 7.3% of them correctly predict the other's destruction rate. Friends do significantly better with 34.3% of them making an accurate prediction ($p = 0.003$). There are no significant differences between treatments if we look at responders who thought the other would not destroy ($p = 0.834$). The better predicting ability of friends is also clear if we look at the correlation coefficients between the actual and the expected destruction rate. It is significantly higher for friends than for strangers (0.518 vs. 0.042, z test, $p = 0.001$).

Lastly, we also find evidence within the paired-strangers sessions indicating that people are better at predicting the destruction rate of their friends. In these sessions, subjects also attended the experiment together as friends but they were assigned to different groups. Nevertheless, they were assigned to the same role.[15] Hence, after informing them of the take rate faced by their friend, we asked responders to predict their friend's destruction rate. If we compare each responder's ability to predict the behavior of their friend versus the behavior of the other responder in their group, we find that they do considerably better when predicting their friend's decision. Specifically, 53.6% of responders correctly predict the destruction rate of the other responder in their group, whereas 73.2% of them correctly predict the destruction rate of their friend. The better accuracy of responders when predicting the behavior of their friends is due to better predictions of positive destruction rates. Only 4.8% of responders correctly predict the positive destruction of the other responder in their group. In contrast, 30.8% of them accurately predict the positive destruction rate of their friend ($p = 0.040$). There are no significant differences for the predictions of zero destruction rates ($p = 0.700$).

Even though friends predict destruction better than strangers, this should not lead to more coordination among friends unless responders within a group care about each other's destruction rate. Looking at the emotional response between responders demonstrates that both friends and strangers care about what the other responder does. However, there is one important difference. Whereas strangers wish to avoid being the one that destroys the most, friends wish to coordinate on the same destruction. This is expressed in the following result.

RESULT 2.7: *The emotional response towards the other responder's behavior facilitates the coordination of destruction among friends but not among strangers*.

Support: To back up this result we compare differences in emotional intensity scores across two sets of responders.[16] The first set consists of responders that destroyed at least 10 percentage points more than the responder they were paired with. For convenience, we will call them the 'punishers'. The second set consists of responders who destroyed at least 10 percentage points less than the other responder, which will be labeled the 'acquitters', for short. In the strangers treatment, punishers reported more anger, disappointment, envy, irritation, and sadness, and less admiration than acquitters ($p < 0.072$). Given this negative emotional response, it stands to reason that, ceteris paribus, strangers would prefer to be an acquitter rather than a punisher. In contrast, in the friends treatment punishers and acquitters reported similar emotional intensities for all the abovementioned emotions ($p > 0.465$). Hence, given the choice, friends unlike strangers might be indifferent between being a

[15] Subjects did not learn that they were assigned to the same role until they reached the debriefing questionnaire. The experiment's instructions simply gave no information concerning the role assigned to their partner (the only information given was that their partner would not be part of their group).

[16] Since we are interested in explaining the difference between friends and strangers, we concentrate on the emotions that, in this analysis, exhibited significantly different patterns across treatments, namely: admiration, anger, disappointment, envy, gratitude, irritation, joy, and sadness.

punisher or an acquitter. Further support is obtained if we compare acquitters with paired responders who destroyed similar amounts (i.e. within 10 percentage points): the 'coordinators'. In the friends treatment, acquitters reported more anger, disappointment, irritation, and sadness, and less joy than coordinators ($p < 0.067$). Therefore, given the choice, friends would presumably prefer to be coordinators rather than acquitters. In contrast, in the strangers treatment, acquitters reported similar emotional intensities as coordinators for all these emotions ($p > 0.140$). Thus, from this point of view, strangers would be indifferent between being a coordinator or an acquitter.[17]

Summarizing, in view of the just discussed differences in emotional responses between responders, if subjects anticipate their emotional response then destroying seems to be more risky for strangers than for friends since it might leave them in the 'punisher' position inducing additional negative emotions. This may explain why, even when very angry, strangers often decide not to destroy. We would also like to point out that friends seem to have a strong preference for coordination. This is evident from the emotional response where friends who coordinate experienced more admiration, joy, and gratitude, and less anger, disappointment, envy, irritation, and sadness than strangers who coordinate ($p < 0.062$). Consequently, angry friends may be much more inclined to destroy, especially if they believe that the other responder will also destroy. Moreover, the fact that they also predict better each other's behavior further facilitates them to obtain the positive emotional boost of coordination.

However, we should also discuss a potentially alternative explanation of destruction behavior. Since in our experiment friends had the possibility of interacting after the experiment, side payments were possible. So, one could perhaps argue or conjecture that the stronger coordination among friends who destroy is due to these side payments and not because of any differences in emotional responses. We have not succeeded ourselves in finding a convincing explanation, sidestepping emotions, in which side payments may lead to more coordination among responders. Nonetheless, in order to test if side payments played a role, we asked subjects in the debriefing questionnaire, first, if they intended to share their earnings after the experiment, and, second, if the possibility of sharing earnings after the experiment affected their decision. If side payments would indeed boost coordination one would expect more coordination among responders who answered positively to one or both of the abovementioned questions. However, we find no significant difference in coordination between those who answered positively or negatively to any of these two questions. This is also true if we look only at angry responders, that is, the responders who acted noticeably different across the two treatments ($p = 0.652$ and $p = 0.351$). Hence, we tentatively conclude that although side payments were possible they played no significant role in the game.

[17] If we compare coordinators to punishers, their emotional response suggests that both friends and strangers prefer to be coordinators rather than punishers.

2.4.5 Expected take rates and fairness perceptions

Expectations about the take rate turned out to have an important influence on the intensity of emotions and destruction behavior. This is further illustrated by Figure 2.3. Responders are divided into two groups: 'optimists', that is, people who expected a lower take rate than the one they faced (observations above the diagonal), and 'pessimists', who expected a higher take rate than the one they faced (observations below the diagonal). It is easy to see from the figure that destruction, and especially high destruction (at least 50%), is carried out almost exclusively by 'optimists'.[18] A plausible explanation is that subjects first form an expectation of what the average proposer will do. This expectation serves as a reference point which influences their emotional reaction. Negative deviations trigger anger-like emotions and are punished with destruction.

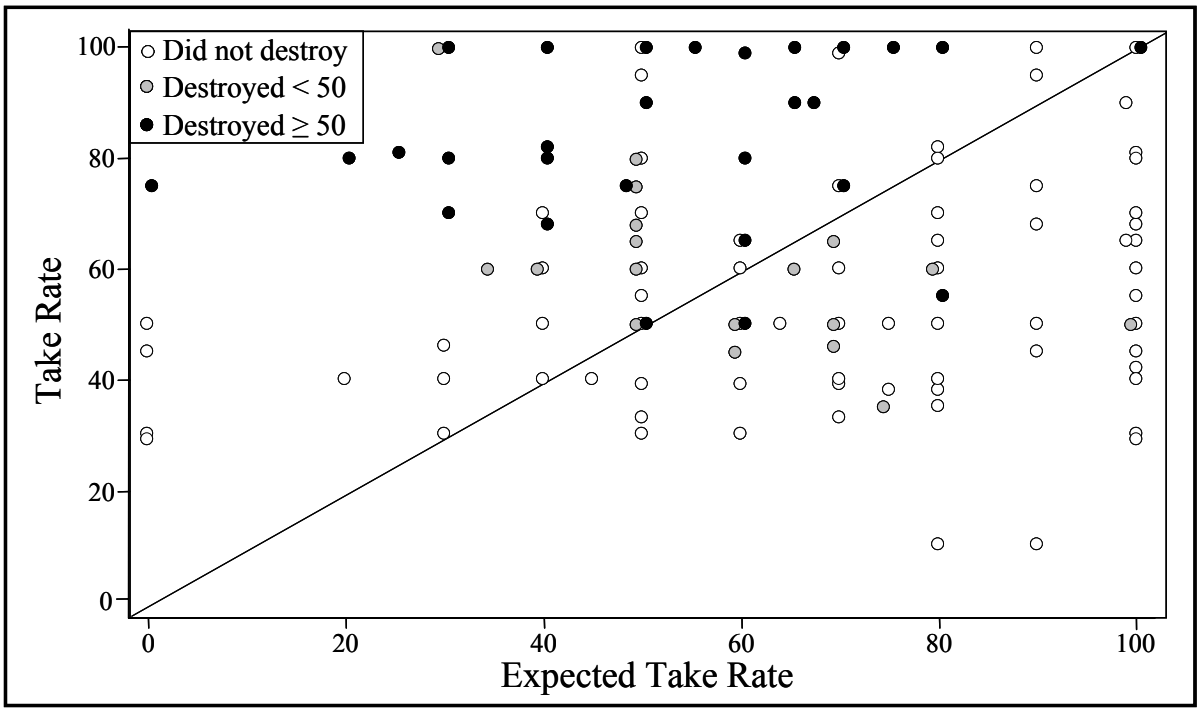

FIGURE 2.3 – SCATTER PLOT OF ACTUAL AND EXPECTED TAKE RATES

Instead of having as a reference point the expectation of what the proposer will do, a responder may (also) be affected by what he thinks the proposer *should* do, that is, the reference point could be the take rate that she considers fair. However, as noted in the discussion of Result 2.3, the responder's fair take rate is not as clearly related to destruction as the expected take rate is. For example, once we control for the effect of the take rate by looking only at take rates in the third quartile, we find that, although responders who experienced an above average difference between take rate and fair take rate destroyed more frequently than the rest, this difference is not statistically significant ($p = 0.449$). In contrast, if

[18] Of the responders who destroyed (at least 50%) 69.1% (87.5%) are optimists.

we do the same test for the difference between the take rate and the expected take rate, we find a higher frequency of destruction among responders who experienced an above average difference ($p = 0.009$).

It appears that deviations from the expected take rate are more important than deviations from the fair take rate. This is the case even though it is argued that equity considerations are especially important in highly asymmetric situations with complete information (Smith, 1976; Fehr et al., 1993). This is not to say that fairness perceptions do not play a role in the responders' decision-making. It may be that fairness plays a more indirect role than usually envisaged. Suggestive in this respect is the following analysis. Focusing on the difference between the expected take rate and the fair take rate shows that in only 4.6% of the cases this difference was negative. In other words, the overwhelming majority of responders expected a higher take rate than the one they considered fair (reflected by the average fair take rate being lower than the average expected take rate). An almost identical pattern is seen if we look at the relationship between the take rate chosen by proposers and the take rate proposers considered fair (only 8.0% of proposers chose a take rate that was lower than their fair take rate). Indeed, a glance at the scatter plot of the fair and chosen take rates or the fair and expected take rates suggests that individuals seem to use their fair take rate as a lower bound for their choice or expectation (see Figure 2.4).

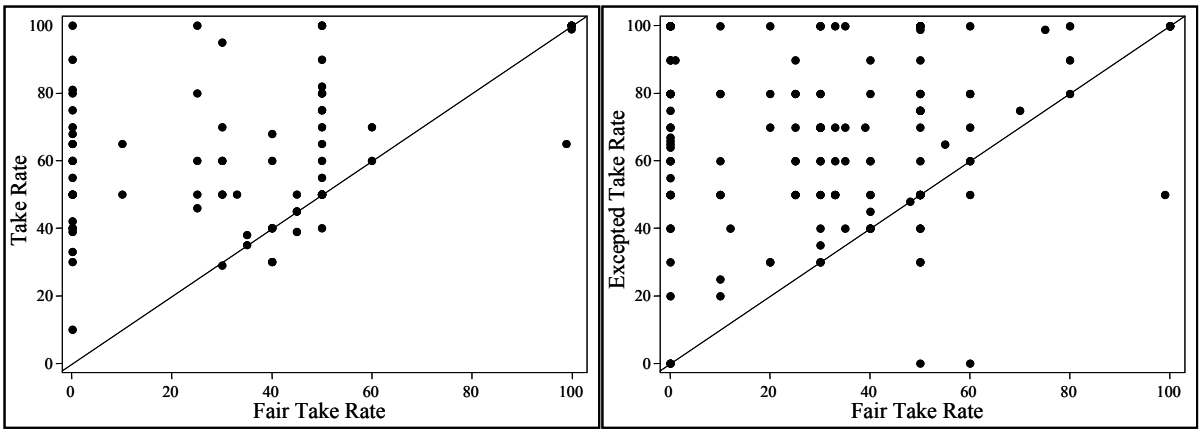

FIGURE 2.4 – RELATIONSHIP BETWEEN FAIR AND ACTUAL OR EXPECTED TAKE RATES

Note: Left (right) is the scatter plot of the actual and fair take rates of proposers (responders).

To conclude, it seems that proposers are using their fair take rate as a reference point for the determination of the optimal take rate. Similarly, responders may use their fair take rate as a reference point to form an expectation of what the real take rate will be. Once this expectation has been formed, it is a deviation from the expected take rate that triggers the high intensities of anger that motivate responders to destroy.

2.5 Discussion

In this section, we further discuss our results and relate them to recent modeling attempts to account for reciprocity in decision-making. First, we emphasize the fact that negative reciprocity seems to be particularly motivated by anger. Second, we draw attention to two important aspects of our results that are missing or unsatisfactorily modeled in the existing approaches. Namely, the role of expectations, and the effects of social ties.

2.5.1 Anger induced punishment

Although it might seem intuitively obvious, it is not until recently that economists have started to realize that punishment of opportunistic behavior may be motivated by anger. Knowing this is important because a model based on the incorrect motivations may lead to incorrect predictions. For instance, if the driving force behind an individual's decision to punish is anger but we incorrectly model it as envy (as suggested in Kirchsteiger, 1994, and Fehr and Schmidt, 2000), we will make wrong inferences regarding the action tendencies and other characteristics of the emotions at stake, and are consequently likely to make wrong predictions. In this case, one might ignore that anger, as opposed to envy, is elicited by acts perceived as intentional (Haidt, 2003) and hence overlook the important role that intentions have on punishment behavior (Falk et al., 2000; Charness and Levine, 2004). Furthermore, one could neglect that anger's action tendency is to attack as opposed to the reduction of disparities (Lazarus, 1991), and hence, expect individuals that were treated unfairly to be in favor of compensation (e.g. from a third party) as opposed to harming the unfair person.

Recent research has demonstrated that individuals who are treated unfairly derive pleasure from successfully punishing the offending party (Quervain et al., 2004). Hence, once the desire to punish exists, one could interpret punishment as simply another good that can be consumed to increase ones utility. This allows us to apply standard theoretical economic analysis to an otherwise puzzling phenomenon (see Carpenter, 2004). In this respect, it is important to point out that, even if anger was triggered by unfair behavior (e.g. deviations from equality or a maximin norm), the goal of angry individuals is to harm the other party, and not, through punishment, to correct unfair material outcomes.[19] This explains why we see individuals punishing, even when it is impossible to reduce income inequalities. For example, in the power-to-take game destruction leads to lower income inequality only when the take rate is above 50.0%. However, we still see 12.8% of subjects destroying a positive amount at take rates that are equal to or below 50.0%. Another less studied aspect of punishment that can be accounted for with an anger-motivated explanation is that some subjects punish more than the amount that is needed to equalize earnings. For instance, in the study presented in Chapter 4 we find that, when given the opportunity, 31.3% of the subjects punished to the

[19] In this respect, as is argued by Carpenter and Matthews (2005), there is an important difference between anger-induced punishment by the affected individual and indignation-induced punishment by an unaffected third party.

point were the offending subject had lower earnings than the punisher did. Similarly, in the public goods game with punishment presented in Chapter 5, we find that 33.3% of subjects, at least once, punished a free rider below their own earnings.[20] In this sense, outcome based models of social preferences such as Fehr and Schmidt (1999), and Bolton and Ockenfels (2000) miss an important characteristic of punishment behavior.

There are various known characteristics of anger that have not been theoretically modeled (e.g. people tend to feel more anger in public than in private, Jakobs et al., 1999). However, here we would like to concentrate on one important component of anger that is not yet modeled satisfactorily, namely, the fact that anger is affected by the expectation of what happened in the past.

2.5.2 Expectations about what happened

Current models that attempt to capture fairness motivations ignore the role of unfulfilled expectations on negative reciprocity. Roughly speaking, models that incorporate fairness notions fit within two approaches, an outcome-based approach, and an intention-based approach. In the former, fairness is related to differences in monetary outcomes (e.g. Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). In the intention-based approach, individuals are considered unfair, or more precisely unkind, depending on how their choices affect the final outcome of a game, and the desirability of this outcome compared to other feasible ones.[21] Though both approaches have proven to be quite successful in explaining a range of different experimental results, they fail to explain the effect of the expected take rate on the decision to destroy in the power-to-take game.

Our experimental results indicate that a responder's expected take rate is an important explanatory variable of whether she destroys or not. From a psychological point of view one can expect that someone would feel higher intensities of anger in the case of high expectations that were proved wrong as opposed to the case of low expectations that were confirmed (Ortony et al., 1988). Nonetheless, this simple and intuitive reaction is not modeled in either the outcome-based or the intention-based approach.

In the outcome-based approach, expectations are not modeled at all. That is to say, in these models, once responders observe the real take rate, the take rate they expected is no longer relevant. Hence, in these models, responders who destroy differ from responders who do not destroy only in how much they care about payoff differences. In other words, their preferences differ. Since we generally think of these preferences as fixed (at least in the short-run), we might think that we are observing stable types of individuals (such as selfish or fair-minded) that will behave as such across games and time. Furthermore, individual behavior would be less susceptible to manipulation through framing or by observing others. If on the

[20] This refers only to the equal treatment, in the unequal treatments the figure is much higher, see Chapter 5 for details.

[21] Papers in this approach include Rabin (1993), Dufwenberg and Kirchsteiger (2005), and Falk and Fischbacher (2005) who combine both the outcome-based and the intention-based approaches.

other hand, expectations explain a large part of the heterogeneity we observe in experiments, then we should be more careful in predicting how subjects will behave across games and time.[22] It would also imply that framing or information that affects expectations could have a big impact on the behavior of individuals. Furthermore, since expectations, unlike preferences, may change substantially in the short-run, behavior might adjust much faster than a model based solely on preferences would predict.[23]

In the intention-based approach, which uses psychological game theory (Geanakoplos et al., 1989), expectations play an important role, but not in the way we have discussed here. As in the outcome-based approach, in these models, once a responder observes the take rate, her expected take rate has no effect on her decision. The only expectation that has an effect is the responder's expectation of what the proposer expects the destruction rate to be (which determines the kindness of the proposer's choice). Furthermore, in these models attention is focused on equilibria when individuals *correctly* anticipate the actions of others. This raises the question, to what extent these models capture experienced emotions (see Elster, 1998).

Some of our results are in line with the theoretical predictions of the models of Rabin (1993) and Dufwenberg and Kirchsteiger (2005). For instance, we find that even when expectations are fulfilled, people that faced above average take rates (more unkindness) experienced higher intensities of anger, disappointment, envy, irritation, and lower intensities of admiration, gratitude, and joy.[24] The reason is that the intensity of emotions does not only depend on the degree of unexpectedness of an event, but also on the extent to which the individual's interests are affected (see e.g. Ortony et al., 1988), which relates to the take rate as such. More research is needed, though, to see whether this effect survives in the longer run. With experience, people not only learn to anticipate what others will do, they may also habituate to situations and may gradually become less emotional about anticipated adverse events.

Although fulfilled expectations might be a plausible assumption for the long run, in many situations there is simply not enough time to learn what others will do. In these circumstances, understanding the emotional reactions to deviations from expected actions might prove very useful for predicting how individuals will behave. Furthermore, in cases in which the long-run outcome is heavily influenced by the initial situation, emotions experienced when expectations are still unfulfilled can have a crucial effect on long-run behavior.

---

[22] We do not dismiss the possibility of consistent differences between people based on how strongly they feel about fairness norms or payoff differences. However, our results suggest that they can explain only a small part of responders' behavior.

[23] For example, one would expect that if a responder consistently sees high take rates, she would eventually update her expectation upwards. In such a case, assuming that the observed relationship between expectations and emotions holds over time, one might observe less destruction.

[24] WMW tests ($p < 0.052$), where 'fulfilled expectations' means that the expected take rate equaled the actual take rate (within a range of plus or minus 5 percentage points).

2.5.3 Social Ties

As shown in the previous section, social ties can have a considerable impact on behavior. Not only did friends react differently to higher take rates, their emotional reaction towards one another was also different. However, none of the models discussed so far incorporate an effect of social ties. Doing so might be important since not all economic interaction occurs between strangers. In cases such as interaction at the work floor, informal credit institutions, scientific research, and political participation, more interaction might actually occur between friends than between strangers. In fact, as argued by Rosenblat and Mobius (2004), technological advances that reduce communication costs, such as the internet, can make interaction among groups of friends even more important.

In our experiment, angry friends managed to coordinate destruction much more frequently than angry strangers. If we consider the emotional reactions between responders, as discussed in Result 2.7, this is not surprising. Note that angry strangers who intend to destroy their income face a situation akin to a collective action problem. Our findings suggest that they would like to see the proposer punished but want to avoid being the player in the game that is left with the lowest payoff. We do not know the precise amount of satisfaction that subjects derive from each of the possible outcomes. Nevertheless, judging by the responders' emotional reactions, it would not be farfetched to model the angry strangers' situation as a prisoner's dilemma or a stag hunt game. In either case, destruction is unlikely.[25] In contrast, in the angry friends case, the observed desire of responders to coordinate on the same action combined with an impulse to destroy makes their situation noticeably different. Angry friends can be modeled as playing a coordination game in which destruction not only gives them the highest payoff but is also the risk-dominant choice, and hence, the most attractive option.

We will discuss two natural ways of allowing for social ties in the two modeling approaches. The first way is to assume that friends are better than strangers at predicting what the other responder will do. The second way is to assume that friends, as opposed to strangers, care for each other's utility.

In the outcome-based approach, knowing what the others will do translates to friends knowing each other's preferences. This might indeed lead to a situation in which strangers destroy less than friends. To see this, note that if the income difference between two responders is important, responders will wish to destroy similar amounts. In this case, uncertainty about the other's preferences makes a responder's destruction decision much more difficult. If, as suggested by Fehr and Schmidt (1999), individuals prefer advantageous inequality to disadvantageous inequality, then destroying less becomes more attractive since it gives a higher payoff if responders fail to coordinate. Hence, since strangers face more uncertainty about the other's actions, on average, they will be likely to destroy less than friends. Nonetheless, this line of thought fails to describe one important aspect of the data,

[25] This is certainly true for the prisoners' dilemma. In the stag hunt game, destroying is the risk-dominated action and hence less likely, especially if subjects are risk adverse.

namely that the difference between friends and strangers occurs at high take rates. In the outcome-based approach, at high take rates the main concern of responders is to lower the income difference between themselves and the proposer. This is because when facing a high take rate unilateral destruction will only create small income differences between the responders. Hence, under these conditions both friends and strangers would destroy very similar amounts.

Assuming that friends care for each other's utility might be a more promising way for the outcome-based approach to explain why friends destroy more than strangers when facing high take rates. If responders care for the *utility* of the other responder, then in addition to receiving disutility because the proposer has a higher income than they do, they will also receive disutility because the proposer has higher income than their friend does. This leads to a stronger desire to destroy, even at high take rates. In this respect, investigating the precise effect of an interdependent utility function might prove a fruitful line of research.

In the intention-based approach, knowing better what the other responder will do does not affect behavior. The reason is that, in all the models within this approach, individuals care about each other's income only if they can affect it through their actions. Since in this experiment responders could not affect each other's income, these theories predict that knowing what the other responder will do has no effect on behavior. Therefore, as with the outcome-based approach, a more helpful way of incorporating social ties might be to assume that friends care for each other's wellbeing. In the current models, an individual evaluates the kindness of others by looking at how their actions lead to a higher or lower payoff for the individual. The friends case could be modeled by allowing individuals to include into their evaluation the way others treat their friends. Unfortunately, due to the complexity of these models an interdependent utility function might produce a model that is very difficult to analyze.

2.6 Conclusion

An important goal of this chapter was to extend the explicit study of emotions in economic decision-making to cases with more than two players. In addition, we looked at how social ties can affect emotional reactions and behavior. For this purpose, we used a three-person power-to-take game. The experimental data demonstrate that high intensities of anger-like emotions and low intensities of joy-like emotions induce responders to punish proposers by destroying their income. We also find that friends and strangers as responders experience different emotional reactions towards one another, which leads to more destruction in the case of angry friends.

Our results indicate that the study of emotions helps explain observed behavior. Anger-like emotions appeared to be the main driving force behind the decision to destroy income. Furthermore, by observing the emotional reaction between subjects we could explain why friends are able to coordinate on destruction more frequently than strangers. Without

investigating these emotional responses, the precise mechanism by which social ties affected the subjects' choices would have remained unclear.

In this chapter, we have also emphasized the role of expectations in determining the subjects' emotional responses. One interesting issue for future research concerns the interaction between expectations and social norms. If social norms are based on the actual behavior of the majority of individuals in a society, then expectations may be largely fulfilled in many well-established situations. However, when faced with new circumstances in which a social norm is not clearly defined, the initial expectations of individuals might have an important effect on the behavior that later becomes a norm. Hence, norms might be very susceptible to the initial conditions in which they are formed.

In addition to expectations, we would like to emphasize the importance of studying social ties. Our experiment showed clear differences in emotional reactions depending on the presence of a social tie. In some situations, this could lead to very different behavior that might be economically relevant. For example, the emotionally boost that friends receive from coordination might be one of the reasons people prefer to interact with friends rather than strangers. Results from the literature on social distance suggest that this type of preferences can lead to segregation, inefficient outcomes, and conflict between groups (Schelling, 1978; Borjas, 1995; Glaeser et al., 1996; Akerlof, 1997).

The behavioral differences induced by social ties could prove especially important if one is thinking on the effectiveness of various policies. It might be the case that a given policy would improve a situation only if individuals (do not) share social ties. Investigating the emotional responses may reveal the precise mechanism through which social ties affect the subjects' choices and could therefore help us predict the effect of a policy. To give an example, the emotional responses of friends indicate a strong desire to coordinate their actions. In this case, a coordination mechanism such as the possibility to make decisions sequentially (as in Potters et al., 2005) might be more effective among friends than among strangers.

So far, the effects of social ties have received little attention in experimental investigations. To some extent, this neglect is due to the difficulty of creating strong social ties in controlled environments. The usual ingredients of complete anonymity, no face-to-face communication, and a short time period of interaction, produce an environment in which meaningful social bonding is difficult. Nevertheless, the design we have used suggests that it is possible to include social ties in experiments and to acquire insights into how to model them.

Appendix 2A – Instructions

These are the instructions for the 'friends' treatment. The instructions for the 'strangers' treatment were very similar, and are available upon request.

Instructions (translation from Dutch)

In order to sign up for this experiment, you had to sign up together with a second participant. For convenience, we will refer to this second participant as your *partner*. In the experiment each of you will be assigned to a 3-person group, that is, you plus two other participants. We will explain how groups are formed later on.

Throughout the experiment, the type of decision you make will depend on your position in your group. Some of you will be positioned to move first, and some of you will be positioned to move second. Participants moving first will be referred to as As while participants moving second will be referred to as Bs.

Before the experiment started each desk was assigned either an A or a B. Therefore, by randomly assigning the yellow cards (in the reception room), each participant was randomly assigned a position. Once you are informed which position has been assigned to you, the corresponding letter will appear on the screen.

The 3-person group that you belong to depends on your position as well as on the position of others in the following way:

- Your group (including yourself) consists of one A and two Bs.
- If you are a B:

  Then, the other B in your group is your partner.

  The other participant will be a randomly chosen A.
- If you are an A:

  Then your partner is also an A and thus he/she is not in your group.

  The other two participants will be a randomly selected pair of Bs that signed up together for the experiment.

Note: Groups, including your own, are formed randomly in the sense that the A in the group does not know who the B's are, and similarly the B's do not know who the A is.

The experiment

At the beginning of the experiment each participant – this includes all A participants and all B participants – will receive 10 euros as his/her initial endowment. The experiment consists of two phases. In phase one, only the A participant must make a decision. Similarly, in phase two, only the B participants must make a decision. Hence, every participant makes only one decision. In addition to the decision, during the experiment you will be asked to answer a few questions.

Phase one: A chooses a percentage

In this phase, A must choose a percentage and type it into the corresponding field on the screen. This percentage determines how much of the money of each B in the group after phase two, will be transferred to A. The percentage chosen by A must be an integer between 0 and 100 (inclusive). If you wish to make any calculations, you can use the calculator located on your desk. Once you are satisfied with your decision, you have to confirm it by clicking on

the button ‘Ready’. Note that all decisions are final; once you have clicked on ‘Ready’ there is no way of changing your choice. Once A has completed phase one, phase two begins.

Phase two: each B chooses a percentage

At the beginning of this phase, each B is informed of the percentage chosen by A. At this point, each B must also choose a percentage and type it into the corresponding field on the screen. This percentage determines how much of his/her initial endowment will B destroy. The percentage chosen by B must be an integer between 0 and 100 (inclusive). Hence, the transfer from each participant B to participant A will be based on the endowment of B that is left.

Once you are satisfied with your decision, you have to confirm it by clicking on the button ‘Ready’. Note that all decisions are final; once you have clicked on ‘Ready’ there is no way of changing your choice. Once each person has made his/her decision, phase two ends.

Payoffs

After phase two, all participants will be informed of the amount of money they have earned during the experiment. You will also be informed of the amount of money earned by the other two participants in your group.

Example of how to calculate your payoffs

We will now give an example for the purpose of illustration. Remember that all participants in your group have an initial endowment of 10 euros. Suppose that in phase one participant A decides that 30% of the endowment of each participant B will be transferred to him/her (participant A). In phase two, each B can destroy part or everything of his/her initial endowment. Suppose that both Bs destroy 0% percent of their initial endowment. The transfer from each B to A is then equal to 3 euros (30% of 10 euros). The earnings of each B are equal to 7 euros (i.e. the initial endowment of 10 euros minus the transfer of 3 euros). The final endowment of A is equal to 16 euros (i.e. the initial endowment of 10 euros plus twice a transfer of 3 euros).

Now suppose that in this example, one of the B participants decides to destroy 50% of his/her initial endowment. In this case, his/her transfer to A is only 1.5 euros (namely, 30% of the endowment that was not destroyed, i.e. is 30% of 5 euros). The earnings of A are equal to 14.5 euros (namely, the initial endowment of 10 euros plus 3 euros transferred from the B who destroyed 0% plus 1.5 euros transferred from the B who destroyed 50%). The earnings of the B who destroyed 0% are again 7 euros, and, finally, the earnings of the B who destroyed 50% are 3.5 euros (namely, 50% of the initial endowment minus the transfer of 1.5 euros).

In summary

In the experiment you will be divided into groups of 3, each consisting of one A and two Bs (who signed up together for the experiment). The roles of A and B where randomly and anonymously assigned by drawing your table number. Each participant receives 10 euros as

an initial endowment. There are two phases. In phase one, A decides on a percentage that indicates how much of the endowments of each B after phase two will be transferred to A. In phase two, each B decides what percentage of his/her initial endowment will be destroyed.

If you have any questions now, please raise your hand. If you do not have any questions, please click on 'Ready'. Note that once you click on 'Ready' you will not be able to go back to the instructions. Next, we will ask you to answer a few questions in order to familiarize you with the calculation of your earnings.

Appendix 2B – Descriptive Statistics

Table 2B.1 presents descriptive statistics for the behavior and beliefs of responders for each of the two treatments.

TABLE 2B.1 – RESPONDER BEHAVIOR

Treatment	Take Rate	Expected Take Rate	Destruction Rate	Freq. of Destruction	Fair Take Rate
Strangers	59.5 (19.6)	66.8 (24.2)	15.6 (33.2)	30.0 (44.1)	32.3 (26.4)
Friends	62.3 (22.7)	66.4 (26.0)	29.4 (43.0)	40.0 (49.3)	30.0 (26.9)

Note: Numbers between brackets are standard deviations.

Table 2B.2 presents descriptive statistics of the behavior and beliefs of responders depending on the action of the responder they were paired with.

TABLE 2B.2 – RESPONDER BEHAVIOR DEPENDING ON THE ACTIONS OF THE OTHER RESPONDER

Treatment: action of the *other* responder	Own Destruction Rate	Expected Destruction Rate	Other's Destruction Rate
Strangers: other did not destroy	16.9 (34.5)	21.8 (35.9)	0.0 (0.0)
Strangers: other did destroy	12.0 (29.5)	22.4 (33.7)	60.2 (39.8)
Friends: other did not destroy	14.3 (30.7)	18.5 (28.1)	0.0 (0.0)
Friends: other did destroy	52.1 (49.0)	51.8 (49.4)	73.6 (36.8)

Note: Numbers between brackets are standard deviations.

Table 2B.3 and Table 2B.4 present a summary of the emotional reaction of responders towards the proposer and to the other responder in their group.

TABLE 2B.3 – MEAN EMOTIONAL INTENSITY OF RESPONDERS REGARDING THE PROPOSER

	Friends		Strangers	
Emotions of *i*	$d_i = 0$	$d_i > 0$	$d_i = 0$	$d_i > 0$
Admiration	2.7 (1.9)	1.3 (0.7)	2.6 (1.8)	2.0 (1.4)
anger	2.3 (1.6)	4.1 (2.2)	2.8 (2.0)	4.1 (1.7)
contempt	2.2 (1.7)	4.5 (2.2)	2.6 (2.0)	3.9 (2.0)
disappointment	2.7 (1.7)	4.3 (2.1)	3.1 (2.0)	4.7 (1.7)
envy	3.3 (1.7)	3.7 (2.3)	3.0 (1.9)	4.2 (1.8)
gratitude	3.5 (2.1)	1.5 (0.9)	3.0 (1.9)	1.8 (1.4)
guilt	1.4 (0.9)	1.4 (0.8)	1.6 (1.0)	1.6 (1.1)
irritation	2.9 (1.9)	5.2 (2.1)	3.3 (2.2)	4.6 (2.0)
joy	3.6 (2.0)	1.6 (1.3)	3.0 (1.9)	1.9 (1.2)
pride	2.3 (1.4)	2.8 (2.3)	2.6 (1.9)	2.8 (1.9)
regret	1.4 (1.0)	1.6 (1.2)	1.5 (1.1)	1.9 (1.2)
sadness	1.6 (1.1)	2.2 (1.5)	1.9 (1.6)	2.4 (1.5)
shame	1.4 (0.9)	1.6 (1.3)	1.3 (0.8)	1.6 (1.3)
Surprise	4.0 (2.1)	4.5 (1.9)	3.6 (2.2)	3.7 (1.8)

Note: Numbers between brackets are standard deviations.

TABLE 2B.4 – MEAN EMOTIONAL INTENSITY OF RESPONDERS REGARDING THE OTHER RESPONDER

	Friends			Strangers		
Emotions of *i*	$d_i > d_j$	$d_i = d_j$	$d_i < d_j$	$d_i > d_j$	$d_i = d_j$	$d_i < d_j$
Admiration	3.1 (2.2)	3.8 (1.9)	3.1 (2.2)	2.3 (1.6)	2.5 (1.7)	3.0 (2.2)
anger	2.3 (2.3)	1.1 (0.3)	1.6 (1.0)	3.0 (1.9)	1.4 (1.2)	1.8 (1.5)
contempt	1.7 (1.3)	1.1 (0.3)	1.3 (0.7)	3.2 (2.0)	1.3 (0.9)	1.6 (0.8)
disappointment	2.6 (2.4)	1.1 (0.3)	1.6 (1.1)	3.7 (1.9)	1.5 (1.3)	1.7 (1.3)
envy	2.3 (2.1)	1.1 (0.3)	1.5 (1.1)	2.7 (1.5)	1.4 (1.0)	2.0 (1.6)
gratitude	2.7 (2.1)	3.2 (2.1)	2.0 (1.3)	2.1 (1.5)	1.9 (1.4)	2.3 (1.7)
guilt	1.5 (1.1)	1.2 (0.8)	1.4 (0.9)	2.2 (1.5)	1.2 (0.7)	1.8 (1.3)
irritation	2.5 (2.3)	1.1 (0.2)	1.6 (1.2)	3.6 (2.0)	1.6 (1.4)	2.1 (1.8)
joy	2.6 (2.0)	5.0 (1.3)	3.2 (2.1)	2.3 (1.6)	3.1 (1.9)	2.5 (2.0)
pride	3.1 (2.1)	4.8 (1.9)	2.5 (1.9)	3.1 (1.8)	3.2 (2.1)	2.3 (1.9)
regret	1.3 (0.7)	1.2 (0.7)	1.8 (1.5)	2.0 (1.4)	1.2 (0.7)	1.7 (1.2)
sadness	1.9 (1.9)	1.1 (0.5)	1.3 (0.6)	2.3 (1.5)	1.4 (1.1)	1.2 (0.6)
shame	1.9 (1.8)	1.1 (0.3)	1.9 (1.6)	2.0 (1.4)	1.2 (0.7)	1.7 (1.1)
surprise	4.7 (1.7)	2.1 (1.4)	4.1 (2.5)	4.7 (2.0)	1.8 (1.4)	4.0 (2.1)

Note: Numbers between brackets are standard deviations.

Appendix 2C – Regressions

Tobit model with the destruction rate as the dependent variable and censured both at $d = 0$ and $d = 100$ (robust standard errors and clustering within groups).

TABLE 2C.1 – TOBIT MODEL ESTIMATING THE AMOUNT OF DESTRUCTION

Variable	Coefficient	Std. Error	*p*-value
Friends × Take Rate	2.672	0.836	0.001
Strangers × Take Rate	1.854	0.754	0.014
Take Rate – Expected Take Rate	1.672	0.544	0.002
Take Rate – Fair Take Rate	0.878	0.453	0.052
Economist	–18.451	26.722	0.427
Female	–61.396	23.222	0.022
Constant	–177.975	55.160	0.001
Number of obs. = 174		LR $\chi^2(6)$ = 30.63	
Log likelihood = –228.411		Prob > χ^2 = 0.000	

Note: Dummy variables: Friends: 1 if friends treatment, 0 otherwise; Strangers: 1 if strangers treatment, 0 otherwise; Economist: 1 if economics mayor, 0 otherwise; Female: 1 if female, 0 if male.

Ordered probit model with the average of the three anger-like emotions (anger, irritation, and contempt) as the dependent variable. Robust standard errors and clustering within groups.

TABLE 2C.2 – ORDERED PROBIT MODEL ESTIMATING THE INTENSITY OF ANGER-LIKE EMOTIONS

Variable	Coefficient	Std. Error	*p*-value
Take Rate	0.0168	0.0061	0.006
Take Rate – Expected Take Rate	0.0105	0.0033	0.002
Take Rate – Fair Take Rate	0.0027	0.0036	0.456
Economist	0.0873	0.1655	0.598
Female	–0.3357	0.1459	0.021
Friends	0.0237	0.1514	0.876
Number of obs. = 174		LR $\chi^2(6)$ = 52.97	
Log likelihood = –450.866		Prob > χ^2 = 0.000	

Note: Dummy variables as in Table 2C.1.

Chapter 3

The Aftermath of Punishment

Emotions, Fairness Norms, and the Reaction of the Punished[*]

In this chapter, we investigate how behavior changes after receiving punishment for performing an unkind action. A repeated version of the power-to-take game is used. The focus is on the effects of fairness perceptions, experienced emotions, and the actions of responders on the way proposers adjust their behavior. Furthermore, we examine whether proposer behavior is consistent over time and role experiences.

3.1 Introduction

By now, it is a well-documented fact that individuals who participate in economic experiments are willing to spend money in order to punish people who have treated them unkindly. Emotions are often cited as the motivating factor behind this type of behavior (e.g. Fehr and Gächter, 2002). As we have seen in the previous chapter, a number of researchers have begun to explicitly investigate the link between emotions and reciprocity.[26] They explain an individual's decision to negatively reciprocate as a tradeoff between an emotional urge to punish unkind behavior and the reward of a monetary gain. However, all these studies concentrate on the motivations and behavior of the individuals who do the punishing.

An important goal of this chapter is to investigate through an experiment the motivations and behavior of individuals who *receive* the punishment. Whereas emotions seem to play an important role in motivating individuals to punish others, it is not clear yet in which way (if at all) emotions affect the decisions of the punished. Another goal of this chapter is to study how an individual's perception of fairness affects her reaction to punishment. Finally we also study whether individuals who behave (un)fairly do so consistently over time and across positions in a game.

For our study, we use a repeated version of the power-to-take game. In this game, the proposer can make a claim on the resources of a responder. Then, the responder can destroy any part (including nothing and everything) of her own resources (Bosman and van Winden,

[*] This chapter is based on Reuben and van Winden (2005a).

[26] See for example Bosman and van Winden (2002), Sanfey et al. (2003), Ben-Shakhar et al. (2004), and Quervain et al. (2004).

). In the experiment, this game was played for two consecutive periods, where some of the subjects kept their role of either proposer or responder while others switched roles.

An important part of the experimental design is the measurement of the emotions of *proposers* after they observed whether responders, by destroying their own income, punished them or not in the first period. This allows us to study how the proposers' emotional reaction (in the first period) affects their decision in the second period. It turns out that emotions play an important role in determining how proposers change their decision from one period to the next. Proposers who were punished and felt high intensities of shame lowered their claims, while proposers who were not punished and felt high intensities of regret increased their claims. Furthermore, we find that the experience of shame and guilt does not simply depend on the size of the proposer's claim but is associated with the proposer's perception of what is fair.

The experimental design also allows us to determine the extent to which fairness perceptions vary among the subjects. Current theories that incorporate a notion of fairness typically assume that people know what is fair or unfair. Although we find support for the idea that fairness matters, we do not find much support for the presence of a clear and stable fairness norm. For example, we find that, compared to subjects who experienced only the role of proposers, subjects who experienced both the role of proposer and responder thought that proposers were entitled to claim more money.

Finally, with our design we are able to observe how the same subjects behave when they are in the proposer role and when they are in the responder role. Most theories predict that subjects who are 'kind' as proposers will also be the ones that negatively reciprocate, in contrast to 'unkind' proposers who are predicted to be unwilling to bear the cost of reciprocity. Our findings are not in line with this prediction. In fact, in some cases we find the opposite result.

The chapter is organized as follows. In Section 3.2 we address some related research. Section 3.3 presents the experimental design and the main predictions that can be derived from the theoretical literature. In Section 3.4 we describe the experimental procedures. Results are presented in Section 3.5, while Section 3.6 concludes.

3.2 Related Research

Our work is related to three different areas of research. First, we hope to contribute to the growing literature on the economic significance of emotions and their role in decision-making. Second, our work is related to research focusing on proposer behavior in ultimatum games. Finally, this study is related to research concerned with how fairness norms affect individual behavior.

There are a number of studies explicitly investigating the role of emotions in punishment behavior. By now, there is strong evidence suggesting that anger-like emotions (such as anger, irritation, and contempt) motivate responders to punish proposers. This line of research is reviewed in Chapter 2. The next logical step is to study in detail how proposers

react to punishment. In this and the following chapter we investigate whether emotions triggered after being punished affect the proposers future actions. This chapter concentrates on whether proposers will act more fairly after punishment, whereas Chapter 4 focuses on retaliation against the punishers.

