Incentives and social preferences

Four essays in experimental economics

Inaugural-Dissertation
zur Erlangung des Grades
Doctor oeconomiae publicae (Dr. oec. publ.)
an der Ludwig-Maximilians-Universität München

2009

vorgelegt von

Christina Strassmair

Referent:	Prof. Dr. Klaus M. Schmidt
Korreferent:	Prof. Dr. Martin G. Kocher
Promotionsabschlussberatung:	18. November 2009

Acknowledgements

First and foremost, I would like to thank my supervisor Klaus M. Schmidt. He always provided me with excellent comments and inspiring encouragement. His support, reaching far beyond the scope of my dissertation, was truly exceptional. I am also thankful to Martin Kocher and Joachim Winter who agreed to serve as second and third supervisor to my dissertation. I benefitted a lot from their advice and guidance.

I am also much indebted to my co-authors Sandra Ludwig, Hannah Schildberg-Hörisch, and Matthias Sutter. I have very much enjoyed working with them and learnt a lot from our joint projects. I thank them especially for their enthusiasm and patience.

I have received insightful comments and helpful support from many other friends, colleagues, and participants at conferences and seminars. I am thankful to all of them. René Cyranek, Florian Englmaier, Georg Gebhardt, Markus Reisinger, Ludwig Reßner, Anno Stolper, Christian Traxler, and Ferdinand von Siemens deserve special mention. I also thank my colleagues at the Seminar for Economic Theory. Many thanks go to my flat mate and good friend Silvia Appelt for making the time so enjoyable.

I also thank the members of CREED at the University of Amsterdam, in particular Frans van Winden, for hosting me as a visiting Ph.D. student. My stay at the center was a very inspiring and productive time. Financial support from ENABLE Marie Curie Research Training Network, funded under the 6^{th} Framework Program of the European Union, and from the SFB Transregio 15 is gratefully acknowledged.

Last but by no means least, I thank Felix and my family. Their love and their confidence have constantly supported me.

Christina Strassmair

Contents

List of Tables

List of Figures

Preface

This dissertation is comprised of four self-contained chapters that contribute to different research areas in the field of behavioral economics. Still, each chapter addresses the topic of incentives and social preferences.

Both monetary and non-monetary incentives are frequently used in order to motivate individuals to behave in a certain way. If individuals are not completely selfish but have some sort of social preferences, as it is often observed in empirical studies, then the effects of these incentives may be considerably different than predicted by standard theory. In the first chapter we empirically study monetary incentives that are intended to deter individuals from taking or stealing. We set up an experiment in which individuals have the possibility to take from another individual's endowment and are punished for taking with a given detection probability and a given fine. By varying deterrent incentives, i.e. the detection probability and the fine, in the experiment we empirically test Becker's (1968) deterrence hypothesis that crime is reduced by deterrent incentives. In an extension we check whether the observed incentive effects depend on the incentives' labeling. In the second chapter we address the question how free-riding in teams can be reduced. In contrast to the first chapter, we do not focus on the effect of monetary incentives, i.e. the team members' remuneration, but on the effect of information flows in teams. If a team member receives information or signals about a colleague's performance prior to his decision, the team member's and his colleague's incentives to free-ride may change. In an experimental study we analyze the effect of signals and how it depends on the signals' types. In the third chapter we experimentally study whether intentions can spoil the kindness of a gift. If so, then the success of a campaign that intends to stimulate individuals to behave in a certain way by distributing gifts or behaving in a way that seems altruistic crucially depends on whether the individuals recognize the intention behind the campaign. In our experimental study a first mover can do a second mover a favor and the second mover can reciprocate with a given probability. The higher the probability that the second mover can reciprocate, the higher is the probability that the first mover can expect a return. We vary the probability that the second mover can reciprocate and analyze whether the second mover returns more the smaller this probability. In the fourth chapter we focus on behavior in hierarchically structured organizations in which the exchange of resources is determined by power relations. We set up an experiment and empirically analyze whether more powerful individuals in these organizations use their power in order to seize resources of less powerful individuals, and whether they finally end up wealthier than their less powerful fellows. If the latter does not hold and less power is a blessing, promotion to upper hierarchies is no suitable instrument to incentivize less powerful individuals.