Even though there seem to be no studies exploring the role of emotions in proposer behavior, there is considerable research on proposer behavior in the ultimatum game. Space constraints allow only a quick overview of the main findings.[27] Broadly speaking, proposers seem to be motivated by a combination of 'strategic' and 'non-strategic' behavior. Strategic behavior, in the restricted sense of going for the highest offer that will not be rejected, is clearly observed since proposers adjust their behavior depending on the likelihood of responders to reject an offer. For example, offers go down in cases where responders are less likely to reject, such as when the total size of the pie is unknown (see e.g. Camerer and Loewenstein, 1993; Straub and Murnighan, 1995; Rapoport et al., 1996a), when there is competition among responders (Roth et al., 1991), or in the extreme case of a dictator game in which responders cannot reject at all (Forsythe et al., 1994). However, the fact that, even in completely anonymous dictator games there are positive offers seems to indicate that there is a degree of non-strategic (perhaps fairness-guided) behavior. Further evidence of non-strategic behavior is provided by Lin and Sunder (2002), who find that (given the responders' reactions) the proposers' offers are slightly higher than the optimal profit-maximizing offer.[28] Moreover, if one thinks that the non-strategic behavior is due to fairness norms, there is growing evidence that these can be subject to self-serving biases (Knez and Camerer, 1995; Schmitt, 2004). By analyzing whether and to what extent emotions play a role in proposer behavior, we hope to contribute to the differentiation of such strategic and non-strategic factors in the decisions of proposers.

Finally, this study is related to research on fairness norms. Over the past decade, numerous authors have been studying behavior that cannot be explained with the standard economic model assuming self-regarding preferences. More importantly, some of the seemingly 'anomalous' behavior has been successfully modeled by theories that try to incorporate such norms into the utility functions of individuals. Different authors have used different notions of what constitutes fair and unfair behavior, which has fostered an extensive debate on which notion best describes experimental results.[29] However, only a few researchers have explicitly asked for the fairness perceptions of individuals and, more importantly, analyzed how their fairness perceptions interact with other variables. Pillutla and Murnighan (1996) measure the fairness perceptions of responders and find that perceived unfairness is related to the rejection of offers. However, as was shown in Chapter 2, once the

[27] For a good summary of the main results see Camerer (2003).

[28] See also Henrich et al. (2001) for clear evidence of such non-optimal offers in various non-western societies.

[29] Examples of different ways of modeling fairness include: equality (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), a combination of efficiency and maximizing the welfare of the poorest individual (Charness and Rabin, 2002), and the midpoint between the best and the worst available alternatives (Rabin, 1993).

effect of the offered amount and the responder's expected offer are taken into account, the perceived unfairness of an offer has no longer a clear effect on destruction. Nonetheless, even though fairness perceptions might not be an important determinant when it comes to responder behavior, they might play a significant role when it comes to proposer behavior. In this chapter, we investigate how and in what way fairness perceptions affect the decisions of proposers.

3.3 Experimental Design and Theoretical Predictions

For our study, we use a repeated version of the power-to-take game utilized by Bosman et al. (2005b). Subjects played the game for two periods. In each period a proposer (with endowment E^{prop}), was matched a responder (with endowment E^{resp}). First, the proposer decides on the 'take rate' $t \in [0,1]$, which determines the proportion of the responder's endowment that is transferred to the proposer. Second, the responder decides on a 'destruction rate' $d \in [0,1]$, which determines the proportion of the responder's endowment that is destroyed. Hence, the proposer's payoff equals her endowment plus the money she took from the responder's remaining endowment, i.e. $E^{prop} + t(1 - d)E^{resp}$. The responder's payoff equals the part of her endowment that she does not destroy minus the amount transferred to the proposer, i.e. $(1 - t)(1 - d)E^{resp}$. In the experiment all endowments were equal ($E^{resp} = E^{prop}$).

In each of the two periods, subjects were randomly assigned to either the proposer's role or the responder's role. Each proposer was randomly paired with a responder using a perfect-strangers matching protocol. Note that this eliminates any incentive to build up a reputation. In addition, this procedure produced a group of subjects that had the same role in both periods and another group that switched roles from one period to the other. Studying any differences between subjects that were proposers in both periods and subjects that were proposers only in the second period, allows us to test whether the role experienced in the first period affects behavior in the second period.

During the experiment, we used self-reports as the research method for measuring emotions, expectations, and fairness perceptions. Since we concentrate on proposer behavior, Figure 3.1 shows the precise order in which the proposers' decisions, emotions, expectations, and fairness perceptions were measured. Expectations were measured by asking subjects to indicate the most likely value for d. As in Chapter 2, subjects' emotions towards the other player were measured through self-reports after the subject observed what the other player did. Similarly, emotions were measured by providing subjects with a list of fourteen emotion names and asking them to report on a 7-point scale with what intensity they experienced each emotion. The scale ranged from "no intensity at all" (1) to "very intensely" (7). The list included the following emotions: pride, envy, anger, guilt, joy, shame, irritation, gratitude, surprise, contempt, disappointment, admiration, regret, and sadness. We asked for the subjects' perceptions of the fair take rate, at the end, in a debriefing questionnaire.

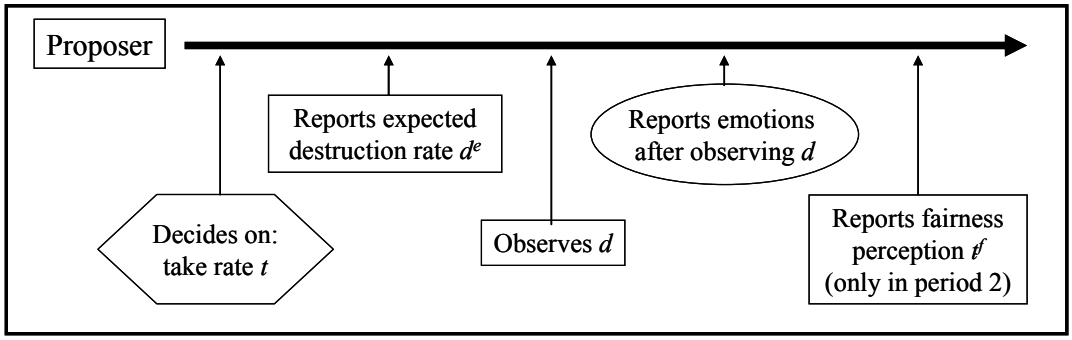

FIGURE 3.1 – SEQUENCE OF EVENTS FOR PROPOSERS IN EACH PERIOD

We now turn to the theoretical predictions of this game. Traditional economic theory, assuming own-payoff maximization, predicts that a proposer will choose to take essentially all of the responder's endowment and that the responder will not destroy any of it. However, previous work has proven that this is not the case. In the power-to-take game, responders consistently destroy some or all of their endowment when faced with high take rates, and proposers hardly ever choose to take all of the responders' endowment. In order to explain behavior in this and similar types of games researchers have constructed models that incorporate different kinds of fairness norms (see footnote 29) or emotions (Kirchsteiger, 1994; Levine, 1998). They do so by altering the utility function of individuals to include the monetary payoff and intentions of others. Some of these models predict remarkably well the behavior of responders.[30] However, to the best of our knowledge, there are no theoretical models than can satisfactorily explain proposer behavior in the power-to-take game. All models, if calibrated to explain proposer behavior in the ultimatum game, predict that, in the power-to-take game, proposers will take considerably less from responders than they actually do (this is discussed further in Chapter 5). It is not the aim of this chapter to test the performance of individual models of social preferences. Instead, we wish to investigate whether proposer behavior provides support for some of the main assumptions in this literature.

Models of social preferences make various common assumptions. A first assumption is that the utility function of individuals has a 'material' part, which represents how much they value their own monetary payoff, and a 'non-material' part, which shows how much individuals value a combination of the distribution of monetary payoffs and the intentions of others. A second assumption is that individuals differ regarding the intensity with which they care about the non-material part of the utility function (relative to the material part). A third assumption is that, although individuals differ with respect to their valuation of the non-material part, everyone shares the same 'type' of preferences and this fact is common knowledge. To put it bluntly, everyone knows what is fair and what is unfair in every situation, but not everyone cares as much about it. Finally, a fourth assumption is that the

[30] Especially: Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Charness and Rabin (2002), and Falk and Fischbacher (2005).

value individuals attach to the non-material part of the utility function is the same irrespective of their position in the game and their past experiences. Although there is considerable evidence that supports to the first two assumptions, the third and fourth assumptions have not been sufficiently tested. In this chapter, we study whether there is support for the latter two assumptions. In order to do so, we first describe the prediction that of models of social preferences make concerning the behavior of proposers in the repeated power-to-take game.

In the models of social preferences that predict responder behavior rather well (see footnote 30), proposers behave in roughly two distinct ways:

First: There is a group of proposers who value the material part of the utility function relatively more than the non-material part. These proposers have a tendency to choose take rates that maximize their monetary payoff (given their beliefs of how responders behave). This usually leads them to choose relatively high take rates. Furthermore, since the choices of these proposers are restricted by the likelihood that responders will destroy, they will increase their take rate from period one to period two if their experience in period one makes them adjust downwards their general belief of the likelihood of destruction. Lastly, if these proposers were playing the role of a responder then they would be less likely to destroy their endowment than other individuals would.[31]

Second: There is another group of proposers who value the non-material part of the utility function relatively more than the material part. These proposers have a tendency to choose take rates that maximize their 'other-regarding' preferences. This usually means that these proposers are not maximizing their own monetary payoff (given their beliefs of how responders behave). This leads them to choose relatively low take rates. Furthermore, since normally the choices of these proposers will not be restricted by the likelihood that responders will destroy, they should not increase their take rate from period one to period two. Finally, if these proposers were playing the role of responders then they would be more likely to destroy their endowment than other individuals would.

In summary, models of social preferences predict that some proposers will be relatively more interested in their material payoff. They will behave more 'strategically' and less in line with the fairness norm. The other proposers will be relatively more interested in their non-material payoff. They will behave less strategically and more in line with the fairness norm.

[31] This need not be the case in the model of Fehr and Schmidt (1999). In their model, responders destroy because they dislike disadvantageous inequality whereas proposers choose low take rates because they dislike advantageous inequality. Hence, proposers who choose high take rates will be less likely to destroy only if the aversion to both types of inequality is correlated. However, the accuracy of some of their predictions does rely on assuming such correlation exists (Fehr and Schmidt, 1999).

3.4 Experimental Procedures

The computerized experiment was run in October 2003 at the CREED laboratory of the University of Amsterdam. The experiment was conducted with z-Tree (Fischbacher, 1999). In total 92 subjects, almost all undergraduate students from the University of Amsterdam, participated in the experiment. About 41% of the subjects were women. Moreover, 41% were students of economics and the other 59% were students from various fields such as biology, political science, law, and psychology. Subjects received a show-up fee of 2.5 euros, independent of their earnings in the experiment, and 10 euros as endowment in each of the two periods. On average, subjects were paid out 21.40 euros. The whole experiment took one and a half hours.

After arrival in the lab's reception room, each subject drew a card to be randomly assigned to a seat in the laboratory. Once everyone was seated, the instructions for the experiment were read aloud (a translation of the instructions is provided in Appendix 3A). Subjects were told that the experiment consisted of two independent parts (each part being one of the periods in the two-period power-to-take game). We emphasized the fact that their choices in the first part of the experiment would not affect their earnings in the second part of the experiment. Furthermore, it was explained that the instructions for each part would be given at the beginning of the respective part. After this, the one-shot power-to-take game was described in the instructions as the first part of the experiment. Thereafter, subjects had to answer a few exercises in order to check their understanding of the procedures. After these exercises the subjects were informed, by opening an envelope on their desk, which role (that of proposer or responder) they had been assigned in the first period of the game. The game was framed as neutral as possible, avoiding suggestive terms (like 'take rate'). Subsequently, the subjects played the first period of the power-to-take game via the computer. Once everyone finished, the instructions for the second part of the experiment were read. This simply consisted in informing subjects that they would play the same game once again. However, we did stress that they would be playing against a different person, and furthermore, that their role in the second part would be randomly determined and did not depend on what their role was in the first part. Subsequently, the subjects played the second period of the two-period power-to-take game.

3.5 Results

In this section, we present and analyze the decisions that were taken by proposers. Furthermore, we investigate whether fairness norms and the emotions reported by the subjects help explain the observed behavior.

3.5.1 Proposer behavior

In general, proposers in this experiment behaved in a similar way to proposers in other power-to-take games. The average take rate, taken over both periods, was 58.8%, and the median

take rate was 60.0%. See Table 3.1 for additional descriptive statistics. As in previous studies, we did not find that demographic variables, such as gender, age, or area of study, had an effect on the chosen take rate.

TABLE 3.1 – DESCRIPTIVE STATISTICS OF PROPOSERS

	Proposer in both periods		**Proposer only in period 1**	**Proposer only in period 2**
Mean	**Period one**	**Period two**	**Period one**	**Period two**
Take Rate	56.5% (21.8)	59.0% (20.0)	56.3% (16.5)	63.7% (18.2)
Destruction Rate	14.8% (31.9)	6.3% (21.2)	6.9% (18.4)	16.1% (34.9)
Expected Destruction Rate	13.8% (24.8)	8.5% (15.4)	16.9% (33.6)	20.3% (29.1)
Frequency of Destruction	25.0% (44.2)	12.5% (33.8)	22.7% (42.9)	22.7% (42.9)
Fair Take Rate	32.0% (29.2)		43.9% (23.7)	40.9% (19.8)

Note: Numbers between brackets are standard deviations.

To start, we analyze how proposers adjust their decision. Namely, how they change their take rate from one period to the next. In order to do so, we must concentrate on the subjects that had the role of proposers in both period one and period two. We will refer to this group of proposers as 'role-keepers'.

On average, the take rate chosen by role-keepers was about the same in both periods (see Table 3.1). However, this hides considerable change at the individual level. Overall, 70.8% of role-keepers changed their take rate from period one to period two (29.2% changed it by more than 10 percentage points). Of the role-keepers that changed their take rate, 29.4% decreased their take rate and the other 70.6% increased it. What is more, the role-keepers' decision to change the take rate was strongly affected by the behavior of the responder they faced in the first period. This leads us to our first result.

RESULT 3.1: *Role-keepers who faced a responder who destroyed lowered their take rates whereas role-keepers who faced a responder who did not destroy increased their take rates*.

Support: The average take rate of role-keepers who did not experience destruction increased from 52.8% to 58.1%. The change in the take rate is significantly different from zero with a

Wilcoxon matched-pairs signed-rank (WSR) test ($p = 0.025$).[32] The average take rate of role-keepers who experienced destruction decreased from 67.5% to 61.7%. This change in the take rate is marginally significant (WSR test, $p = 0.087$).

However, experiencing destruction explains only part of the role-keepers' decision to change the take rate. For instance, only 66.7% of the role-keepers who experienced no destruction increased their take rates. Similarly, only 50.0% of role-keepers who experienced destruction decreased their take rates. Furthermore, the amount destroyed did not seem to play an important role in determining by how much the role-keeper changes the take rate. For example, if we concentrate on the role-keepers that experienced some destruction, we find the following. On average, role-keepers who experienced a destruction rate above the median decreased their take rate by 6.7 percentage points. Moreover, role-keepers who experienced a destruction rate below the median decreased their take rate by a very similar 5.0 percentage points. We cannot reject equality using a Wilcoxon-Mann-Whitney (WMW) test ($p = 0.817$).

The next step in our analysis is to try to explain why, when faced with a similar situation, some role-keepers decide to change their take rate and some do not. In order to do so, we divide role-keepers depending on whether or not they experienced destruction. Then, we compare the role-keepers who changed their take rate to the role-keepers who did not.

We start by looking at role-keepers who faced a responder who did not destroy. A possible reason why some of these role-keepers increased their take rates while others did not is that their expected destruction rates might have been different. It is reasonable to imagine that role-keepers who expected some destruction and observed no destruction would be more likely to increase their take rate than role-keepers who correctly anticipated no destruction. On average, this seems to be the case. However, the relationship is weak. Of the role-keepers who correctly anticipated no destruction, 63.6% increased their take rate, whereas, of the role-keepers who expected some destruction (but experienced no destruction), 71.4% increased their take rates. There is not a significant difference between the two groups ($p = 0.740$).

In order to find a possible explanation for why some role-keepers change their take rate and some do not, we turn to the role-keepers' emotional response. We find the following result (for some descriptive statistics concerning the emotional response of proposers, see Appendix 3B).

RESULT 3.2: *Role-keepers who increased their take rates after experiencing no destruction were role-keepers that reported high intensities of regret.*

Support: WMW tests reveal that, among role-keepers who experienced no destruction, role-keepers who increased their take rate reported significantly higher intensities of regret than role-keepers that did not change or decreased their take rate (regret intensity scores of 2.58 vs. 1.00, $p = 0.006$).

[32] Throughout the chapter, unless it is otherwise noted, we always use a two-sided Wilcoxon-Mann-Whitney test.

Result 3.2 is quite intuitive. As one would expect, if a proposer reported feeling regret after observing that the responder did not destroy, this is because the proposer realized that he or she could have chosen a higher take rate. Interestingly, feeling regret does not seem to be related to the proposers' expectations. One would think that proposers that reported high intensities of regret were proposers that expected responders would destroy and then experienced no destruction. However, if we look at the role-keepers' expectations we find that this is not the case. Role-keepers who expected some destruction and experienced no destruction reported an average intensity of regret of 2.00. This is actually lower than the 2.09 average intensity of regret reported by the role-keepers who expected and experienced no destruction (the difference is not significant, $p = 0.845$).[33] This result might be explained by a possible confounding effect that is hinted at by models of social preferences. Namely, proposers may behave in a norm-abiding way and therefore, since they do not want to take more than the amount they are already taking, they do not feel regret when they realize they could have chosen a higher take rate. However, if this is the case then, contrary to the predictions of the models of social preferences, norm-abiding proposers do not necessarily choose lower take rates than more strategic proposers. Role-keepers that increased their take rate after observing no destruction (i.e. behaved more strategically) had actually chosen lower average take rates than role-keepers that did not change or decreased their take rate (50.0% vs. 58.3%). However, this difference is not a significant ($p = 0.189$).

We now turn to role-keepers who faced a responder who chose a positive destruction rate. As in the previous case, it is possible that expectations could explain why some of these role-keepers reduced their take rate while others did not. Role-keepers who experienced a destruction rate that was higher than expected would be more likely to decrease their take rate than role-keepers who experienced a lower than expected destruction rate. Unfortunately, we cannot test whether this is true or not since none of role-keepers who experienced destruction expected a destruction rate that was higher than the one they confronted. Again, in order to get additional insights on the role-keepers' behavior, we analyze their emotional response. We find the following result.

RESULT 3.3: *Role-keepers who decreased their take rates after experiencing some destruction were role-keepers that reported high intensities of shame*.

Support: A WMW test shows that, among role-keepers who experienced some destruction, role-keepers who decreased their take rate reported higher intensities of shame than role-keepers that did not change their take rate (shame intensity scores of 4.67 vs. 1.33, $p = 0.043$).

[33] Similarly, the change in the take rate for role-keepers who expected some destruction and faced no destruction was not significantly different from the change in the take rate of role-keepers who correctly anticipated no destruction ($p = 0.817$). This might suggest that expectations do not have a strong impact on a proposer's decision to change the take rate. However, since we do not have information on what proposers expected responders would do at take rates other than the chosen one, it would be premature to conclude that expectations do not play a role.

We also find the same qualitative pattern for the related emotion of guilt. However, in this case the difference is not significant ($p = 0.261$). Result 3.3 gives us an important insight into why proposers decrease their take rates from one period to the next. We explore this in the following subsection.

3.5.2 Social norms

The emotions of shame and guilt are triggered when an individual violates an internalized social norm. Furthermore, in the case of shame, the disapproval of others plays an important role (Tangney and Dearing, 2002). As we would expect, if the actions of responders make the proposers feel bad by triggering these negative emotions, one would expect proposers to adjust their behavior in order to feel better. Presumably, this leads proposers to lower their take rates. Naturally, this opens up the question of why some proposers feel shame and guilt while others do not. A casual look at the data reveals that, among role-keepers who observed destruction, shame and guilt are not simply triggered by high take rates: role-keepers with take rates above the median (on average, a take rate of 77.5% and an intensity of shame and guilt of 2.75 and 2.00) did not experience more shame and guilt than role-keepers with take rates below the median (on average, a take rate of 47.5% and an intensity of shame and guilt of 3.50 and 3.00). Thus, if shame and guilt are indeed triggered by deviations from a norm (presumably a fairness norm), it appears as though not all proposers think that choosing a high take rate is norm violation. On closer inspection, this seems to be the case. Once we take into account what proposers perceive to be the fair take rate t^f, we get a clear result.

RESULT 3.4: *Role-keepers who chose take rates that they considered unfair experienced higher intensities of shame and guilt*.

Support: If we divide the role-keepers into role-keepers that chose an unfair take rate (i.e. a take rate above what they considered fair, $t > t^f$), and role-keepers that chose a fair take rate (i.e. a take rate that equaled or was below what they considered fair $t \leq t^f$), we find that in both the first and the second period, role-keepers who chose unfair take rates reported higher intensities of shame and guilt than role-keepers who chose fair take rates ($p < 0.039$ for shame and $p < 0.074$ for guilt). Similarly, among the role-keepers who experienced destruction, role-keepers who chose unfair take rates reported higher intensities of shame than role-keepers who chose fair take rates ($p = 0.043$).[34]

[34] This part of the result is valid only for the first period. Unfortunately, in the second period all the role-keepers who experienced destruction happened to be role-keepers that considered they made an unfair offer. Hence, we cannot test whether they experienced more shame or not. Furthermore, we again find a similar pattern for guilt that nevertheless is not significant.

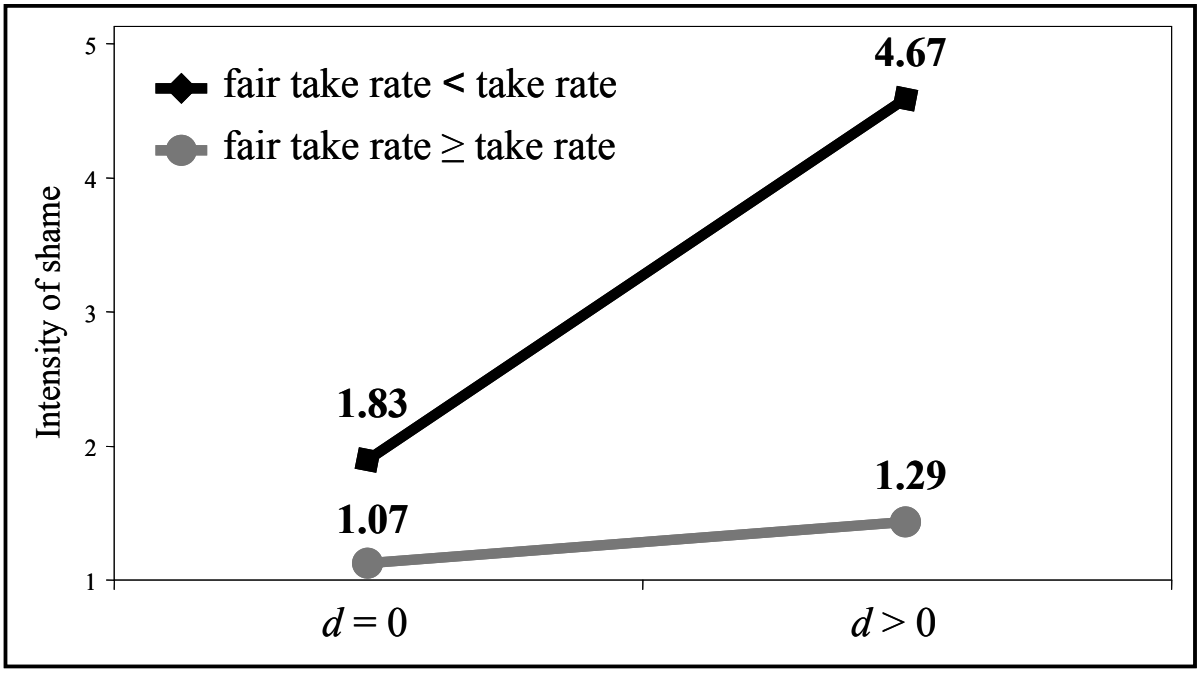

FIGURE 3.2 – SHAME AND DESTRUCTION

Note: Role-keepers mean intensity of shame depending on their fairness perception and on whether the responder destroyed (period 1).

In summary, as Figure 3.2 illustrates, we seem to have a group of role-keepers who acknowledge they made an unfair choice and hence feel high intensities of shame (especially if they faced destruction) and guilt, and another group of role-keepers who believe they made a fair choice and therefore feel low intensities of both emotions (even after facing destruction). Furthermore, as Result 3.3 shows, feeling high intensities of shame might be an important determinant on whether role-keepers lower their take rates. This would make role-keepers in the first group more likely to lower their take rates than role-keepers in the second.[35]

Combining the last two results we can arrive to the plausible conclusion that fairness perceptions, by triggering feelings of shame and perhaps guilt, have an impact on proposer behavior. However, what is perceived as fair varies from one person to another. In fact, there is more variation in fair take rates than in take rates (see the standard deviations in Table 3.1).[36] Moreover, it is not necessarily true that, individuals that, in their opinion, are acting fairly are being considerably nicer to others. The main difference between role-keepers who chose a take rate they thought was fair and role-keepers who chose a take rate they thought was unfair is their fairness perception and not their chosen take rate. For example, in the first period, role-keepers that thought they were unfair chose a take rate that was 13.0 percentage

[35] Result 3.3 refers only to role-keepers who faced destruction. However, we find a similar pattern for role-keepers who faced no destruction. Among these, role-keepers who reduced their take rates reported above average intensities of shame.

[36] This is also true in Chapter 2, see Table 2B.1.

points higher than role-keepers who thought they were fair, but at the same time, they reported a fair take rate that was 30.3 percentage points lower than role-keepers who thought they were fair.

We know from the literature on self-serving biases that groups of people in different circumstances can evaluate what is fair is different ways. With our results, we can add that even under the same circumstances there can be considerable variation in fairness perceptions. If fairness perceptions are indeed so diverse, studying whether past experiences affect the fairness perception of individuals becomes an important question. In order to answer it, we examine whether experiencing the role of a responder in the first period has an effect on the way proposers behave in the second period.

Again, in Table 3.1 we present descriptive statistics for the group of proposers who had the responder role in the first period and the proposer role in the second. We will refer to this group of proposers as 'role-switchers'. Comparing the choices of role-keepers and role-switchers in the second period, we note that role-switchers choose slightly higher take rates. However, the most striking difference between the two groups is actually in the fair take rates. This is stated in the following result.

RESULT 3.5: *Role-switchers reported higher fair take rates than role-keepers*.

Support: On average, role-switchers chose a fair take rate equal to 40.9% while role-keepers chose a fair take rate of 32.0% (the former is higher, $p = 0.097$).

It appears that being a responder in the first period has a noticeable effect on the proposers' fairness perceptions. However, we find the same effect on the responders' side. Subjects that had the responder role for periods one and two reported lower fair take rates than subjects that were first a proposer and then a responder (28.8% vs. 43.9%, $p = 0.054$). Hence, it seems that experiencing both roles instead of just one is what produces an important effect on what is considered fair.[37] More specifically, the shift in fair take rates is caused by more individuals stating it is fair to take 50.0% instead of 0.0% (see Figure 3.3).

This difference in fair take rates has a noticeable effect on the emotions of shame and guilt. As role-keepers did, role-switchers who chose take rates they considered unfair experienced more shame than role-switchers who chose take rates they considered fair ($p = 0.067$). However, the higher fair take rates produce a lower proportion of role-switchers that think they made an unfair choice. Specifically, whereas 75.0% of role-keepers considered they made an unfair choice, only 59.1% of role-switchers considered they did. Consequently, on average, role-switchers experienced lower intensities of shame and guilt than the role-keepers. If, as is true for role-keepers, high intensities of shame induce proposers to take less,

[37] Perhaps, experiencing both roles makes more salient the fact that there is mobility between positions in the game. This may induce a belief that everyone can be in an advantageous position at some point, and hence consider it acceptable for people to take advantage of those occasions.

one could suppose that role-switchers would be less inclined to decrease their already high take rates.

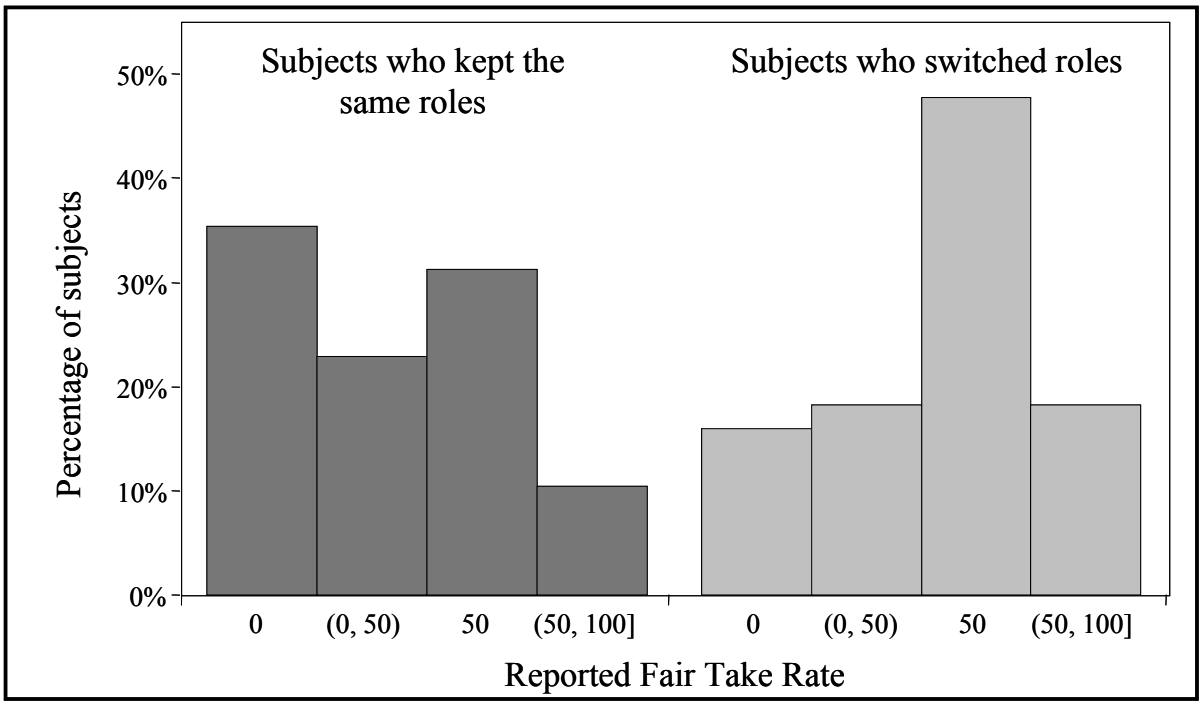

FIGURE 3.3 – FAIR TAKE RATES AND ROLE SWITCHING

Note: Histogram of the fair take rates of all subjects depending on whether they switched roles or not.

3.5.3 Social preferences

We now turn to see whether alternating roles has also an impact on individual behavior. As we have seen, alternating roles has a considerable impact on subjects' fairness perceptions and consequently on their emotional reaction. However, on average there is not a significant effect on their behavior. Role-keepers chose lower take rates than role-switchers, but this difference is not significant ($p = 0.765$). Similarly, on the responder side, responders that kept the same role destroyed less frequently than responders that were first a proposer (12.5% vs. 22.7%). However, again the difference is not significant ($p = 0.370$). Nevertheless, we do find some interesting results once we look at how individuals choose across periods.

As we mentioned, models of social preferences predict that individuals should be consistent across periods (since preferences are assumed to be constant). More specifically, role-keepers that chose high (low) take rates in the first period should, on average, choose high (low) take rates in the second period. Similarly, responders that chose high (low) destruction rates in the first period should, on average, choose high (low) destruction rates in the second period. When it comes to role-switchers, models of social preferences predict that subjects who, as responders, destroyed in the first period, should on average select low take rates in the second period. In the same way, subjects that, as proposers, decided on high take rates in the first period should, on average, destroy less in the second period (see footnote 31).

As it turns out, we find that these predictions are only consistent with subjects that did not switch roles.

RESULT 3.6: *In line with models of social preferences, role-keepers who chose low take rates in the first period also chose low take rates in the second period*.

Support: Role-keepers who chose a low take rate (below the median) in period one chose lower take rates in period two compared to role-keepers who chose a high take rate (above the median) in period one (48.3% vs. 69.6%, $p = 0.006$).

This result is consistent with models of social preferences. One could argue that role-keepers that consistently chose low take rates are revealing a preference for fair outcomes. After all, they are forgoing money with their choice. On average, role-keepers who chose low take rates earned less than role-keepers who chose high take rates ($p = 0.051$). The behavior of subjects who were a responder in both periods is also consistent with models of social preferences. Among these responders, those who chose a positive destruction rate in the first period also destroyed more frequently in the second period (33.3% vs. 5.6%, $p = 0.081$). This is remarkable given that, in the second period, both groups of responders faced very similar take rates (57.5% vs. 57.2%). It is evident that responders who destroyed earned less than responders who did not destroy. Hence, this gives some support to the idea that some responders have preferences for punishing opportunistic behavior. We now turn to the subjects who switched roles and our next result.

RESULT 3.7: *Contrary to models of social preferences, role-switchers who destroyed some of their income in the first period chose high take rates in the second period*.

Support: Role-switchers who destroyed a positive amount in period one (as responders) chose significantly higher take rates in period two (as proposers) than role switchers who did not destroy in period one (77.0% vs. 59.8%, $p = 0.053$).

This result is in conflict with the predictions of the models of social preferences. Individuals that destroy a positive amount show, according to these models, that they value considerably the non-material part of their utility function. It follows that, in period two, these same individuals should choose low take rates. On the responders' side we find a similar, albeit weaker, result. Among subjects who switched from the proposer role to the responder role, subjects who chose high take rates in period one destroyed as frequently in period two as subjects who chose low take rates (18.2% vs. 27.3%). However, this difference is not significant ($p = 0.619$). Nevertheless, we do not find, as predicted, that subjects who chose high take rates as proposers also destroy less as responders. In the following section, we discuss how these findings question the viability of models of social preferences as they are now conceived, and what are some of the promising ways of improving them.

3.6 Conclusion

In this chapter we have investigated how proposers in the power-to-take game adjust their behavior depending on their interaction with responders, their fairness perceptions, felt emotions, and role experiences. Our main results can be summarized by the following statements:

- Fairness considerations appear to play an important role in the behavior of proposers. Specifically, deviations from a perceived fairness norm trigger feelings of shame and guilt, which induce proposers to lower their take rates.
- The perceived fairness norm varies considerably between individuals. In fact, there is more variation in the perception of the fairness norm than variation in the behavior of the individuals.
- The different types of individuals predicted by models of social preferences can be traced among the subjects that played the same role in both periods, but fail to describe the behavior of subjects who switched from one role to the other.

Our results show that fairness norms, through the emotions of shame and guilt, play a significant role in the proposers' decision-making process. Since shame and guilt are emotions that are experienced when individuals think they have violated a social norm (Lazarus, 1991), our findings suggest that an internalized social norm influences proposers in this type of games. This is suggested by the observation that if responders destroy, they provoke high intensities of shame among proposers. This is not surprising since the emotion of shame is strongly associated with the perceived disapproval of others (Tangney and Dearing, 2002). As a consequence, the punishment of proposers becomes cheaper since destroying income not only reduces the income of the other but it also makes proposers feel bad.[38] Furthermore, it highlights the importance for responders of being able to signal their displeasure. The bare existence of a fairness norm might not be enough to restrain the behavior of proposers.

Knowing that shame affects the decision-making process of proposers can help us improve our models in order to make predictions that are more accurate across different situations.[39] For example, research on emotions tells us that people feel more shame and guilt in situations in which others can clearly observe their actions and can show their disapproval (Tangney and Dearing, 2002). This would be consistent with proposers asking more in ultimatum games in which the amount to be divided is unknown. Uncertainty over the size of the pie prevents responders from clearly judging the actions of proposers. This might make proposers feel less shame. It would also be consistent with a lower proportion of proposers

[38] It would also explain why simply making your displeasure known has a similar effect to monetary punishment in public goods settings (Masclet et al., 2003).

[39] At this point, the only model that incorporates both shame and guilt is Bowles and Gintis (2001).

choosing the equal split in the dictator game as opposed to the ultimatum game.[40] Since, in dictator games responders cannot signal their displeasure, proposers are less exposed to feeling high intensities of shame. In fact, the emotion of shame would also explain why proposers in dictator games take more as we increase the level of anonymity (as in double-blind experiments; Hoffman et al., 1996). Of course, for shame or guilt to have an effect in one-shot games, it must be the case that subjects are able to anticipate their emotional response. In this experiment, we cannot test whether this is true. However, recent work in Lazear et al. (2005) provides some support for this idea. If proposers make positive offers in dictator games in order to avoid feeling shame or guilt, then it is possible that, if given a choice, proposers would like to avoid playing the dictator game in the first place. Lazear et al. (2005) study precisely this situation and find that proposers who are most generous when forced to play a dictator game are also willing to pay the highest amount in order to avoid playing it. This is exactly what would be predicted by a model that incorporates anticipated feelings of shame and guilt.

Although fairness norms appear to have an effect on proposer behavior, we also observe that the perception of what is fair varies substantially among proposers. Considering that the concept of fairness is vague in many situations, it is not surprising that not all proposers agree on what is fair in the power-to-take game. This means that, even if proposers want to be fair when playing the game, they first have to figure out what fairness means in that specific context. Clearly, this opens the door to self-serving biases. However, we find that the disagreement among proposers on what is fair is far greater than the disagreement between proposers and responders.[41] This raises the question whether we are indeed observing the effect of a social norm, which is necessarily linked to what other people think, or rather a personal value. Further research is needed to differentiate between the two possibilities. However, we do feel that the prominence of an emotion like shame points in the direction of there being a social norm that is simply perceived differently by different people. In this case, in order to act optimally, individuals must not only learn how others behave but also the appropriate interpretation of fairness. As individuals interact, they can adjust their beliefs of what is fair. What turns out to be fair in the long run could vary considerably depending on the experiences of those involved in the process. Suggestive in this respect is the observed shift in fair take rates between role-keepers and role-switchers.

---

[40] This difference between dictator and ultimatum games cannot be explained by most models of social preferences (the exception being Levine, 1998). In these models, a proposer who chooses the equal split in the ultimatum game does so because that allocation provides her with the highest utility and not because of the possibility that the responder might reject. Hence, in the dictator game, these proposers should also choose an equal split.

[41] We do not find significant evidence of a self-serving bias. On average, subjects who were only proposers considered that the fair take rate was 32.0% whereas subjects who were only responders considered the fair take rate to be 28.8%. The difference is not significant ($p = 0.923$).

This brings us to our final set of results. That is, individuals that are willing to punish others for treating them badly are not necessarily willing to treat others nicely. So far, the literature on social preferences views negative and positive reciprocity as two sides of the same coin. However, this assumption has not been tested exhaustively. Evidence from public good games with punishment does suggest that individuals who cooperate are indeed also the individuals who punish others for not cooperating (Fehr and Gächter, 2000b). However, it is not clear whether this relation will survive when individuals are playing in different roles across different games. Our results seem to indicate that people behave in accordance with models of social preferences when playing the same role (as is also the case in public good experiments). Yet, they do not behave as these models predict when changing from one role to another.[42]

Our research seems to indicate that the situation of a responder is actually quite different from the situation of a proposer. Whereas the decisions of proposers seem to be influenced by fairness norms and emotions such as shame and guilt, responders seem to react in a different way. As we saw in Chapter 2, responders destroy because they are angry. Furthermore, their anger seems to be triggered by the difference between the take rate and their expected take rate. Their fairness perception seems to play an indirect role by influencing the responders' expectations. To conclude, it appears that *the motivations behind the behavior of responders are different from the motivations behind the behavior of proposers*. More generally, modeling kind and unkind behavior as two separate phenomena might be a fruitful line of research.[43]

Finally, we wish to emphasize that measuring the emotional reaction of subjects can help us understand what is motivating them to make certain decisions. In this case, we identify shame, and possibly guilt, as an important motivation for proposers to reduce their take rates. We do not argue that it is always necessary to know the precise emotional and cognitive processes by which subjects arrive at a decision. However, whenever we have a situation in which our theories are not providing us with good predictions, a better understanding of the motivations of individuals can help improve the modeling of behavior.

[42] Coricelli (2002) finds similar results where, depending on the sequence that games are played, individuals sometimes exhibit negative reciprocity but no positive reciprocity and sometimes the reverse.

[43] See also Loewenstein et al. (1989) for further discussion on the qualitative difference between the reactions of individuals depending on whether they have a positive of negative relationship with others.

Appendix 3A – Instructions

These are the instructions used in the experiment.

Instructions (translation from Dutch)

Welcome to this experiment on decision-making. In this experiment, you can earn money. How much you earn depends on your decisions and the decisions of other participants. In addition to your earnings, you will also receive a show-up fee of 2.50 euros.

The experiment consists of two parts, *Part I* and *Part II*. In each part, you can earn money. Note that, the two parts of the experiment are completely independent of each other. In other words, what you decide in Part I will not affect your earnings in Part II. At the end of Part II, you will be paid privately in cash the total amount you have earned plus your show-up fee. During the experiment, you are not allowed to communicate with other participants. If you have a question, please raise your hand. We will then come to you to answer it. We will begin now with the instructions for Part I of the experiment. You will receive the instructions for Part II after Part I has been completed.

During the experiment, you will be asked to fill in a few questionnaires. The answers you provide in these questionnaires are completely anonymous. They will not be revealed to anyone neither during the experiment nor thereafter. Furthermore, your answers will not affect your earnings during the experiment. If you have any questions now, please raise your hand. If you do not have any questions, please click on 'Ready'.

Instructions - Part I

In Part I of the experiment each of you will be paired to another participant. This other participant will be chosen at random from among the other participants in the experiment.

In this part of the experiment, some of you will be positioned to move first and some of you will be positioned to move second. Participants moving first will be referred to with the letter A, while participants moving second will be referred to with the letter B. Before the experiment started each desk was assigned either an A or a B. Therefore, by randomly picking a yellow card (in the reception room), each participant was randomly assigned to a position in the experiment. The letter that you were assigned is written inside the envelope located on your desk. You will be asked to open the envelope once we finish reading the instructions. The corresponding letter will also appear on the top-right part of the computer screen. Note that each A is paired with a B. Moreover, since the pairing is random, the identities of both participants will remain anonymous.

At the beginning of Part I, all participants (both A and B) receive 10 euros. We will refer to this amount as the *endowment* of each participant. Part I consists of two phases. In phase one, only A must make a decision. Similarly, in phase two, only B must make a decision. Hence, every participant makes only one decision. We will now describe the decision of each A and B.

Phase 1: A chooses a percentage

In this phase, A must choose a percentage and type it into the corresponding field in the computer screen. This percentage determines *how much of the endowment of B after phase two will be transferred to A*. The percentage chosen must be an integer between 0 and 100 (inclusive). If you wish to make any calculations, you can use the calculator located on your desk.

Once you are satisfied with your decision, you have to confirm it by clicking on the button 'Ready'. Note that all decisions are final, once you have clicked on 'Ready' you cannot change your choice. Once A has completed phase 1, phase 2 begins.

Phase 2: B chooses a percentage

At the beginning of this phase, B is informed of the percentage chosen by A. Then, B must also choose a percentage and type it into the corresponding field in the computer screen. This percentage determines *what percentage of B's endowment (of the 10 euros) will be destroyed*. Again, the percentage must be an integer between 0 and 100 inclusive. Note that, the transfer from B to A will be based only on the endowment of B that is not destroyed. Again, if you wish to make any calculations, you can use the calculator located on your desk.

Once you are satisfied with your decision, you have to confirm it by clicking on the button 'Ready'. Note that all decisions are final, once you have clicked on 'Ready' you cannot change your choice. Once B has made his or her decision phase 2 ends.

Earnings

After phase 2, all participants will be informed of the amount of money they have earned. You will also be informed of the amount of money earned by the participant you are paired with.

Example of how to calculate you earnings

We will now give an example for the purpose of illustration. Remember that both A and B have an endowment of 10 euros. Suppose that in phase 1, A decides that 30% of the endowment of B will be transferred to him or her (participant A). In phase 2, B can destroy part or everything of his or her 10 euros. Suppose B decides to destroy 0% percent of his or her endowment. The transfer from B to A is then equal to 3 euros (30% of 10 euros). The earnings of B are equal to 7 euros (namely, the endowment of 10 euros minus the transfer of 3 euros). The earnings of A are equal to 13 euros (namely, the endowment of 10 euros plus the transfer of 3 euros).

Now suppose that in this example B decides to destroy 50% of his or her endowment. In this case, the transfer to A is only 1.50 euros (namely, 30% of the remaining endowment after phase 2, that is 30% of 5 euros). The earnings of A are equal to 11.50 euros (namely, the endowment of 10 euros plus the transfer of 1.5 euros). The earnings of B are equal to 3.50 euro (namely, 50% of the endowment of 10 euros minus the transfer of 1.50 euros).

In summary

In this part of the experiment, each A is randomly and anonymously paired with a B, and each participant receives an endowment of 10 euros. There are two phases. In phase 1, A decides on a percentage that indicates how much of the endowment of B after phase 2 will be transferred to A. In phase 2, B decides what percentage of his or her endowment will be destroyed.

Next, we will ask you to answer a few questions in order to familiarize you with the calculation of your earnings. If you have any questions now, please raise your hand. If you do not have any questions, please click on 'Ready'. Note that once you click on 'Ready' you will not be able to go back to the instructions.

Instructions - Part II

In Part II of the experiment, you will face a situation that is similar to Part I. Each participant will receive an additional 10 euros (which we will call again your endowment). Please note that Part I and Part II are independent so that earnings in Part I will not be affected by your earnings in Part II.

Two differences with respect to Part I

There are two differences between Part I and Part II. One is that your position (A or B) might not be the same, and the other is the participant you are paired with. Again, before the experiment started, each desk was assigned either an A or a B for Part II as well as Part I. Therefore, by randomly assigning the cards; each participant was also randomly assigned to a position in Part II. The position to which you were assigned in Part II will be displayed in the computer screen. Note that, *whichever position you are assigned does not depend on the position you were assigned in Part I*. Furthermore, in Part II, the participant you will be paired with will not be the same participant with whom you were paired in Part I of the experiment. Your new pair will be chosen at random by the computer from among the other participants. In other words, you might be paired with anyone except the participant with whom you were paired in Part I. The rest of the experiment is as in Part I.

In summary

In this part of the experiment, each participant receives an endowment of 10 euros. There are two phases. In phase 1 A decides on a percentage that indicates how much of B's endowment (of Part II) after phase 2 will be transferred to A. In phase 2, B decides what percentage of his or her endowment (of Part II) will be destroyed.

If you have any questions now, please raise your hand. If you do not have any questions, please click on 'Ready'. Note that once you click on 'Ready' you will not be able to go back to the instructions.

Appendix 3B – Descriptive Statistics

TABLE 3B.1 – EMOTIONAL REACTION OF PROPOSERS DEPENDING ON DESTRUCTION

	Proposer in both periods		Proposer only in period 1	Proposer only in period 2
Emotion	**Period one**	**Period two**	**Period one**	**Period two**
	Responder did not destroy			
admiration	3.4 (2.2)	3.6 (2.1)	2.6 (2.0)	2.9 (2.1)
anger	1.0 (0.0)	1.1 (0.3)	1.5 (1.4)	1.6 (1.6)
contempt	1.1 (0.2)	1.1 (0.4)	1.4 (1.0)	1.6 (1.1)
disappointment	1.2 (0.7)	1.2 (0.7)	1.5 (1.3)	1.6 (1.2)
envy	1.1 (0.2)	1.1 (0.3)	1.4 (0.9)	1.6 (0.9)
gratitude	5.1 (1.7)	5.1 (1.6)	3.8 (1.9)	3.6 (2.1)
guilt	1.9 (1.0)	2.8 (1.5)	2.1 (1.8)	2.2 (1.8)
irritation	1.2 (0.5)	1.1 (0.4)	1.5 (1.4)	2.4 (2.0)
joy	5.0 (1.2)	5.1 (1.6)	4.4 (1.5)	3.9 (2.1)
pride	4.1 (1.2)	4.5 (1.7)	3.9 (1.6)	3.4 (2.1)
regret	2.1 (1.4)	2.1 (1.7)	2.2 (1.9)	2.1 (1.9)
sadness	1.1 (0.2)	1.2 (0.7)	1.5 (1.4)	1.7 (1.6)
shame	1.6 (0.9)	2.2 (1.2)	1.8 (1.6)	2.1 (1.4)
surprise	3.6 (2.0)	3.8 (1.6)	2.7 (2.2)	2.8 (1.8)
	Responder destroyed a positive amount			
admiration	2.3 (1.4)	2.0 (1.7)	1.8 (1.1)	3.2 (2.5)
anger	4.7 (2.6)	4.3 (1.2)	2.2 (1.6)	2.8 (2.5)
contempt	2.7 (2.1)	3.7 (2.3)	3.4 (2.1)	2.6 (1.5)
disappointment	4.7 (2.2)	4.7 (2.5)	3.4 (1.5)	3.8 (2.6)
envy	3.3 (1.8)	2.3 (2.3)	1.2 (0.4)	3.2 (2.2)
gratitude	2.7 (1.5)	2.3 (2.3)	3.0 (2.7)	3.6 (2.8)
guilt	2.3 (1.5)	2.3 (2.3)	2.2 (1.8)	1.4 (0.5)
irritation	4.7 (1.9)	5.0 (1.0)	3.4 (2.4)	2.6 (2.2)
joy	3.2 (2.0)	2.0 (1.7)	4.8 (2.3)	4.6 (1.9)
pride	3.0 (1.4)	2.0 (1.7)	3.4 (2.1)	5.0 (2.5)
regret	3.2 (1.7)	3.7 (2.5)	1.0 (0.0)	1.6 (0.9)
sadness	3.0 (2.3)	3.3 (2.5)	1.2 (0.4)	1.8 (0.8)
shame	3.0 (2.0)	2.3 (2.3)	1.4 (0.5)	2.2 (1.8)
surprise	4.0 (2.4)	5.3 (2.1)	4.4 (0.5)	3.2 (3.0)

Note: Numbers between brackets are standard deviations.

Chapter 4

The Revenge of the Shameless

Emotions and the Cost of Social Punishment[*]

This chapter studies the effects of punishment on cooperation when it is possible to retaliate against the punishers. The goal is to investigate when and for what reason the punished retaliate or refrain from doing so, and to explore the consequences thereof.

4.1 Introduction

Cooperation in social dilemmas is a phenomenon that is important to understand. Contrary to the predictions of theories that assume rational own-payoff-maximizing individuals, people cooperate with each other in many situations (e.g. see Ostrom, 1998). Social norms and their enforcement through informal sanctions seem to be an important mechanism for the promotion of cooperation. As shown by Fehr and Gächter (2000b), cooperative behavior can persist when there is an opportunity to punish defectors. However, although punishment can have desirable consequences, it can also have a negative effect on welfare (Fehr and Rockenbach, 2003; Egas and Riedl, 2005; Gächter and Herrmann, 2005). To correctly predict when punishment will have positive results, we must understand the behavior of individuals who punish as well as that of individuals who are punished. To do this, one must realize that emotions play an important role in decision-making (Damasio, 1994; Loewenstein, 1996; Elster, 1999; Thaler, 2000).

The goal of this chapter is to understand the type of motivations that must be present, among both the punishers and the punished, for punishment to be an effective institution for the promotion of cooperation. We concentrate on the role of social emotions, such as shame and guilt, as an essential component for the successful enforcement of cooperative norms. In particular, we are interested in their role as inhibitors of retaliatory behavior by the punished individuals.

Although it has attracted little attention, antisocial behavior such as retaliation or the punishment of cooperative individuals has been observed in various kinds of laboratory experiments, including, for example, public good games (Fehr and Gächter, 2000b), prisoner dilemma games (Falk et al., 2005), and moonlighting games (Abbink et al., 2000). Furthermore, this type of behavior is widespread, it is observed in around one quarter of all subjects (e.g. Falk et al., 2000; Cinyabuguma et al., 2004). It is important to understand the

[*] This chapter is based on Hopfensitz and Reuben (2005).

motivations behind antisocial behavior since, it is not only quite common but, in some cases it can make punishment an inefficient (Egas and Riedl, 2005) and even ineffective (Nikiforakis, 2004; Gächter and Herrmann, 2005) institution for sustaining cooperation.

As was shown in Chapter 2, emotions influence an individual's decision to punish opportunistic behavior. In particular, unkind behavior induces anger and the angrier people are, the more likely they are to incur costs in order to penalize such behavior (see also Bosman and van Winden, 2002; Quervain et al., 2004). However, anger cannot explain whether punishment will effectively promote prosocial behavior. The effectiveness of punishment depends on the reaction of the individuals who are punished. If individuals feel anger after being punished, they may be motivated to retaliate towards the punisher. Therefore, anger alone may induce multiple rounds of punishment and retaliation and consequently a significant destruction of resources. What is missing to make punishment effective is a 'moral' reaction of the punished. Namely, after receiving punishment the punished should act more cooperatively and abstain from retaliation. We will show that the social emotions of shame and guilt motivate individuals to react in precisely this way.

Moral behavior has been shown to be critically linked to the ability for emotional reactions (Anderson et al., 1999; Moll et al., 2002). While this is true for emotional reactivity in general, of particular importance are emotions that facilitate prosocial behavior (i.e. prosocial emotions such as shame, guilt and empathy; see Bowles and Gintis, 2001). They do so by inducing a feeling of discomfort when doing something that violates one's values or norms, or those of other agents whose opinion one cares about. Shame and guilt are both 'self-reproach' emotions elicited by the individuals' own blameworthy actions (Ortony et al., 1988). While they differ in multiple dimensions concerning elicitation and action tendency (Tangney and Dearing, 2002), they are two very similar emotions and are often elicited at the same time.

The influence of prosocial emotions on behavior might be twofold. First, the anticipation and wish of avoidance of shame and guilt might induce norm-abiding behavior. Second, the experience of shame or guilt, after an action, might lead to behaviors to diminish the feeling. This can be through repayment, future cooperation or avoidance of contact with the interaction partner. If these emotions are elicited through punishment of selfish behavior, they might inhibit retaliation and encourage individuals to act more cooperatively in the future.

To test whether this true, we study, by means of an experiment, cooperation and punishment behavior in a social dilemma game. We introduce a more realistic punishment institution where individuals who are punished always have the opportunity to retaliate. After all, if a punishment technology exists, it is likely that both the punisher and the punished have access to it. Indeed, we find that many individuals punish back after being punished. In various cases, this escalates as individuals punish each other in turns, resulting in considerable welfare losses. Nevertheless, this punishment institution is still effective for sustaining cooperation.

In order to explain the behavior of both punishers and punished, we control for the emotional experience of 'punishment-inducing' emotions such as anger and irritation and 'norm-enforcing' emotions like shame and guilt. An important finding is that individuals that act unkindly do nevertheless feel considerably angry when punished. Consequently, individuals retaliate unless feelings of shame restrain the anger-induced desire to fight back. Finally, in order to observe the effect of retaliation on future cooperative and punishment behavior, we had individuals play the game twice. We find that although retaliation considerably increases the cost of punishing opportunistic behavior, it does not deter future cooperation or punishment. Hence, its effect seems to be restricted to welfare loses caused by the destruction of resources.

The chapter is organized as follows. In Section 4.2 we describe the design of the experiment. Section 4.3 describes the subjects' behavior. In Sections Result 4.1 and 4.5, we analyze the relationship between the emotions and the behavior of the punishers and the punished. Section 4.6 discusses the results and concludes.

4.2 The Experiment

Lately, punishment mechanisms have been analyzed in the context of public good games (using the design of Fehr and Gächter, 2000b). However, in this study we use a simpler setting where the causes and effects of emotions can be easily observed and analyzed. To study the impact of social emotions, we used a two-person social dilemma game with and without punishment opportunities. Our game is similar to many of the social dilemma games in the literature, such as, the sequential prisoners' dilemma, the investment game (Berg et al., 1995), the gift exchange game (Akerlof, 1982; Fehr et al., 1993), etc.

4.2.1 The game

We first describe the game without punishment opportunities and then we explain how punishment is introduced. The game consists of two players taking part in a one-shot game. We will refer to these players as the 'first mover' and the 'second mover'. At the start of the game, the first mover receives 150 points whereas the second mover receives 100 points (see Figure 4.1 for the game tree). In the first stage, the first mover decides to either defect or cooperate. If the first mover defects, he keeps his 150 points, the second mover keeps her 100 points, and the game ends. If the first mover cooperates, 50 of his 150 points are multiplied by six and transferred to the second mover. Thus, the second mover receives 300 points while the first mover loses 50 points. In the second stage, the second mover returns an amount of points (r) back to the first mover. Specifically, she could return 150 points (an equal split of the gains), 50 points (returning exactly the points lost by the first mover), or 0 points. After the decision of the second mover the game ends. Hence, if the first mover cooperates his payoff is $\pi_1 = 100 + r$ and the payoff of the second mover is $\pi_2 = 100 + 6 \times 50 - r$. This describes the game without punishment.

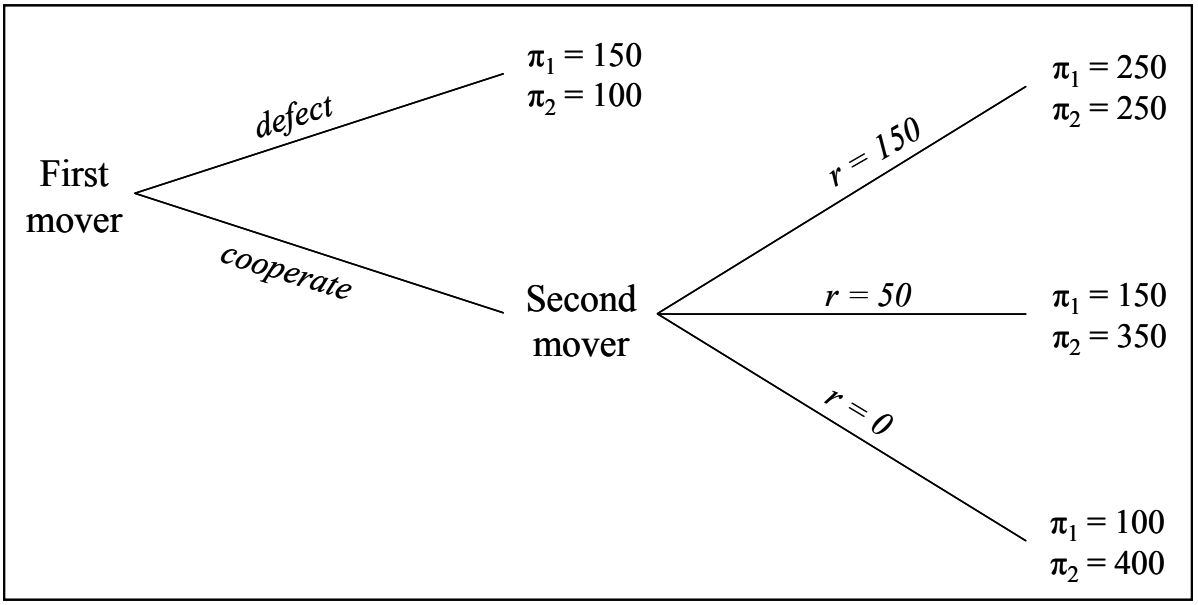

FIGURE 4.1 – GAME TREE IN THE CASE OF NO PUNISHMENT OPPORTUNITIES

In the game with punishment both players can assign punishment points. Doing so is costly for both players. We denote p_{it} as the amount of points assigned by player i (for $i \in \{1,2\}$) in punishment round t. After the second mover decides how much to return, the first round of punishment starts. First, the first mover gets the opportunity to assign a nonnegative amount of punishment points to the second mover (p_{11}). The first mover looses p_{11} points and the second mover looses $4 \times p_{11}$ points. In order to avoid losses during the experiment, the first mover could assign punishment points only as long as the second mover had a positive number of points (i.e. $¼(100 + 6 \times 50 - r) \geq p_{11} \geq 0$). If the first mover chooses $p_{11} = 0$ the game ends. However, if the first mover chooses $p_{11} > 0$ the second mover gets the opportunity to assign punishment points to the first mover (p_{21}). In order to avoid confusion, we will refer to punishment by the second mover as retaliation. Punishment by first movers and retaliation by second movers had the same cost and did the same amount of harm. Thus for each retaliation point assigned, the first mover looses four points. Once more, the second mover could assign retaliation points only as long the first mover had a positive number of points (i.e. $¼(100 + r - p_{11}) \geq p_{21} \geq 0$). If $p_{21} = 0$ the game ends, but if $p_{21} > 0$ the game continues to a second round of punishment. That is, the first mover gets the opportunity to assign additional punishment points to the second mover (p_{12}). As before, if $p_{12} = 0$ the game ends but if $p_{12} > 0$, the second mover gets the opportunity to assign additional retaliation points (p_{22}), and so on. The process repeats itself until either one of the players has zero points and cannot be punished further, or one of the players assigns zero punishment points. Therefore, if

the first mover cooperates his payoff is $\pi_1 = 100 + r - \Sigma_t p_{1t} - 4 \times \Sigma_t p_{2t}$ and the payoff of the second mover is $\pi_2 = 100 + 6 \times 50 - r - \Sigma_t p_{2t} - 4 \times \Sigma_t p_{1t}$.[44]

If we use the standard assumption of rational individuals with self-regarding preferences, the unique subgame-perfect Nash equilibrium of the game with and without punishment, is for second movers to return zero points and thus for first movers not to cooperate.[45] The predictions can change if individuals possess other-regarding preferences such as a concern for unequal payoffs, efficient outcomes, and/or reciprocating kind and unkind actions.[46] In the game without punishment, if the frequency of selfish individuals is sufficiently low then there can be equilibria where some second movers return positive amounts and some first movers cooperate. In the game with punishment, in addition to individuals who are willing to act kindly, there might be individuals who are willing to punish selfish behavior. If punishment leads to higher returns from the second movers, it gives first movers an additional incentive to cooperate.[47] Certainly, the first movers' willingness to punish depends on the willingness of second movers to retaliate, which in turn depends on the willingness of first movers to punish once again, and so on. This, in our opinion is a more realistic way of modeling social punishment. If both the punisher and the punished have access to the punishment technology, the punished will always have the opportunity to retaliate. Moreover, both players have the option to avoid further interaction by deciding not to punish and thus ending the game. To our knowledge, no other study examines punishment behavior in such a setting.[48]

4.2.2 Experimental design and procedures

The computerized experiment was conducted in March 2005 in the CREED laboratory at the University of Amsterdam. In total 162 students from the University of Amsterdam participated in the experiment. Approximately 54% were students of economics and the rest came from a variety of fields such as biology, political science, law, and psychology. Moreover, 58% of the participants were male.

---

[44] Note that players can have negative earnings if by assigning punishment points to the other player they reduce their own earnings below zero. This way a subject cannot avoid punishment or retaliation by reducing the earnings of the other to zero. A show-up fee was given to cover any losses incurred during the experiment.

[45] Note that since punishment is always costly, it is never credible at any stage.

[46] See Rabin (1993), Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Charness and Rabin (2002), Dufwenberg and Kirchsteiger (2005), and Falk and Fischbacher (2005).

[47] For example, using the same assumptions they use about the distribution of types, the model of Fehr and Schmidt (1999) predicts that, in the case of no punishment, 40% of second movers would return 150 points. In this situation only 30% of first movers cooperate (the other 70% prefers to avoid the chance of ending up with extreme disadvantageous inequality). In the case of punishment, there are enough first movers that would punish so that all second movers return 150 points and hence all first movers cooperate.

[48] Nikiforakis (2004) studies punishment in a public good game in which retaliation was possible. However, in this case the punishment phase automatically ended after retaliation. As we will see, this restriction might have limited the amount of initial punishment.

Each subject played *twice* the social dilemma game described in the previous section. We used a perfect strangers matching protocol to avoid any reputation effects. In total, 26 subjects participated in the baseline treatment, that is, the game without punishment opportunities. The remaining 136 subjects participated in the punishment treatment. Earnings were calculated in points and points were exchanged for money at a rate of 40 points for 1 euro. The average earnings were 10.55 euros (this includes a show-up fee of 1 euro). The whole experiment lasted about one hour. Subjects were recruited through the CREED recruitment website and the experiment was programmed with z-Tree (Fischbacher, 1999).

After arrival in the reception room, subjects were randomly assigned to a table in the lab. Once everyone was seated, subjects were given the instructions for the experiment (see Appendix 4A). Subjects were told that the experiment consisted of two independent parts. We emphasized the fact that they will interact with different individuals in each part, and that, their choices in the first part will not affect their earnings in the second part. After this, the one-shot social dilemma game was described as the first part of the experiment. When everybody had finished reading the instructions, subjects had to answer a few questions to ensure their understanding of the game. Subsequently, the subjects played the social dilemma game via the computer (period 1). At the end of the first part, instructions were distributed concerning the second part of the experiment. The instructions informed subjects that they were about to play the same game once again. Furthermore, they would be in the same position as in the first part (i.e. first or second mover), and with certainty, their partner would not be the same partner they had played with in the first part. After they played the second part of the experiment (period 2), subjects filled in a debriefing questionnaire and thereafter they were paid out their earnings in private and dismissed.

To observe if emotional reactions of shame and guilt influence behavior, we used self-reports to measure these and other emotions during the game. We also measured fairness perceptions and expectations concerning the behavior of the other player. Emotions were always measured after subjects observed the choice of the other player but before they made their own choice. Expectations about the behavior of the other player were asked after the subjects made their choice but before they observed the other player's actual choice. Finally, fairness perceptions were measured at the end of the experiment in the debriefing questionnaire.

As in previous chapters, we used self-reported measures of emotions, expectations, and fairness perceptions. Emotions and fairness perceptions were measured using seven-point scales, and expectations were measured by asking for a point estimate of the most likely action.[49] The measured emotions were anger, gratitude, guilt, happiness, irritation, shame, and surprise.

[49] Emotional intensity was measured from: 1 = 'not at all' to 7 = 'very intensely'. The fairness of an action was measured from: 1 = 'very unfair' to 7 = 'very fair'. The questions used are available in Appendix 4A.

4.3 Observed Behavior

In this section, we give an overview and a brief discussion of the behavior of first and second movers. A summary of the behavioral data can be found in Appendix 4B. We start by investigating how often first movers cooperate and, when given the opportunity, how much second movers return. Comparing the baseline and the punishment treatments allows us to observe the effect of the punishment institution on the subjects' behavior. Then, in order to explain any differences induced by punishment, we analyze the punishment behavior of first movers as well as the retaliatory behavior of second movers. Finally, we examine whether punishment and retaliation in period 1 have an effect on their behavior in period 2.

4.3.1 Cooperation and Returns

Figure 4.2 summarizes the main differences between the baseline and the punishment treatment. First movers cooperate more often and second movers return more in the presence of punishment.

As can be seen in Figure 4.2A, in both treatments, almost all first movers cooperate in the first period (more than 84.6%). However, in the absence of punishment, cooperation decreases substantially in the second period. In contrast, if there are punishment opportunities, first movers cooperate equally often in both periods. Testing for differences between treatments confirms this observation. There is no significant difference in the frequency of cooperation in the first period ($p = 0.837$) but a highly significant difference in the second period ($p = 0.001$).[50] There is an even starker difference between treatments when we consider the behavior of second movers. That is, in each period, second movers return noticeably less in the absence of punishment ($p < 0.044$). Given the behavior of second movers, it is easy to understand the decrease in cooperation in the baseline treatment. Remember that first movers who cooperate send 50 points. In the baseline treatment, they receive on average a smaller amount in return. In contrast, first movers who cooperate in the punishment treatment receive back roughly twice the sent amount. It is clear that, even when it is possible to retaliate, punishment limits the opportunistic behavior of second movers.

In spite of this, punishment did not lead to overall higher earnings. In period 1 average earnings are actually higher in the baseline treatment (230.8 vs. 189.0 points), whereas in period 2, average earnings are higher in the punishment treatment (187.3 vs. 182.7 points). However, in neither case is the difference significant ($p > 0.198$). In the following paragraphs, we examine how subjects punish and retaliate.

---

[50] Throughout the chapter, unless it is otherwise noted, we always use a two-sided Wilcoxon-Mann-Whitney test. We use each subject as an independent observation for tests concerning either period 1 or period 2. If we combine the data of both periods, for each subject we first calculate the mean for the variable in question and then compute the test using these means as the independent observations. There are subjects from whom we have data from only one of the periods for various variables (e.g. a second mover who faces a first mover who cooperates in period 1 and a first mover who defects in period 2). In these cases, we take the data from the period for which we have information as that subject's mean.

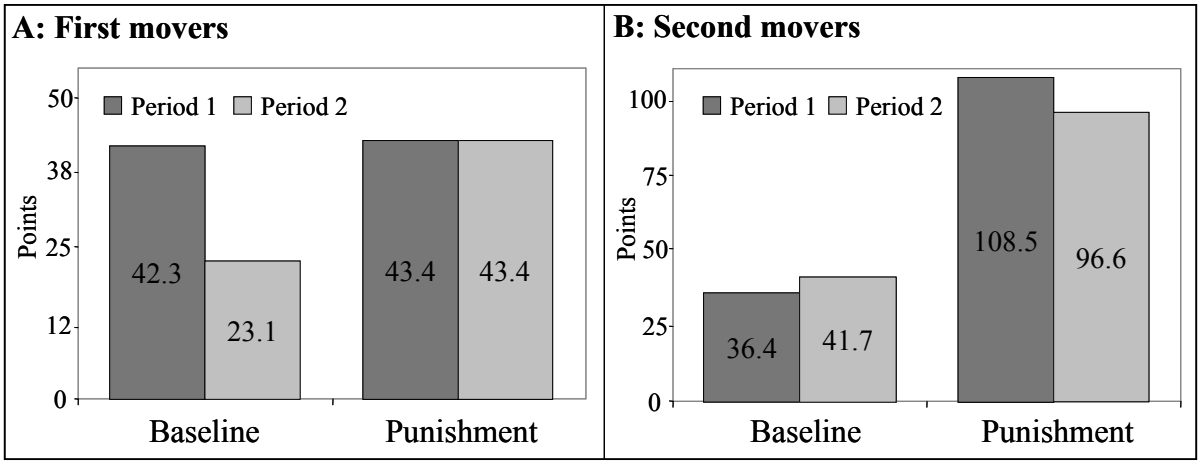

FIGURE 4.2 – COOPERATION BY FIRST MOVERS AND RETURNS BY SECOND MOVERS

Note: A) Mean number of points sent by first movers in each period and treatment. Note that, since first movers could send only 0 or 50 points, if multiplied by two, this is equals the frequency of first movers who cooperate. B) Mean number of points returned by second movers in each period and treatment. For the frequency of second movers sending 0, 50 or 150 points see Appendix 4B.

4.3.2 Punishment and Retaliation

As Figure 4.3A illustrates (see also Table 4B.1), a large number of subjects are willing to spend some or all of their earnings in order to punish second movers or to retaliate against first movers. In fact, around one third of the cases in which first and second movers interact wind up in punishment by the first movers. If returns were less than 150 points, about two thirds of the interactions end up in punishment (68.1%). When given the opportunity, retaliation by second movers is somewhat less frequent (40.0%). We even observe that, of the first movers who had the chance to punish second movers who retaliated, 55.6% decided to do so (we refer to this as 'additional punishment').[51]

Figure 4.3B shows that the amount spent on punishment by first movers who got back less than 150 points was clearly higher than the amount spent on retaliation by second movers who got punished (p = 0.002). Surely, since the earnings of first movers when they faced retaliation were lower than the earnings of second movers when they faced punishment, this is partly explained by the ability of first movers to spend more on reducing the other's payoff. Still, if we normalize both punishment and retaliation using the maximum amount of points that an individual could assign to the other, we see that first movers are more aggressive punishers than second movers (p = 0.080).

[51] We only observe one case in which the second mover retaliated once again ($p_{22} > 0$). However, this is probably because in all the other pairs where the first mover punished a second time ($p_{12} > 0$) at least one of the players ended up with zero points and hence the punishment stage ended automatically.

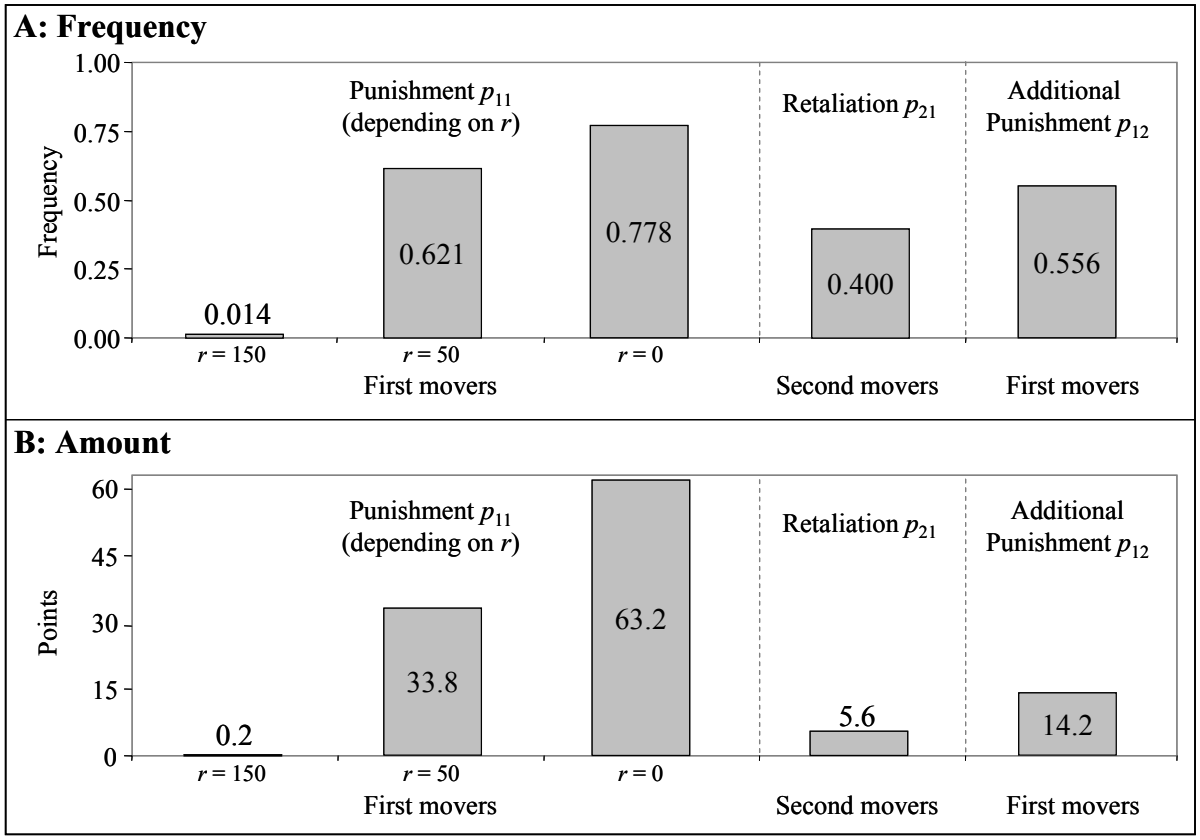

FIGURE 4.3 – PUNISHMENT AND RETALIATION

Note: A) Frequency of punishment (p_{11}), retaliation (p_{21}), and additional punishment (p_{12}) over both periods. B) Mean amount of points spent on punishment (p_{11}), retaliation (p_{21}), and additional punishment (p_{12}) over both periods.

Although it is not predicted by traditional economic theory (assuming own-payoff maximization), the punishment behavior of first movers is not surprising given that similar behavior has been observed in numerous experiments (see Camerer, 2003). Similarly, it is expected that the amount and frequency of punishment increases as the amount returned decreases. First movers who received 150 points punish less and less often than first movers who received 50 or 0 points (in each period $p < 0.001$). If we compare first movers who received 50 points with those who received 0 points, we find that the latter punish significantly more only in the second period ($p = 0.020$, and in the other cases $p > 0.193$).

We find more intriguing the willingness of second movers to retaliate. After all, these subjects had behaved unkindly by returning less than 150 points. Furthermore, when they had to decide whether they wanted to retaliate, 65.0% of the second movers had earnings that were actually higher or equal to the earnings of the first mover. It is remarkable that 7 (i.e. 53.8%) of these 13 second movers chose a positive amount of retaliation.[52] Unlike for first movers, the retaliatory behavior of second movers does not seem to depend on the actions of

[52] If one thinks that low contributors anticipate punishment from high contributors, this behavior is akin to 'misdirected' punishment in public good games (Cinyabuguma et al., 2004; Gächter and Herrmann, 2005).

the other player. For instance, there is no significant difference in the amount or the frequency of retaliation between second movers who received a large amount of punishment and second movers who received a small amount (punishment above and below the median, $p > 0.355$).

It is instructive to calculate how retaliation affects the first movers' 'real' cost of punishment. Whenever first movers punish, they not only incur the cost of reducing the second mover's earnings, but they also risk further losses if the second mover decides to retaliate.[53] If there is no retaliation, the cost of punishment is 0.250 points per point reduced. Including the actual losses due to retaliation increases the average costs of punishment by 0.149 points per point reduced. Nonetheless, even though this is a substantial increase of 59.4%, punishment remains an inexpensive tool for the reduction of the second mover's earnings. This might explain why cooperation is sustained in spite of frequent retaliation. However, more generally the impact of retaliation on the costs of punishment will depend on the game played and its parameters. It is possible that in some cases retaliation will drive the costs of punishment to the point where punishment fails to sustain cooperation.[54] A similar analysis for the real cost of retaliation (given losses due to additional punishment) gives that second movers incur an additional 0.763 points per point reduced. This remarkable 305.6% increase might explain why second movers punish less aggressively than first movers do. We now turn to how first and second movers adjust their behavior from period 1 to period 2.

4.3.3 Dynamics

As already noted, the starkest difference between treatments concerning the behavior of first movers is the large decrease in cooperation from period 1 to period 2 in the baseline treatment compared to the punishment treatment. On closer inspection, this difference is due to two reasons. First, as shown in Figure 4.4A, in the baseline treatment 66.7% of the first movers who got back less than 150 points in period 1 defected in period 2. In contrast, in the punishment treatment it was only 19.0% (the difference is significant, $p = 0.013$). Second, in the baseline treatment more second movers chose to return less than 150 points (81.8% in the baseline treatment vs. 35.6% in the punishment treatment, $p = 0.005$). Hence, it appears that punishment has two desirable effects. On one hand, second movers anticipate punishment and as a result increase the amount returned. On the other hand, after experiencing opportunistic behavior, first movers are more willing to keep on cooperating if they have the opportunity to punish. In fact, if we examine how first movers in the punishment treatment adjust their behavior, we find that, among the first movers who received less than 150 points, those who

[53] The only case in which second movers cannot retaliate after being punished occurs when first movers who get back 0 points spend all of their remaining earnings punishing the second mover. In this case, both subjects end up with 0 points and no further retaliation is possible. Overall, 24.3% of the cases in which there was positive punishment fit this description.

[54] In public good settings, punishment stops sustaining cooperation when the cost of punishing increases over 0.500 per point reduced (Nikiforakis and Normann, 2005).

actually punished are less likely to stop cooperating than those who did not punish ($p = 0.087$, Figure 4.4B).

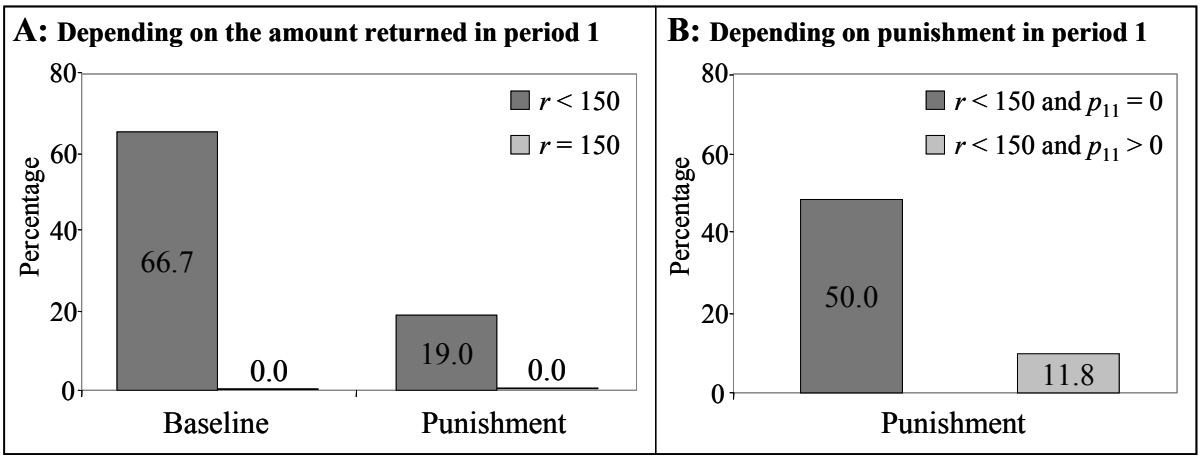

FIGURE 4.4– DEFECTION IN PERIOD 2 DEPENDING ON THE EVENTS IN PERIOD 1

Note: A) Percentage of first movers who defect in period 2 depending on the amount returned by the second mover of period 1 in each treatment. B) Percentage of first movers who defect in period 2 depending on whether or not they punished the second mover of period 1 for returning less than 150 points.

We now turn to the effects of punishment on the future behavior of second movers. If we concentrate on second movers who had a good chance of being punished (i.e. those who returned less than 150), we find that, on average, second movers who were not punished decrease their returned amount by 25.0 points whereas those who were punished increase it by 10.0 points ($p = 0.113$). Hence, although actual punishment does promote prosocial behavior, the effect is not particularly strong. In other words, punishment has a bigger impact by deterring second movers from returning low amounts in the first place than by increasing the returns of those who behave opportunistically in spite of the threat of punishment. For example, if none of the second movers who returned a low amount had been punished in period 1, the average return in period 2 would have been 87.7 points (instead of 96.6 points). In contrast, if the threat of punishment were not there at all then the average return would have been as low as in the baseline treatment (i.e. 41.7 points).

Lastly, we analyze the impact of retaliation on both future cooperation and punishment by first movers. In general, retaliation in period 1 does not deter first movers from cooperating in period 2. For instance, among first movers who punished a low return in period 1, those who received retaliation were as likely to cooperate in period 2 as those who received no retaliation ($p = 0.480$). It is also the case that retaliation does not deter first movers from punishing. Among the first movers who punished in period 1 and then received a low return in period 2, those who had received positive retaliation punished in period 2 as often as those who had received no retaliation ($p = 0.414$). In fact, they punished as often as those who received a low return in period 2 after they had received a high return in period 1 ($p = 0.228$). The main findings from the behavioral data are summarized in the following result:

RESULT 4.1: *In the presence of punishment opportunities, cooperation is sustained at high levels. This is because, second movers return more, and first movers who punish do not stop cooperating after experiencing opportunistic behavior. Punishment of opportunistic behavior is common and persistent despite the fact that in numerous cases punishment leads to retaliation by second movers*.

4.4 Emotions and Punishment

In this section, we first examine which of the first movers' emotions are related to punishment. We find that anger-like emotions explain why some first movers punish while others do not. Subsequently, we concentrate on anger and analyze what triggers first movers to feel high intensities of this emotion.

4.4.1 Anger and Punishment

Throughout the experiment, first movers experienced a variety of emotions. However, we find that anger-like emotions are the only ones that are clearly related to the punishment decision. First movers that felt high intensities of anger-like emotions punish more than those who felt low intensities of these emotions. Furthermore, we also find that differences in anger-like emotions can explain why, after receiving retaliation, some first movers punish again while others do not.

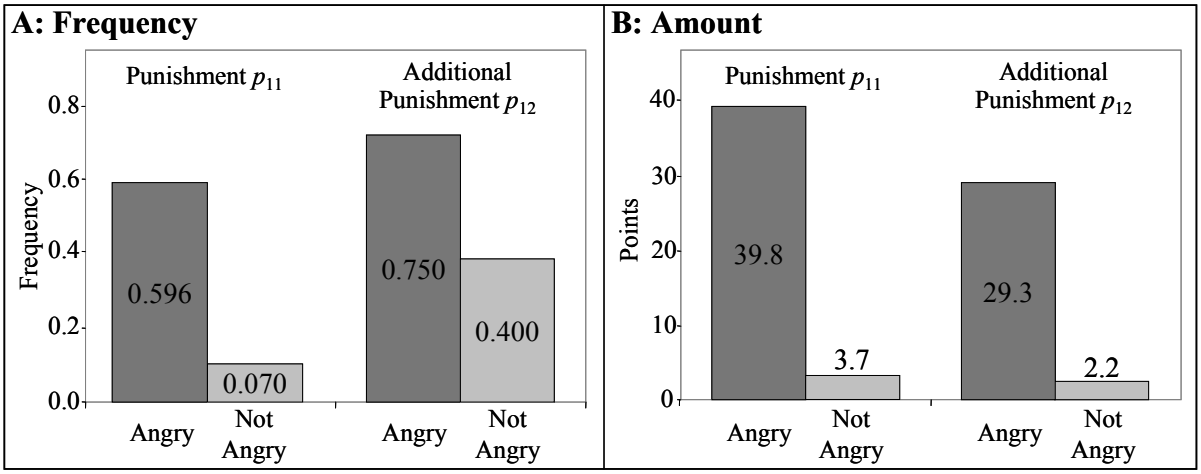

FIGURE 4.5 – ANGER AND PUNISHMENT

Note: A) Frequency of punishment by first movers depending on anger. B) Mean amount of points spent on punishment by first movers depending on anger.