In all four chapters we use laboratory experiments in order to empirically address our research questions. In the last decades laboratory experiments have been established as a common tool in empirical economic analysis. A crucial role in this progress has been played by Vernon L. Smith – one of the first experimenters in economics – who was awarded the Nobel Prize in economics in 2002. Experimental studies complement field studies. On the one hand, laboratory experiments allow more control, on the other hand, the environment in the laboratory is less natural than in the field. Given the research questions we address in all four chapters of this dissertation, there are either no field data available or field data that typically lack of control. Hence, laboratory experiments are an appropriate way to shed more light on these questions.

All four chapters contribute to the literature on social preferences. Field as well as experimental data have provided ample evidence that human behavior often systematically deviates from the predictions of standard theory but is consistent with the predictions of models of social preferences (for an overview see Camerer, 2003; Fehr and Schmidt, 2006). This empirical evidence suggests that (i) not all individuals are completely selfish as typically assumed in standard theory but at least some individuals have social preferences and care about fairness or reciprocity, and (ii) social preferences may have real behavioral effects. Of course, this is not always the case and there are a lot of situations in which the self-interest model explains behavior well, e.g. in large competitive markets. In the settings that are analyzed in the four chapters of this dissertation (some) individuals can influence the payoffs of those they interact with. Typically, social preferences play a crucial role in such settings.

In the first chapter – a joint study with Hannah Schildberg-Hörisch – we focus on the question how to combat crime. Different professions have discussed this topic and have come up with various answers. The standard economic contribution, which was first formalized by Becker (1968), states that punishment, i.e. the probability of detection and the amount of the fine, reduces crime. In the last decades, however, empirical evidence that shows that monetary incentives can be detrimental in certain contexts (for an overview see Frey and Jegen, 2001) has accumulated. On the one hand, monetary incentives affect material (extrinsic) motivation to undertake or not undertake an action, on the other hand, they may crowd out intrinsic motivation. Empirical studies on the deterrence hypothesis typically rely on aggregate field data and find mixed evidence that is largely in line with Becker's (1968) deterrence hypothesis. Yet, the detection probability and the severity of punishment explain only a small part of criminal activity (for an overview see Eide, 2000; Glaeser, 1999). Serious drawbacks of these studies are the methodological problems they have to tackle, e.g. simultaneity bias, omitted variable problems and measurement errors. We go down a different road and gather individual data from an experiment in which we exogenously vary incentives and are able to directly test Becker's (1968) deterrence hypothesis. We ask a very basic but important question: Do deterrent incentives work?

In order to answer this question, we set up one of the simplest experimental designs in which person B has the possibility to take or steal from person A's endowment. B's theft is detected with a given probability and – if detected – B has to return the stolen amount to A and is punished with a given fine. We exogenously vary the detection probability and the fine across treatments from no deterrent incentives up to very strong deterrent incentives such that stealing does not pay off in expectation. As punishment possibilities are usually restricted in real life, we devote special attention to treatments with small and intermediate incentives. The experiment is neutrally framed and does not include terms such as "steal", "theft", or "fine".

In contrast to Becker's (1968) hypothesis, we find an inverted-U shape relationship between the average taken amount and deterrent incentives: Intermediate incentives strictly increase B's average taken amount compared to the setting with no incentives, and only very strong incentives deter taking. Hence, intermediate incentives backfire. This observation contradicts predictions of standard theory as well as of the most frequently used models of social preferences. Our data can be explained by a model in which about 50 % of individuals are selfish and act according to the deterrence hypothesis and the rest has fairness concerns which are crowded out by monetary incentives. In an extension we check whether our observed incentive effects change when incentives have a strong moral connotation. We run additional experiments in which we use the terms "steal" and "fine", and still find evidence for backfiring of incentives.

Our results suggest that monetary incentives may be detrimental. Taking our observations literally would imply to punish criminal activities either very hard or not at all to efficiently deter crime and avoid backfiring of incentives. We are, however, well aware that the laboratory is a less natural environment than the field and abstracts from social norms and stigmata that could be the driving forces behind punishment in reducing criminal behavior in real life. Nevertheless, our data reject Becker's (1968) deterrence hypothesis that punishment deters crime independent of all other factors. Thus, monetary incentives should be used with great care and without neglecting other (social) factors that may be at work.