As is illustrated in Figure 4.5, first movers who, after observing the amount returned by the second mover, felt high intensities of anger punish more and more often than first movers who felt low intensities of anger ($p < 0.001$ for both period 1 and 2).[55] Similarly, on average,

[55] In the following analysis we will refer to a person feeling 'angry' if the reported value for anger was above the median, and as feeling 'not angry' if the value was below the median. Likewise for other emotions.

after observing the amount of retaliation assigned to them by the second mover, first movers who felt angry punish more and more often than first movers who did not feel as angry (the difference is marginally significant for the amount of additional punishment $p = 0.096$, but not for its frequency $p = 0.322$).[56]

Having found that punishment is related to experienced anger, the question arises what explains the different intensities of anger. We answer this question in the following subsection.

4.4.2 Causes of Anger

Parallel to our findings in Chapter 2, anger experienced after observing the amount sent back by the second mover is caused by returns of less than 150 points, especially if they were unexpected or considered unfair (the emotional reaction of first movers to the amount returned can be found in Appendix 4B).

In both treatments, the most important trigger of high intensities of anger is simply receiving back less than 150 points. First movers in the punishment treatment who received 150 points felt lower intensities of anger than first movers who received either 50 or 0 points back ($p < 0.001$, see Table 4B.3). Moreover, although on average first movers who received 0 points were angrier than those who received 50 points, the difference is marginally significant only in the second period ($p = 0.328$ for period 1 and $p = 0.075$ for period 2).

In addition to the returned amount, the first movers' expectations have an effect on the intensity of anger. In particular, first movers who overestimated the amount returned by the second mover tended to be angrier than first movers who underestimated it. For example, if we control for the amount that was actually returned by concentrating on first movers who got back 50 points, we find that first movers who were expecting back 150 points were angrier than first movers who were expecting back 50 or 0 points (in each period $p < 0.043$).

Lastly, we also observe that fairness perceptions influence the amount of anger experienced by first movers. First movers who thought it is unfair to return low amounts were angrier than those who thought that it is fair to return low amounts (below or above median fairness). For instance, if we look again only at first movers who got back 50 points, we find that first movers who thought returning 50 was unfair were angrier than first movers who thought returning 50 was fair ($p = 0.004$).

We get similar results in a regression. Specifically, we estimate anger using the returned amount, the expected returned amount, the perceived fairness of returning 50 points, and some demographic variables. We find first movers feel angrier the less is returned, especially if they were expecting a return of 150 points or considered low returns very unfair (see Table 4C.1 in Appendix 4C).

---

[56] Throughout this section, we report the results of tests done with the emotion of anger. However, we find very similar results and significance levels if we use irritation or (lack of) happiness.

Focusing on the emotional reaction of first movers to the amount of retaliation received from the second mover gives a comparable finding. Namely, first movers who faced no retaliation experienced lower intensities of anger than first movers who faced positive retaliation (p = 0.037, see Table 4B.4). Unfortunately, in this case we do not have enough observations to test for the effects of expectations and fairness perceptions. The findings of this section are summarized in the following result.

RESULT 4.2: *First movers who punish do so because they are angry. High intensities of anger are triggered by opportunistic behavior by the second mover, especially if it is unexpected and considered unfair. Retaliation by second movers also makes first movers angry and leads to additional punishment*.

4.5 Social Emotions and Retaliation

We now turn to the relationship between the emotions and behavior of second movers. To begin with, we investigate the relationship between the emotions of second movers and their decision to retaliate. We also analyze whether emotions influence how second movers adjust their behavior over time. Next, we try to explain the difference in the emotional reactions of second movers.

4.5.1 Shame and Retaliation

As for first movers, the emotional reaction of second movers is clearly related to their behavior (the emotional reaction of second movers can be found in Table 4B.5). In particular, second movers who felt no shame are more likely to retaliate than other second movers. Furthermore, we also find that, for second movers who were punished, experiencing shame induces them to correct their behavior.

As can be seen in Figure 4.6A, second movers who felt no shame after being punished are more likely to retaliate than second movers who felt shame (p = 0.045).[57] We also get a similar result if we test for differences in the amount of points spent on retaliation (p = 0.091).

Interestingly, we also find that anger has an effect on the second movers' decision to retaliate. However, in this case the effect is not as straightforward. A simple look at the relationship between anger and retaliation, suggests that second movers who are angry retaliate more and more often than second movers who are not angry (see Figure 4.6). However, these differences are not significant (p = 0.739 when testing for differences in the amount of retaliation and p = 0.965 for differences in frequency).

[57] We only report the results of tests using shame. However, for all findings in this section, we get very similar results and significance levels if we use guilt instead of shame.

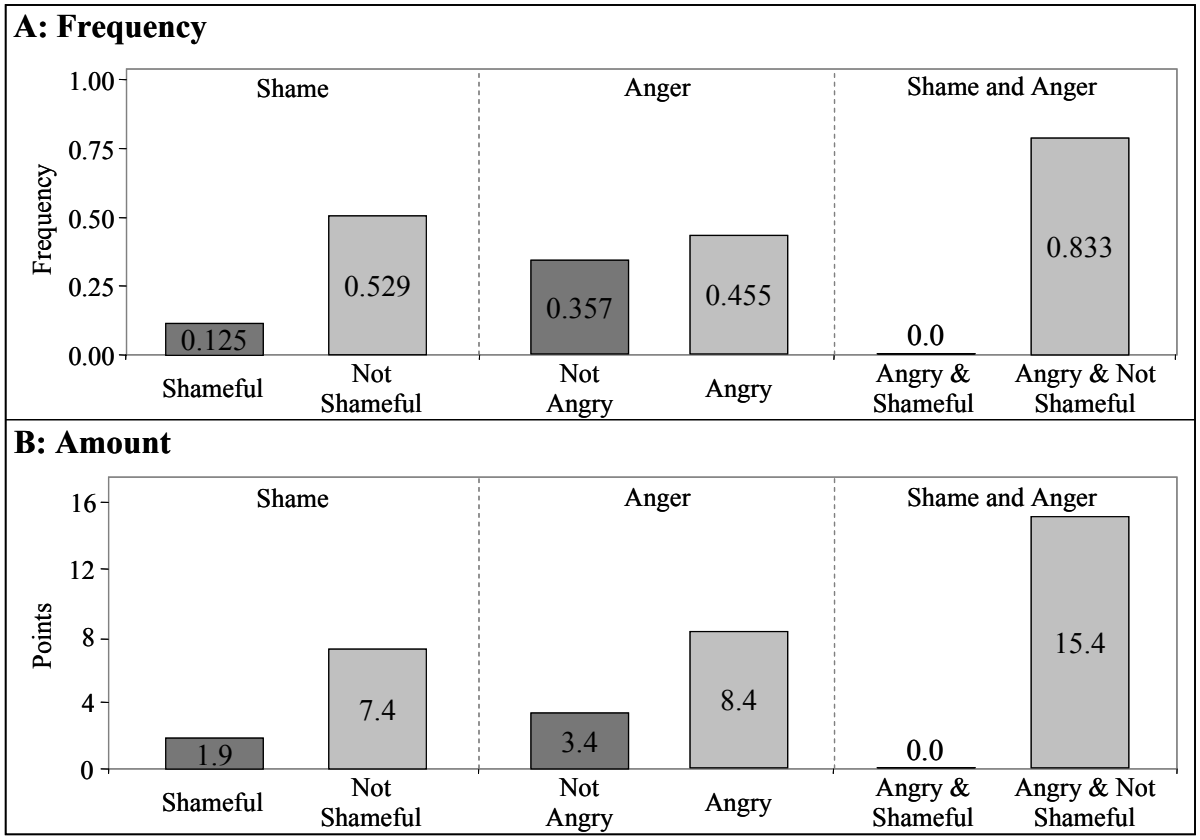

FIGURE 4.6 – SHAME, ANGER, AND RETALIATION

Note: A) Frequency of retaliation by second movers depending on anger and shame. B) Mean amount of points spent on retaliation by second movers depending on anger and shameful.

The effect of anger becomes obvious once we examine the interaction of anger and shame. In this case, a clear result is obtained. Namely, second movers who were angry and felt no shame retaliate more and more frequently than second movers who were angry and felt shame (p = 0.032 and p = 0.024). For second movers who were not angry, there are no significant differences between those who felt no shame and those who did (p > 0.637).

As in Chapter 3, shame is also related to how individuals adjust their behavior from period 1 to period 2. In Section 4.3.3 it was shown that second movers who were punished tend to return more in the subsequent period than second movers who were not punished. However, this difference is not significant. The emotional reaction of second movers reveals that punishment induces higher returns only if it is complemented by feelings of shame. On average, second movers who felt shame after being punished increase the amount returned by 35.7 points whereas those who did not feel shame decrease the amount returned by 12.5 points (p = 0.053).[58]

In conclusion, our results suggest that high intensities of anger provide second movers with a motivation to retaliate and high intensities of shame restrain them from doing so.

[58] Since most second movers who returned less then 150 points were punished, we do not have enough observations to test the effects of shame on subjects that were not punished.

Furthermore, shame seems to be necessary for punishment to have an effect on how second movers adjust their behavior. Next, we explain the differences in the intensities of anger and shame experienced by second movers.

4.5.2 Causes of anger and shame

The experience of anger among second movers depends on how many points they sent back to the first mover and on the amount of points the first mover spent punishing them. That is, second movers felt high intensities of anger if they received a high amount of punishment from the first mover. Furthermore, the intensity of anger is stronger the higher the amount they had returned before being punished.

The most important reason why second movers get angry is simply receiving a positive amount of punishment (see Table 4B.5). For example, second movers who were punished at least once reported significantly more anger than those who were never punished ($p = 0.001$).[59] Interestingly, if we examine whether the amount of punishment has an effect on anger we do not find a significant result. For example, second movers who were punished by a very large amount were not significantly angrier than those who were punished by a very small amount (top versus bottom quartile, $p = 0.624$). However, once we take into account the amount the second mover returned, we find a clearer effect. Among second movers who returned 50 points, those who were punished by a very large amount were angrier than those who were punished by a very small amount (top versus bottom quartile, $p = 0.133$). The same pattern exists for second movers who returned 0 points (this time, $p = 0.168$).

These effects are more clearly captured in a regression. We estimate anger using the following independent variables: the amount returned, the expected amount of punishment, the perceived fairness of returning 50 points, some demographic variables, and three variables capturing the interaction between the amount of punishment and the amount returned.[60] The regression is available in Table 4C.2. We find that punishment triggers high intensities of anger. Moreover, the increase in anger is bigger the higher the return of the second mover. More concisely, second movers became angry whenever they were punished, but if they had returned 50 instead of 0 points, they got angry at lower punishment amounts. This is understandable given that second movers who returned 50 points, not only behaved somewhat nicer than those who returned less, they also had lower earnings. Unlike for first movers, we do not find that fairness perceptions or expectations (about the amount of punishment) have an effect on anger.

---

[59] This is also true if we restrict ourselves to second movers who returned less than 150 points ($p = 0.002$).

[60] We use three variables I^r with $r \in \{0, 50, 150\}$. $I^r = 0$ if the amount returned was different from r and I^r = the amount of punishment received if the amount returned was r. We obtain positive and significant coefficients for I^0, I^{50}, and I^{150} ($p < 0.001$). Furthermore, the coefficients are all significantly different from each other, with the coefficient for I^0 being the smallest and the one for I^{150} being the largest (Wald tests, $p < 0.012$). See Table 4C.3 in Appendix 4C for details.

As for anger, the intensity of shame is related to the amount returned and to the amount of punishment received from the first mover. The clearest trigger of high intensities of shame is the amount the second mover returned. Second movers who returned 150 points reported lower intensities of shame than those who returned less (in all treatments $p < 0.001$). In the punishment treatment, this is true even when we control for whether or not the second mover faced punishment. Specifically, second movers who returned 150 points and were not punished felt lower intensities of shame than second movers who returned less and were not punished ($p = 0.001$). If anticipated, this type of emotional reaction supports the idea that some individuals will avoid opportunistic acts in order not to feel high intensities of shame.

The effect of punishment on shame is not as straightforward. Among second movers who returned less than 150 points, second movers who were punished by a low amount (below the median) reported significantly lower intensities of shame than second movers who were punished by a high amount (1.21 vs. 1.83, $p = 0.035$). Hence, it looks like punishment triggers higher intensities of shame. However, if we compare second movers who were punished by a low amount to second movers who were not punished at all, we find that those who were not punished reported slightly higher intensities of shame (1.21 vs. 1.50, $p = 0.154$). Second movers who were not punished reported the same intensities of shame than those who where punished by a high amount ($p = 0.534$). Thus, punishment seems to have a nonlinear effect so that a small amount of punishment actually triggers lower intensities of shame.

This is also seen in a regression. We estimate shame using the following independent variables: the expected amount of punishment, the perceived fairness of returning 50 points, some demographic variables, and three variables representing second movers who returned a low amount and received zero, a small, or a big amount of punishment. The regression is available in Table 4C.3. We find that returns below 150 points trigger high intensities of shame only if the second mover was either not punished or punished by a big amount. The regression also shows a gender effect, that is, women report lower intensities of shame. The findings of this section are summarized in the following result.

RESULT 4.3: *Second movers who retaliate do so because they are angry and do not feel shame. In addition, following the feeling of shame, second movers rectify their opportunistic behavior. High intensities of anger are triggered by punishment, especially if the second mover had returned a positive amount. High intensities of shame are triggered by opportunistic behavior and by substantial amounts of punishment.*

4.6 Discussion and Conclusions

In this chapter, we have shown that a realistic punishment institution, in which multiple rounds of punishment and retaliation are possible, is an effective tool for the support of cooperative behavior. However, retaliation is a commonly observed behavior that often results in the extreme reduction of the payoffs of the individuals involved. Furthermore, we confirm

some of the findings of Chapter 2, that is, anger-like emotions are an important motivation for punishment. Opportunistic behavior induces anger and thus increases the likelihood of punishment. Lastly, we have shown that the experience of prosocial emotions, namely shame and guilt, restrain angry individuals from retaliating. Therefore, prosocial emotions can be seen as a mechanism managing the behavioral reactions of anger.

Given that costly punishment has been shown to be an effective way of enforcing cooperative behavior, it is important to have a good understanding of the motivations and reactions of both the punishers and the punished. We find interesting that individuals who are willing to punish are also willing to keep on cooperating (see Result 4.1). This guaranties that, as long as these individuals have the opportunity to punish, cooperation can be sustained. This kind of behavior is essential for cooperation to thrive, especially if it was initially rare. In addition, the same type of behavior is necessary to support punishment in the presence of retaliation. If retaliation deters individuals from using the punishment mechanism, cooperation can unravel (Nikiforakis, 2004). However, if the opportunity to punish back always exists, this could prevent retaliation from limiting the punishment of opportunistic behavior.[61]

An important and yet overlooked aspect of punishment is the emotional reaction of the punished. As was shown in this chapter, prosocial emotions such as shame play a crucial role for the viability of punishment for the enforcement of social norms. In Section 4.5 we have shown that feeling shameful helps explain why some individuals who acted selfishly adjust their behavior whereas others do not. It has been observed that in public good games, the use of non-monetary punishment has a positive effect on contribution levels.[62] However, our results indicate that it is the combination of feeling shame and receiving substantial monetary punishment that has a significant effect on behavior. This suggests that shame alone will not have an effect if the cooperative norm is not vigorously enforced. Hence, although non-monetary punishment has the desirable property that it can affect behavior without destroying resources, the lack of real consequences for free-riders make this effect deteriorate over time (Masclet et al., 2003). In this sense, as is shown by Noussair and Tucker (2005), the best performing punishment institution is one in which both symbolic and monetary punishments are available.

Another essential role for shame is the prevention of retaliation by punished individuals. As was shown in Section Result 4.1, even if they acted unkindly, individuals do feel angry when they are punished. However, it is only those individuals who are angry and do not feel shame that decide to retaliate. Therefore, if it were not for some individuals experiencing shame, retaliation would be much more common and punishment of selfish

[61] As we have shown in Section 4.3, retaliation did not subdue punishment of low returns. Unfortunately, we do not have enough observations to determine if retaliation deters additional punishment.

[62] For instance, Masclet et al. (2003) use symbolic punishment points and find that, in the short run, they work almost as well as real punishment points. Barr (2001) reports that the public blaming of the free-rider can increase cooperation in future rounds.

behavior much more costly. For example, if second movers who felt shame had behaved as second movers who felt no shame (controlling for anger) then retaliation would have been 42.6% more frequent and 50.6% higher. Furthermore, the decrease in the amount returned from period 1 to period 2 would have been 48.8% more severe. Social emotions like shame are thus essential for the effectiveness of a punishment institution. This fits the assumption that social emotions coevolved with institutions and anger-like emotions in order to limit antisocial actions (Bowles and Gintis, 2001). An interesting question for further exploration is the specific evolutionary mechanisms that lead to this situation.

Finally, even though we did not differentiate in our analysis between shame and guilt, we would like to stress that the action tendencies of the two emotions can be different (Tangney and Dearing, 2002). Guilt is more related to the blameworthiness of an act and is thus more likely to result in reparation and action. Shame is related to a devaluation of the self. Therefore the action tendency of shame is withdrawal and avoidance of further contact.[63] Therefore, increasing feelings of shame (e.g. through framing) might not always lead to an increase in prosocial behavior. For instance, if individuals have the possibility to avoid contact altogether, they might prefer to do so instead of participating in an activity where feelings of shame 'force' them to act prosocially (see Lazear et al., 2005). In other words, when trying to decrease the frequency of selfish behavior, the attempt to explicitly induce shame, might result in avoidance of further interaction instead of in more cooperation.

Appendix 4A – Instructions

These are the instructions for the first movers used in the punishment treatment. The instructions for the second mover and for the baseline treatment are available upon request.

Instructions for Part 1

There are two types of participants in this part, participants A and participants B. Half of the persons participating in the experiment will be in the role of participant A, and the other half in that of participant B. *You are a participant A*.

In part 1 of the experiment, you will be randomly assigned a participant B. During this part, you will interact only with this participant B. Moreover, you will *not* interact again with this participant in part 2 of the experiment. Part 1 consists of three steps. In step one, you must decide whether you will transfer points to participant B or if you will retain the points for yourself. In step two, participant B will decide if he will transfer points to you or if he will keep them himself. In step three, both of you must again make a decision. There are various

[63] Economists usually distinguish shame and guilt by the visibility of behavior. Shame is said to be triggered in social situations in which actions are seen by others, whereas guilt is more related to internalized values and hence is not influenced by the presence of others (e.g. Kandel and Lazear, 1992). However, research by psychologists has shown that people feel shame even when their actions are unobserved (Tangney et al., 1996), and that the experience of guilt varies considerably depending on the interpersonal context (Baumeister et al., 1994).

options in step three, which will be explained below. We will also describe the exact experimental procedure on the next pages.

Procedure for the three steps

At the beginning of part 1 you and participant B will each receive 100 points as earnings.

Step one

At the beginning of the first step you will receive 50 decision points. Participant B will receive no decision points. In step one, you must decide whether you want to transfer your 50 decision points to participant B or transfer no points to participant B. If you transfer the 50 points, they will be multiplied by six, meaning that participant B will receive $6 \times 50 = 300$ points. Then, step two begins. If you decide to transfer nothing part 1 will end here.

Step two

In step two, participant B has to decide whether he will transfer 150, 50 or 0 points to you. You will then receive exactly the number of points B transferred.

Therefore, four possibilities exist after the first two steps:

	Your additional earnings	**B's additional earnings**
You retain your decision points.	50 points	0 points
You transfer your decision points and B transfers 150 points.	150 points	150 points
You transfer your decision points and B transfers 50 points.	50 points	250 points
You transfer your decision points and B transfers nothing.	0 points	300 points

Hence, after step two your total earnings will be:

100 + the additional earnings from the table above.

Step three

In step three, you will be informed how many points participant B transferred to you. Now, you can assign penalty points to participant B. The assignment of penalty points has financial consequences for both participants, A and B. Each penalty point which you assign costs you one point, while four points are deducted from your participant B. If you assign three penalty points to participant B, this will cost you three points and participant B will have twelve points deducted.

You cannot deduct more points from participant B than his total earnings in that part (i.e. 100 + B's additional earnings). If participant B has 250 points after step 2, then with your assignment of penalty points you can reduce his earnings by at most 250 points. Hence, as long as your participant B has positive earnings, you can assign him as many penalty points as you want. You can also assign him no penalty points.

Participant B will then be informed how many penalty points you assigned him and how many points were deducted from his earnings. If you decided not to assign penalty points, part 1 will end here. If you assigned penalty points to participant B, he can decide to assign penalty points to you. The assignment of penalty points has the same financial consequences as described above. Each penalty point that participant B assigns to you costs him one point, while four points are deducted from your earnings. You cannot be deducted more points than the total earnings you own at that moment. If participant B decides to assign no penalty points to you, part 1 will end here. Note: Participant B can assign penalty points even if his earnings at that point are zero. If he does so, he will lose points in part 1 of the experiment.

If participant B assigned you penalty points, you and participant B will have the option to assign penalty points to each other in turns. Part 1 will end when either you or participant B decides to assign no penalty points, or if either you or participant B can not be assigned penalty points because your or his earnings are zero or less. In other words, as long as one of you assigns a positive amount of penalty points, the other will have the opportunity to assign penalty points back. Note that, you will be able to assign penalty points *even if your earnings at that point are zero*. Furthermore, you *cannot* be assigned penalty points if your *own* earnings are zero.

Finally

Remember that you participate in part 1 only once. Therefore consider your decisions carefully. At the end of part 1 you will receive instructions for part 2 of the experiment.

Instructions for Part 2

We will now give you the instructions for part 2 of the experiment.

In this part there will also be two types of participants, participants A and participants B. Every person participating in the experiment will be in the role they had in part 1. Therefore, *you are a participant A*. As in part 1 you will be randomly assigned a participant B. During this part, you will interact only with this participant B. You can be certain that *this participant B is not the same person as in part 1*.

This part will consist of the same three steps as part 1. Therefore exactly the same instructions apply for part 2 as for part 1. Remember that you will participate in this part only *once*. Therefore consider your decisions carefully.

Examples of questions in the self-reports

To measure emotions:
Indicate how intensely you feel each of the following emotions right now, *after knowing the amount that B transferred to you*?
The subject then filled in a series of seven-point scales that ranged from 'not at all' (1) to 'very intensely' (7).

To measure expectations:
Player A can now assign you penalty points. How many penalty points do you think A will assign to you?
The subject then entered a point estimate.

To measure fairness perceptions:
Suppose that participant A transfers the 50 decision points to participant B. Participant B has to choose to transfer back either 150 points, 50 points or 0 points. In your opinion, how *fair* do you believe is each of these choices:
1. If participant B transfers back 150 points this choice is ... ?
2. If participant B transfers back 50 points this choice is ... ?
3. If participant B transfers back 0 points this choice is ... ?

The subject then filled in three seven-point scales (one for each choice) that ranged from 'very unfair' (1) to 'very fair' (7).

Appendix 4B – Descriptive Statistics

Table 4B.1 and Table 4B.2 summarize of the behavioral data for each of the two treatments. Note that the numbers presented the columns titled 'Both periods' are the mean behavior of each subject across both periods. In other words, first we take the mean behavior across periods for each subject and then we take the mean across all subjects. In the cases where a subject had only one opportunity to take an action, we take the data from that period as that subject's mean.

TABLE 4B.1 – SUMMARY OF THE BEHAVIORAL DATA IN THE PUNISHMENT TREATMENT

Mean	**Period 1**	**Period 2**	**Both periods**
Points sent (cooperation) standard deviation	43.4 (17.1)	43.4 (17.1)	43.4 (14.7)
Frequency of cooperation	86.4	86.4	86.4
Number of observations	68	68	68
Points returned standard deviation	108.5 (58.1)	96.6 (62.9)	103.4 (57.5)
Frequency of returning 150	0.644	0.559	0.614
Frequency of returning 50	0.237	0.254	0.227
Frequency of returning 0	0.119	0.186	0.159
Number of observations	59	59	66
Points spent on punishment standard deviation	17.3 (31.4)	18.7 (35.5)	18.1 (26.2)
Frequency of punishment	0.305	0.254	0.278
Number of observations	59	59	63
Points spent on retaliation standard deviation	5.5 (8.7)	5.9 (10.0)	5.2 (8.2)
Frequency of retaliation	0.375	0.444	0.4
Number of observations	16	9	20
Points spent on additional punishment standard deviation	6.2 (8.8)	24.3 (28.0)	14.2 (20.6)
Frequency of additional punishment	0.6	0.5	0.556
Number of observations	5	4	9

TABLE 4B.2 – SUMMARY OF THE BEHAVIORAL DATA IN THE BASELINE TREATMENT

Mean	Period 1	Period 2	Both periods
Points sent (cooperation) standard deviation	42.3 (18.8)	23.1 (25.9)	32.7 (15.8)
Frequency of cooperation	84.6	46.2	65.4
Number of observations	13	13	13
Points returned standard deviation	36.4 (59.5)	41.7 (58.5)	35.4 (56.9)
Frequency of returning 150	0.182	0.167	0.167
Frequency of returning 50	0.182	0.333	0.208
Frequency of returning 0	0.636	0.5	0.625
Number of observations	11	6	12

The emotional reaction of first movers in the punishment treatment is summarized in Table 4B.3 and Table 4B.4. In the baseline treatment, the emotional reaction of first movers was statistically indistinguishable from the one in the punishment treatment. It seems that the opportunity to punish does not affect how first movers feel about the amount returned to them by second movers.

TABLE 4B.3 – MEAN EMOTIONAL INTENSITY OF FIRST MOVERS AFTER OBSERVING THE AMOUNT RETURNED BY THE SECOND MOVER IN THE PUNISHMENT TREATMENT

Emotions	Got back 150	Got back 50	Got back 0
Anger	1.1 (0.5)	4.5 (1.9)	5.8 (1.5)
Irritation	1.2 (0.7)	5.0 (1.5)	6.1 (1.5)
Happiness	6.1 (1.0)	2.3 (1.4)	1.8 (1.1)
Gratitude	4.9 (1.8)	2.4 (1.7)	1.6 (1.1)
Shame	1.2 (0.5)	1.9 (1.6)	2.9 (2.3)
Guilt	1.1 (0.5)	1.3 (0.9)	1.8 (1.7)
Surprise	4.2 (1.6)	3.9 (1.7)	4.5 (2.5)
Number of observations	53	27	17

Note: Numbers between brackets are standard deviations.

TABLE 4B.4 – MEAN EMOTIONAL INTENSITY OF FIRST MOVERS AFTER OBSERVING THE AMOUNT OF RETALIATION THEY RECEIVED FROM THE SECOND MOVER

Emotions	No Retaliation	Positive Retaliation
Anger	1.9 (1.5)	3.6 (2.2)
Irritation	2.2 (1.7)	4.7 (2.2)
Happiness	3.4 (1.8)	2.6 (1.3)
Gratitude	2.4 (2.0)	2.7 (1.9)
Shame	2.1 (1.8)	1.3 (0.9)
Guilt	2.1 (1.9)	1.5 (1.1)
Surprise	4.8 (1.9)	2.3 (1.6)
Number of observations	14	10

Note: Numbers between brackets are standard deviations.

The emotional reaction of second movers is summarized in Table 4B.5.

TABLE 4B.5 – MEAN EMOTIONAL INTENSITY OF SECOND MOVERS AFTER OBSERVING THE AMOUNT OF PUNISHMENT THEY RECEIVED FROM THE FIRST MOVER

Emotions	Not Punished	Below Median Punishment	Above Median Punishment
Anger	1.1 (0.8)	3.6 (2.2)	3.9 (1.9)
Irritation	1.3 (1.2)	3.5 (2.3)	4.8 (2.3)
Happiness	5.0 (1.6)	2.4 (1.4)	1.5 (0.8)
Gratitude	4.0 (2.0)	2.5 (1.5)	2.3 (1.7)
Shame	1.2 (0.9)	1.3 (0.6)	1.7 (1.1)
Guilt	1.4 (1.1)	1.8 (1.3)	1.9 (1.3)
Surprise	2.5 (1.9)	4.0 (2.1)	5.2 (2.1)
Number of observations	55	14	13

Note: Numbers between brackets are standard deviations.

Appendix 4C – Regressions

Model estimating the intensity of anger experienced by first movers after they observed the amount of points returned by the second mover in the punishment treatment. Ordered probit estimates using robust standard errors and clustering on each subject.

TABLE 4C.1 – ORDERED PROBIT MODEL ESTIMATING FIRST MOVERS' ANGER

Variable	Coefficient	Std. Error	*p*-value
Return = 50	2.648	0.337	0.000
Return = 0	3.352	0.438	0.000
Expected Return = 50	–0.368	0.338	0.276
Expected Return = 0	–0.891	0.473	0.059
Fairness of Returning 50	–0.226	0.115	0.049
Economist	–0.043	0.302	0.888
Female	–0.322	0.290	0.267
Number of obs. = 118		LR $\chi^2(7)$ = 111.03	
Log likelihood = –96.765		Prob > χ^2 = 0.000	

Note: The variables 'Return = *x*' = 1 if the return was *x*, and 0 otherwise. The variable 'Fairness of returning 50' ranges from 1 = 'very unfair' to 7 = 'very fair'. Dummy variables: Economist: 1 if economics mayor, 0 otherwise; Female: 1 if female, 0 if male.

Note that the coefficients of 'Return = 50' and "Return = 0' are significantly different from each other (Wald test, $p = 0.047$). That is, anger is highest when the return is zero.

Model estimating the intensity of anger experienced by second movers after they observe the amount of punishment given to them by the first mover. Ordered probit estimates using robust standard errors and clustering on each subject.

TABLE 4C.2 – ORDERED PROBIT MODEL ESTIMATING SECOND MOVERS' ANGER

Variable	Coefficient	Std. Error	*p*-value
Return	–0.349	0.263	0.185
Punishment if Return = 150	0.228	0.062	0.000
Punishment if Return = 50	0.038	0.006	0.000
Punishment if Return = 0	0.024	0.005	0.000
Expected Punishment	–0.004	0.003	0.185
Fairness of Returning 50	–0.101	0.137	0.460
Economist	–0.199	0.315	0.528
Female	0.272	0.353	0.441
Number of obs. = 118		LR $\chi^2(8)$ = 132.23	
Log likelihood = –82.549		Prob > χ^2 = 0.000	

Note: The variable Return = 0 if the return was 0 points, 1 if the return was 50 points, and 2 if the return was 150 points. The variables 'Punishment if Return = *x*' = amount of punishment if the return was *x*, and 0 otherwise. The other variables are the same as in Table 4C.1.

Note that the coefficients of the variables 'Punishment if Return = *x*' are significantly different from each other (Wald tests, $p < 0.012$), indicating that for a given amount of punishment second movers get angrier the more they had returned to the first mover.

Model estimating the intensity of shame experienced by second movers after they observe the amount of punishment given to them by the first mover. Ordered probit estimates using robust standard errors and clustering on each subject.

TABLE 4C.3 – ORDERED PROBIT MODEL ESTIMATING SECOND MOVERS' SHAME

Variable	**Coefficient**	**Std. Error**	***p*-value**
Return < 150 & No Punishment	1.599	0.434	0.000
Return < 150 & Low Punishment	0.891	0.661	0.178
Return < 150 & High Punishment	1.745	0.481	0.000
Expected Punishment	0.004	0.004	0.253
Fairness of Returning 50	–0.010	0.120	0.933
Economist	–0.134	0.333	0.687
Female	–0.741	0.381	0.052
Number of obs. = 118		LR $\chi^2(7)$ = 47.79	
Log likelihood = –52.592		Prob > χ^2 = 0.000	

Note: The variables 'Return < 150 & No/Low/High Punishment' = 1 if returns are less than 150 points and punishment is 0 (No), between 0 and 50 (Low), or greater than 50 (High), and 0 otherwise. The other variables are the same as in Table 4C.1.

Note that in all regressions we take into account the effect of perceived fairness norms, by estimating the models using the variable 'Fairness of returning 50 points'. The reason being that this variable exhibited the most variance among the three variables measuring fairness perceptions. For the variable 'Fairness of returning 150 points', 85.3% of subjects agreed that is was very fair. For the variable 'Fairness of returning 0 points', 83.1% of subjects agreed that is was very unfair.

Chapter 5

Defining What is Fair

On the Enforcement of Different Cooperation Norms in Public Good Games[*]

In this chapter, we investigate the effects of endowment heterogeneity on public goods games with punishment opportunities. In particular, we study the differences between cooperation norms enforced in homogenous and in heterogeneous groups. We also look at whether the enforcement of a particular cooperation norm depends on the cooperation possibilities of individuals.

5.1 Introduction

An important objective of the social sciences is to increase our understanding of cooperative behavior in social dilemmas. This line of research has been recently revitalized by the finding that costly punishment can help sustain cooperation in public good games (Fehr and Gächter, 2000b; Masclet et al., 2003; Bochet et al., 2005; Egas and Riedl, 2005; Nikiforakis and Normann, 2005; Noussair and Tucker, 2005). However, most of these results come from experiments in which all individuals have the same endowment. This raises the question whether behavior in these experiments can be generalized to other more unequal situations. In this chapter, we answer this question by studying the effects of heterogeneous endowments on cooperation and punishment in public good games. In particular, we focus on differences in the punishment behavior of individuals depending on their endowment and the endowment of others. This allows us to determine whether different 'cooperation norms' are enforced depending on the degree of endowment heterogeneity.

By now, there is considerable experimental evidence that individuals are willing to incur costs in order to punish those who deviate from an established social norm (e.g. Güth et al., 2001; Fehr and Gächter, 2000b; Bosman and van Winden, 2002; and previous chapters). In public good games, punishment seems to be motivated by anger-like emotions triggered by unfair behavior (Fehr and Gächter, 2002). This means that, as long as notions of fairness and reciprocity are affected by income differences,[64] punishment behavior will depend on a

[*] This chapter is partly based on Reuben and Riedl (2005).

[64] See Blount (1995) and Cox (2004) for evidence that shows income differences can motivate subjects to punish and reward others, even when these differences are not the result of intentional acts.

group's income distribution. Given that changes in punishment behavior change the incentives of individuals to cooperate, endowment heterogeneity could result in different cooperation levels.

For our study, we use a linear public good game with punishment opportunities. Subjects interact for ten periods in fixed groups of three. We use one control and two *unequal* treatments. In the control or *baseline* treatment, all subjects receive the same endowment. In the unequal treatments, endowment heterogeneity is introduced by providing one subject per group (the rich subject) an endowment that is twice the endowment of other group members (the poor subjects). In the first unequal treatment, the *restricted* treatment, rich subjects receive a higher endowment but their contributions are restricted to the maximum amount poor subjects can contribute. In this way, we introduce inequality in endowments without affecting the actions that subjects can make. In the second unequal treatment, the *unrestricted* treatment, rich subjects receive a higher endowment and they can contribute any fraction of it to the public good. Hence, the unrestricted treatment has the same endowment distribution as the restricted treatment and differs only in the maximum amount that rich subjects can contribute. Analyzing the differences between the three treatments allows us to separate the effect of inequality in endowments and the effect of inequality in contribution possibilities.

Our results indicate that endowment heterogeneity does not affect the effectiveness of the punishment institution but it does affect its efficiency. Furthermore, we also find that the surplus generated by cooperation is distributed differently in the different treatments. Our findings are largely explained by differences in punishment behavior. In the baseline treatment, subjects enforce a cooperation norm in which all group members contribute the same amount. In the restricted treatment, the same cooperation norm is enforced. Hence, despite the difference in endowments, rich and poor subjects end up contributing similar amounts. In the unrestricted treatment, subjects enforce a cooperation norm in which rich subjects contribute twice as much as poor subjects. The higher contributions of rich subjects in the unrestricted treatment lead to a redistribution of earnings from rich to poor that does not occur in the restricted treatment. Lastly, we find that, irrespective of their contributions, rich subjects are heavily punished by poor subjects in the restricted treatment. Compared to other treatments, this makes groups in the restricted treatment less efficient.

The chapter is organized as follows. In Section 6.2 we describe the design of the experiment in relation to theoretical models of fairness and reciprocity. Section 6.3 analyzes the subjects' cooperation and punishment behavior, and Section 6.4 discusses the main results and concludes.

5.2 Related Literature

Our work is related to, on one hand, studies that investigate changes in punishment behavior due to asymmetries in the game, and on the other hand, to studies that look at the effects of heterogeneous endowments on cooperation in public good settings.

There is substantial evidence from the literature on ultimatum bargaining games that shows that introducing asymmetries in the game can affect the subjects' behavior (for an overview see Camerer, 2003). For example, using asymmetric payoffs Kagel et al. (1996) report that responders who have a high payoff conversion rate receive higher offers. Furthermore, responders who have a low payoff conversion rate reject more often (this is attributed to conflicting fairness norms). Similarly, Knez and Camerer (1995) find that outside options that produce different self-serving interpretations of what constitutes a fair offer, substantially increase rejection rates. Introducing asymmetries in pie size in combination with incomplete information generally leads to significantly lower offers and to lower rejection rates (e.g. Straub and Murnighan, 1995; Rapoport et al., 1996a; Rapoport et al., 1996b). Limiting the offers that proposers can make can also have a considerable effect on behavior. For example, Güth et al. (2001) replace equal-split offers with near-equal-split offers and find that proposers make low offers more often (see also Falk et al., 2000).

Also relevant to this study is the clear change in behavior between the ultimatum game and the power-to-take game of Bosman and van Winden (2002). In both games, a proposer makes an offer to split an amount of money. Thereafter, a responder can accept the offer or reject it by destroying (some of) the money at stake. An important difference between the two games is that in the power-to-take game the proposer receives additional income. Thus if proposers wish to split the total joint-income in half, they have to offer the responder 100% of the money at stake. If we compare behavior across the games, it is easy to see that any theory that wishes to predict behavior in both games will have to take into account the possibility of changing reference points. As reported in Camerer (2003), in ultimatum games, the proposers' mean offer is usually between 40% and 30% and the modal offer is usually 50%. Furthermore, there is barely any rejection if proposers offer more than 40%, while on aggregate about 50% of the income is destroyed if proposers offer less than 20%. In the power-to-take game reported in Chapter 3, the mean offer is 41.2% and the modal offer is 50%. Moreover, whereas only 4.3% of the responders' income is destroyed at offers above 40%, 50.7% of their income is destroyed at offers below 20%. On the surface, the results are strikingly similar. However, since in the power-to-take game proposers receive additional income, they walk away with about twice the earnings of proposers in the ultimatum game. In summary, whereas some asymmetries, such as carefully chosen outside options, can induce conflicting fairness perceptions that generate more punishment, other asymmetries, such as giving proposers a fixed payment, produce a general shift in fairness perceptions that has little effect on the amount of punishment but a large effect on the distribution of earnings.

A few experiments have investigated the effect of heterogeneous endowments on public good provision. Most of these studies do not incorporate punishment opportunities and hence are not directly comparable. However, it is of interest to observe if the effects of endowment heterogeneity are similar in slightly different settings. Some of this earlier work is reviewed in Ledyard (1995). Bagnoli and McKee (1991), Rapoport and Suleiman (1993), and van Dijk et al. (2002) find that inequality reduces contributions to the public good. Chan et al.

(1996), using a non-linear public goods game, report that increasing the degree of endowment inequality leads to higher levels of cooperation. In similar experiments, van Dijk and Grodzka (1992) and Chan et al. (1999) find that introducing unequal endowments does not change the amount contributed to the public good.[65] Anderson et al. (2004), introduce heterogeneity by varying show-up fees. They find that inequality tends to reduce contributions. However, since their conclusions are based on only one independent observation per treatment, their results should be interpreted with some care. Sadrieh and Verbon (2005) induce heterogeneity by allowing endowments to accumulate over time. They find that the degree of inequality does not affect cooperation. Cherry et al. (2005), using a one-shot linear public good game, report that heterogeneous endowments reduce contributions to the public good.

Most of these studies vary in considerable ways, and hence, they are not easily compared. Thus, the effects of unequal endowments on public good provision are still ambiguous. Additional work is needed to clarify under what conditions inequality affects cooperation. In this sense, the experiments presented in this chapter can contribute to this line of research. Since individuals tend to punish unfair behavior, by observing punishment patterns, we can gain insights into what subjects consider unfair in these situations.

Even though there is mixed evidence concerning the effect on overall contributions of endowment heterogeneity, a more robust finding is the difference between the contributions of rich and poor subjects. Although, in absolute terms, rich subjects usually contribute more than poor subjects do, if one looks at contributions relative to their endowment, poor subjects turn out to be the highest contributors (Chan et al., 1996; Chan et al., 1999; van Dijk et al. (2002); Cherry et al., 2005).

The study that comes closest to the one in this chapter is a field experiment conducted with subjects from South African fishing communities (Visser and Burns, 2005). In their design, Visser and Burns (2005) use a linear public goods game with or without punishment opportunities. Then they compare groups in which all subjects have the same endowment and groups in which there are two rich and two poor subjects. They find that, irrespective of punishment, unequal groups contribute more. They also confirm that poor subjects contribute relatively more than rich subjects.

5.3 Experimental Design and Theoretical Predictions

The experiment consists of a repeated public-good game with punishment opportunities. Subjects are divided into groups of three and then they play the one-shot version of the game for ten consecutive periods. Group composition remains the same during the whole experiment. Furthermore, both the number of periods and group composition are common knowledge. Each period of the game consists of two stages: a contribution stage and a punishment stage.

[65] Chan et al. (1999) also report that simultaneously introducing unequal endowments and unequal valuations increases contributions to the public good.

In the contribution stage, each subject i receives an endowment of y_i tokens. Thereafter, subjects simultaneously decide what amount c_i they wish to contribute to their group's public good. Contributions to the public good are multiplied by 1.5 and then returned equally to the three members of the group. In other words, the marginal per capita return from contributions to the public good is 0.5. Similar parameters are commonly used in experiments since they provide individuals an incentive to free ride while, from the groups' perspective, the best outcome is attained when everyone contributes as much as possible.

In the punishment stage, subjects are first informed of the endowments and individual contributions of other group members. Next, subjects decide how many punishment points to assign to other subjects in their group. Each punishment point costs the punisher one token and reduces the earnings of the punished subject by three tokens. Subjects can assign up to ten punishment points to each other subject. In order to avoid large losses during the experiment, subjects are not allowed to punish others below zero earnings.[66] After subjects make their punishment decision, they are informed of the total number of punishment points assigned to them by other subjects in their group. As in Fehr and Gächter (2000b) subjects are not informed which subject assigned them punishment points. In summary, if earnings are positive, each subject i earns in each period the following amount.[67] We denote p_{ij} as the number of punishment points i assigns to j.

$$\pi_i = y_i - c_i + 0.5\Sigma_j c_j - \Sigma_{j\neq i} p_{ij} - 3\Sigma_{j\neq i} p_{ji}$$

Subjects participate in one of the three different treatments. In the baseline treatment all subjects receive the same endowment, whereas in the two unequal treatments one subject receives a higher endowment. We now describe each treatment in detail.

- The *baseline* treatment: In this treatment all subjects receive an endowment of $y_i = 20$ tokens per period. Furthermore, they can contribute any amount of their endowment to the public good, $c_i \in [0,20]$.
- The *restricted* treatment: This is the first of the unequal treatments. In this treatment, one subject per group is randomly selected to be the rich subject. The other two subjects in the group are therefore the poor subjects. Subjects are rich or poor for the duration of the experiment. Rich subjects receive an endowment of $y_R = 40$ tokes per period. Poor subjects receive an endowment of $y_P = 20$ tokens per period. Lastly, rich subjects face the same action set as poor subjects. In other words, both rich and poor subjects can contribute at most 20 tokens per period, $c_R \in [0,20]$ and $c_P \in [0,20]$.
- The *unrestricted* treatment: This is the second of the unequal treatments. This treatment is the same as the restricted treatment except that in this case, both rich and

[66] Whereas subjects can never be punished below zero tokens, they can incur losses if they decide to punish others (see footnote 67).

[67] If earnings are negative then they equal: $\pi_i = \max[0, y_i - c_i + 0.5\Sigma_j c_j - 3\Sigma_{j\neq i} p_{ji}] - \Sigma_{j\neq i} p_{ij}$.

poor subjects can contribute any amount of their endowment to the public good. In other words, $c_R \in [0,40]$ and $c_P \in [0,20]$.

The restricted treatment differs from the baseline treatment only in the endowment of one subject per group. This allows us to observe the effect of introducing heterogeneous endowments without affecting the choices that subjects can make. The unrestricted treatment differs from the restricted treatment only in the maximum amount that rich subjects can contribute to the public good. Thus, by comparing the two, we can determine the effect of the change in the contribution possibilities of rich subjects without changing the degree of endowment inequality. Lastly, note that across the three treatments, poor subjects and subjects in the baseline treatment receive the same endowment and face the same marginal per capita return. Hence, any differences between the behaviors of these subjects must be due to the endowment or contribution possibilities of the rich subject.

Given that the game has a known end and that punishment is costly, own-profit-maximizing individuals do not have an incentive either to punish or to cooperate. Indeed, we know from previous experiments that, in the absence of punishment, contributions to the public good quickly decline (Ledyard, 1995). However, this is not the case when subjects are allowed to impose sanctions on each other. In this case, subjects use punishment to enforce high levels of cooperation (Fehr and Gächter, 2000b). It is not clear why this is the case. In theory, if individuals can credibly commit to punish certain kind of behavior, punishment can be used to enforce a positive amount of cooperation. However, this amount of cooperation is not necessarily high (Hirshleifer and Rasmusen, 1989; Boyd and Richerson, 1992). In groups with equal endowments, it is possible that high and equal contributions are a natural focal point that serves as a coordination device (Fehr and Schmidt, 1999). After all, enforcing such behavior gives all individuals the same earnings, increases efficiency, and requires all individuals to contribute the same absolute amount. Since in groups with unequal endowments this is no longer the case, subjects might have more difficulties coordinating on a specific contribution level.

Given this multiplicity of equilibria, different models can be used to account for cooperation in public good games with punishment. However, a model that explains why subjects cooperate in these games should also explain the way subjects punish. Various recent theories are able to predict the punishment patters observed in various experiments. In general, they do so by assuming that individuals care not only for their own earnings but also for the earnings and intentions of others (e.g. see footnote 5). In the following paragraphs, we describe how individuals are assumed to punish in some of these theoretical models.

Theories that assume individuals dislike income differences (e.g. Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) predict that, in the baseline treatment, individuals who contribute high amounts and care enough about inequity will punish individuals who contribute less than they do. In the restricted and unrestricted treatments, punishment depends on whether an individual is rich or poor. In particular, since rich individuals have a higher

endowment, they will punish a poor individual only in the unlikely event that after contributing, the poor individual ends up with higher earnings. In this sense, the restricted treatment provides a strong test for this prediction. In this treatment, irrespective of the amount contributed, the rich individual will never have a lower income than the poor individuals will. Hence, we should never see a poor individual being punished by a rich individual. Poor individuals that care enough about inequity will punish individuals who earn more than they do. This means that poor individuals punish each other as in the baseline treatment. It also means that, since rich individuals will typically have higher earnings, punishment of rich individuals should be quite common. Lastly, individuals will punish in the same way in the unrestricted and in the restricted treatments.

Motivations for punishment are similar in the intention-based model of Falk and Fischbacher (2005). Although in this model, individuals do not actually dislike income differences as such. They do use income differences to judge the kindness and intentions of others. Thus, given their higher earnings, rich individuals will not interpret a low contribution from a poor individual as being unkind. In other words, they consider poor individuals have a legitimate reason to free ride. Therefore, we should see little if any punishment of poor individuals by rich individuals (none in the restricted treatment). Conversely, unless rich individuals contribute considerably (20 tokens more than the poor individual), poor individuals will interpret their action as intentionally unkind, and hence, they might be willing to punish them. This implies that, on average, rich individuals act more unkindly towards poor individuals than poor individuals act towards other poor individuals. Therefore, poor individuals punish rich individuals more than poor individuals punish each other. As before, in this model individuals punish in the same way in the restricted and in the unrestricted treatments.

Motivations for punishment do change if we consider the model presented in Charness and Rabin (2002). In this model, individuals care for both the earnings of the least well off as well as for overall efficiency. In all treatments, these two preferences motivate individuals to punish low contributors.[68] Hence, in this model, rich individuals will punish poor individuals who contribute less than other poor individuals. Also, note that, compared to poor individuals, rich individuals can contribute more and therefore increase efficiency more in the unrestricted treatment. This can translate into more severe punishment for rich individuals who do not cooperate compared to poor individuals who do not cooperate. Finally, given that in the restricted treatment, rich individuals have the same contribution possibilities than poor individuals, in this model, individuals punish in the same way in the restricted treatment and in the baseline treatment.

Punishment is qualitatively the same in the intention-based fairness models of Rabin (1993), and Dufwenberg and Kirchsteiger (2005). In fact, these models assume that

[68] A low contributor (irrespective of whether she is rich or poor) is usually not the individual with the lowest earnings, and their low contribution fails to promote efficiency.

individuals do not choose inefficient allocations, which rules out the possibility of punishment. However, as in Bolton and Ockenfels (2005), we can still gain insights if we concentrate on their definition of unfair behavior and simply assume that individuals punish unfairness. In this case, since these models assume the kindness of an action is defined only over feasible outcomes, the 20 extra tokens received by rich individuals in the restricted treatment do not affect fairness evaluations and thus do not affect punishment behavior. Consequently, motivations for punishment are the same in the baseline and the restricted treatment. In the unrestricted treatment, the fact that the rich individuals can contribute more to the public good implies that, for a given low contribution, rich individuals are seen as more unkind than poor individuals, and therefore, they are more likely to be punished. In other words, ceteris paribus, poor individuals will punish rich individuals more heavily than they punish other poor individuals. In the next section, we present and analyze the results of the experiment.

5.4 Results

In total 57 subjects participated in the experiment, 18 participated in the baseline treatment, 21 in the restricted treatment, and 18 in the unrestricted treatment. About 40% of the subjects were women. The experimental procedures and the instructions are found in Appendix 5A. Descriptive statistics of the data are available in Appendix 5B.

5.4.1 Contributions to the public good

In all treatments, we observe the familiar contribution pattern that has been reported in similar studies (e.g. Fehr and Gächter, 2002; Egas and Riedl, 2005). This is illustrated in Figure 5.1. As we can see, subjects start contributing just over half of their maximum contribution. Thereafter, contributions either remain constant or increase slightly over time. On average, contributions are highest in the unrestricted treatment (18.43), then in the baseline treatment (15.73), and lowest in the restricted treatment (14.23). However, the higher contribution level in the unrestricted treatment is due to the ability of the rich subjects to contribute a higher amount. If we control for this by looking at *relative* contributions (i.e. relative to the maximum contribution), we find that the baseline treatment has the highest relative contribution (0.79), followed by the restricted treatment (0.71), and then by the unrestricted treatment (0.70). However, we do not find these differences to be statistically significant. This is stated in our first result.

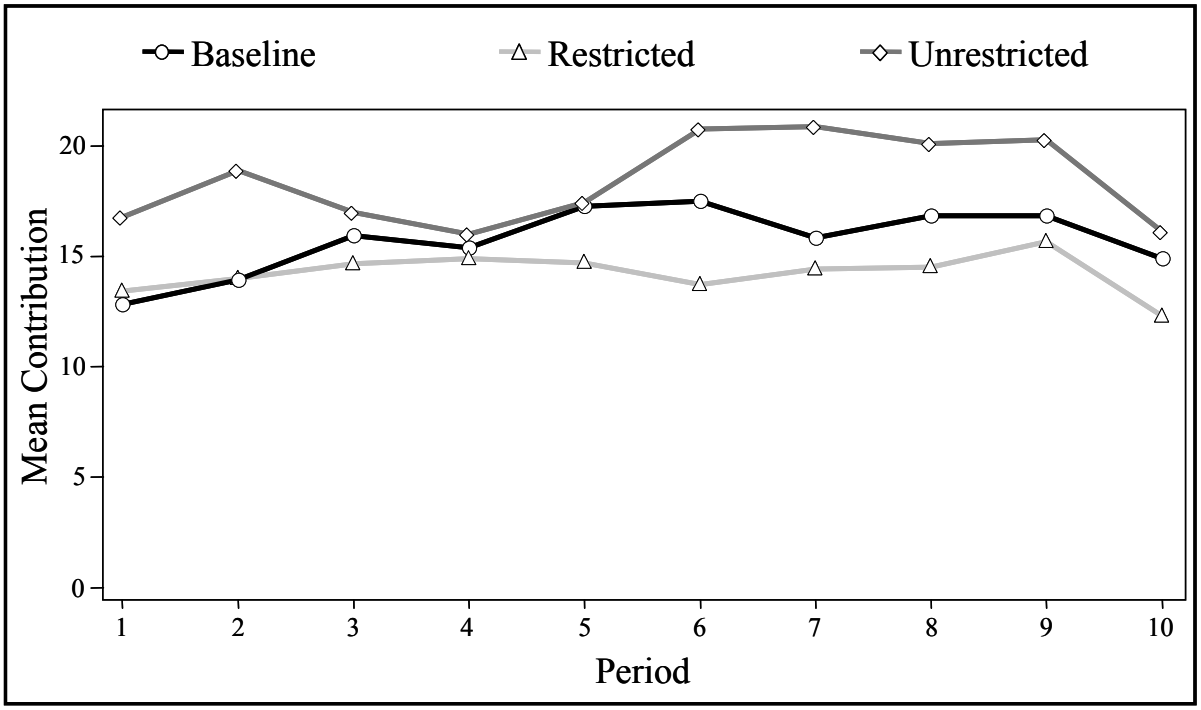

FIGURE 5.1 – MEAN CONTRIBUTIONS

Note: Mean amount contributed to the public good in each treatment over the ten periods.

RESULT 5.1: *Contribution levels are the same in treatments with equally and unequally distributed endowments*.

Support: We cannot reject the null hypothesis that the average contribution level over all ten periods is the same in the baseline treatment compared to the restricted treatment ($p = 0.520$), or compared to the unrestricted treatment ($p = 0.\ 337$).[69] The same is true if we use relative contribution levels ($p > 0.520$). In addition, we find no significant differences between the baseline and the unequal treatments when we compare, separately, each period's contribution level ($p > 0.107$). We should note, however, that contributions in the restricted treatment are significantly lower than contributions in the unrestricted treatment in periods two, six, and seven ($p < 0.099$).

At the group level, the only difference between the baseline treatment and the unequal treatments is the trend of contributions over time. Using Cuzick's trend test (Cuzick, 1985) to see whether contributions increase over periods gives a significant result for the baseline treatment ($p = 0.057$, $p > 0.471$ for the other treatments). Hence, over a longer period of time, contributions in the baseline treatment could diverge from contributions in the unequal treatments.

[69] Throughout the chapter, unless it is otherwise noted, we always use a two-sided Wilcoxon-Mann-Whitney test, and group averages as independent observations. Given the low number of observations, we refer to a difference as being statistically significant if the *p*-value of the test is below 0.100.

Result 5.1 concentrates only on average contributions. We are also interested in the effects of inequality on individual behavior, specifically, in the differences between the contributions of rich and poor subjects. This leads is to the next result.

RESULT 5.2: *Rich individuals contribute more than poor individuals in the unrestricted treatment but not in the restricted treatment*.

Support: This result can be clearly observed in Figure 5.2. In the unrestricted treatment, rich subjects contribute significantly more than poor subjects ($p = 0.037$). They also contribute significantly more than subjects in the baseline treatment ($p = 0.054$). In fact, rich subjects contribute about twice as much as poor subjects. Hence, as a proportion of their endowment, rich and poor contribute at very similar rates (0.67 by the rich vs. 0.72 by the poor, $p = 0.936$). In contrast, the opposite is true in the restricted treatment. In this treatment, rich subjects contribute the same amount as poor subjects ($p = 0.565$), and the same amount as subjects in the baseline treatment ($p = 0.253$). Thus, as a proportion of their total endowment, rich subjects contribute significantly less than poor subjects do (0.33 vs. 0.74, $p = 0.003$). Hence, it appears that the possibility to contribute 40 tokens instead of 20 tokens has a big effect on the behavior of the rich subjects. Contributions by rich subjects are significantly higher in the unrestricted treatment ($p = 0.032$). In contrast, there are no significant differences between the contributions of poor subjects in the unequal treatments, or between poor subjects and subjects in the baseline treatment ($p > 0.721$).

5.4.2 Punishment behavior

On average, the amount of tokens spent on punishment is around the same in all three treatments. In the baseline treatment, each subject allocated on average 1.08 punishment points on every period. In the unequal treatments, there is slightly more punishment. In the restricted treatment, the average amount of punishment is 1.57 whereas in the unrestricted treatment it is 1.20 (see Appendix 5B). However, these differences are not statistically significant ($p > 0.391$). If we look at how punishment evolves, we find that the amount of punishment decreases significantly over time in both the baseline and the unrestricted treatments (Cuzick trend tests, $p < 0.035$). In the restricted treatment punishment remains constant (Cuzick trend test, $p = 0.925$).

In the baseline treatment, subjects commonly punish free riders and to a lesser extent subjects who contribute more than they do. They rarely punish subjects who contribute the same amount. For example, if we use as the reference point the contribution of the punishing subject, we find the following significant differences: first, subjects punish more and more often those who contribute less than those who contribute more ($p < 0.054$); second, they punish more and more often those who contribute less than those who contribute the same amount ($p < 0.003$); and third, they punish more and more often those who contribute more than those who contribute the same amount ($p < 0.074$). This U-shaped punishment pattern is commonly observed in laboratory as well as field experiments (Cinyabuguma et al., 2004;

Egas and Riedl, 2005). Punishment of low contributors is commonly interpreted as punishment of unfair (or unkind) behavior. However, punishment of high contributors could be due to spiteful behavior (Falk et al., 2005) or to retaliation for anticipated punishment (Gächter and Herrmann, 2005).

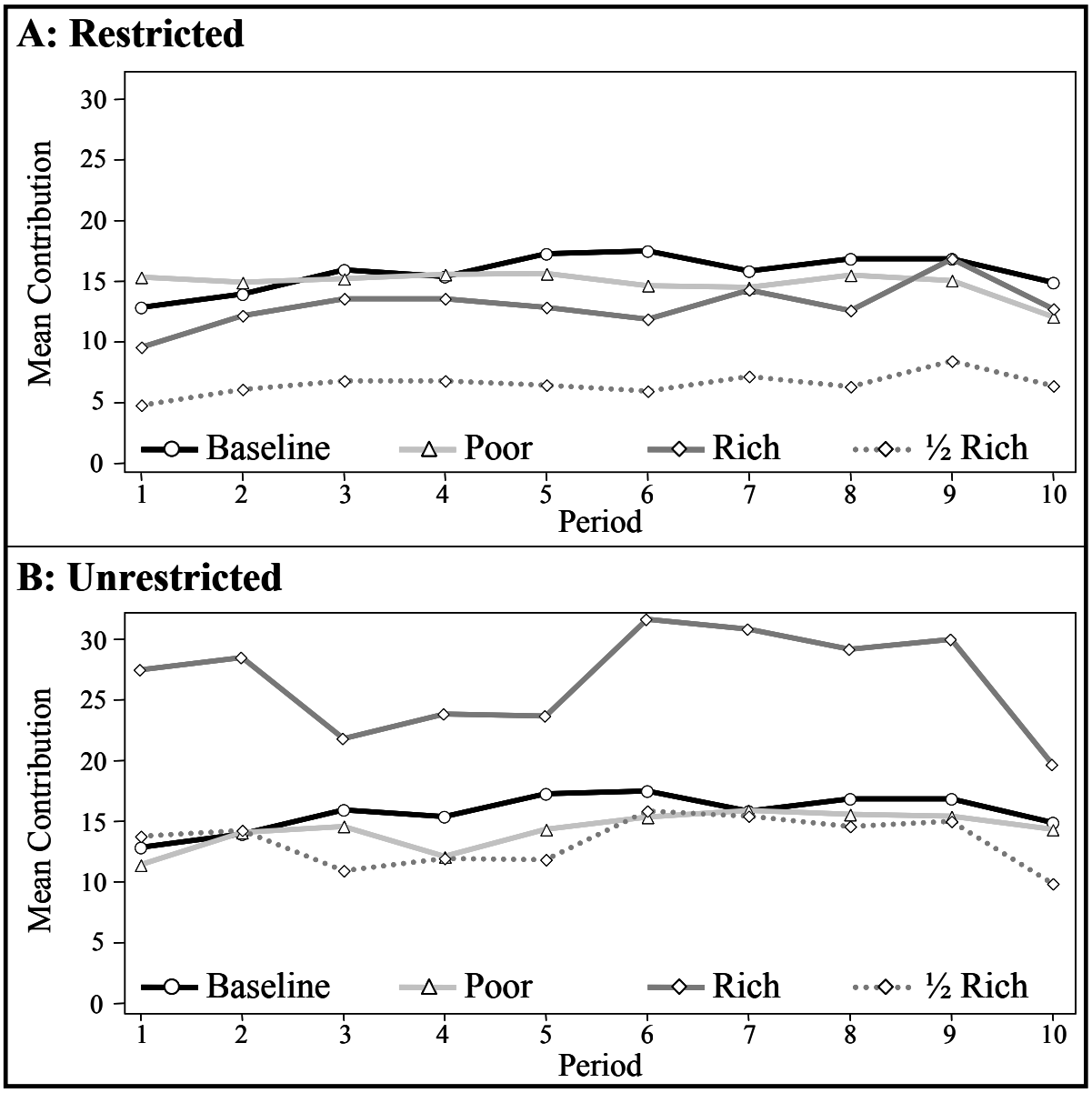

FIGURE 5.2 – CONTRIBUTIONS BY RICH AND POOR SUBJECTS

Note: A) Mean contributions by rich and poor subjects in the restricted treatment, and by subjects in the baseline treatment. B) Mean contributions by rich and poor subjects in the unrestricted treatment, and by subjects in the baseline treatment. The dotted lines, which are half the mean contributions of rich subjects, allow us to compare the contributions of rich and poor relative to their endowment.

Whereas in the baseline treatment it is intuitive to look at the way subjects punish deviations from their own contribution, in the unequal treatments, it is unclear how to compare the

contributions of rich and poor subjects. However, if we assume that the punishment pattern observed in the baseline treatment also exists in the other treatments, we can use the punishment data to find what 'cooperation norm' is being enforced in each of the unequal treatments. More specifically, we assume that subjects will not punish those who contribute what they think is a 'fair' amount relative to their own contribution. Low contributors are punished for acting unfairly, and high contributors are punished as retaliation for expected punishment. Thus, if the cooperation norm were to contribute as much as others do, then we would observe the punishment pattern of the baseline treatment. If, on the other hand, the cooperation norm were for rich subjects to contribute twice the amount of poor subjects, then a poor subject who contributes 15 tokens will not punish a rich subject who contributes 30 tokens but will punish a rich subject who contributes 40 or 20 tokens. In order to derive the cooperation norm applied in each of the unequal treatments we analyzed separately each of the following cases: poor subjects punishing rich subjects, rich subjects punishing poor subjects, and poor subjects punishing other poor subjects.[70] In each case, we use the following model.

$$NO\ PUN = \beta_0 + \beta_1 FAIR + \beta_2 PERIOD + \beta_3 PUN^{-1} + \varepsilon$$

Where *NO PUN* is a dummy variable that equals one when subject *i* does not punish subject *j*, and equals zero otherwise. Given that our experiment used a partners matching protocol, there might be incentives for reputation formation and for retaliation of past punishment. To control for these two effects we introduce a variable with the period number (*PERIOD*), and a variable indicating whether subject *i* was punished in the previous period (PUN^{-1}). Lastly, *FAIR* is a dummy variable that equals one if subject *j* contributed a fair amount from subject *i*'s perspective. More specifically,

$$FAIR = 1 \text{ if } | \phi c_i - c_j | < 2 \text{ and 0 otherwise}$$

Where ϕ is the reference point or cooperation norm that defines the way subjects compare their own contribution to the contribution of others. For example, if $\phi = 0.5$ this implies subjects think it is fair for the other person to contribute half the amount they contributed. Thus, if subject *i* contributed 10 tokens, she would consider it fair for the other to contribute between 3 and 7 tokens. In order to find out the reference point used by subjects in each of the abovementioned cases, we estimated the model using different values of ϕ and restricting β_1 to a nonnegative number (this ensures a punishment pattern as the one present in the baseline treatment).[71] We then selected the value of ϕ that gave the best fit for the data.[72] The resulting

[70] As a control, we use the same model to analyze the way subjects punish each other in the baseline treatment.

[71] Specifically, we varied ϕ in the [0,3] interval using steps of 0.01.

[72] We used probit estimates with robust standard errors and clustering on each group. We used the pseudo R-squared to select the best fit.

ϕs and the corresponding model are presented in Table 5.1. The main finding is summarized in the following result.

RESULT 5.3: *In the baseline and restricted treatments, individuals enforce a cooperation norm in which everyone contributes the same amount. In the unrestricted treatment, individuals enforce a cooperation norm in which rich individuals contribute twice as much as poor individuals*.

Support: By looking at Table 5.1, we see that in the unrestricted treatment, the models that best fit the observed punishment behavior, are models where rich subjects punish poor subjects who deviate from half their own contribution, and where poor subjects punish rich subjects who deviate from twice their own contribution. In the restricted treatment, the models that best fit the punishment data are those that indicate poor and rich subjects punish deviations from their own contribution irrespective of the endowment of the other subject. This result is robust to changes in the structure of the models used. In particular, we get similar values for ϕ if we differentiate between positive and negative deviations,[73] if we use deviations from the group's average contribution instead of the subject's own contribution, if we drop the variables that control for reputation effects and retaliation of past punishment, or if we use slightly different interval for the definition of a fair contribution (variable *FAIR*).

The enforcement of different cooperation norms explains why we see such a big difference between the contributions of rich subjects in the restricted and in the unrestricted treatments. It also helps explain why rich subjects punish the poor even though their earnings are generally greater (86.9% of punishment by rich subjects occurs when rich subjects have higher earnings). On average, rich subjects spend 0.62 (0.76) tokens per person in each round punishing poor subjects in the restricted (unrestricted) treatment. Poor subjects spend 0.56 (0.33) tokens per person in each round punishing the other poor subject (the differences between rich and poor subjects are not significant, $p > 0.462$). Hence, it appears that rich subjects punish poor subjects in a similar way as poor subjects punish each other. This is in line with the motivations for punishment presented in Rabin (1993), Charness and Rabin (2002) and Dufwenberg and Kirchsteiger (2005), and it does not support the idea of Fehr and Schmidt (1999), Bolton and Ockenfels (2000) and Falk and Fischbacher (2005) that rich subjects punish less simply because they have a higher endowment.

[73] More specifically, we find similar values for ϕ and we also see that negative deviations are punished more often than positive deviations (although this difference is not significant in the case of rich punishing poor in the unrestricted treatment).

TABLE 5.1 – ESTIMATION OF THE COOPERATION NORM ENFORCED IN EACH TREATMENT

	ϕ	*FAIR*	*PERIOD*	PUN^{-1}	**Predicted** ***NO PUN***
Baseline treatment	1.08	0.395** (0.047)	0.015** (0.007)	–0.094** (0.032)	0.860
Restricted treatment poor to poor	1.00	0.280** (0.083)	–0.002 (0.005)	–0.108** (0.047)	0.920
Restricted treatment poor to rich	1.09	0.440** (0.047)	0.011 (0.014)	–0.160 (0.134)	0.757
Restricted treatment rich to poor	1.08	0.242** (0.072)	0.008 (0.014)	–0.341** (0.156)	0.826
Unrestricted treatment poor to poor	1.05	0.238** (0.045)	–0.001 (0.007)	–0.119** (0.083)	0.945
Unrestricted treatment poor to rich	2.05	0.248** (0.071)	0.010 (0.009)	–0.003 (0.058)	0.858
Unrestricted treatment rich to poor	0.54	0.339** (0.094)	0.025 (0.017)	–0.019 (0.065)	0.870

Note: Probit estimates for the value of ϕ that gave the best fit to the subjects' punishment behavior. For each variable, we report the predicted change in the probability of not punishing. Furthermore, the rightmost column displays the predicted probability of not punishing at the average values of the dependent variables. Numbers between brackets are robust standard errors. ** Significant at the 1 percent level. * Significant at the 5 percent level.

Whereas rich and poor subjects spend similar amounts punishing other poor subjects, poor subjects themselves spend around twice as much punishing rich subjects than punishing the other poor subjects. In the restricted (unrestricted) treatment, poor subjects spend 2.08 (2.13) times more punishing rich subjects than punishing poor subjects. This combined with the fact that there are two poor subjects in every group means that rich subjects receive an above average share of the total amount of punishment (49.7% in the restricted treatment and 39.4% in the unrestricted treatment). Thus, subjects seem to enforce a cooperation norm and punish deviations from it, but poor subjects do differentiate between other poor and rich subjects, and they punish the latter more harshly.[74] In the restricted treatment, this is compatible with punishment motivations in Fehr and Schmidt (1999), Bolton and Ockenfels (2000) and Falk

[74] This is clearly seen in Table 5.1, where for both unequal treatments, the predicted probability of not punishing (at average values) is higher in the regression of poor punishing poor than in the regression of poor punishing rich. The difference is statistically significant in the restricted treatment but not in the unrestricted treatment (p = 0.037 and p = 0.211). In the restricted treatment it appears that the poor also punish the rich more severely for deviating from the cooperation norm.

and Fischbacher (2005), and it fails to support the idea that when poor subjects punish, they will completely ignore the additional endowment of rich subjects (Rabin, 1993; Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2005).

5.4.3 Efficiency and Inequality

Although punishment has been shown to be an effective institution to promote cooperation, the effect of punishment on group earnings is sometimes detrimental (especially in cases where groups are randomly re-matched every round; e.g. Egas and Riedl, 2005). The amount of income destroyed through punishment is sometimes greater than the benefits derived from cooperation. In this subsection, we analyze the effect of heterogeneous endowments on, first, the profitability of the punishment institution, and second, on the distribution of earnings within the group. As was previously shown, compared to the baseline treatment, cooperation levels in the unequal treatments are somewhat similar. There is slightly more cooperation in the unrestricted treatment and slightly less cooperation in the restricted treatment. Furthermore, punishment also differed slightly. In both the restricted and the unrestricted treatments, subjects punish more than in the baseline treatment. These differences have a direct effect on the final earnings of groups. This is stated in the following result.

RESULT 5.4: *Compared to their endowments, individuals in the baseline and unrestricted treatments increase their earnings. Furthermore, the increase is larger over time. In the restricted treatment, individuals do not increase their earnings, and their situation does not improve over time.*

Support: In order to compare earnings across treatments, we normalize group earnings so that they are equal to zero if there is no cooperation and no punishment, and they are equal to one if there is full cooperation and no punishment. The average normalized earnings are highest in the baseline treatment (0.37), slightly lower in the unrestricted treatment (0.33), and considerably lower in the restricted treatment (0.10).[75] If we compare the subjects' earnings after they interact in the game with their endowments, we find that subjects are significantly better off in the baseline and in the unrestricted treatments but not in the restricted treatment (p = 0.023, p = 0.087, and p = 0.368).[76] Moreover, whereas in the baseline and the unrestricted treatments we observe earnings increasing over time, in the restricted treatment they do not improve (this can be seen in Figure 5.3). Earnings increase significantly over the first nine periods in both the baseline and the unrestricted treatments but not in the restricted treatment (Cuzick trend tests, p = 0.006, p = 0.074, and p = 0.502).[77] Hence, it is unclear

[75] The difference between the baseline treatment and the restricted treatment is significant if we consider each period independently (p = 0.035). Other differences are not statistically significant.

[76] One-sided Wilcoxon matched-pairs sign-rank tests, where the alternative hypothesis is that normalized earnings are more than zero.

[77] If we run the tests including the last period, the trend is still significant in the baseline treatment but not significant in the unrestricted treatment (p = 0.013 and p = 0.276).

whether subjects in the restricted treatment will benefit from playing the game, even if it were to be played for more periods.

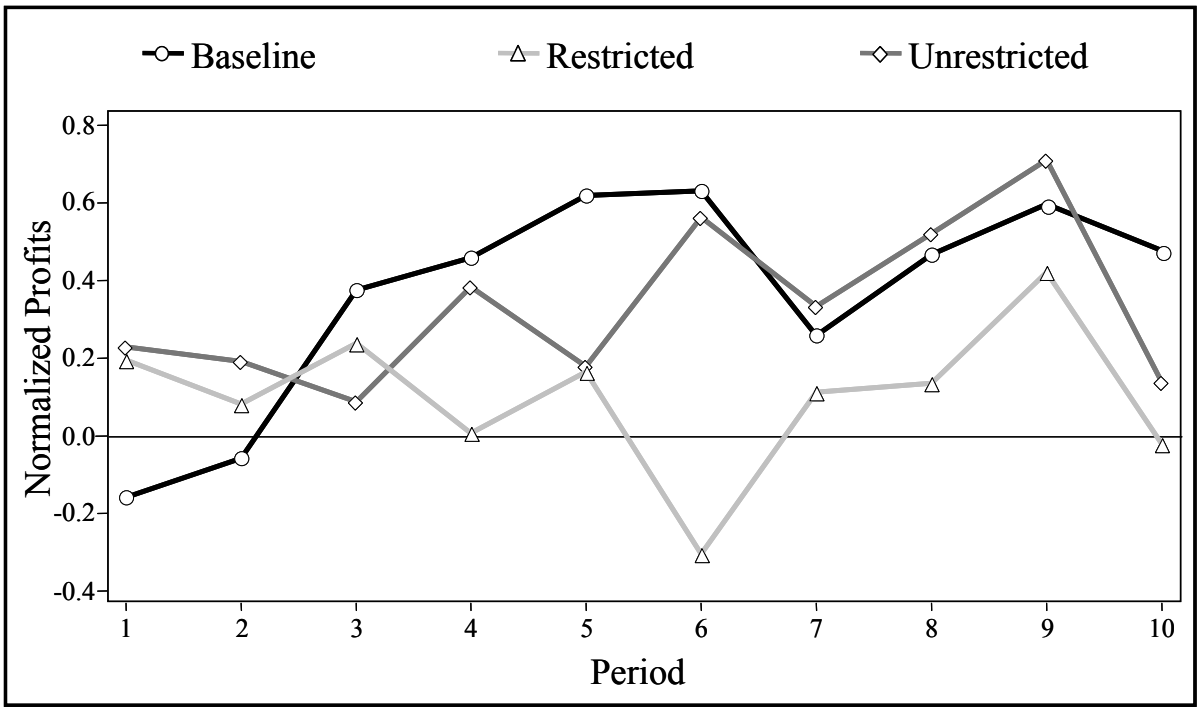

FIGURE 5.3 – GROUP EARNINGS

Note: Mean earnings normalized so that, when subjects do not contribute and do not punish, earnings are equal to zero, and when they contribute everything and do not punish, earnings are equal to one.

Although groups in both the baseline and the unrestricted treatments benefit from playing the game, the number of subjects who benefit within a group differ across the two treatments. To see this, consider the number of 'unsuccessful' subjects per group. That is, subjects that, at the end of the ten periods, have lower average earnings than their endowment. In the baseline treatment, the frequency of unsuccessful subjects per group is 0.11 whereas in the unrestricted treatment it is 0.33 (the difference is significant, $p = 0.033$). In fact, this difference is due to the lack of success of rich subjects. In all groups in the unrestricted treatment, rich subjects earned on average less than their endowment. This means that, all the group's net gains are made exclusively by poor subjects. In contrast, in the restricted treatment, the frequency of unsuccessful subjects per group is around the same for both rich and poor subjects (0.57 vs. 0.50). These findings point to our final result, namely, the difference between treatments concerning the distribution of earnings within groups.

RESULT 5.5: *Compared to the distribution of endowments, earnings are similarly distributed in the baseline and restricted treatments, and more equally distributed in the unrestricted treatment*.

Support: In order to compare the degree of inequality across treatments, we look at the share of each group's total earnings obtained by the rich subject (in the baseline treatment the 'rich'

subject is the subject with the highest earnings in that group).[78] In the baseline treatment, the highest-earning subject obtained, on average, 34.8% of the group's total earnings. This share is remarkably close to the 33.3% share that ensures full equality and that is obtained if nobody cooperates.[79] In the restricted treatment, we observe a similar result. Namely, the share of earnings obtained by rich subjects (48.5%) is similar to the share they would obtain if nobody cooperates (i.e. 50.0%).[80] In contrast, in the unrestricted treatment rich subjects end up with only 38.5% of their group's total earnings. See Figure 5.4.

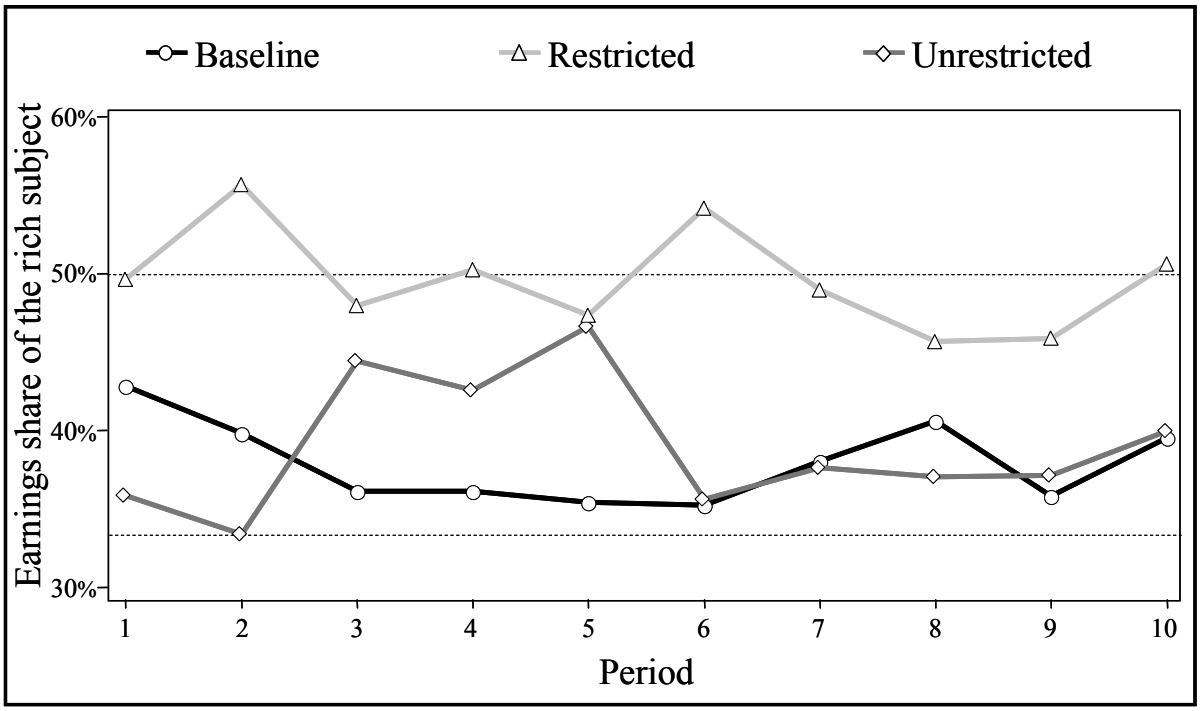

FIGURE 5.4 – EARNINGS INEQUALITY

Note: Mean share of group earnings received by rich subjects per treatment. In baseline treatment, the 'rich' subject is the subject with the highest average earnings per group. The dotted lines correspond to a share of 50.0% and a share of 33.3%.

In summary, in the baseline treatment subjects are able to increase their average earnings and at the same time keep inequality at low levels. In the restricted treatment, subjects are unable to increase their average earnings and inequality remains at high levels. In the unrestricted treatment, subjects increase their average earnings and reduce the level of inequality.

[78] The same results are obtained with other measures of inequality such as the Gini coefficient or the variance of earnings within groups.

[79] Although the difference between full equality and the 34.8% share obtained by the highest earner is very small it is statistically significantly (Wilcoxon matched-pairs sign-rank test, $p = 0.028$).

[80] The difference is not significant (Wilcoxon matched-pairs sign-rank test, $p = 0.176$).

5.5 Conclusion

In this chapter, we have studied the effects of heterogeneous endowments on the provision of public goods when there are punishment opportunities. We find that rich subjects contribute and punish differently than poor subjects do. The difference can be interpreted as the enforcement of different cooperation norms. Furthermore, we show that the behavior of both rich and poor is affected to a large extent by the amount that rich subjects are allowed to contribute instead of the amount they receive as their endowment.

Since models that assume self-regarding preferences fail to predict punishment behavior, various models have emerged that assume individuals take into consideration the income and intentions of others. A highly successful branch of this literature assumes people dislike unequal income distributions (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000). However, although in numerous experiments these models predict behavior remarkably well, they fail to predict the punishment behavior of rich subjects in the unequal treatments. As was pointed out in Section 5.3, these models predict that, given their advantageous position, in most cases rich subjects will not punish poor subjects. However, punishment of poor subjects by rich subjects is commonly observed.

In order to explain the different contribution levels and punishment patterns presented in this chapter, these models must consider the possibility that subjects do not evaluate income differences in the same way across the different treatments. To see this, note that even if a model allows rich individuals to punish poor individuals (e.g. Levine, 1998), it will incorrectly predict the same punishment pattern in the unrestricted and the restricted treatments. In order to provide a more satisfactory explanation of the data, it is necessary to model a reference point that shifts as the contribution possibilities of the rich individuals goes from 40 to 20 tokens. Suggestive in this sense is the relationship between fairness and expectations that is highlighted in Chapter 2. If the expected contributions of rich subjects change with the change in their contribution possibilities, this will affect the subjects' emotional reaction and hence their punishment behavior.

Fairness evaluations shifting with changes in the subjects' endowments are also seen in other experiments. For example, as was previously mentioned, another clear case of a shift in fairness evaluations is observed if we compare the behavior of proposers in the ultimatum game and the power-to-take game (see Chapter 3). Given the behavior of responders, proposers who choose a fifty-fifty split in the ultimatum game forgo money in order to implement an equal distribution of earnings (Lin and Sunder, 2002). In other words, these proposers seem to reveal a preference for the equal split over more unequal earnings distributions. If this is the case, these proposers will also choose the equal split in the power-to-take game. However, even though in ultimatum games about half of the subjects choose to offer a fifty-fifty split, in the power-to-take game reported in Chapter 3, only one of the sixty-eight proposers offered to split total earnings equally. Again, this difference between the two games cannot be explained by a theory that assumes subjects focus solely on an experiment's total income distribution.

As suggested by Bolton and Ockenfels (2005), in these and similar cases, distribution-based fairness models work rather well if one simply adjusts them to take into account only the 'relevant' income distribution. However, if we are to correctly predict behavior across a wide range of situations, we must incorporate into these theories an explicit way in which individuals adjust their reference incomes. In this respect, intention-based fairness models (e.g. Charness and Rabin, 2002 and Dufwenberg and Kirchsteiger, 2005) perform better than the distribution-based models. These models assume that individuals use a fairness benchmark that depends solely on feasible outcomes. Therefore, in the restricted treatment, since rich subjects cannot contribute more than 20 tokens, their extra endowment is not taken into account when evaluating fairness. This helps explain why a different cooperation norm is enforced in each of the unequal treatments.

Nevertheless, we should point out that, although treatments differ with respect to the cooperation norm that is enforced, there are other important differences between the treatments. In line with intention-based models, the punishment data suggests that the same cooperation norm is enforced in both the baseline and the restricted treatments. However, in contrast to these models' predictions, there are some differences in punishment behavior between the two treatments. In particular, the high amount of punishment in the restricted treatment (a lot of it directed to rich subjects) and the fact that it does not decline over time. In the baseline treatment, cooperation increases and punishment decreases, whereas in the restricted treatment both cooperation and punishment remain at the same level. This translates into lower overall earnings. Furthermore, it suggests that, although poor subjects enforce a one-to-one cooperation norm, they might not be satisfied with the situation and therefore constantly punish the rich subjects independently of how much they contribute (as predicted by the distribution-based models).[81]

Lastly, we point out that although groups seem to benefit from interacting in the baseline and the unrestricted treatments, not everyone benefits from cooperation in the unrestricted treatment. Specifically, poor subjects benefit at the expense of rich subjects. The fact that a group of subjects does not benefit from playing the game can have serious consequences if participation is voluntary. In this case, rich subjects could avoid the game altogether and the group would loose the efficiency gains attained through their contributions. In contrast, in the baseline treatment playing the game is ex ante a Pareto improvement over not playing it. Therefore, all individuals have an incentive to participate.[82]

In this study, we have shown that different reference points can lead to the enforcement, through punishment, of different cooperation norms. In theory, punishment can be used to enforce any kind of behavior (Boyd and Richerson, 1992). Remarkably, in the

[81] In addition, poor subjects seem to hold the rich subjects more strictly to the norm and punish them more severely if they deviate from it.

[82] Interestingly, this advantage of homogenous groups suggests that rich individuals would prefer to interact only with other rich individuals, and if groups were to form endogenously, segregation along income levels would occur.

baseline treatment (and other experiments based on Fehr and Gächter, 2000b), subjects seem to concentrate on enforcing a particular cooperation norm. It is still debated why this is the case (see Henrich, 2004, and Boyd and Richerson, 2005). However, as is evident from the unequal treatments and the results to be presented in Chapter 6, different behavior can be enforced depending on the composition of groups in the game.

Appendix 5A – Experimental Procedures and Instructions

Experimental procedures

The computerized experiment was run in October 2004 at the CREED laboratory of the University of Amsterdam. Subjects were recruited from the student population in the university through emails and through CREED's website. Subjects that had taken part in other public good experiments were not allowed to participate. The experiment was conducted with z-Tree (Fischbacher, 1999). On average, subjects were paid out 14.27 euros. The whole experiment took around one hour.