In the second chapter – a joint study with Sandra Ludwig – we focus on the question how to reduce free-riding in teams. As agents in a team are rewarded according to the team output but the costs of contributing to the team output accrue to the individual member only, agents work inefficiently little. One way to stimulate agents to work more is the implementation of information flows. If an agent receives a signal on the colleague's performance prior to his decision, the agent's as well as the colleague's incentives to free-ride are affected. According to standard theory the overall effect depends on the team production function. If agents' contributions are, for example, substitutes (i.e. the more one agent contributes, the lower are the incentives for the other agent to work), the first moving agent is predicted to work very little in order to induce the second moving agent to work hard. However, if agents are concerned about fairness and reciprocity, the implementation of information flows may have a positive effect for all agents, even if agents' contributions are substitutes. In our study we raise the question how the structure of information flows affects team behavior and which types of signals are especially effective.

In order to empirically study the effect of signals in teams, we set up an experiment in which two agents sequentially exert effort to contribute to the team output. The second mover either observes no signal, the first mover's effort, the first mover's contribution, or both prior to his decision. At the end of the experiment both agents earn the same wage, which positively depends on the team output, minus their individual effort costs. The wage scheme is exogenously fixed, e.g. by a market or a principal that we do not model in our experiment, such that agents' contributions are substitutes.

We observe that signals in teams raise both the average effort of the first mover and the second mover. In general, the second mover reacts positively to signals. For example, if the second mover observes the first mover's effort, he works more the more the first mover worked. Furthermore, the first mover does not lean back if signals on his performance will be available for the second mover prior to the second mover's decision. Especially the first mover's effort seems to be a stimulating signal. In an extension we check whether our results also hold for the case where agents' contributions are complements. Our findings in these treatments are similar. Overall, our results are in contradiction to the predictions of standard theory but largely in line with those of a simple model of social preferences. The latter assumes that individuals avoid payoff inequalities generated by different effort levels.

We conclude that on average information flows reduce free-riding, irrespective of whether agents' contributions are substitutes or complements. Especially the provision of information on the team mate's effort seems to be an effective way to achieve more efficient outcomes. Our results suggest that social comparison plays a crucial role in team work and shapes behavior: The reaction to signals may be contrary to the predictions of the self-interest model and signals that are considered to have no effect may, in fact, be influential.

In the third chapter we study whether intentions can spoil the kindness of a gift. Campaigns or business strategies often rely on distributing small gifts or behaving in a way that seems altruistic when they try to gain the future support or cooperation of individuals. We ask, however, whether the kindness of these gifts and their actual future returns are spoiled by the expectation of future rewards.

We address this question in an experiment in which individual A can transfer an amount of money to individual B. B receives the tripled amount transferred and, with a given probability, he can return a part of it to A. We vary the probability that B can return money and, thereby, that A can expect a return. Models of intention-based reciprocity predict that A's expected rewards spoil the kindness of A's transfer and that, therefore, B returns less for a given transfer the higher the probability that he can return money.

In contrast to standard theory, we observe that gifts generate future cooperation. On average, B returns more (given he is asked to decide) the higher A's transfer. Intentions, however, do not seem to spoil the kindness of the gifts and their returns. B's average return for a given transfer of A does not vary in the probability that B can return money. This empirical evidence contradicts models of intention-based reciprocity but is in line with models of social preferences based on outcomes. Our results suggest that gifts are rewarded (if possible), irrespective of the donor's intentions.

In the fourth chapter – a joint study with Matthias Sutter – we address the question whether more powerful agents in organizations where power governs the exchange of assets (e.g. in vertical hierarchies) finally end up less wealthy than their less powerful fellows. Piccione and Rubinstein (2004) have shown that this may actually happen if agents are selfish and more powerful agents are initially wealthier. In this case some agents seize the wealth of other – less powerful and initially less wealthy – agents and relatively powerful agents become an attractive target due to their initial wealth and their power which enables them to attack others and accumulate more wealth. In the presence of the abundant literature on the importance of social preferences like fairness and reciprocity (for an overview see Camerer, 2003; Fehr and Schmidt, 2006), we raise the questions whether individuals seize others' assets and, thereby, create extremely unequal distributions of final wealth, and whether less power may, indeed, be a blessing.