After arrival in the lab's reception room, each subject drew a card to be randomly assigned to a seat in the laboratory. Once everyone was seated, the instructions for the experiment were read aloud (a translation of the instructions is provided below). In both the restricted and the unrestricted treatments, subjects were informed in the instructions whether they would receive a high endowment or a low endowment. Thereafter, subjects had to answer a few exercises in order to check their understanding of the game. Next, the subjects played the repeated pubic goods game with punishment via the computer. Once the game ended subjects answered a debriefing questionnaire after which they were paid in private and dismissed.

Instructions

These are the instructions given to rich subjects in the unrestricted treatment. The instructions given to poor subjects and to subjects in other treatments are available upon request.

Introduction

This experiment is divided into different periods. There will be 10 periods in total. During all 10 periods, the participants are divided into groups of three. Therefore, you will be in a group with 2 other participants. The composition of the groups will remain the same during all of the experiment.

Each period consists of two stages. In the first stage, you have to decide how many tokens you contribute to a group project. In the second stage, you will learn how much the other members of your group contributed to the project.

The first stage

At the beginning of *each* period, each participant in your group receives a number of tokens. In each group: one of the participants will receive 40 tokens per period and the other two participants will receive 20 tokens per period. Before the experiment started, each desk was assigned to receive either 20 or 40 tokens. Therefore, by randomly assigning the yellow cards (given in the reception room), each one of you was randomly assigned one of these amounts. Across rounds, the amount of tokens that each participant receives will be the same. Hence, either you receive 20 tokens at the beginning of each of the 10 rounds or you receive 40 tokens at the beginning of each of the 10 rounds. *You will be the participant who receives 40 tokens per period*. We will refer to these tokens as the initial endowment.

In the first stage, you decide how to use your initial endowment. You have to choose how many tokens you want to contribute to a group project and how many of them to keep for yourself. You can contribute any amount of your initial endowment to the group project. How many tokens you contribute is up to you. Each other group member will also make such a decision. All decisions are made simultaneously. That is, nobody will be informed about the decision of the other group members before everyone made his or her decision.

Earnings in the first stage

Your earnings in tokens, in each period, are the sum of two parts:

- The number of tokens that you kept for yourself.
- Your income from the group project. This income equals:

$$0.5 \times \text{sum of contributions of all group members to the project}$$

Notice that, for each token that you keep for yourself you earn 1 token. If instead you contribute this token to the group project, then the total contribution to the project will rise by one token. Your income from the group project will rise by 0.5 tokens. Moreover, the other group members' income from the project will also rise by 0.5 tokens. Your contribution to the group project therefore also raises the income of the other group members. For each token contributed to the project, the total earnings of the group will rise by 1.5 tokens. Note that, you also earn tokens for each token contributed to the group project by the *other* group members. For each token contributed by any member, you earn 0.5 tokens. We can call 0.5 the group project's multiplication factor.

In summary, your earnings in tokens at the first stage of a period is equal to:
Your initial endowment – your contribution + 0.5 × (sum of contributions)

After everyone has made his or her decision the first stage ends.

Example for the first stage

Here is an example that illustrates how the earnings in tokens are calculated in the first stage of each period. The numbers used in the example are arbitrarily chosen.

You are in a group with two other people (group member 1 and group member 2). The initial endowments are equal to: you = 40 tokens, group member 1 = 20 tokens, and group member 2 = 20 tokens. Suppose that, you contribute 15 tokens to the group project, group member 1 contributes 5 tokens to the group project, and group member 2 contributes 10 tokens to the group project. The earnings in tokens of each of the participants are then given by:

$$\text{Initial endowment} - \text{tokens contributed} + 0.5 \times \text{sum of all contributions}$$

In your case this equals: $40 - 15 + 0.5 \times (15 + 5 + 10) = 40$ tokens.
For group member 1 this equals: $20 - 5 + 0.5 \times (15 + 5 + 10) = 30$ tokens.
For group member 2 this equals: $20 - 10 + 0.5 \times (15 + 5 + 10) = 25$ tokens.

The second stage

At the beginning of the second stage, everyone in the group will see how much each of the other group members contributed to the project as well as their earnings from the first stage. The decision each group member has to make in the second stage is to either reduce or leave equal the earnings of each other group member. Reducing other group members' earnings can be done by spending tokens. The other group members can also reduce your earnings if they wish to. All decisions are made simultaneously. That is, nobody will be informed about the decision of the other group members before everyone made his or her decision.

More concisely, in this stage, you must decide whether and if yes how many tokens you want to spend to reduce the earnings of the other two group members. If you want to reduce another member's earnings, you do that by allocating deduction points. For each deduction point that you allocate to another group member his or her earnings are reduced by *3 tokens* and your own earnings are reduced by *1 token*. If you do not wish to change the earnings of another group member then you must allocate 0 deduction points to him or her. Note, that you will not be allowed to reduce the earnings of a group member to less than zero.

Remember that, for every deduction point you receive from other group members, your earnings will be reduced by *3 tokens* (but never below zero). Every participant can spend up to a maximum of *10 tokens* (i.e. allocate 10 deduction points) on each group member in each period.

After everyone has made a decision, you will be informed how many deduction points you received from the other group members and also what your total earnings in tokens for that period are. Note that you do not get to know how individual group members spend their deduction points. In other words, you will only be informed of the total amount of deduction points allocated to you by the other two group members. You will not know how many deduction points each individual group member allocated to you.

Examples for the second stage

Here are some arbitrarily chosen examples that illustrate how your final earnings are calculated. You, group member 1 and group member 2 are all members of the same group.

Example 1:

Suppose that after the first stage you have earnings that are equal to 30 tokens. In the second stage you decide to allocate 3 deductions points to group member 1 (this reduces group member 1's earnings by 9 tokens) and 0 deduction points to group member 2 (this does not change group member 2's earnings). After all have made their decision, you learn that the others allocated you a total of 4 deduction points. In this case, your total earnings in tokens in this period are given by:

(Your first stage earnings – 3 × deduction points allocated to you)*

– deduction points you allocated

* If the number between brackets is negative then replace it with zero.

In this example, your earnings are equal to: $(30 – 3 \times 4) – 3 = 18 – 3 = 15$ tokens.

Example 2:

Suppose that after the first stage you have earnings that are equal to 18 tokens. In the second stage you decide to allocate 4 deductions points to group member 1 (this reduces group member 1's earnings by 12 tokens) and 6 deduction points to group member 2 (this reduces group member 2's earnings by 18 tokens). After all have made their decision, you learn that the others allocated you a total of 8 deduction points.

In this case, your earnings are equal to: $(18 – 3 \times 8) – 10 = 0 – 10 = –10$ tokens.

Note that $18 – 3 \times 8 = –6$, since this is a negative number it is replaced by zero.

Negative earnings

It is, in principle, possible that you make negative earnings in a period. However, you can always avoid this by not spending any tokens in the second stage (that is, by not allocating any deduction points to the other members). Hence, you can always avoid negative earnings with certainty through your own choices.

Summary

In summary, your earnings in tokens in each period are equal to:

(Your initial endowment – your contribution to the project
+ 0.5 × (sum of contributions)
– 3 × total deduction points received from others)*
– amount of deductions points you allocated to others

* If your earnings up to this point are negative then replace them with zero

Appendix 5B – Descriptive Statistics

Table 5B.1 summarizes the contributions to the public good per period in each treatment.

TABLE 5B.1 – MEAN CONTRIBUTIONS IN EACH TREATMENT

Period	Baseline	Restricted	Unrestricted
1	12.83 (5.97)	13.43 (6.86)	16.78 (12.74)
2	13.94 (6.22)	14.00 (6.32)	18.89 (12.18)
3	15.94 (4.92)	14.67 (5.59)	17.00 (11.94)
4	15.39 (6.02)	14.90 (6.28)	16.00 (12.70)
5	17.28 (3.46)	14.71 (5.55)	17.44 (11.40)
6	17.50 (3.49)	13.71 (6.53)	20.78 (10.81)
7	15.83 (6.61)	14.43 (5.86)	20.89 (10.55)
8	16.83 (4.58)	14.52 (6.20)	20.11 (10.90)
9	16.83 (4.97)	15.67 (5.26)	20.28 (10.64)
10	14.89 (6.91)	12.29 (8.49)	16.11 (12.33)
Total	15.72 (5.49)	14.23 (6.28)	18.43 (11.50)

Note: Mean contribution per subject per period. Numbers between brackets are standard deviations.

Table 5B.2 summarizes the contributions to the public good per period by poor and rich subjects in each of the unequal treatments.

TABLE 5B.2 – MEAN CONTRIBUTIONS OF RICH AND POOR SUBJECTS

Period	Restricted Poor	Restricted Rich	Unrestricted Poor	Unrestricted Rich
1	15.36 (5.87)	9.57 (7.50)	11.42 (7.29)	27.50 (15.08)
2	14.93 (6.58)	12.14 (5.79)	14.08 (6.83)	28.50 (15.35)
3	15.21 (5.81)	13.57 (5.38)	14.58 (7.53)	21.83 (17.84)
4	15.57 (5.79)	13.57 (7.48)	12.08 (8.56)	23.83 (16.62)
5	15.64 (4.80)	12.86 (6.84)	14.33 (7.00)	23.67 (16.27)
6	14.64 (5.72)	11.86 (8.07)	15.33 (5.91)	31.67 (10.33)
7	14.50 (5.65)	14.29 (6.73)	15.92 (5.88)	30.83 (11.14)
8	15.50 (4.82)	12.57 (8.44)	15.58 (6.11)	29.17 (13.20)
9	15.07 (6.06)	16.86 (3.18)	15.42 (6.20)	30.00 (11.40)
10	12.07 (8.79)	12.71 (8.50)	14.33 (7.56)	19.67 (19.20)
Total	14.85 (5.97)	13.00 (6.74)	14.31 (6.82)	26.67 (14.29)

Note: Mean contribution per subject per period. Data corresponds to the unequal treatments only. Numbers between brackets are standard deviations.

Table 5B.3 and Table 5B.4 summarize punishment per period in each treatment.

TABLE 5B.3 – PUNISHMENT GIVEN BY RICH AND POOR SUBJECTS

Period	Baseline	Restricted Poor	Restricted Rich	Unrestricted Poor	Unrestricted Rich
1	2.00 (2.50)	1.36 (1.55)	0.86 (1.57)	1.08 (1.56)	1.83 (1.83)
2	1.89 (2.52)	1.21 (2.72)	2.71 (3.09)	1.75 (3.11)	1.67 (2.88)
3	1.06 (1.55)	1.64 (2.68)	0.43 (0.79)	1.25 (2.01)	3.00 (4.82)
4	0.78 (1.11)	1.79 (2.83)	2.14 (3.76)	0.42 (1.00)	1.33 (1.63)
5	0.61 (1.29)	1.86 (3.37)	0.57 (1.51)	1.75 (2.96)	1.50 (2.35)
6	0.61 (1.04)	2.93 (4.53)	1.57 (2.15)	0.75 (1.86)	0.67 (1.63)
7	1.61 (3.16)	1.79 (3.04)	1.00 (1.83)	1.00 (2.37)	2.50 (3.89)
8	0.94 (1.98)	1.93 (3.08)	0.57 (1.51)	0.67 (1.50)	1.00 (2.45)
9	0.61 (0.98)	1.00 (2.80)	0.71 (1.50)	0.25 (0.62)	0.00 (0.00)
10	0.67 (2.35)	1.86 (3.06)	1.86 (3.76)	1.50 (3.12)	1.67 (4.08)
Total	1.08 (2.00)	1.74 (2.99)	1.24 (2.32)	1.04 (2.15)	1.52 (2.78)

Note: Mean amount of punishment points assigned per subject per period. Numbers between brackets are standard deviations.

TABLE 5B.4 – PUNISHMENT RECEIVED BY RICH AND POOR SUBJECTS

Period	Baseline	Restricted Poor	Restricted Rich	Unrestricted Poor	Unrestricted Rich
1	2.00 (2.50)	0.50 (1.02)	2.57 (2.51)	1.25 (1.60)	1.50 (1.76)
2	1.89 (2.52)	2.21 (4.77)	0.71 (1.11)	1.08 (2.54)	3.00 (4.69)
3	1.06 (1.55)	0.79 (1.37)	2.14 (2.12)	1.83 (2.44)	1.83 (3.25)
4	0.78 (1.11)	1.64 (4.03)	2.43 (3.82)	1.08 (1.78)	0.00 (0.00)
5	0.61 (1.29)	0.71 (1.27)	2.86 (3.63)	1.75 (4.14)	1.50 (1.76)
6	0.61 (1.04)	1.43 (1.70)	4.57 (5.77)	0.58 (1.38)	1.00 (1.55)
7	1.61 (3.16)	1.29 (1.90)	2.00 (3.65)	1.67 (2.23)	1.17 (2.04)
8	0.94 (1.98)	0.50 (1.16)	3.43 (3.82)	0.67 (1.50)	1.00 (2.00)
9	0.61 (0.98)	0.86 (1.56)	1.00 (2.65)	0.17 (0.39)	0.17 (0.41)
10	0.67 (2.35)	1.93 (5.53)	1.71 (2.98)	0.83 (1.95)	3.00 (4.00)
Total	1.08 (2.00)	1.19 (2.87)	1.09 (2.18)	2.34 (3.37)	1.42 (2.55)

Note: Mean amount of punishment points received per subject per period. Note that in the baseline treatment, punishment points given and received are equal. Numbers between brackets are standard deviations.

Chapter 6

The Disadvantage of Privileged Groups

The Importance of Reference Groups and Interpersonal Comparisons for Social Punishment[*]

In this chapter, we consider a different public goods experiment. This experiment is of particular interest since it investigates contributions to a public good when cooperation is not supported by social norms such as reciprocity and equity.

6.1 Introduction

Social scientists have found that, when it comes to explaining the provision of public goods, it is much easier to explain failure than success. As is clearly explained by Olson (1965), individual incentives to free ride on the effort of others lead to a sub-optimal provision of public goods. Hence, unless a group has very specific characteristics, the provision of the collective good is doomed to fail. Although there is a large literature addressing this problem, we still do not have a complete and satisfactory explanation of public good provision. Undoubtedly, important characteristics have been identified that help groups overcome the free rider problem, in particular, face-to-face communication (Ostrom et al., 1994; Ostrom and Walker, 1997), and as has been discussed throughout this thesis, decentralized punishment (see also Fehr and Gächter, 2000b, and Fehr and Fischbacher, 2004).

In his work, Olson (1965) identifies a type of group in which provision of the public good is not a serious problem. Correspondingly, he calls them privileged groups. Such groups are characterized by having one or more individuals who receive a disproportionately high utility from the consumption of the public good. Thus, they are willing to pay for the public good to be provided at a significant level. The other group members simply benefit from the public good without having to pay for it. Perhaps because public good provision in these groups can be seen as trivial, privileged groups have not received much attention in the literature. However, we argue that recent findings concerning cooperative behavior call for a better understanding of privileged groups. First, the role of fairness norms and reciprocity can have big effects on the willingness of individuals to tolerate free riding (e.g. Fehr and

[*] This chapter is partly based on Reuben and Riedl (2005).

Gächter, 2000b). Second, the asymmetric nature of privileged groups can cause additional conflict as different notions of fairness can come into play (see Chapter 5).[83]

In this chapter, we experimentally investigate cooperative behavior in privileged and non-privileged groups. In particular, we compare contributions to a public good in situations when decentralized punishment is available and when it is not. In this respect, we find that, whereas in non-privileged groups punishment is clearly beneficial, in privileged groups this is not the case. In fact, in spite of the reduced free-riding incentives, privileged groups fail to outperform non-privileged groups in cases when punishment is possible. To explain this, we analyze how individuals punish and how they react to punishment depending on their benefit from contributions to the public good.

For our study, we use a public goods experiment in the line of Isaac et al. (1984) where in half of the treatments subjects are allowed to punish each other (as in Fehr and Gächter, 2002). Subjects participate in either a privileged or a baseline (non-privileged) treatment. In the baseline treatment, all subjects in a group have an incentive to free ride. In contrast, in the privileged treatment, although most subjects face the same incentives as in the baseline treatment, there are subjects who do not benefit from free riding. Irrespective of what other subjects do, these subjects get higher earnings by contributing to the public good.[84] We refer to these subjects as high-value subjects and to the others as low-value subjects. Traditional economic theory (assuming self-regarding preferences) predicts that, in all treatments, only high-value subjects will contribute to the public good.

Punishment has been shown to be an effective way of increasing cooperation in non-privileged groups.[85] This is attributed to the willingness of high contributors to punish low contributors, which, in turn, makes free riding unprofitable. This type of behavior might be supported by two important characteristics of non-privileged groups. First, since contributions to the public good decreases the contributor's earnings and increases the earnings of others, it is clear that a high contributor is being kind. Second, contributions by a low contributor reduce the income difference between the low contributor and a high contributor. Hence, cooperation is supported by both reciprocity-based and equity-based fairness norms.

In privileged groups, fairness norms are not necessarily compatible with high levels of cooperation. First, contributions by high-value subjects can be due to kindness towards others

---

[83] Naturally, studying the behavior of individuals in asymmetric situations can also help us gain a better understanding of fairness norms and reciprocal behavior.

[84] Individuals may value differently the consumption of the public good for numerous reasons. It could be simply a difference in monetary benefits. For example, an individual that owes a large plot of land would benefit much more from a regional irrigation system than an individual that owes a small plot of land. However, it can also be that people perceive differently the importance of the public good. For instance, some neighbors might enjoy more than others the existence of a neighborhood swimming pool.

[85] A few examples of studies on decentralized punishment are Ostrom et al. (1992), Fehr and Gächter (2000b), Masclet et al. (2003), Carpenter (2004), Bochet et al. (2005), Egas and Riedl (2005), Gächter and Herrmann (2005), and the previous chapter.

or to the maximization of their own payoff. Therefore, it is no longer clear what the intentions behind their contributions are. Since intentions have been shown to be of importance for reciprocal behavior (Falk et al., 2000; McCabe et al., 2003), this can make low-value subjects unwilling to reciprocate high contributions by high-value subjects, and high-value subjects unwilling to punish low contributions by low-value subjects. Second, contributions by low-value subjects can actually increase income differences between themselves and the high-value subject (this is the case in our design). In other words, equity-based notions of fairness actually support a situation in which low-value subjects do not contribute to the public good.[86] Hence, as long as subjects dislike transferring income from the poorer subjects to the richest subject, they will accept low contributions by low-value subjects and might even use punishment to enforce them.[87]

An analysis of the way subjects punish in the privileged treatment supports some of these assertions. On one hand, we find that low-value subjects do not reciprocate the high contributions of high-value subjects. On the other hand, we find that punishment is still an effective way of inducing low-value subjects to contribute. In fact, the lack of positive reciprocity makes punishment the only tool high-value subjects can use to increase cooperation. It also explains why low-value subjects contribute much less than subjects in non-privileged groups, even though they are punished by similar amounts.

To the best of our knowledge, no other experiment investigates cooperation and punishment in privileged groups. In public good experiments, it is clear that changing the incentive to cooperate (i.e. the marginal per capita return to cooperation) has a strong effect on the willingness of subjects to contribute to the public good (see Ledyard, 1995). Closer to our experiment is the study by Fisher et al. (1995), in which they analyze the effect of unequal incentives to cooperate within subjects in a group.[88] As we do, they use a linear public good framework. However, in their case all subjects still have an incentive to free ride. This does not create the conflict between cooperation and fairness norms that was just described. They find that contributions to the public good are higher when the marginal per capita return to cooperation differs between group members. Furthermore, unlike our findings, they observe that low-value subjects contribute slightly more than subjects in homogenous groups.

The chapter is organized as follows. In Section 6.2 we describe the design of the experiment in relation to theoretical models of fairness and reciprocity. Section 6.3 analyzes

[86] The same is true for other types of distributional fairness notions such as a concern for the income of the least well off (Rawls, 1971).

[87] On the other hand, if subjects care for overall efficiency, low contributions by low-value subjects are clearly undesirable. For more discussion on distributional preferences, see Charness and Rabin (2002) and Engelmann and Strobel (2004).

[88] Other experimental work that addresses the effects of heterogeneity in preferences for the public good include Marwell and Ames (1979), Brookshire et al. (1993), Rapoport and Suleiman (1993), Chan et al. (1999), and Chan et al. (2003).

the subjects' cooperation and punishment behavior, while Section 6.4 discusses the main results and concludes.

6.2 Experimental Design and Theoretical Predictions

The experiment consists of a repeated public-good game with or without punishment opportunities. Subjects are divided into groups of three and then they play the one-shot version of the game for ten consecutive periods. Group composition remains the same during the whole experiment. Furthermore, both the number of periods and group composition are common knowledge. In the treatments without punishment, each period of the game consists of one stage, that is, a contribution stage. In the treatments with punishment, each period consists of two stages: a contribution stage and a punishment stage.

In the contribution stage, each subject i receives an endowment of 20 tokens. Thereafter, subjects simultaneously decide what amount $c_i \in [0,20]$ they wish to contribute to their group's public good. For each token contributed to the public good by any group member, subject i receives α_i tokens as earnings. In other words, subjects receive a marginal per capita return (MPCR) from contributions to the public good of α_i. As long as $\alpha_i < 1$ individuals have an incentive to free ride. Moreover, from the groups' perspective, if $\Sigma_i \alpha_i > 1$, the best outcome is attained if everyone contributes all their endowment.

In the punishment stage subjects are first informed of the MPCR and the individual contributions of other group members. Next, each subject i decides how many punishment points to assign to each other subject j in their group, $p_{ij} \in [0,10]$. Each punishment point costs the punisher one token and reduces the earnings of the punished subject by three tokens. In order to avoid large losses during the experiment, subjects are not allowed to punish others below zero earnings.[89] After subjects make their punishment decision, they are informed of the total number of punishment points assigned to them by other subjects in their group. As in Fehr and Gächter (2000b) subjects are not informed which subject assigned them punishment points. In summary, if earnings are positive, each subject i receives in each period the following amount.[90]

$$\pi_i = 20 - c_i + \alpha_i \Sigma_j c_j - \Sigma_{j \neq i} p_{ij} - 3\Sigma_{j \neq i} p_{ji}$$

Subjects participate in one of four different treatments. The treatments differ in the availability of punishment and in the value of α_i received by one member in each group. Next, each treatment is described in detail. In addition, the differences are summarized in Table 6.1.

- The *baseline* treatment *without* punishment: In this treatment all subjects receive the same MPCR, $\alpha_i = 0.5$. Furthermore, they play only the contribution stage in each period.

[89] As in Chapter 5, subjects can never be punished below zero tokens, but they can incur losses if they punish others.

[90] If earnings are negative then they equal: $\pi_i = \max[0, 20 - c_i + \alpha_i \Sigma_j c_j - 3\Sigma_{j \neq i} p_{ji}] - \Sigma_{j \neq i} p_{ij}$.

- The *baseline* treatment *with* punishment: As above, in this treatment all subjects receive the same MPCR, $\alpha_i = 0.5$. However, they play both the contribution stage and the punishment stage in each period. Actually, this treatment is identical to the baseline treatment in Chapter 5. Therefore, instead of running additional sessions, we use the data obtained in Chapter 5's experiment.
- The *privileged* treatment *without* punishment: In this treatment, one subject per group is randomly selected to be the high-value subject. The other two subjects in the group are therefore the low-value subjects. Subjects are either high-value or low-value for the duration of the experiment. High-value subjects receive an MPCR of $\alpha_H = 1.5$. Low-value subjects receive an MPCR of $\alpha_L = 0.5$. Lastly, subjects play only the contribution stage in each period..
- The *privileged* treatment *with* punishment: This treatment is the same as the privileged treatment without punishment except that subjects play both the contribution stage and the punishment stage in each period.

TABLE 6.1 – SUMMARY OF THE PARAMETERS USED IN EACH TREATMENT

	Without Punishment	**With Punishment**
Baseline	$\alpha_i = 0.5 \; \forall \; i$ $p_{ij} = 0 \; \forall \; i, j \neq i$	$\alpha_i = 0.5 \; \forall \; i$ $p_{ij} \in [0,10] \; \forall \; i, j \neq i$
Privileged	$\alpha_1 = 1.5$ and $\alpha_i = 0.5 \; \forall \; i \neq 1$ $p_{ij} = 0 \; \forall \; i, j \neq i$	$\alpha_1 = 1.5$ and $\alpha_i = 0.5 \; \forall \; i \neq 1$ $p_{ij} \in [0,10] \; \forall \; i, j \neq i$

Note: In each group there are three subjects, that is $i, j \in \{1,2,3\}$.

Note that in the privileged treatments, high-value subjects receive 1.5 tokens for every token they contribute to the public good. Thus, they do not have an incentive to free ride. In fact, they have a dominant strategy to contribute all their endowment in every period. As the treatments' names suggest, this mirrors Olson's definition of privileged groups (Olson, 1965) as being groups in which the public good is supplied because some individuals receive a net benefit from doing so.

The baseline treatments and the privileged treatments differ only in the presence of a high-value subject instead of a low-value subject. Thus, by comparing the two, we can determine the high-value subject's effect on overall contribution levels when punishment is or is not available. Furthermore, note that, low-value subjects and subjects in the baseline treatments receive the same endowment and face the same MPCR. Hence, any differences between these subjects' behavior must be due to the presence of the high-value subject, and not due to a change in the incentives provided by the parameters of the game.

Since the game has a known end, in the baseline treatment without punishment, individuals with self-regarding preferences have no incentive to cooperate. Hence, if we make this assumption, nobody is predicted to contribute a positive amount to the public good.

Indeed, public good experiments show that, with repetition, cooperation declines to very low levels (see Ledyard, 1995). In the privileged treatment without punishment, low-value individuals face the same incentives for cooperation, and therefore they are not expected to contribute. On the contrary, high-value individuals earn more money the more they contribute, and hence, they should contribute all their endowment. In summary, as suggested by Olson (1965), individuals with self-regarding preferences are better off in the privileged treatment, where low-value individuals can free ride on the contributions of the high-value individual.

Given that punishment is costly and hence not credible, these predictions do not change for the treatments in which punishment is available. However, in this case, the experimental evidence does not conform to these theoretical predictions. If punishment is possible, subjects frequently sanction each other and contributions to the public good do not decline with repetition (Fehr and Gächter, 2000b). In order to explain this and similar results, various models have been proposed that are better able to explain punishment in a range of experiments (e.g. see footnote 5). They commonly assume that individuals possess other-regarding (social) preferences. In the following paragraphs, we describe the incentives for contribution and punishment that individuals face in some of these theoretical models.

In models that assume individuals dislike income differences (e.g. Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), some individuals have an incentive to cooperate even if their MPCR is below one. In particular, they are willing to contribute a positive amount only if others do the same. This means that in the baseline treatment without punishment, a group will cooperate only if all its members care enough about income differences. In the baseline treatment with punishment, the situation is quite different. In this case, individuals who contribute high amounts and care enough about inequity will punish individuals who contribute less. Punishment gives both selfish and non-selfish individuals an incentive to cooperate. Therefore, cooperation can be sustained even if a group has some individuals who only care about their own earnings.

In the privileged treatment without punishment, individuals face a different situation. First, note that contributions by the high-value individual have no effect on income differences within a group. Hence, since contributing makes them richer, they will contribute all their endowment irrespective of their preferences over income differences. In contrast, positive contributions by low-value individuals only *increase* income differences. Consequently, no low-value individuals cooperate. The situation does not change if punishment is introduced. If high-value individuals contribute all their endowment and low value individuals contribute nothing, everyone in the group earns 30 tokens. Since punishment is motivated by income differences and any contribution by a low-value individual increases inequality, there is no reason to force low-value individuals, through

punishment, to cooperate more.[91] Thus, unlike in the baseline treatments, punishment does not increase cooperation in the privileged treatments.

Incentives for contribution and punishment are similar in the intention-based model of Falk and Fischbacher (2005). In this model, individuals use income differences to judge the kindness and intentions of others. In the baseline treatments, individuals will interpret contributions higher than theirs as kind and contributions lower than theirs as unkind. In the absence of punishment, contributions can be high only if everyone cares enough to reward the high contributions of others. If punishment is available, high contributors might be willing to punish low contributors. This gives all individuals an incentive to cooperate. In the privileged treatments, the earnings of high-value individuals are never lower than the earnings of low-value individuals. This means that low-value individuals will not consider high contributions by high-value individuals as being kind. Moreover, for the same reason, high-value individuals will not consider low contributions by low-value individuals as unkind. This implies that high-value individuals will contribute all their endowment and never punish low-value individuals. It also implies that, as long as they contribute all their endowment, their behavior will be ignored by the low-value individuals. Thus, from the perspective of low-value individuals, they are in the same position as individuals in the baseline treatments, except that they are in a group of two instead of three. That is, if there is no punishment, their contributions will be high only if both care enough to reward the high contributions of the other. If there is punishment, higher levels of cooperation can be enforced by a low-value individual who cares enough for reciprocation.

Incentives to contribute and punish do change if we consider the model presented in Charness and Rabin (2002). In this model, individuals care for both the earnings of the least well off and for overall efficiency. In all treatments, these two preferences can motivate individuals to contribute a positive amount. In fact, if the efficiency motive is strong enough, individuals will contribute unilaterally to the public good. Furthermore, since low contributors are usually not the worst off in their group and their low contribution fails to promote efficiency, all other individuals might be willing to punish them. Hence, as in previous models, punishment gives all individuals an incentive to cooperate. However, unlike in previous models, high-value individuals can punish low-value individuals for contributing too little. In fact, contributions by high-value and low-value individuals have similar effects on the earnings of the least well off and on overall efficiency. Therefore, for a given contribution, high-value individuals will be punished as severely as low-value individuals. Lastly, like in Falk and Fischbacher (2005), but unlike the models of Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), it is never the case that high-value individuals will receive punishment

[91] Certainly, as long as earnings end up equal, punishment can be used to enforce other cooperation levels. In our design, this would imply that, in equilibrium, low-value individuals punish high-value individuals simply to reduce income differences. However, earnings in these equilibria are always Pareto-dominated by the case in which low-value individuals do not contribute.

if they contribute all their endowment. In the following section, we analyze the results of the experiment.

6.3 Results

In total 81 subjects participated in the experiment, 21 (18) participated in the baseline treatment without (with) punishment, and 24 (18) in the privileged treatment without (with) punishment. About 50% of the subjects were women. The precise experimental procedures and the instructions are found in Appendix 6A. Furthermore, descriptive statistics of the data are available in Appendix 6B.

6.3.1 Cooperation

Overall, in the two treatments where there are no punishment opportunities, average contributions to the public good are low and decrease over time. Moreover, compared to the baseline treatment, contributions are significantly higher in the privileged treatment (8.68 vs. 4.21, $p = 0.004$).[92] See Figure 6.1.

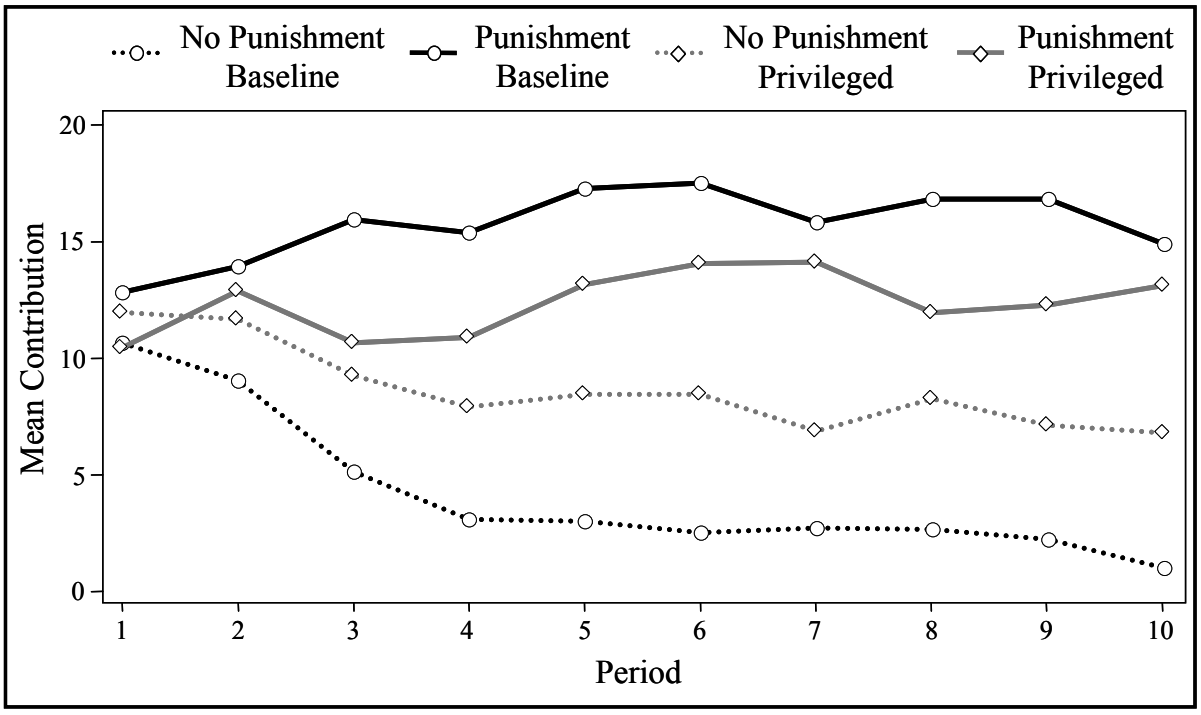

FIGURE 6.1 – MEAN CONTRIBUTIONS

Note: Mean contribution per treatment over the ten periods.

[92] Throughout the chapter, unless it is otherwise noted, we always use a two-sided Wilcoxon-Mann-Whitney test, and group averages as independent observations. Furthermore, as in Chapter 5, given the low number of observations, we refer to a difference as being statistically significant if the p-value of the test is below 0.100.

Even though in both of these treatments the decline of contributions is highly significant (Cuzick trend tests, $p < 0.001$), in the baseline treatment contributions decline by a larger amount. Certainly, this difference can be attributed to convergence towards different Nash equilibria. In the baseline treatment, we observe the familiar decline to zero contributions. In the privileged treatment, we see that contributions converge towards 6.67 tokens, which is the average contribution obtained when the high-value subject contributes 20 tokens and the two low-value subjects do not contribute.

In the two treatments where subjects have the opportunity to punish, contributions do not decline over time. However, it is no longer true that subjects contribute more in the privileged treatment. On average, subjects contribute 15.73 tokens per period in the baseline treatment, whereas in the privileged treatment, they contribute 12.36 tokens. This difference is not statistically significant ($p = 0.200$). We also note that whereas contributions remain constant in the privileged treatment (Cuzick trend test, $p = 0.254$), they show an upward trend in the baseline treatment (Cuzick trend test, $p = 0.057$).

If we look at the effect of the punishment institution, we find a large and significant difference between the amount contributed in the baseline treatment without punishment and the baseline treatment with punishment ($p = 0.003$). In the privileged treatments, the difference between groups with and without punishment is smaller but is nevertheless statistically significant ($p = 0.071$). It appears that punishment increases contributions in both treatments. However, it does so much more dramatically in cases where all subjects benefit equally from the public good. These findings are summarized in our first result.

RESULT 6.1: *In the absence of punishment, contributions to the public good are higher in the privileged treatment. In contrast, if punishment is possible, the difference between treatments disappears. Hence, although punishment supports cooperative behavior, it has a larger effect in the baseline treatment*.

Undoubtedly, given the different incentives faced by high-value and low-value subjects, it is not surprising that the amounts they contribute are remarkably different. In both privileged treatments, with and without punishment, high-value subjects contribute significantly more than low-value subjects (on average, 17.03 vs. 4.50 when there is no punishment and 18.43 vs. 9.32 when there is punishment, $p < 0.009$).

We find more surprising that although the contributions of high-value subjects are high, they are below 20 tokens, particularly in the treatment without punishment. Recall that high-value subjects have a dominant strategy to contribute all their endowment even if they dislike income differences or inefficient outcomes. Hence, if this reduction in contributions is due to the low level of cooperation by low-value subjects, this supports a notion of conditional cooperation that is not motivated by final outcomes but instead by reciprocity. It appears that high-value subjects dislike helping low-value subjects if the latter do not reciprocate. The behavior of high-value subjects can then be interpreted as punishment

towards the low-value subjects for not contributing. In fact, if this is the case, it explains why contributions in treatments without punishment are lower than in treatments with punishment (17.03 vs. 18.43). That is, since there is a more targeted and cheaper punishing mechanism, there is no need for high-value subjects to reduce their contribution in order to punish. If we test whether average contributions are different from 20 tokens, we obtain a significant difference in the treatment without punishment but not in the treatment with punishment (Wilcoxon matched-pairs signed-rank tests, $p = 0.019$ and $p = 0.159$). Lastly, the way contributions evolve over time, provides more evidence that high-value subjects punish by lowering their contributions. In the treatment without punishment, contributions by high-value subjects initially decrease, but towards the end of the game, they converge towards 20 tokens (see Figure 6.2A). It looks like high-value subjects try to punish at the beginning of the game, but as the number of future rounds decreases, they stop doing so. This 'end-game effect' is similar to the one observed in public good games where punishment is not very effective (Nikiforakis and Normann, 2005).[93]

In the treatments with no punishment, low-value subjects behave in the same way as subjects in the baseline treatment. On average, they contribute a very similar amount (4.50 vs. 4.21, $p = 0.643$) and, in addition, they reduce significantly their contributions over time (Cuzick trend test, $p = 0.001$). In contrast, in the treatments with punishment, the contributions of subjects in the baseline treatment are closer to the contributions of high-value subjects, and significantly higher than the contributions of low-value subjects (15.73 vs. 9.32, $p = 0.055$). Thus, although punishment does increase the contributions of low-value subjects ($p = 0.071$), the increase is considerably smaller compared to the increase in baseline treatment.[94] We summarize these findings in the following result.

[93] We shortly discuss two possible explanations of low contributions by high-value subjects, namely, the possibility of mistakes and spiteful behavior. Since high-value subjects can make mistakes only in the direction of lower contributions, this can explain contributions below 20 tokens. However, it fails to explain the difference between the treatments with and without punishment. Alternatively, if subjects are punishing because of spitefulness, this could explain both the low contributions and the difference caused by the punishment institution. However, spiteful punishment does not explain the convergence, at the end of the game, towards contributing 20 tokens.

[94] If we calculate the difference between the average contribution of each group in the punishment treatments and the average contribution in the non-punishment treatments, we find that the difference between punishment and no punishment is significantly bigger for subjects in the baseline treatment compared to low-value subjects in the privileged treatment (11.52 vs. 4.82, $p = 0.037$).

RESULT 6.2: *High-value subjects always contribute more than low-value subjects and subjects in the baseline treatment. Furthermore, if punishment is not available, they seem to be willing to contribute less than their full endowment in order to punish free riders. Low value subjects increase their contributions if punishment is introduced. However, compared to subjects in the baseline treatment, they do so by a small amount*.

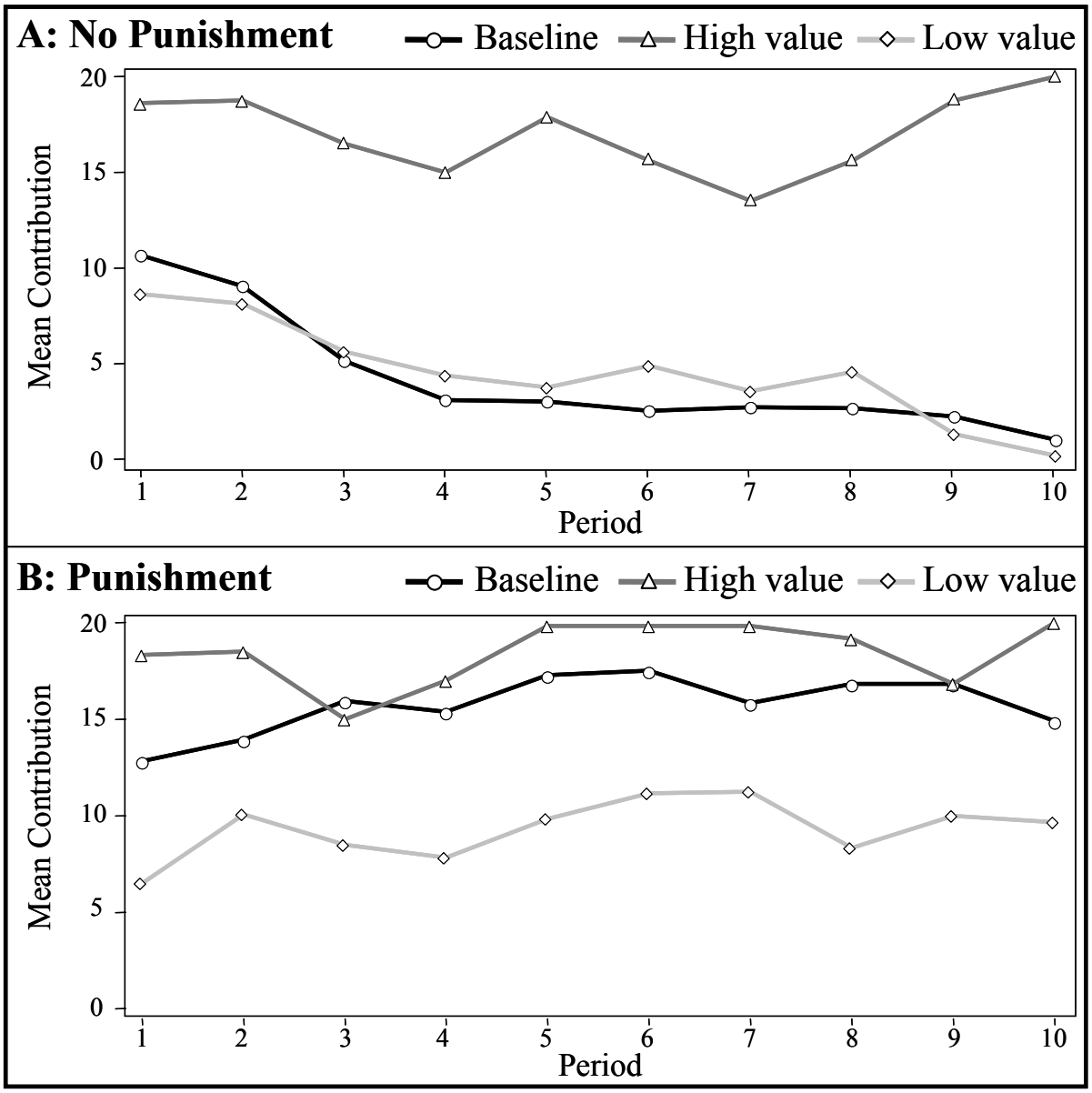

FIGURE 6.2 – CONTRIBUTIONS BY HIGH-VALUE AND LOW-VALUE SUBJECTS

Note: A) Mean contributions by subject type in treatments without punishment opportunities. B) Mean contributions by subject type in treatments with punishment opportunities.

6.3.2 Punishment

On average, the amount of tokens spent on punishment by each subject per period was similar in the privileged and the baseline treatments (1.24 and 1.08, $p = 0.810$). Within the privileged treatment, we find that, on average, high-value subjects punish more than low-value subjects (1.68 vs. 1.07). However, this difference is not statistically significant ($p = 0.296$). If we look at how punishment changes over time, we find that the amount of punishment decreases significantly in both treatments (Cuzick trend tests, $p < 0.005$).

As in Chapter 5, we find that subjects punish differently depending on their treatment and position. This can be seen by running regressions with the amount of punishment given by subject i to subject j (p_{ij}) as the dependent variable. As independent variables we use the contributions of subject i (the punisher), subject j (the punished), and subject k (the third subject in the group), in addition, we control for the period number and the amount of punishment received by i in the previous period. We analyzed separately the following cases: low-value subjects punishing low-value subjects, low-value subjects punishing high-value subjects, and high-value subjects punishing low-value subjects.

The results of the regressions are presented in Appendix 6C. However, they are better illustrated by means of Table 6.2. For each treatment and depending on whether subject i punished subject j, the table shows the average contribution of subject i, subject j, and subject k.[95]

In the baseline treatment, we can see that subjects punish in cases when overall contributions are low, and more importantly, when their own contribution and the contribution of the third subject are higher than the contribution of the subject they punish. In other words, subjects tend to punish the subject who contributed the least. Since a detailed analysis of punishment in the baseline treatment is available in the previous chapter (Section 5.4.2), we now turn to punishment in the privileged treatment.

Low-value subjects punish each other in a similar way as subjects in the baseline treatment. That is, low-value subjects punish other low-value subjects when they contribute more than the other does. However, in this case the contributions of the third subject do not seem to affect their punishment decision. Given the difference in incentives, the contributions of high-value subjects do not provide a clear reference point from which to judge the contributions of low-value subjects. This explains why, low-value subjects seem to ignore the contribution of the third (high-value) subject when deciding on the punishment of the other low-value subject.

---

[95] Specifically, the average contribution of i when i does not punish j is given by $\Sigma_i\Sigma_{i\neq j} f_{ij}c_i \,/\, \Sigma_i\Sigma_{i\neq j} f_{ij}$ where $f_{ij} = 1$ if $p_{ij} = 0$ and $f_{ij} = 0$ otherwise. Correspondingly, the average contribution of j is given by $\Sigma_i\Sigma_{i\neq j} f_{ij}c_j \,/\, \Sigma_i\Sigma_{i\neq j} f_{ij}$, and the average contribution of k is given by $\Sigma_i\Sigma_{i\neq j} f_{ij}c_k \,/\, \Sigma_i\Sigma_{i\neq j} f_{ij}$ where $k \neq i \neq j$. Average contributions when i does punish j are calculated by changing f_{ij}.

TABLE 6.2 – MEAN CONTRIBUTIONS DEPENDING ON PUNISHMENT

	c_i	c_j	c_k
Baseline $p_{ij} = 0$	16.36 (5.35)	17.14 (4.53)	15.60 (5.94)
Baseline $p_{ij} > 0$	13.72 (5.45)	11.24 (5.86)	16.14 (3.69)
Low-value to low-value $p_{ij} = 0$	8.54 (7.66)	9.84 (7.76)	18.52 (4.15)
Low-value to low-value $p_{ij} > 0$	12.00 (5.58)	7.52 (5.54)	18.15 (4.48)
Low-value to high-value $p_{ij} = 0$	8.88 (7.34)	18.68 (3.90)	8.77 (7.35)
Low-value to high-value $p_{ij} > 0$	11.63 (7.24)	17.11 (5.53)	12.21 (6.90)
High-value to low-value $p_{ij} = 0$	18.65 (4.28)	9.99 (7.47)	9.30 (7.55)
High-value to low-value $p_{ij} > 0$	18.00 (4.09)	7.98 (7.04)	9.35 (7.07)

Note: Mean contributions of subject *i*, *j*, and *k* depending on whether *i* punished *j* or not (see footnote 95). Numbers between brackets are standard deviations.

Low-value subjects punish high-value subjects when the latter contribute relatively low amounts, and notably, when low-value subjects contribute relatively high amounts. High-value subjects usually contribute all their endowment to the public good. Hence, a relatively low contribution in their case is a contribution below 20 tokens. As can be seen in Table 6.2, the average contributions of high-value subjects that are punished are somewhat lower than the contributions of those who are not punished. However, judging by the significance of the coefficients in the regression (see Table 6C.1), it appears that the contribution of high-value subjects is not the main determinant of whether they are punished or not. More important are the contributions of low-value subjects. Specifically, low-value subjects punish high-value subjects when they, as well as the other low-value subject, contribute a high amount, in other words, when the average contribution of the low-value subjects is relatively high. For example, if we look at periods in which high-value subjects contribute all their endowment, we find that when the average contribution of low-value subjects is relatively high (top quartile), low-value subjects assign 2.92 punishment points to high-value subjects. In contrast, when their average contribution is relatively low (bottom quartile), low-value subjects assign only 0.42 punishment points to high-value subjects ($p = 0.031$). This behavior is somewhat

inconsistent since by contributing a high amount, low-value subjects increase overall efficiency, inequality, and the earnings of high-value subjects. However, at the same time, by punishing in the subsequent stage, they reduce efficiency, inequality, and earnings of high-value subjects. A possible reason for this type of behavior is that low-value subjects contribute to avoid punishment from the high-value subject. However, they dislike this situation and therefore, the more they are 'forced' to contribute, the more they punish the high-value subject.[96]

High-value subjects punish the low-value subjects who contribute a relatively low amount. More specifically, as can be seen in Table 6.2, low-value subjects that receive punishment from high-value subjects are those who contribute less than the other low-value subject in the group. Furthermore, the contribution of high-value subjects has no apparent effect on their punishment decision. Therefore, in the same way as low-value subjects, high-value subjects seems to judge whether the contribution of a low-value subject is high or low by comparing it solely to the contribution of the other low-value subject.[97] Lastly, unlike low-value subjects and subjects in the baseline treatment, high-value subjects punish less if they were punished in the previous period. The following result summarizes these findings.

RESULT 6.3: *In the baseline treatment, subjects punish those who contribute relatively low amounts compared to all group members. In the privileged treatment, both low-value and high-value subjects punish low-value subjects who contribute relatively low amounts compared to the other low-value subject. High-value subjects are punished irrespective of their very high contributions, and they are punished by low-value subjects that contribute high amounts.*

Although punishment behavior differs between treatments, this does not explain why punishment increases contributions more in the baseline treatment than in the privileged treatment. In particular, it does not explain the difference in contributions between subjects in the baseline treatment and low-value subjects in the privileged treatment. In treatments with no punishment, low-value subjects and subjects in the baseline treatment contribute very similar amounts (4.50 vs. 4.21). However, in treatments with punishment opportunities, low-value subjects contribute considerably less than subjects in the baseline treatment (9.32 vs. 15.73; see Result 6.2). Moreover, this is despite the fact that low-value subjects were assigned similar amounts of punishment points (1.34 vs. 1.08, $p = 0.873$). Punishment opportunities also slightly increase the contributions of high-value subjects. However, since this difference

[96] Given that contributions increase income differences, this might motivate low-value subjects to punish more the more they are forced to contribute (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000).

[97] The model of Charness and Rabin (2002) can explain why high-value subjects might punish low-value subjects. However, in this model, the punishment of a low-value subject does not depend on the actions of the other low-value subject.

is small, we concentrate on the effects of being punished on the contributions of low-value subjects.

Given that the reaction of subjects to punishment depends on the amount they contributed before being punished (Cinyabuguma et al., 2004), we first classify subjects into different types depending on their contribution relative to the contribution of others. We use three types: low contributors, median contributors, and high contributors. A subject is classified as a low contributor in a period if her contribution is strictly lower than the contributions of others in the group. Similarly, a subject is classified as a high contributor if her contribution is strictly greater than the contributions of others in the group. Consequently, median contributors are subjects who contribute the intermediate amount in periods where there are three different contributions or subjects who contribute the same amount as at least one other subject. Next, we run a regression with the amount contributed in period t as the dependent variable. As independent variables, we use: the amount contributed in period $t - 1$, the period number, dummy variables for the subject's type in period $t - 1$, and the amount of punishment received by the subject in period $t - 1$ depending on her type in that period. We ran a separate regression for subjects in the baseline treatment and for low-value subjects in the privileged treatment. The results are presented in Table 6.3.[98]

Judging by the signs of the significant coefficients, in the baseline treatment, subjects contribute more in the next period if in the current period they are either a low contributor or a median contributor. Furthermore, if they are punished when they are a low contributor, they contribute even more in the next period. Note that, median contributors who are punished do not contribute more than those who are not punished. Moreover, high contributors who are punished decrease their contribution. In other words, punishment has desirable effects only when targeted towards low contributors.

Punishment of low-value subjects has very similar effects as punishment in the baseline treatment. Namely, low contributors who are punished contribute more in the next period than those who are not punished. Moreover, punishment of median or high contributors does not lead to higher contributions. The main difference between treatments is thus the reaction of subjects to being a low or a median contributor, irrespective of how much punishment they receive.

[98] Running the same regression for high-value subjects, results in insignificant coefficients for all variables except for the amount contributed in period $1 - t$. The lack of significance is probably due to very little variation in the contributions of high-value subjects. In 81.48% of the cases, they contribute all their endowment.

TABLE 6.3 – REGRESSION RESULTS FOR THE AMOUNT CONTRIBUTED PER PERIOD

Independent Variable	Baseline	Low-value
Contribution in the previous period	1.516** (0.202)	1.330** (0.115)
Period	–0.305 (0.242)	–0.024 (0.329)
Punishment after being a low contributor in the previous period	1.753** (0.540)	1.913** (0.415)
Punishment after being a median contributor in the previous period	–0.039 (0.536)	0.243 (0.875)
Punishment after being a high contributor in the previous period	–1.045* (0.625)	0.258 (0.522)
Low contributor in the previous period	5.688** (2.126)	3.445** (1.627)
Median contributor in the previous period	6.965** (1.574)	0.216 (1.715)
Constant	–9.357** (3.240)	–6.482** (2.408)
$\chi^2(7)$	93.82**	192.64**
Observations	162	108

Note: The dependent variable is the amount contributed by each subject in each period excluding the first period. We use Tobit estimates censured both at zero and at twenty tokens. Number between brackets are robust standard errors. ** Significant at the 5 percent level. * Significant at the 10 percent level.

In both treatments, being a low contributor leads to higher contributions in the following period. Furthermore, although this effect is somewhat stronger in the baseline treatment, the difference between the coefficients is not statistically significant (Wald test, $p = 0.377$).[99] More dissimilar is the reaction of median contributors. In the baseline treatment, being a median contributor leads to higher contributions in the next period. In contrast, for low-value subjects, being a median contributor has no effect at all (the coefficient for median contributors is significantly higher in the baseline treatment, Wald test $p = 0.003$). In

[99] In order to test differences between coefficients we use a single regression including subjects in the baseline treatment and low-value subjects. Each independent variable interacts with a dummy variable that indicates the treatment.

summary, it appears that, compared to subjects in the baseline treatment,[100] low-value subjects are more reluctant to raise their contributions unless they are punished. This can explain why, even though they receive similar amounts of punishment, subjects in the baseline treatment contribute more than low-value subjects in the privileged treatment. This is stated in our fourth result.

RESULT 6.4: *In the baseline treatment, subjects contribute more if they contributed a low or an intermediate amount in the previous period. In the privileged treatment, low-value subjects contribute more only if they contributed a low amount in the previous period. In both treatments, only low contributors increase their contributions after receiving punishment.*

6.3.3 Efficiency and Inequality

Decentralized punishment has the desirable characteristic that it increases the amount of contributions to the public good, which has a positive effect on the group's earnings. However, as long as some subjects choose to punish, earnings are also negatively affected. In order to easily compare the different treatments, we normalize earnings such that if subjects do not contribute and do not punish, earnings are equal to zero, and if subjects contribute all their endowment and do not punish, earnings are equal to one. In the baseline treatment without punishment opportunities, average normalized earnings equal 0.21. In contrast, when punishment is possible, they equal 0.37. Although this difference is not statistically significant ($p = 0.668$), there is a clear difference if we look at earnings over the ten periods. Whereas in the treatment without punishment earnings decrease over time, in the treatment with punishment earnings tend to increase (Cuzick trend tests, $p = 0.001$ and $p = 0.013$). As can be seen in the Figure 6.3, which shows the difference between earnings in the punishment and the non-punishment treatments, from period three onwards, earnings are higher in the punishment treatment (excepting period seven, earnings are significantly higher in all periods after period five, $p < 0.084$).

In the privileged treatments, it is not as clear whether punishment leads to higher overall earnings. In the treatment without punishment, normalized earnings equal 0.43. Similarly, in the treatment with punishment, they equal 0.45 (the difference is not significant $p = 0.796$). As in the baseline treatment, we find that average earnings show a decreasing trend in the treatment without punishment, and a increasing trend in the treatment with punishment (Cuzick trend tests, $p = 0.001$ and $p = 0.051$). However, unlike the baseline treatment, towards the end of the experiment earnings are not clearly higher when there are punishment opportunities. Only in periods five and seven, are earnings in the punishment treatment significantly higher than in the non-punishment treatment ($p < 0.091$; for other

[100] The subjects' behavior in the baseline treatment has been observed in various public good games (e.g. Keser and van Winden, 2000).

periods, $p > 0.153$). This follows form the fact that high-value subjects contribute similar amounts in both the punishment and the non-punishment treatments. Thus, if punishment is to increase earnings, it must increase the contributions of low-value subjects. However, as is shown in Result 6.2, low-value subjects do not increase their contributions as much as subjects in the baseline treatment do.

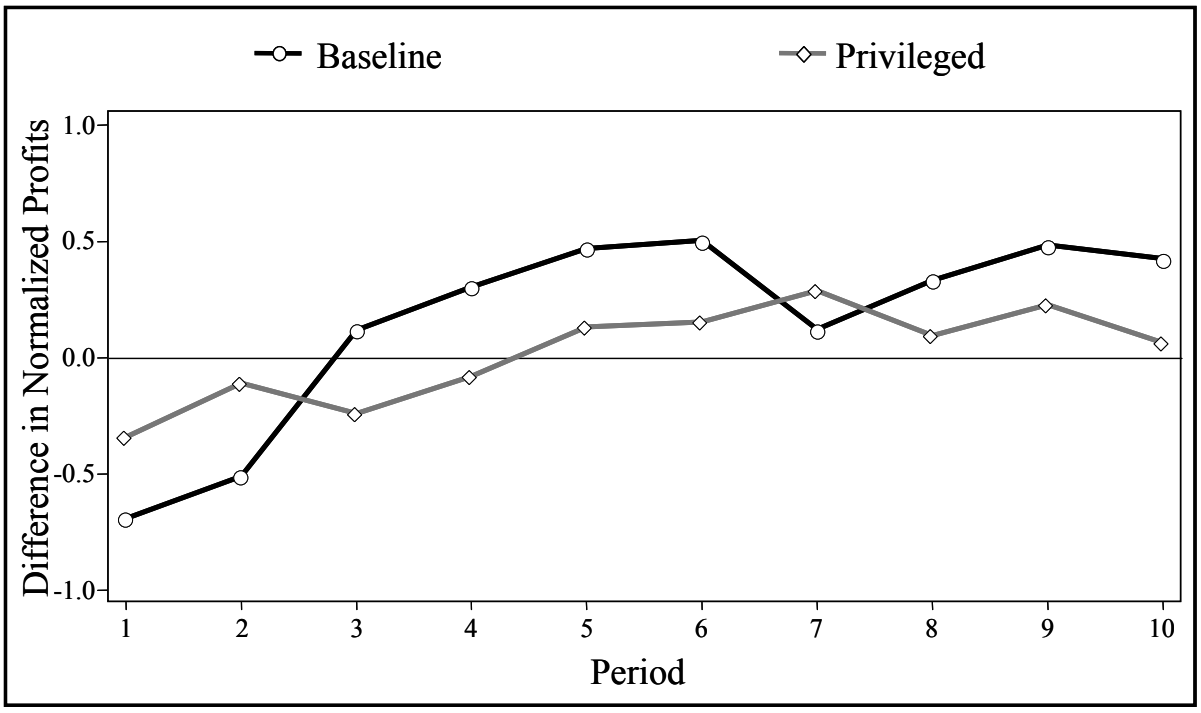

FIGURE 6.3 – EARNINGS GAINED WITH THE PUNISHMENT INSTITUTION

Note: Difference between mean group earnings in treatments with punishment and treatments without punishment. Earnings are normalized so that, when subjects do not contribute and do not punish, earnings are equal to zero, and when subjects contribute everything and do not punish, earnings are equal to one.

Even though low-value subjects free ride on the contributions of high-value subjects, they earn more than their counterparts in the baseline treatment only in the absence of punishment. If punishment is unavailable, the earnings of low-value subjects are significantly higher than those of subjects in the baseline treatment ($p = 0.001$). However, when punishment is available, their earnings do not differ ($p = 0.749$).[101] Hence, the higher contributions induced by punishment in the baseline treatment offsets any benefits obtained by the free riding of low-value subjects in the privileged treatment.

Punishment not only affects the earnings of subjects, but also the way these earnings are distributed. An intuitive way of looking at the distribution of earnings within a group is to calculate the share of the group's total earnings (after punsihment) that is received by the group member who earned the most. In the baseline treatment without punishment, the

[101] Due to the high MPCR, the earnings of high-value subjects are always higher than the earnings of low-value subjects and subjects in the baseline treatment ($p < 0.010$).

highest earner receives on average 35.73% of the group's earnings. When punishment is available, the highest earner receives 34.77%. Thus, on average, earnings are more equally distributed in the punishment treatment. However, the difference is not statistically significant ($p = 0.317$). In the privileged treatments, we find the opposite result. Whereas in the treatment without punishment, the highest earner receives 42.12% of the group's earnings, in the treatment with punishment, the highest earner receives 52.52%. In this case, the difference between treatments is statistically significant ($p = 0.071$). Hence, whereas punishment does not affect the distribution of earnings in the baseline treatments, it leads to a more unequal distribution in the privileged treatments.

We arrive at a stronger result if we look at the effect of punishment on the distribution of earnings within the punishment treatments. More specifically, we calculate the share of earnings received by the highest earner before taking into account the effects of punishment. In the baseline treatment, the highest earner before punishment receives 35.98% of total earnings (after punishment the share is 34.77%). In the privileged treatment, the share of the highest earner before punishment is 48.47% (after punishment the share is 52.52%). Thus, in the baseline treatment, the way subjects punish leads to a more equal distribution of earnings, whereas in the privileged treatment, it produces a more unequal distribution (both differences are significant, Wilcoxon matched-pairs signed-rank tests $p = 0.028$ and $p = 0.046$). Our final result summarizes these findings.

RESULT 6.5: *In the baseline treatments, after a few initial periods, punishment leads to higher overall earnings. Moreover, punishment has little effect on how earnings are distributed. In the privileged treatments, punishment does not lead to higher earnings, and it produces a more unequal earnings distribution*.

6.4 Conclusion

In this chapter, we have seen that, in the absence of punishment, privileged groups (i.e. groups in which one subject lacks an incentive to free ride) enjoy higher levels of cooperation than groups in which all subjects have an incentive to free ride. However, if subjects are allowed to punish each other, it is no longer the case that contributions to the public good are higher in privileged groups. The difference can be attributed to a very small increase in contributions by low-value subjects when punishment is introduced. Surprisingly, this low level of contributions is not due to subjects giving less punishment or to subjects reacting less when they receive punishment. In fact, the low contributions of low-value subjects are due to their reluctance to increase their contributions when they are not punished.

In this study, we confirm that decentralized punishment is an effective way of sustaining cooperative behavior. However, our results indicate that the level of cooperation that is enforced depends on the characteristics of group members (see also Chapter 5). To a great extent, understanding individuals' perception of fair or kind behavior can help us predict

the effects of punishment in different situations.[102] It seems that people are able to judge whether an action is unfair or unkind and are willing to punish such behavior. However, it is not yet understood how individuals reach their judgment. In the previous chapter, it was shown that individuals punish deviations from different cooperation norms depending on the endowment of others and on their contribution possibilities. It is clear, that individuals are using different reference points when considering what behavior deserves to be punished. Although this does not conform to some of the predictions of current theoretical models, it does fall in line with their notion of reciprocity. Specifically, that judgment of kind or unkind behavior by an individual is done independently of the behavior of others in the group.[103] In this chapter, we show that punishment depends not only on the actions of the punisher and the punished, but also on the actions of other *comparable* individuals.

Most models that assume the existence of social preferences do not take into account that behavior might be considered fair or unfair depending on the actions of others. For example, the model of Fehr and Schmidt (1999) assumes that individual *i*'s decision to punish *j* depends solely on the incomes of *i* and *j*, and not on the income of another individual *k*. Even models that incorporate intentions such as Rabin (1993), Dufwenberg and Kirchsteiger (2005), and Falk and Fischbacher (2005), assume that *i* will judge *j*'s behavior to be unkind if it negatively affects *i*, but this judgment does not depend on how *j*'s behavior compares to *k*'s. The way high-value subjects punish low-value subjects clearly contradicts these assumptions. Our results indicate that, when making their punishment decision, high-value subjects look at the difference between the contributions of low-value subjects and punish more heavily the one who contributed the least.[104]

More generally, when subjects in the privileged treatment decide on the punishment of low-value subjects, they concentrate solely on the differences between the contributions of low-value subjects, and they ignore the contribution of the high-value subject. Given that they have different incentives to contribute, comparing the contributions of high-value subjects to those of low-value subjects could be considered irrelevant and even unfair (like comparing apples and oranges). Thus, it makes sense for subjects to ignore the contributions of the high-value subject when punishing a low-value subject.

---

[102] In this sense, models that incorporate notions of fairness or reciprocity have provided useful insights into punishment behavior (e.g. Levine, 1998; Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Falk and Fischbacher, 2005).

[103] Bolton and Ockenfels (2000) is an exception. In their model, other individuals affect the average earnings of the group. Thus, in a public good setting, individual *i* will punish individual *j* more severely if individual *k* contributes more to the public good.

[104] This type of behavior has also been observed in experiments that allow social comparisons in ultimatum-bargaining games (Knez and Camerer, 1995; Duffy and Feltovich, 1999; Bohnet and Zeckhauser, 2004). In our experiment, we observe social comparisons without explicitly introducing a reference group. That is, the reference group arises endogenously.

In addition to punishment behavior, the evaluation of fairness using different reference groups can help explain the differences in contributions between low-value subjects and subjects in the baseline treatment. If low-value subjects compare their contributions only among themselves, then low-value subjects who contribute more than the other low-value subject (but less than the high-value subject) are in fact the highest contributor in their reference group. For this reason, they would not consider it necessary to further increase their contribution. In contrast, in the baseline treatment, subjects who contribute an intermediate amount are not the highest contributor in their reference group, and hence they might feel obliged to contribute more in the future. This highlights the fact that punishment alone is not enough to raise contributions to high levels. In the baseline treatment, a subject that contributes a high amount induces others to contribute more even if she does not punish. In contrast, in the privileged treatment, high contributions by high-value subjects do not produce the same response. Since punishment is the high-value individuals' only tool to increase cooperation, it is more difficult for them to induce low-value subjects to contribute more.

In conclusion, not only do individuals use different reference points to evaluate fairness (as in Chapter 5), they also evaluate fairness using different reference groups. This is consistent with happiness studies that report that the income of individuals relative to others in their reference group is a better predictor of happiness than the income of individuals relative to the whole population (van de Stadt et al., 1985; Ferrer-i-Carbonell, 2005). Our experiment shows that reference groups can easily emerge within an experiment and that they are important not only for subjective measures of wellbeing but also for reciprocal behavior.

These differences in reciprocal behavior can translate into differences in the efficiency of institutions such as decentralized punishment. If one considers that individuals care for their earnings but also for equality (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) and efficiency (Charness and Rabin, 2002), then decentralized punishment has the desirable property that it has the potential to increase all three. In the baseline treatment, punishment indeed increases the subjects' earnings, as well as equality and overall efficiency (in the long run). However, in the privileged treatment, punishment has a positive effect only on the earnings of high-value subjects. It has an ambiguous effect on the groups' overall efficiency, and a negative effect on both equality and the earnings of low-value subjects. This questions Olson's assertion that individuals in privileged groups are more likely to enjoy the benefits of a public good and are therefore better off than individuals in a homogenous group where everyone has an incentive to free ride (Olson, 1965).

Appendix 6A – Experimental Procedures and Instructions

Experimental procedures

The computerized experiment was run in October 2004 at the CREED laboratory of the University of Amsterdam. The experiment was conducted with z-Tree (Fischbacher, 1999). On average, subjects were paid out 13.76 euros. The whole experiment took around one hour.

After arrival in the lab's reception room, each subject drew a card to be randomly assigned to a seat in the laboratory. Once everyone was seated, the instructions for the experiment were read aloud (a translation of the instructions is provided below). In the privileged treatments (with and without punishment), subjects were informed in the instructions whether they would be a high-value subject or a low-value subject. Thereafter, subjects had to answer a few exercises in order to check their understanding of the game. Next, the subjects played the repeated pubic goods game with or without punishment via the computer. Once the game ended subjects answered a debriefing questionnaire after which they were paid in private and dismissed.

Instructions

The instructions for this experiment are (almost) identical to the instructions of the experiment used in Chapter 5. See Appendix 5A for details. The specific instructions for the treatments in this chapter are available upon request.

Appendix 6B – Descriptive Statistics

Table 6B.1 summarizes the average amount contributed to the public good per period in each treatment.

TABLE 6B.1 – MEAN CONTRIBUTIONS IN EACH TREATMENT

Period	Baseline No Punishment	Baseline Punishment	Privileged No Punishment	Privileged Punishment
1	10.67 (7.79)	12.83 (5.97)	11.96 (6.99)	10.44 (7.15)
2	9.05 (7.24)	13.94 (6.22)	11.67 (7.25)	12.89 (7.11)
3	5.14 (5.33)	15.94 (4.92)	9.25 (7.61)	10.67 (8.10)
4	3.10 (3.60)	15.39 (6.02)	7.92 (7.61)	10.89 (7.27)
5	3.00 (3.48)	17.28 (3.46)	8.46 (8.16)	13.17 (7.69)
6	2.52 (3.70)	17.50 (3.49)	8.46 (7.68)	14.06 (7.91)
7	2.71 (4.01)	15.83 (6.61)	6.88 (7.64)	14.11 (7.03)
8	2.67 (3.32)	16.83 (4.58)	8.25 (8.52)	11.94 (8.56)
9	2.24 (3.19)	16.83 (4.97)	7.13 (8.82)	12.28 (8.78)
10	1.00 (2.47)	14.89 (6.91)	6.79 (9.56)	13.11 (8.85)
Total	4.21 (5.52)	15.73 (5.49)	8.68 (8.05)	12.36 (7.78)

Note: Mean contribution per subject per period. Numbers between brackets are standard deviations.

Table 6B.2 breaks down the average contribution in the privileged treatments by high-value and low-value subjects.

TABLE 6B.2 – MEAN CONTRIBUTIONS OF HIGH-VALUE AND LOW-VALUE SUBJECTS

Period	Low-Value No Punishment	Low-Value Punishment	High-Value No Punishment	High-Value Punishment
1	8.63 (5.80)	6.50 (4.83)	18.63 (3.50)	18.33 (3.20)
2	8.13 (5.88)	10.08 (6.80)	18.75 (3.54)	18.50 (3.67)
3	5.63 (4.72)	8.50 (7.35)	16.50 (7.23)	15.00 (8.37)
4	4.38 (4.32)	7.83 (6.41)	15.00 (8.02)	17.00 (4.69)
5	3.75 (4.71)	9.83 (7.42)	17.88 (4.36)	19.83 (0.41)
6	4.88 (4.95)	11.17 (8.33)	15.63 (7.29)	19.83 (0.41)
7	3.56 (3.74)	11.25 (7.03)	13.50 (9.30)	19.83 (0.41)
8	4.56 (6.03)	8.33 (8.29)	15.63 (8.21)	19.17 (2.04)
9	1.31 (2.33)	10.00 (8.64)	18.75 (3.54)	16.83 (7.76)
10	0.19 (0.75)	9.67 (9.07)	20.00 (0.00)	20.00 (0.00)
Total	4.50 (5.12)	9.32 (7.36)	17.03 (6.13)	18.43 (4.23)

Note: Mean contribution per subject per period. Data corresponds to the privileged treatments only. Numbers between brackets are standard deviations.

Table 6B.3 shows the average amount of punishment points given and received in each period depending on the treatment and type.

TABLE 6B.3 – PUNISHMENT GIVEN AND RECEIVED

Period	Baseline	Punishment Given by Low-Value Subjects	Punishment Received by Low-Value Subjects	Punishment Given by High-Value Subjects	Punishment Received by Low-Value Subjects
1	2.00 (2.50)	1.17 (1.85)	2.17 (1.34)	3.67 (2.73)	1.67 (2.42)
2	1.89 (2.52)	0.92 (1.24)	1.42 (2.27)	2.00 (2.45)	1.00 (1.67)
3	1.06 (1.55)	1.08 (1.51)	3.08 (3.73)	4.83 (6.01)	0.83 (1.17)
4	0.78 (1.11)	1.25 (2.01)	1.67 (3.03)	2.67 (3.93)	1.83 (2.48)
5	0.61 (1.29)	0.83 (1.47)	0.67 (0.78)	0.67 (1.03)	1.00 (2.00)
6	0.61 (1.04)	1.25 (2.34)	0.67 (0.98)	0.33 (0.82)	1.50 (3.21)
7	1.61 (3.16)	0.67 (1.61)	0.42 (0.90)	0.33 (0.82)	0.83 (2.04)
8	0.94 (1.98)	0.75 (1.48)	1.00 (1.60)	0.50 (0.84)	0.00 (0.00)
9	0.61 (0.98)	0.25 (0.45)	0.33 (0.65)	0.17 (0.41)	0.00 (0.00)
10	0.67 (2.35)	2.00 (5.78)	2.00 (3.07)	1.67 (1.97)	1.67 (4.08)
Total	1.08 (2.00)	1.02 (2.34)	1.34 (2.21)	1.68 (2.91)	1.03 (2.18)

Note: Mean amount of punishment points given and received per subject per period. Note that in the baseline treatment, punishment given and received are equal. Numbers between brackets are standard deviations.

Appendix 6C – Regressions

In the following regressions, the dependent variable is p_{ij}, the amount of punishment points subject i (the punisher) assigns to subject j (the punished) in the privileged treatment. In all regressions, there are 120 observations.[105] We use Tobit estimates censured both at zero and at ten.

TABLE 6C.1 – REGRESSIONS WITH THE AMOUNT OF PUNISHMENT AS THE DEPENDENT VARIABLE

Independent Variable	Low-value to low-value	Low-value to high-value	High-value to low-value
Contribution of the punisher	0.284**	0.197	–0.013
	(0.104)	(0.155)	(0.089)
Contribution of the punished	–0.212**	–0.111	–0.179**
	(0.070)	(0.144)	(0.085)
Contribution of the other group member	–0.013	0.263*	0.190**
	(0.086)	(0.136)	(0.091)
Punishment received in the previous period	0.039	0.278	–0.809**
	(0.230)	(0.377)	(0.345)
Period	0.035	–0.909**	–0.541**
	(0.174)	(0.249)	(0.160)
Constant	–3.869	–4.754	1.887
	(2.507)	(3.565)	(1.761)
χ^2	13.54**	28.77**	14.78**

Note: Numbers between brackets are robust standard errors. ** Significant at the 5 percent level. * Significant at the 10 percent level.

[105] In the regressions for 'low-value to low-value' and 'low-value to high-value', each low-value subject has one data point per round. In the regression for 'high-value to low-value' each high-value subject has two data points per round.

Chapter 7

Conclusions

In this chapter, we review some of this thesis' results. Given the variety of the studied games and behavior, we do not discuss all of our findings. Instead, it is more instructive to concentrate on a topic present throughout the thesis. Specifically, we address the question, which set of ingredients provides a better account of the enforcement, through punishment, of fairness norms. We highlight the contributions made by each chapter to this issue and make a few suggestions for future research. In Section 7.1, we describe the motivations behind enforcing and conforming to fairness norms. Section 7.2 highlights the findings related to the evaluation of fairness. Lastly, a brief discussion on inter-personal comparisons and the heterogeneity of fairness norms is presented in Section 7.3.

7.1 Motivations Behind Social Punishment

Throughout the thesis, we study different aspects of norm enforcement. In all the studied games, some subjects have the opportunity to increase their earnings at the expense of others. Subsequently, subjects get the chance to spend some of their money in order to punish other subjects. In line with comparable studies, we find that subjects are willing to spend considerable amounts of money punishing others even though they do not benefit from doing so (Camerer, 2003). Furthermore, in some, but not all cases, subjects stop behaving opportunistically after being punished. Analyzing the subjects' emotional response suggests that anger-like emotions motivate the punishment of others whereas shame-like emotions motivate individuals to act less selfishly.

In both Chapter 2 and Chapter 4, we show that individuals who punish opportunistic behavior are those who report high intensities of anger-like emotions. This confirms the results of previous studies that find a strong correlation between self-reported anger and the destruction of income in power-to-take games (Bosman and van Winden, 2002; Bosman et al., 2005b). Similar results have been found with physiological and neurological measures of anger (Sanfey et al., 2003; Ben-Shakhar et al., 2004; Quervain et al., 2004).

In relation to the causes of anger, we find that a subject *i* feels angry when the choices of another subject *j* reduces subject *i*'s earnings. This is true when *j*'s choice is an opportunistic act, but also when *j* is punishing *i* for behaving selfishly (see Chapter 4). In addition, we find that the intensity of anger is also affected by two other factors. One is, the unexpectedness of the other's choice. Opportunistic behavior generates more anger when it is unexpected than when it is expected. The second factor concerns the subject's perceived

fairness norm. Subjects who think it is very unfair to act opportunistically report higher intensities of anger when confronted with such behavior.

It is important to point out that, although anger is triggered by unfair behavior, the goal of angry subjects is to harm the other party, and not, through punishment, to correct unfair material outcomes. That is why subjects punish even when it is impossible to reduce income inequalities. For example, in Chapter 2, 12.8% of the subjects punish in situations in which punishment does not decrease income differences. Anger-motivated punishment also explains why some subjects punish more than the amount needed to equalize earnings. For instance, in Chapter 4 we find that around one in three subjects punish others to the point where the punished subject has lower earnings than the punisher.[106]

In Chapter 3 and Chapter 4 we analyze the motivations behind the behavior of subjects who are (or are not) punished. We show that punishment does not always generate prosocial behavior. In some cases, it even provokes antisocial behavior such as retaliation. Further analysis indicates that punishment induces subjects to act fairly in the future only if the punished subject feels shameful. It is also the case that feeling shame prevents punished subjects from retaliating against the punishers. Specifically, after being punished, some subjects feel angry but if, in addition to anger, they also feel shameful, they do not retaliate. Given that retaliation can considerably increase the cost of enforcing fair behavior (see Chapter 4), the emotion of shame is essential for the effectiveness of a punishment institution.[107]

It is less clear what causes feelings of shame. In Chapter 3, we find that subjects feel high intensities of shame when they are punished. This is especially true for subjects who acknowledge that they behaved unfairly. In Chapter 4, we find again that punishment leads to high intensities of shame. However, in this case, we find that this is only true when subjects are punished substantially. Receiving a small amount of punishment actually produces lower intensities of shame than no punishment at all.[108] The finding that punishment induces high intensities of shame is unsurprising since the emotion of shame is strongly associated with the perceived disapproval of others (Lazarus, 1991; Tangney and Dearing, 2002). It also suggests that, in order for people to act fairly, punishment must be available. The existence of a fairness norm that is not enforced might not trigger intensities of shame that are high enough to restrain opportunistic behavior. This helps explains why a lower proportion of proposers

[106] This is also true in the public good games with punishment analyzed in Chapter 5 and Chapter 6. There we find that subjects in advantageous positions (higher earnings) routinely punish subjects in disadvantageous positions.

[107] This supports the hypothesis that emotions such as shame, coevolved with punishment institutions and anger-like emotions in order to limit antisocial behavior in groups (Bowles and Gintis, 2001).

[108] In Chapter 4, unlike in Chapter 3, we do not find that shame is affected by fairness perceptions. We think this might be due to the different games and the way fairness perceptions are measured. In Chapter 3, the subjects choice and fairness perception are on the same scale (i.e. both are a take rate) and hence it is easy to calculate the difference between the two.

choose the equal split in the dictator game compared to the ultimatum game. Since, in dictator games responders cannot punish, proposers are less exposed to feeling high intensities of shame (see Chapter 3).[109] It also explains why symbolic punishment can increase cooperation (by triggering feelings of shame), but its effect deteriorates over time (Masclet et al., 2003; Noussair and Tucker, 2005). The lack of real consequences for punished free riders probably triggers lower intensities of shame than real punishment does.

Knowing that emotions such as anger and shame motivate behavior in punishment institutions can help explain many of the observed experimental results and can help direct future research. For instance, the fact that anger is elicited by intentional acts (Haidt, 2003) explains why intentions have an important effect on punishment behavior (Falk et al., 2000; Charness and Levine, 2004). Furthermore, the fact that people feel more shame in situations where others can clearly observe their actions (Tangney and Dearing, 2002), explains the effects of uncertainty in ultimatum games. Uncertainty over the size of the pie prevents responders from clearly judging whether proposers are acting fairnly or unfairly. If this makes proposers feel less shame, we should expect them to take a bigger slice of the pie (Camerer and Loewenstein, 1993; Straub and Murnighan, 1995; Rapoport et al., 1996a; Schmitt, 2004). Future research could explore the effects of various known characteristics of anger-like and shame-like emotions on the enforcement of fairness norms. For example, Jakobs et al. (1999) find that people tend to feel more anger in public than in private. This suggests that more people will wish to punish unfair behavior if they are playing in a group as opposed to individually (as in Bosman et al., 2005a). In addition, research on apologies suggests that displaying feelings of shame or guilt can satisfy an injured party and might lead to reduced punishment (Ohbuchi et al., 1989). This is important if it leads to a still effective but less wasteful enforcement of fair behavior.

Another question that should be answered in future research is whether individuals possess stable propensities to enforce and to comply with fairness norms, and if so, whether they are correlated. Our research shows that anger motivates punishment whereas shame motivates individuals to comply with fairness norms. Given that a large number of studies have found evidence of stable personality traits (for an overview see Carstensen et al., 2003), we should expect some subjects to have a propensity to enforce or to comply with fairness norms. Indeed, as we report in Chapter 3, in the repeated power-to-take game, responders who punish in the first period are more likely to punish in the second period. Similarly, proposers who act fairly in the first period are more likely to act fairly in the second period.

There is less reason to think that subjects that are more susceptible to feel anger are also more susceptible to feel shame. Evidence from public good games with punishment suggests that individuals who cooperate also punish those who do not cooperate (Fehr and Gächter, 2000b). However, in Chapter 3 we find no evidence indicating that subjects who

[109] The emotion of shame also explains why proposers in dictator games take more as the level of anonymity increases (e.g. in double-blind experiments, Hoffman et al., 1996).

punish (as responders) are more likely to act fairly (as proposers). Future research will be needed to explore this issue in more detail.

Lastly, we address an important question concerning punishment and the enforcement of cooperative behavior, namely, whether subjects free ride on the punishment of others. As argued by Oliver (1980), if individuals free ride on the punishment of uncooperative behavior, punishment becomes a second-order public good and therefore it cannot sustain cooperative behavior. However, this is contradicted by the evidence reported in the previous two chapters as well as numerous other studies (see footnote 85).

The results presented in Chapter 2 can help us determine whether punishment is indeed a second-order public good. In the three-person power-to-take game studied in that chapter, we find that punishment of unfair behavior is more like a coordination problem rather than a social dilemma. Specifically, we find that responders, who do not free ride on the punishment of others, feel more positively valued emotions and less negatively valued emotions than responders who do free ride.[110] Thus, judging by the responders' emotional response, it appears that subjects do not enjoy free riding on punishment. However, this does not mean that they are unaffected by the actions of others. We find that responders dislike punishing if the other responder does not do the same. Hence, responders face a problem akin to a coordination game in which they have to decide to coordinate on punishment or on no-punishment. In Chapter 2, we also show that, to a large extent, the type of social tie between responders determines how easy it is for subjects to coordinate on punishment.

7.2 Evaluating Fairness

Even though there is a lot of evidence indicating that people are willing to punish unfair behavior, we still do not have a clear understanding of how individuals define what is fair. Observing the punishment behavior of individuals in the public good games presented in Chapter 5 and Chapter 6, can shed some light on this issue. The introduction of heterogeneity into the public good game provides a rich environment in which to analyze punishment behavior and test the assumptions of various theoretical models. We find that, although most models can explain some of the punishment patterns, no model captures the subjects' punishment behavior in all treatments and roles.

Models that motivate punishment using income differences are able to explain why high-contributors punish low-contributors in homogenous groups. Furthermore, they also account for punishment by poor subjects in groups with heterogeneous endowments (Chapter 5). In particular, they explain why poor subjects who contribute, punish poor subjects who do not contribute, and why poor subjects punish rich subjects more heavily than they punish other poor subjects. However, these models fail to explain the punishment behavior of subjects who are in an advantageous position (either rich subjects in Chapter 5 or high-value

[110] In the case of friends, this difference is significant but not so in the case of strangers. By free riding, we refer to responders who do not punish and are paired with a responder who does.

subjects in Chapter 6). In these models, individuals with high earnings do not punish individuals with low earnings. However, this is not what we find.

Models that use a notion of fairness that incorporates intentions can also explain why high-contributors punish low-contributors in homogenous groups. In addition, they also provide subjects in advantageous positions with a motivation to punish other subjects. However, these models do not explain the difference between the punishment behavior in the baseline treatment and the restricted treatment in Chapter 5. Specifically, they do not explain why poor subjects punish rich subjects more than they punish other poor subjects.

In Chapter 5, we show that depending on the contribution possibilities of rich subjects, groups tend to enforce either a cooperation norm in which all (rich and poor) subjects contribute the same amount or a cooperation norm in which rich subjects contribute twice as much as poor subjects. As suggested by Bolton and Ockenfels (2005), these differences can be explained by an inequity-aversion model that incorporates a reference point that can shift in order to make the relevant income comparisons. Such a model would improve the predictive power of these models and at the same time be simple enough to derive predictions in complex games. Accurately understanding how individuals set this reference point when making fairness evaluations is certainly a fruitful line of research. Ariely et al. (2003) have shown that individuals can use arbitrary reference points to evaluate how much they value a certain good. Moreover, once this reference point is set, they are able to coherently compare the value of the good to the value of other goods. This gives the impression that individuals possess stable preferences when in reality they are susceptible to reference-point shifts (e.g. induced by framing). It would be of great value to know if social preferences are affected in a similar way. If so, cues used by individuals to define fairness can have significant effects on the behavior that is eventually enforced.