In order to empirically address these questions, we set up a simple experiment in which three agents are ordered by power. Each agent can – but needs not – attack at most one less powerful agent at given transaction costs. If several agents attack the same victim, then only the most powerful attacker prevails and receives the victim's initial as well as accumulated assets, while the other attackers do not receive anything and still have to bear transaction costs.

We observe that individuals try to seize others' assets more – instead of less – often than selfish individuals are expected to do. This behavior creates an extremely unequal distribution of final wealth. Furthermore, the second most powerful agent ends up less wealthy than the least powerful agent almost always when it is predicted by Piccione and Rubinstein (2004) and even sometimes when it is not. Our results are in line with the predictions of standard theory but, at first sight, they seem to be at odds with those of models of social preferences. Yet, we show that some of our results that seem to be unsocial can be reconciled with a frequently used model of social preferences.

Our results suggest that individuals do not shy away from using their power in hierarchies and seize others' assets. As a consequence, agents may even end up less wealthy than their weaker fellows. This creates disincentive effects. First, weak agents in hierarchies may be unwilling to invest ex ante, e.g. in firm specific knowledge. Second, promotion to upper-hierarchies does not serve as an instrument to incentivize weak agents since less power may be a blessing.

The following four chapters all have their own introduction and appendix and each can be read independently of the other three chapters.

Chapter 1

An experimental test of the deterrence hypothesis[*]

1.1 Introduction

That crime has to be punished seems to be universally accepted. The purpose and the level of punishment, however, are controversial. Immanuel Kant advocated punishment to re-establish justice, Georg Friedrich Wilhelm Hegel stressed that ill has to be retaliated with ill. Both philosophers regard punishment as a mean to establish justice. In contrast, there exist schools of thoughts which stress that punishment shall prevent (future) crime. Becker's (1968) deterrence hypothesis is the classic economic contribution to the debate on punishment. According to Becker (1968) the purpose of punishment is to (efficiently) deter individuals from committing crimes. To achieve deterrence, Becker (1968) relies on the power of deterrent incentives such as the severity and the probability of punishment. The deterrence hypothesis states that crime (weakly) decreases in the severity and in the probability of punishment.

Our laboratory experiment tests the deterrence hypothesis in a controlled environment that permits to exogenously vary deterrent incentives, i.e. the detection probability and the level of punishment. For this purpose, we use a very straightforward context, namely subjects have the possibility to steal from another subject. They can-

[*] This chapter is joint work with Hannah Schildberg-Hörisch from the University of Bonn.

not only decide whether they steal or not, but also how much they steal. We ask a very basic but important question: Do deterrent incentives work?

In order to answer this question, we have chosen one of the simplest possible designs: a modified dictator game. Two agents, A and B, are randomly matched. Agent A is a passive agent and has a higher initial endowment than agent B. Agent B can decide how much he takes away (steals) from A's initial endowment. With probability $1-p$, this amount is transferred from A to B. With probability p (the detection probability), however, this amount is not transferred and a fixed fine f is deducted from B's initial endowment if B has chosen a strictly positive amount.

We conduct six different treatments in which we vary the detection probability p and the fine f. Our benchmark treatment T1 sets $p = f = 0$. Treatments T2, T3, and T4 implement small and intermediate incentives such that taking agent A's whole initial endowment still pays off in expectation. Treatment T5 is characterized by a combination of p and f such that taking everything generates about the same expected payoff as taking nothing. In treatment T6, however, the expected payoff from taking everything is substantially smaller than the one from taking nothing. Each subject participates in two different treatments sequentially. This design permits both an across and a within subject analysis of taking behavior. In other words, we can analyze different regimes and regime changes with the data at hand.

Our experimental design has three main advantages. First, the game is very simple and easy to understand for the subjects. Second, our design allows testing the isolated effect of monetary incentives. This is because the implemented incentives themselves do not contain a message from agent A, e.g. about A's trust, expectations, intentions, costs of a theft, or future actions, as incentives are determined exogenously (by the experimenter) and not by agent A, the payoff table is common knowledge, and A is passive. Third, our design captures some crucial features of many crimes: The victim is rather passive. It cannot affect the severity of punishment and – to a large extent – the detection probability. The distribution of costs and benefits of the committed crime is relatively clear. In case of a theft, for example, the stolen amount is a good predictor of the thief's benefit and the victim's cost.

The results obtained in our across subjects analysis clearly reject the deterrence hypothesis: The average taken amount does not monotonically (weakly) decrease in deterrent incentives. In contrast, we find that incentives may backfire: Subjects take significantly more in the treatment with intermediate incentives than in the absence of incentives. Only very strong incentives deter subjects from taking. Backfiring of small incentives and deterrence of strong incentives can also be observed in our within subjects analysis. These results can be explained by a model of two types of subjects: selfish subjects who react to deterrent incentives as predicted by the deterrence hypothesis and fair-minded subjects who take more when incentives are introduced or raised until incentives reach a very high level. Possible explanations for the behavior of the second type of subjects are crowding out of fairness concerns by extrinsic incentives or fairness concerns regarding expected outcomes. Only lasting crowding out of fairness concerns can explain the sequence effects in our data: Many fair-minded subjects take more in a given treatment if this treatment was preceded by a treatment with stronger incentives than if it was not preceded by any treatment, i.e. played in the first part. Furthermore, we find that p and f seem to be interchangeable instruments in achieving deterrence.

Since we obtain our data from neutrally framed experiments, one may question our results and their relevance for real life stealing. In real life crime and deterrent incentives often have a strong moral connotation, and policy makers may make use of that. Still, we consciously use a neutral frame because our primary aim is to test the (standard) economic approach to crime. Its core, the deterrence hypothesis, relies on incentive effects that are independent of all other factors which may influence crime. In Becker's (1968) model framing might ceteris paribus affect crime, but not the comparative statics with respect to deterrent incentives. Whatever the frame, the taken amount should monotonically decrease in deterrent incentives. In order to measure the effect of moral costs evoked by a non-neutral, moral framing, we run some additional sessions in which we label B's decision as "stealing" if $x > 0$ and the fixed fine f as "penalty" instead of "minus points". In these sessions we still observe backfiring of incentives.

Becker's (1968) seminal paper has triggered numerous theoretical extensions as well as field studies testing its external validity.[1] At large, the empirical evidence from field studies implies that punishment reduces crime, but variations in the detection probability and the severity of punishment explain only a small part of the variation in crime (see Glaeser, 1999). This may be caused by methodological problems that arise when using field data. Usually, only aggregate data are available which results in simultaneity bias and omitted variable problems.[2] Field data often report the behavior of offenders only and not that of the general population. Furthermore, measurement error is widespread as not all crime is reported. All these problems do not exist in the laboratory.

There already exist experimental studies focusing on criminal behavior. The experimental literatures on tax evasion and on corruption explicitly address deterrence.[3] Corruption and tax evasion setups clearly differ from ours, though. In experimental corruption games the cooperation of both the briber and the bribee is typically necessary to successfully bribe. In tax evasion experiments subjects' behavior often does not influence other subjects' payoffs at all, or only indirectly when the collected taxes are used for public good provision or redistribution of resources among a group of subjects. In other tax evasion experiments the tax authority itself is a player and can strategically interact with the taxpayer. In our setup, in contrast, a stealing subject directly hurts another subject, which seems to be a crucial feature of many crimes, the victim is passive, and incentives are set exogenously. Laboratory experiments on criminal behavior other than tax evasion and corruption are scarce. Falk and Fischbacher (2002) explore the influence of social interaction phenomena on committing a crime. Bohnet and Cooter (2001), Galbiati and Vertova (2005), and Tyran and Feld (2006) investigate whether law can act as expressive law, i.e. prevent crime by activating norms that prohibit committing a crime. Tyran and Feld (2006) also compare the effects of exogenously imposed and endogenously chosen incentives. While Falk and Fischbacher

[1] Garoupa (1997) and Polinsky and Shavell (2000a) provide comprehensive overviews on the economic theory of optimal law enforcement. Eide (2000) and Glaeser (1999) survey empirical studies on the deterrence hypothesis.