Also important when evaluating fairness is the way the actions of an individual compares to the actions of others. Chapter 6 provides evidence that subjects make inter-personal comparisons when making their punishment decision. Moreover, it shows that subjects make comparisons only among similar individuals. Specifically, when subjects in the privileged treatment decide on the punishment of low-value subjects, they compare contributions among low-value subjects and ignore the contribution of the high-value subject. In addition to punishment behavior, the evaluation of fairness using different reference groups explains the differences in contributions between low-value subjects and subjects in the baseline treatment. Low-value subjects who contribute an intermediate amount do not increase their contribution even though there is another subject who contributed more that they did (i.e. the high-value subject). In contrast, subjects in the baseline treatment who contribute an intermediate amount raise their contributions if there is another subject who contributed more. This shows that, differences in reference groups can have significant effects on behavior. More research is needed to understand in which cases subjects ignore the presence of others (as in Chapter 6 and Güth and van Damme, 1998) and in which cases they take them into account (as in Chapter 5).

As is discussed in Section 7.3, when evaluating whether someone acted fairly or not, individuals look at how their actions compare to the actions of others. Thus, it is important to understand how these comparisons are made. The choice of the appropriate reference point to make income comparisons or the selection of a reference group of comparable individuals can have important effects on the welfare of those involved. In Chapter 5, the different cooperation norms enforced in the unequal treatments determine whether participants in the game increased their earnings or not (subjects benefit from the game only in the unrestricted treatment). Similarly, in Chapter 6, the different reference groups used by subjects in the baseline and the privileged treatments make the punishment institution more profitable in homogeneous groups.

7.3 Fairness as a Social Norm

Although social norms are often taught as behavior you should conform to, irrespective of what others do, in most instances, the behavior of those around you will determine how rigorously you enforce and follow the social norm. It is not the same to accept a bribe in a country with little corruption as in a country where bribes are commonplace. Theories that model fairness norms do not generally consider this. Perhaps this is due to the modeling of norms as social preferences. Most of the work on social preferences assumes that when individual *i* evaluates whether individual *j* has been fair our unfair, *i* will take into account only the actions or income of *j*, and will ignore those of other individuals (Rabin, 1993; Fehr and Schmidt, 1999; Charness and Rabin, 2002; Dufwenberg and Kirchsteiger, 2005; Falk and Fischbacher, 2005). However, we must not forget that, as a social norm, what is considered fair is bound to be influenced by the behavior of others.

Results in this thesis as well as in other experiments indicate that fairness evaluations are affected by the behavior of third parties (Knez and Camerer, 1995; Duffy and Feltovich, 1999; Bohnet and Zeckhauser, 2004). In the experiment of Chapter 6, it is clear that, when making their punishment decision, high-value subjects compare the contributions of the two low-value subjects, and then they punish more heavily the one who contributed the least. Given that high-value subjects disregard their own contribution, this demonstrates that fairness evaluations can be based solely on comparisons with third parties. In Chapter 2, the subjects' emotional reaction reveals that responders are indeed affected by the behavior of other responders, even when their earnings are not affected by their action. In this case, the different emotional reactions between pairs of friends and pairs of strangers translate into more punishment by pairs of friends.

In Chapter 3 we find that fairness perceptions have an effect on the emotions experienced by subjects and therefore on how they react to punishment. However, it is important to point out that the perception of what is fair varies substantially among subjects. This illustrates an added complication to studying behavior that is affected by fairness norms. So far, economists have attributed differences in punishment (or rewarding) behavior to differences in social preferences (controlling for expectations). In other words, differences in

how much individuals value money relative to enforcing or complying with a fairness norm. However, this interpretation can lead us to wrong conclusions. Suppose that, in the power-to-take game from Chapter 3, we observe a proposer who chooses a take rate of 75%. Given that this is quite high, we might conclude that this proposer is interested in maximizing her earnings. Thus, we would predict that she would adjust her take rate depending on the responder's willingness to destroy. However, it is possible that the proposer does care for fairness, but she believes that proposers are entitled to a very large share of the pie. Thus, when confronted by responders who are willing to destroy, she does not adjust her decision. If we are to correctly identify types of individuals (i.e. people with different social preferences) that will behave consistently across different situations, we will have to control for their fairness perceptions.

Finally, since social norms are based on a mutual understanding among group members, we should expect that, when individuals are confronted with a new situation (like an experiment), they will adjust their belief of what is the right way to behave. This means that as individuals interact they might change their fairness perception. Accordingly, what turns out to be fair in the long run can vary considerably depending on the experiences of those involved in the process. It also means that, for some time, individuals can disagree on what is and what is not fair. Generally, this will provoke more conflict and will lead to welfare losses (e.g. Knez and Camerer, 1995). Hence, different types of institutions or policies might be necessary to regulate behavior in different groups depending on how much they agree on what fair behavior is. Measuring the heterogeneity of fairness perceptions might help us implement the right policies in situations that on the surface appear to be similar.

The enforcement of fairness norms is clearly a complex phenomenon. Much more research is needed to address all the unanswered questions. Nevertheless, this thesis shows that studying this type of behavior with the use of a variety of tools (such as measures of emotions, expectations, and fairness perceptions) can increase our understanding of otherwise puzzling results.

Samenvatting in het Nederlands

Summary in Dutch [*]

Het bestaan van normen voor rechtvaardigheid is een belangrijk onderdeel van onze samenleving. Vanaf de geboorte wordt ons het belang van eerlijk handelen bijgebracht. Rechtvaardig gedrag wordt geprezen en onrechtvaardig gedrag wordt afgekeurd. Ondanks het onomstreden belang van rechtvaardigheidsnormen begrijpen we er nog weinig van. Nog steeds weten we niet hoe zulke normen ontstaan, hoe hun inhoud wordt vastgesteld en in welke situaties ze een rol spelen. In dit proefschrift wordt het opleggen van normen voor rechtvaardigheid bestudeerd, om een antwoord te kunnen geven op bovenstaande vragen.

In deze samenvatting gaan we in op een aantal van de belangrijkste resultaten van het proefschrift. Vanwege het uiteenlopende karakter van de situtaties die zijn bestudeerd worden niet alle bevindingen gepresenteerd. We concentreren ons op een onderwerp dat door het proefschrift heen een rol speelt: welke combinatie van ingrediënten kan een betere verklaring verschaffen voor het opleggen van rechtvaardigheidsnormen door middel van straf. Eerst behandelen we de beweegredenen voor het opleggen en het conformeren aan normen. Daarna presenteren we de bevindingen die verband houden met de beoordeling van rechtvaardigheid, en ten slotte presenteren we een beschouwing over interpersoonlijke vergelijkingen en de heterogeniteit van normen voor rechtvaardigheid.

Motivaties voor sociaal straffen

In dit proefschrift bestuderen we verschillende aspecten van het opleggen van normen. In alle experimentele spellen die we beschouwen, krijgen sommige proefpersonen de mogelijkheid om hun eigen verdiensten te verhogen ten koste van die van anderen. Vervolgens krijgen de proefpersonen de kans om hun eigen geld uit te geven om andere proefpersonen te straffen. We zien dat proefpersonen bereid zijn om aanzienlijke hoeveelheden geld uit te geven aan het straffen van anderen, zelfs als ze daar niet zelf (geldelijk) van profiteren. Daarnaast zien we dat in sommige gevallen de proefpersonen hun opportunistische gedrag staken nadat ze werden gestraft. Uit de analyse van de emotionele respons van de proefperonen lijkt het dat woede-gerelateerde emoties het straffen van anderen motiveren, terwijl schaamte-gerelateerde emoties de individuen motiveren om minder zelfzuchtig te handelen.

In hoofdstuk 2 en 4 laten we zien dat individuen die opportunistisch gedrag bestraffen, diegene zijn die een hoge intensiteit van woedegerelateerde emoties rapporteren. In verband met de oorzaken van woede vinden we dat een proefpersoon zich boos voelt als de keuzes

[*] I would like to thank Eva van de Broek for this translation.

van iemand anders zijn verdiensten verminderen. Daarnaast vinden we dat de intensiteit van de woede door twee andere factoren wordt bepaald. Een daarvan is de mate van onverwachtheid van de keuze van de ander. Opportunistisch gedrag brengt meer woede teweeg als het onverwacht is dan wanneer het verwacht is. De tweede factor heeft met de proefpersoons opvatting van rechtvaardigheid te maken. Proefpersonen die denken dat het onrechtvaardig is om opportunistisch te handelen, rapporteren hevigere woede als ze met zulk gedrag worden geconfronteerd.

In hoofdstuk 3 en 4 analyseren we de drijfveren achter het gedrag van de proefpersonnen die wel of niet worden gestraft. We laten zien dat straffen niet altijd 'prosociaal' gedrag oproept. In sommige gevallen roept het zelfs antisociaal gedrag op zoals vergelding. Verdere analyse wijst erop dat straffen alleen dan proefpersonen aanzet om rechtvaardig te handelen in de toekomst, als de gestrafte schaamte ervaart. Daarnaast blijkt dat schaamte de gestraften ervan weerhoudt om wraak te nemen op de strafgevers. Oftewel, mensen ervaren woede als ze gestraft zijn, maar als ze daarnaast ook schaamte ervaren zullen ze geen wraak nemen. Gezien het feit dat vergelding de kosten van het opleggen van rechtvaardig gedrag aanzienlijk kan verhogen, is de emotie schaamte essentieel voor de effectiviteit van een straf-institutie.

Ten slotte gaan we kort in op een belangrijke kwestie die met straf en het afdwingen van samenwerking te maken heeft, namelijk of proefpersonen meeliften op het strafgedrag van anderen. De resultaten uit hoofdstuk 2 kunnen ons helpen te bepalen in welke gevallen straf inderdaad een tweede-orde public good is. Bij het driepersoons power-to-take-spel dat in dat hoofdstuk wordt bestudeerd, vinden we dat proefpersonen die niet meeliften op het strafgedrag van anderen, meer positieve emoties en minder negatieve emoties ervaren dan personen die wel meeliften. Afgaande op de emotionele reactie van de personen lijkt het dat proefpersonen meeliften niet aangenaam vinden. We vinden dat personen het uitdelen van straf onaangenaam vinden als de andere persoon niet hetzelfde doet. Blijkbaar stuiten personen op een probleem dat op het coordinatiespel lijkt, waarin men moet besluiten om te coördineren om te straffen of niet.

De beoordeling van rechtvaardigheid

Hoewel er veel aanwijzingen zijn dat mensen bereid zijn om onrechtvaardig gedrag te straffen, hebben we nog steeds geen duidelijk begrip van hoe individuen vaststellen wat rechtvaardig is. Door het strafgedrag van individuen in de public good-spellen in hoofdstuk 5 en 6 te observeren, kunnen we dit vraagstuk nader bekijken. Het introduceren van heterogeniteit in deze spellen verschaft ons een rijke omgeving om strafgedrag te analyseren.

In hoofdstuk 5 laten we zien dat afhankelijk van verschillende factoren, groepen verschillende normen voor samenwerken zullen afdwingen. Deze verschillen kunnen worden geïnterpreteerd als een opschuiving in het referentiepunt voor rechtvaardigheid. Daarnaast is het belangrijk voor de beoordeling van rechtvaardigheid hoe handelingen tussen verschillende personen te vergelijken zijn. Hoofstuk 6 levert bewijs dat proefpersonen interpersoonlijke

vergelijkingen maken voor de beslissing om iemand te straffen. We laten zien dat proefpersonen alleen vergelijkingen maken tussen vergelijkbare individuen. De keuze van een geschikt referentiepunt om inkomenseffecten te vergelijken, of de selectie van een referentiegroep bestaande uit vergelijkbare individuen, kunnen een belangrijk effect hebben op de welvaart. In hoofdstuk 5 en 6 bepalen die verschillende normen voor samenwerking of referentiegroepen of de spelers hun inkomsten vergroten of niet.

Interpersoonlijke vergelijkingen en de perceptie van rechtvaardigheid

Hoewel sociale normen vaak worden onderwezen als gedragingen waaraan je je onafhankelijk van want anderen doen moet conformeren, zullen in de meeste gevallen juist die anderen bepalen hoe rigoreus je je zult conformeren en de norm zal opleggen aan anderen.
De resultaten uit dit proefschrift wijzen erop dat ervaringen worden beïnvloed door het gedrag van derden. In het experiment in hoofdstuk 6 wordt duidelijk dat sommige proefpersonen rechtvaardigheid alleen baseren op vergelijkingen met derden. In hoofstuk 2 zagen we dat de emotionele reacties van de proefpersonen onthulden dat ze inderdaad worden aangedaan door het gedrag van anderen, zelfs als hun verdiensten niet te lijden hadden onder de daden van de ander. In dit geval resulteerden de verschillen tussen de emotionele reacties onder vreemden en onder vrienden in meer straf bij vriendenparen.

In hoofdstuk 3 vinden we dat percepties van rechtvaardigheid een effect hebben op de ervaren emoties en dus hoe personen reageren op straf. Hier is het belangrijk te benadrukken dat de perceptie van wat rechtvaardig is aanzienlijk varieert onder proefpersonen. Dit illustreert een tweede complicatie bij het bestuderen van gedrag dat beïnvloed wordt door rechtvaardigheidsnormen. Tot dusverre hebben economen verschillen in strafgedrag (of beloning) toegeschreven aan verschillen in sociale voorkeur. Met andere woorden, verschillen in hoezeer een persoon geld waardeert in vergelijking met het opleggen of conformeren aan een rechtvaardigheidsnorm. Maar deze interpretatie kan ons aanzetten tot verkeerde conclusies. Veronderstel dat we een persoon geld zien afnemen van een ander. Daaruit zouden we kunnen conluderen dat hij geïnteresseerd is in het maximaliseren van zijn eigen verdiensten, en niet wordt beïnvloed door rechtvaardigheidsoverwegingen. Toch is het mogelijk dat hij wel degelijk om rechtvaardigheid geeft, maar gelooft dat individuen in zijn positie gerechtvaardigd zijn om geld weg te nemen.

Omdat sociale normen gebaseerd zijn op overeenstemming tussen groepsgenoten zouden we verwachten dat in nieuwe situaties (zoals een experiment) individuen hun overtuiging van wat de juiste manier is om je te gedragen zouden aanpassen. Dit betekent dat tijdens de interactie perceptie van rechtvaardigheid van de proefpersonen kan veranderen. Daaruit volgt dat wat uiteindelijk rechtvaardig wordt gevonden behoorlijk kan variëren, afhankelijk van de ervaringen van de spelers. Het betekent ook dat gedurende een periode individuen van mening kunnen verschillen over wat rechtvaardig is en wat niet. Dit roept meer conflicten op en kan tot welvaartsverlies lijden.

Het is duidelijk dat het opleggen van rechtvaardigheidsnormen een complex fenomeen is. Meer onderzoek is nodig om op alle nog onbeantwoorde vragen in te gaan. Dit proefschrift laat zien dat het bestuderen van dit type gedrag aan de hand van verschillende hulpmiddelen (zoals het meten van emoties, verwachtingen en rechtvaardigheidspercepties) ons begrip kan vergroten van deze anders zo raadselachtige vragen.

Bibliography

Abbink, K., B. Irlenbusch, and E. Renner (2000). The moonlighting game: An experimental study on reciprocity and retribution. *Journal of Economic Behavior and Organization* 42: 265-277.

Abbink, K., B. Irlenbusch, and E. Renner (2002). Group size and social ties in microfinance institutions. Discussion Paper. University of Erfurt.

Akerlof, G.A. (1982). Labor contracts as partial gift-exchange. *The Quarterly Journal of Economics* 97: 543-569.

Akerlof, G.A. (1997). Social distance and social decisions. *Econometrica* 65: 1005-1027.

Alm, J., I. Sanchez, and A. de Juan (1995). Economic and noneconomic factors in tax compliance. *KYKLOS* 48: 3-18.

Anderson, L.R., J.M. Mellor, and J. Milyo (2004). Inequality and public good provision: An experimental analysis. Working Paper. College of William and Mary.

Anderson, S., A. Bechara, H. Damasio, D. Tranel, and A. R. Damasio (1999). Impairment of social and moral behavior related to early damage in human prefrontal cortex. *Nature Neuroscience* 2: 1032-1037.

Andreoni, J., B. Erard, and J. Feinstein (1998). Tax compliance. *Economic Journal* 36: 818-860.

Ariely, D., G. Loewenstein, and D. Prelec (2003). Coherent arbitrariness: Stable demand curves without stable preferences. *The Quarterly Journal of Economics* 118: 73-106.

Bagnoli, M. and M. McKee (1991). Voluntary contribution games: Efficient private provision of public goods. *Economic Inquiry* 29: 351-366.

Barr, A. (2001). Social Dilemmas and Shame-based Sanctions: Experimental results from rural Zimbabwe. Working paper. University of Oxford.

Baumeister, R. F., A. M. Stillwell, and T. F. Heatherton (1994). Guilt: An interpersonal approach. *Psychological Bulletin* 115: 243-267.

Ben-Shakhar, G., G. Bornstein, A. Hopfensitz, and F. van Winden (2004). Reciprocity and emotions: Arousal, self-reports, and expectations. Working paper. University of Amsterdam.

Bereby-Meyer, Y. and M. Niederle (2005). Fairness in bargaining. *Journal of Economic Behavior and Organization* 56: 173-186.

Berg, J., J. Dickhaut, and K. McCabe (1995). Trust, reciprocity, and social history. *Games and Economic Behavior* 10: 122-142.

Blount, S. (1995). When social outcomes aren't fair: The effect of casual attributions on preferences. *Organizational Behavior and Human Decision Processes* 63: 131-144.

Bochet O., T. Page, and L. Putterman (2005). Communication and Punishment in Voluntary Contribution Experiments. *Journal of Economic Behavior and Organization* forthcoming.

Bohnet, I. and R. Zeckhauser (2004). Social comparisons in ultimatum bargaining. *Scandinavian Journal of Economics* 106: 495-510.

Bolton, G. and A. Ockenfels (2000). A theory of equity, reciprocity, and competition. *American Economic Review* 90: 166-193.

Bolton, G. and A. Ockenfels (2005). A stress test of fairness measures in models of social utility. *Economic Theory* 25: 957-982.

Borjas, G.J. (1995). Ethnicity, neighborhoods, and human capital externalities. *American Economic Review* 85: 365-389.

Bosman, R., and F. van Winden (2002). Emotional hazard in a power-to-take game experiment. *The Economic Journal* 112: 147-169.

Bosman, R., H. Hennig-Schmidt, and F. van Winden (2005a). Exploring group decision making in a power-to-take experiment. *Experimental Economics* forthcoming.

Bosman, R., M. Sutter, and F. van Winden (2005b). On the impact of real effort and emotions in power-to-take experiments. *Journal of Economic Psychology* 26: 407-429.

Bower, G. (1992). How might emotions affect learning? In Christianson, S.A. (Ed.) *Handbook of Emotion and Memory: Research and Theory*. Hillsdale: Erlbaum.

Bowles, S. and H. Gintis (2001). The economics of shame and punishment. Working paper.

Boyd, R. and P.J. Richerson (1992). Punishment allows the evolution of cooperation (or anything else) in sizable groups. *Ethology and Sociobiology* 13: 171-195.

Boyd, R. and P.J. Richerson (2005). Solving the Puzzle of Human Cooperation. In Levinson S.C. and P. Jaisson (Eds.) *Evolution and Culture*. Cambridge: MIT Press.

Brookshire, D., D. Coursey, and D. Redington (1993). Special interests and the voluntary provision of public goods. Discussion Paper. University of New Mexico.

Camerer, C. (2003). *Behavioral Game Theory*. New Jersey: Princeton University Press.

Camerer, C. and G. Loewenstein (1993). Information, fairness, and efficiency in bargaining. In Mellers B. and J. Baron (Eds.) *Psychological Perspectives on Justice: Theory and Applications*. Cambridge: Cambridge University Press.

Carpenter, J. P. (2004). The Demand for Punishment. Working paper. Middlebury College.

Carpenter, J. P. and P. Matthews (2005). Norm Enforcement: Anger, Indignation or Reciprocity. Working paper. Middlebury College.

Carstensen, L.L., S.T. Charles, D.M. Isaacowitz, and Q. Kennedy (2003). Emotion and life-span personality development. In Davidson, R. J., K. R. Scherer, and H. H. Goldsmith (Eds.) *Handbook of Affective Sciences*. Oxford: Oxford University Press.

Chan, K.S., S. Mestelman, R. Moir, and R.A. Muller (1996). The voluntary provision of public goods under varying income distributions. *The Canadian Journal of Economics* 29: 54-59.

Chan, K.S., S. Mestelman, R. Moir, and R.A. Muller (1999). Heterogeneity and the voluntary provision of public goods. *Experimental Economics* 2: 5-30.

Chan, K.S., S. Mestelman, R. Moir, and R.A. Muller (2003). Heterogeneity, communication, information and voluntary contributions towards the provision of a public good. Working paper. McMaster University.

Charness, G. and D. Levine (2004). The road to hell: An experimental study of intentions. Working paper No. 11. Center for Responsible Business, University of California.

Charness, G. and M. Rabin (2002). Understanding social preferences with simple tests. *The Quarterly Journal of Economics* 117: 817-869.

Charness, G., E. Haruvy, and D. Sonsino (2003). Social distance and reciprocity: The Internet vs. the laboratory. Working paper. University of California, Santa Barbara.

Cherry, T.L., S. Kroll, and J.F. Shogren (2005). The impact of endowment heterogeneity and origin on public good contributions: Evidence from the lab. *Journal of Economic Behavior and Organization* 57: 357-365.

Chong, D. (1991). *Collective Action and the Civil Rights Movement*. Chicago: University of Chicago Press.

Cinyabuguma, M., T. Page, and L. Putterman (2004). On perverse and second-order punishment in public goods experiments with decentralized sanctioning. Working paper. Brown University.

Clore, G.L. and M.D. Robinson (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin* 128: 934-960.

Coricelli, G. (2002). Sequence matters: An experimental study of the effects of experiencing positive and negative reciprocity. Working paper. University of Siena.

Cox, J.C. (2004). How to identify trust and reciprocity. *Games and Economic Behavior* 46: 260-281.

Cuzick, J. (1985). A Wilcoxon-type test for trend. *Statistics in Medicine* 4: 87-89.

Damasio, A. (1994). *Descartes' Error - Emotion, Reason and the Human Brain*. Harper Collins.

Dawes, R.M. (1980). Social dilemmas. *Annual Review of Psycology* 31: 169-193.

Duffy, J. and Feltovich, N. (1999). Does observation of others affect learning in strategic environments? An experimental study. *International Journal of Game Theory* 28: 131-152.

Dufwenberg, M. and U. Gneezy (2000). Measuring beliefs in an experimental lost wallet game. *Games and Economic Behavior* 30: 163-182.

Dufwenberg, M. and G. Kirchsteiger (2005). A theory of sequential reciprocity. *Games and Economic Behavior* 47: 268-298.

Egas, M. and A. Riedl (2005). The economics of altruistic punishment and the demise of cooperation. Working paper. University of Amsterdam.

Ellickson, R. (1994). *Order Without Law - How Neighbors Settle Disputes*. Cambridge: Harvard University Press.

Elster, J. (1989). *The Cement of Society - A Study of Social Order*. Cambridge: Cambridge University Press.

Elster, J. (1999). *Strong Feelings: Emotion, Addiction and Human Behavior*. MIT Press.

Elster, J. (1998). Emotions and economic theory. *Journal of Economic Literature* 36: 47-74.

Engelmann, D. and M. Strobel (2004). Inequality aversion, efficiency and maximim preferences in simple distribution experiments. *American Economic Review* 94: 857-869.

Erard, B. and J.S. Feinstein (1994). Honesty and evasion in the tax compliance game. *RAND Journal of Economics* 25: 1-19.

Falk, A. and U. Fischbacher (2005). A theory of reciprocity. *Games and Economic Behavior* forthcomming.

Falk, A., E. Fehr, and U. Fischbacher (2000). Testing theories of fairness: Intentions matter. Working paper No. 63. University of Zürich.

Falk, A., E. Fehr, and U. Fischbacher (2005). Driving forces behind informal sanctions. *Econometrica*, forthcoming.

Fehr, E. and U. Fischbacher (2004). Social norms and human cooperation. *TRENDS in Cognitive Sciences* 8: 185-190.

Fehr, E. and S. Gächter (2000a). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives* 14: 159-181.

Fehr, E. and S. Gächter (2000b). Cooperation and punishment in public goods experiments. *American Economic Review* 90: 980-994.

Fehr, E. and S. Gächter (2002). Altruistic punishment in humans. *Nature* 415: 137-140.

Fehr, E. and J. Henrich (2004). Is strong reciprocity a maladaptation? On the evolutionary foundations of human altruism. In Hammerstein, P. (Ed.) *The Genetic and Cultural Evolution of Cooperation*. Cambridge: MIT Press.

Fehr, E. and B. Rockenbach (2003). The detrimental effects of sanctions on human altruism. *Nature* 422: 137-140.

Fehr, E. and K. Schmidt (1999). A Theory of fairness, competition and cooperation. *The Quarterly Journal of Economics* 114: 817-868.

Fehr, E. and K. Schmidt (2000). Theories of fairness and reciprocity: Evidence and economic applications. In Dewatripont, M., L. Hansen, and S.T. Turnovsky (Eds.) *Advances in Economics and Econometrics - 8th World Congress, Econometric Society Monographs*. Cambridge: Cambridge University Press.

Fehr, E., U. Fischbacher, and M. Kosfeld (2005). Neuroeconomic foundations of trust and social preferences. *American Economic Review* forthcomming.

Fehr, E., G. Kirchsteiger, and A. Riedl (1993). Does fairness prevent market clearing? An experimental investigation. *The Quarterly Journal of Economics* 108: 437-459.

Fehr, E., G. Kirchsteiger, and A. Riedl (1998). Gift exchange and reciprocity in competitive experimental markets. *European Economic Review* 42: 1-34.

Ferrer-i-Carbonell, A. (2005). Income and well-being: An empirical analysis of the comparison income effect. *Journal of Public Economics* 89: 997-1019.

Fischbacher U. (1999). Zurich toolbox for readymade economic experiments, experimenter's manual. Working Paper No. 21. Institute for Empirical Research in Economics, University of Zürich.

Fisher, J., R.M. Isaac, J. Schatzberg, and J.M. Walker (1995). Heterogeneous demand for public goods: Effects on the voluntary contributions mechanism. *Public Choice* 85: 249-266.

Forsythe, R., J.L. Horowitz, N.E. Savin, and M. Sefton (1994). Fairness in simple bargaining experiments. *Games and Economic Behavior* 6: 347-369.

Frank, R. H. (1987). If Homo Economicus could choose his own utility function, would he want one with a conscience. *American Economic Review* 77: 593-604.

Frijda, N.H. (1986). *The emotions*. Cambridge: Cambridge University Press.

Frijda, N.H. (1988). The laws of emotion. *American Psychologist* 43: 349-358.

Gächter, S. and B. Herrmann (2005). Norms of cooperation among urban and rural dwellers: Experimental evidence from Russia. Working paper. University of Nottingham.

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

Glaeser, E. (2005). The political economy of hatred. *The Quaterly Journal of Economics* 120: 45-86.

Glaeser, E.L., B. Sacerdote, and J.A. Scheinkman (1996). Crime and social interactions. *The Quarterly Journal of Economics* 111: 507-548.

Goleman, D. (1995). *Emotional intelligence: Why it can matter more than IQ*. New York: Bantam Books.

Güth W. and E. van Damme (1998). Information, strategic behavior and fairness in ultimatum bargaining: An experimental study. *Journal of Mathematical Psychology* 42: 227-247.

Güth, W., S. Huck, and W. Müller (2001). The relevance of equal splits in ultimatum games. *Games and Economic Behavior* 37: 161-169.

Güth, W., R. Schmittberger, and B. Schwarze (1982). An experimental analysis of ultimatium bargaining. *Journal of Economic Behavior and Organization* 3: 367-88.

Haidt, J. (2003). The moral emotions. In Davidson, R. J., K. R. Scherer, and H. H. Goldsmith (Eds.) *Handbook of Affective Sciences*. Oxford: Oxford University Press.

Henrich, J. (2004). Cultural group selection, coevolutionary processes and large-scale cooperation. *Journal of Economic Behavior and Organization* 53: 3-35.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, H. Gintis, and R. McElreath (2001). In search of homo economicus: Behavioral experiments in 15 small-scale societies. *American Economic Review* 91: 73-78.

Hirshleifer, J. (1987). On emotions as guarantors of threats and promises. In Dupre, J. (Ed.) *Latest on the Best: Essays on Evolutions and Optimality*. Cambridge: MIT Press.

Hirshleifer, D. and E. Rasmusen (1989). Cooperation in a repeated prisoner's dilemma with ostracism. *Journal of Economic Behavior and Organization* 12: 87-106.

Hoffman, E., K. McCabe, and V.L. Smith (1996). Social Distance and Other-Regarding Behavior in Dictator Games. *American Economic Review* 3: 653-660.

Hopfensitz, A. and E. Reuben (2005). The Importance of Emotions for the Effectiveness of Social Punishment. Discussion paper 05-075/1. Tinbergen Institute.

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

Jakobs, E., A.S.R. Manstead, and A.H. Fischer (1996). Social context and the experience of emotion. *Journal of Nonverbal Behavior* 20: 123-142.

Jakobs, E., A.S.R. Manstead, and A.H. Fischer (1999). Social motives, emotional feelings, and smiling. *Cognition and Emotion* 13: 321-345.

Kagel, J.H. and K.W. Wolfe (2001). Tests of fairness models based on equity considerations in a three-person ultimatum game. *Experimental Economics* 4: 203-219.

Kagel, J., C. Kim, and D. Moser (1996). Fairness in ultimatum games with asymmetric information and asymmetric payoffs. *Games and Economic Behavior* 13: 100-110.

Kandel, E. and E. P. Lazear (1992). Peer pressure and partnerships. *Journal of Political Economy* 100: 801-817.

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

Kirchsteiger, G. (1994). The role of envy in ultimatum games. *Journal of Economic Behavior and Organization* 25: 373-389.

Knez, M. and C.F. Camerer (1995). Outside options and social comparison in a three-player ultimatum game experiments. *Games and Economic Behavior* 10: 65-94.

Lazarus, R.S. (1991). *Emotion and Adaptation*. New York: Oxford University Press.

Lazear, E. P., U. Malmendier, and R. A. Weber (2005). Sorting in experiments. Working paper. Stanford University.

Lerner, J.S. and D. Keltner (2000). Beyond valence: Toward a model of emotion-specific influences on judgment and choice. *Cognition and Emotion* 14: 473-492.

Levine, D. (1998). Modeling altruism and spitefulness in experiments. *Review of Economic Dynamics* 1: 593-622.

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

Lin, H. and S. Sunder (2002). Using experimental data to model bargaining behavior in ultimatum games. In Zwick R. and A. Rapoport (Eds.) *Experimental Business Research*. Dordrecht: Kluwer.

Loewenstein, G.F. (1996). Out of control: Visceral influences on behavior. *Organizational Behavior and Human Decision Processes* 65: 272-292.

Loewenstein, G. F. (2000). Emotions in economic theory and economic behavior. *American Economic Review, Papers and Proceedings* 90: 426-432.

Loewenstein, G., L. Thompson, and M. Bazerman (1989). Social utility and decision making in interpersonal contexts. *Journal of Personality and Social Psychology* 57: 426-441.

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

McCabe, K.A., M.L. Rigdon, and V.L. Smith (2003). Positive reciprocity and intentions in trust games. *Journal of Economic Behavior and Organization* 52: 267-275.

Masclet, D., C. Noussair, S. Tucker, and M. C. Villeval (2003). Monetary and non-monetary punishment in the voluntary contribution mechanism. *American Economic Review* 93: 366-380.

Moll, J., R. de Oliveira-Souza, P. J. Eslinger, I. E. Bramati, J. Mourao-Miranda, P. A. Andreiuolo, and L. Pessoa (2002). The neural correlates of moral sensitivity: A functional magnetic resonance imaging investigation of basic and moral emotions. *The Journal of Neuroscience* 22: 2730-2736.

Nikiforakis, N. S. (2004). Punishment and Counter-punishment in Public Good Games. Working paper. Royal Holloway University of London.

Nikiforakis, N.S. and H.T. Normann (2005). A comparative statics analysis of punishment in public-good experiments. Working paper. Royal Holloway University of London.

Noussair, C. and S. Tucker (2005). Combining Monetary and Social Sanctions to Promote Cooperation. *Economic Inquiry* forthcoming.

Ohbuchi, K., M. Kameda, and N. Agarie (1989). Apology as aggression control: Its role in mediating appraisal of and response to harm. *Journal of Personality and Social Psychology* 56: 219-227.

Oliver, P. (1980). Rewards and punishments as selective incentives for collective action. *American Journal of Sociology* 86: 1356-1375.

Olson, M. (1965). *The logic of collective action*. Cambridge: Harvard University Press.

Ortony, A., G.L. Clore, and A. Collins (1988). *The cognitive structure of emotions*. Cambridge: Cambridge University Press.

Ostrom, E. (1998). A behavioral approach to the rational choice theory of collective action: Presidential address, American Political Science Association, 1997. *American Political Science Review* 92: 1-22.

Ostrom, E. and J. Walker (1997). Neither, markets nor states: Linking trasnformation processes in collective action arenas. In Mueller, D.C. (ed.) *Perspectives on Public Choice: A Handbook*. Cambridge: Cambridge University Press.

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

Ostrom, E., R. Gardner, and J. Walker (1994). *Rules, Games, and Common-Pool Resources*. Ann Arbor: University of Michigan Press.

Potters, J., M. Sefton, and L. Vesterlund (2005). After you: Endogenous sequencing in voluntary contribution games. *Journal of Public Economics* 89: 1399-1419.

Pillutla, M.M. and J.K. Murnighan (1996). Unfairness, anger, and spite: Emotional rejections of ultimatum offers. *Organizational Behavior and Human Decision Processes* 68: 208-224.

Quervain, D.J.F., U. Fischbacher, V. Treyer, M. Schellhammer, U. Schnyder, A. Buck, and E. Fehr (2004). The neural basis of altruistic punishment. *Science* 305: 1254-1258.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review* 83: 1281-1302.

Rapoport, A. and R. Suleiman (1993). Incremental contribution in step-level public goods games with asymmetric players. *Organizational Behavior and Human Decision Processes* 55: 171-194.

Rapoport, A., J.A. Sundali, and R.E. Potter (1996a). Ultimatums in two-person bargaining with one-sided uncertainty: Offer games. *International Journal of Game Theory* 25: 475-494.

Rapoport, A., J.A. Sundali, and D.A. Seale (1996b). Ultimatums in two-person bargaining with one-sided uncertainty: Demand games. *Journal of Economic Behavior and Organization* 30: 173-196.

Rawls, J. (1971). *A Theory of Justice*. Cambridge: Harvard University Press.

Reuben, E. and A. Riedl (2005). The Role of Inequality for Cooperation and Punishment in Public Goods Experiments: A Comprehensive Study. Work in progress.

Reuben, E. and F. van Winden (2004). Reciprocity and emotions when reciprocators know each other. Discussion paper 04-098/1. Tinbergen Institute.

Reuben, E. and F. van Winden (2005a). Negative Reciprocity and the Interaction of Emotions and Fairness Norms. Discussion paper 05-014/1. Tinbergen Institute.

Reuben, E. and F. van Winden (2005b). Social ties and negative reciprocity: The role of affect. Discussion paper 04-098/2. Tinbergen Institute.

Riedl, A. and J. Vyrastekova (2003). Responder behavior in three-person ultimatum game experiments. Working paper. University of Amsterdam.

Robinson, M. and G. Clore (2002). Belief and feeling: Evidence for an accessibility model of emotional self-report. *Psychological Bulletin* 128: 934-960.

Rosenblat, T.S. and M.M. Mobius (2004). Getting closer or drifting apart? *The Quarterly Journal of Economics* 119: 971-1009.

Roth, A.E., V. Prasnikar, M. Okuno-Fujiware, and S. Zamir (1991). Bargaining and market behavior in Jerusalem, Ljubljana, Pittsburgh and Tokyo: An experimental study. *American Economic Review* 81: 1068-1095.

Sadrieh, A. and H. Verbon (2005). Inequality, cooperation, and growth: An experimental study. *European Economic Review* forthcoming.

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* 300: 1755-1758.

Schelling, T.C. (1978). *Micromotives and Macrobehavior*. New York: Norton.

Schmitt, P. (2004). On perceptions of fairness: The role of valuations, outside options, and information in ultimatum bargaining games. *Experimental Economics* 7: 49-73.

Schwarz, N. (1990). Feelings as information: Informational and motivational functions of affective states. In Higgins E.T. and R. Sorrentino (Eds.) *Handbook of Motivation and Cognition: Foundations of Social Behavior*. New York: Guilford Press.

Smith, V.L. (1976). Experimental economics: Induced value theory. *American Economic Review* 66: 274-279.

Straub, P. and K. Murnighan (1995). An experimental investigation of ultimatum games: Information, fairness, expectations, and lowest acceptable offers. *Journal of Economic Behavior and Organization* 27: 345-364.

Tangney, J. P. and R. L. Dearing (2002). *Shame and Guilt*. The Guilford Press.

Tangney, J. P., R. S. Miller, L. Flicker, and D. H. Barlow (1996). Are shame, guilt and embarrassment distinct emotions? *Journal of Personality and Social Psychology* 70: 1256-1269.

Thaler, R. (2000). From homo economicus to homo sapiens. *Journal of Economic Perspectives* 14: 133-141.

van de Stadt, H., A. Kapteyn, and S. van de Geer (1985). The relativity of utility: Evidence from panel data. *Review of Economics and Statistics* 67: 179-187.

van Dijk, E. and M. Grodzka (1992). The influence of endowments asymmetry and information level on the contribution to a public step good. *Journal of Economic Psychology* 13: 329-342.

van Dijk, F., J. Sonnemans, and F. van Winden (2002). Social ties in a public good experiment. *Journal of Public Economics* 85: 275-99.

van Winden, F. (2001). Emotional hazard exemplified by taxation-induced anger. *KYKLOS* 54: 491-506.

Visser, M. and J. Burns (2005). The great divide: Inequality and punishment in public goods settings. Mimeo.

The Tinbergen Institute is the Institute for Economic Research, which was founded in 1987 by the Faculties of Economics and Econometrics of the Erasmus Universiteit Rotterdam, Universiteit van Amsterdam and Vrije Universiteit Amsterdam. The Institute is named after the late Professor Jan Tinbergen, Dutch Nobel Prize laureate in economics in 1969. The Tinbergen Institute is located in Amsterdam and Rotterdam. The following books recently appeared in the Tinbergen Institute Research Series:

319. H.J. HORBEEK, *The elastic workfloor. About the implementation of internal flexibility arrangements.*
320. P. HOUWELING, *Empirical studies on credit markets.*
321. E. MENDYS, *Essays on network economics.*
322. J. NOAILLY, *Coevolutionary modeling for sustainable economic development.*
323. Y.V. VELD-MERKOULOVA, *Essays on futures markets and corporate spin-offs.*
324. F.J. WILLE, *Auditing using Bayesian decision analysis.*
325. D. HUISMAN, *Integrated and dynamic vehicle and crew scheduling*.
326. S. WANG, *Global climate change policies: An analysis of CDM policies with an adapted GTAP model.*
327. M.W. VAN GELDEREN, *Opportunity entry performance. Studies of entrepreneurship and small business*.
328. W. VAN WINDEN, *Essays on urban ICT policies.*
329. G.J. KULA, *Optimal retirement decision.*
330. R.J. IMESON, *Economic analysis and modeling of fisheries management in complex marine ecosystems.*
331. M. DEKKER, *Risk, resettlement and relations: Social security in rural Zimbabwe.*
332. MULATU, *Relative stringency of environmental regulation and international competitiveness.*
333. C.M. VAN VEELEN, *Survival of the fair: Modelling the evolution of altruism, fairness and morality.*
334. R. PHISALAPHONG, *The impact of economic integration programs on inward foreign direct investment.*
335. A.H. NÖTEBERG, *The medium matters: The impact of electronic communication media and evidence strength on belief revision during auditor-client inquiry.*
336. M. MASTROGIACOMO, *Retirement, expectations and realizations. Essays on the Netherlands and Italy.*
337. E. KENJOH, *Balancing work and family life in Japan and four European countries: Econometric analyses on mothers' employment and timing of maternity.*
338. A.H. BRUMMANS, *Adoption and diffusion of EDI in multilateral networks of organizations.*
339. K. STAAL, *Voting, public goods and violence.*
340. R.H.J. MOSCH, *The economic effects of trust. Theory and empirical evidence.*

341. F. ESCHENBACH, *The impact of banks and asset markets on economic growth and fiscal stability.*
342. D. LI, *On extreme value approximation to tails of distribution functions.*
343. S. VAN DER HOOG, *Micro-economic disequilibrium dynamics.*
344. B. BRYS, *Tax-arbitrage in the Netherlands evaluation of the capital income tax reform of January 1, 2001.*
345. V. PRUZHANSKY*, Topics in game theory.*
346. P.D.M.L. CARDOSO*, The future of old-age pensions: Its implosion and explosion.*
347. C.J.H. BOSSINK, *To go or not to go…? International relocation willingness of dual-career couples.*
348. R.D. VAN OEST, *Essays on quantitative marketing models and Monte Carlo integration methods.*
349. H.A. ROJAS-ROMAGOSA, *Essays on trade and equity.*
350. A.J. VAN STEL, *Entrepreneurship and economic growth: Some empirical studies.*
351. R. ANGLINGKUSUMO, *Preparatory studies for inflation targeting in post crisis Indonesia.*
352. GALEOTTI, *On social and economic networks.*
353. Y.C. CHEUNG, *Essays on European bond markets.*
354. ULE, *Exclusion and cooperation in networks.*
355. I.S. SCHINDELE, *Three essays on venture capital contracting*.
356. C.M. VAN DER HEIDE, *An economic analysis of nature policy.*
357. Y. HU, *Essays on labour economics: Empirical studies on wage differentials across categories of working hours, employment contracts, gender and cohorts.*
358. S. LONGHI, *Open regional labour markets and socio-economic developments: Studies on adjustment and spatial interaction.*
359. K.J. BENIERS, *The quality of political decision making: Information and motivation.*
360. R.J.A. LAEVEN, *Essays on risk measures and stochastic dependence: With applications to insurance and finance.*
361. N. VAN HOREN, *Economic effects of financial integration for developing countries.*
362. J.J.A. KAMPHORST, *Networks and learning.*
363. E. PORRAS MUSALEM, *Inventory theory in practice: Joint replenishments and spare parts control.*
364. M. ABREU, *Spatial determinants of economic growth and technology diffusion.*
365. S.M. BAJDECHI-RAITA, *The risk of investment in human capital.*
366. A.P.C. VAN DER PLOEG, *Stochastic volatility and the pricing of financial derivatives.*
367. R. VAN DER KRUK, *Hedonic valuation of Dutch Wetlands.*
368. P. WRASAI, *Agency problems in political decision making.*
369. B.K. BIERUT, *Essays on the making and implementation of monetary policy decisions.*