[2] See Levitt (1997) for a convincing example of how to address the simultaneity problem.

[3] Torgler (2002) reviews the experimental literature on tax evasion and concludes that evidence on the effectiveness of deterrent incentives is rather mixed (p. 662). While Abbink et al. (2002) find in an experiment that deterrent incentives reduce corruption, Schulze and Frank (2003) observe that deterrent incentives in an experimental corruption game can backfire.

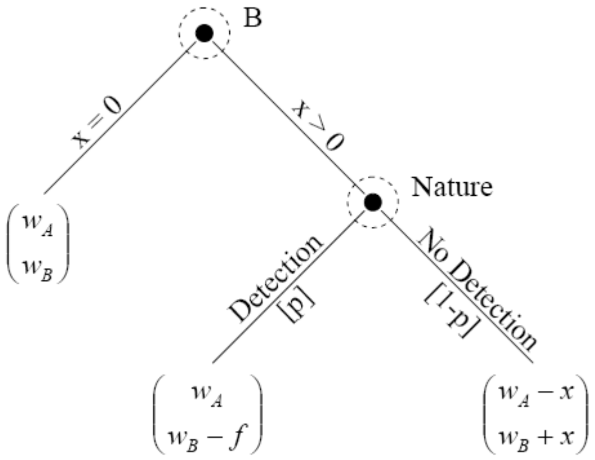
(2002) and Bohnet and Cooter (2001) do not vary incentives and, therefore, cannot test for incentive effects, Galbiati and Vertova (2005) and Tyran and Feld (2006) do, however in a context that differs from our setup. In Galbiati and Vertova (2005) the average contribution in a public good game increases the higher the obligatory minimum contribution. In Tyran and Feld (2006) the average contribution in a public good game increases the higher the punishment for not contributing.

In addition, there is a growing economic literature that investigates the effectiveness of incentives in different contexts, e.g. in labor market relations. These setups are usually richer than ours: Incentives are often determined endogenously by a principal and, therefore, may signal the principal's trust or intentions (Ellingsen and Johannesson, 2008), or norms (Sliwka, 2007), or an action's costs and benefits (Bénabou and Tirole, 2003). Some laboratory and field experiments on such contexts document that (small) incentives backfire and thus challenge the belief in the effectiveness of incentives.[4] Frey and Jegen (2001) stress that introducing incentives has two countervailing effects: Besides the standard relative price effect, incentives may crowd out intrinsic motivation. With small incentives the relative price effect is small and the latter, counterproductive effect may dominate.

This chapter proceeds as follows. Section 1.2 presents the experimental design and procedure, Section 1.3 the behavioral predictions and hypotheses. The across and within subjects analyses are summarized and discussed in Section 1.4. In Section 1.5 we check the robustness of our results by presenting results from sessions with a moral frame. Section 1.6 concludes.

[4] Bowles (2007), Fehr and Falk (2002), and Frey and Jegen (2001) survey the economic literature on crowding out of intrinsic motivation. The origins of this literature are in psychology, see for example Deci (1971) and Lepper et al. (1973). Deci et al. (1999) provide a meta-analysis of more than 100 psychological studies on the effect of extrinsic rewards on intrinsic motivation.

Figure 1.1: Structure of the stealing game


B
x = 0
x > 0
Nature
$\begin{pmatrix} w_A \\ w_B \end{pmatrix}$
Detection
[p]
No Detection
[1-p]
$\begin{pmatrix} w_A \\ w_B - f \end{pmatrix}$
$\begin{pmatrix} w_A - x \\ w_B + x \end{pmatrix}$

1.2 Experimental design and procedure

Consider the simplest possible stealing game with two agents, A and B. Agent A is initially endowed with w_A, and agent B is initially endowed with w_B, with $w_A > w_B$.[5] While agent A is passive, agent B can take any amount $x \in [0, w_A]$ from agent A's initial endowment. If B does not take anything, i.e. $x = 0$, agents A and B both receive their initial endowments w_A and w_B, respectively. If B takes a strictly positive amount, i.e. $x > 0$, with probability $(1 - p) \in [0, 1]$ the taken amount x is transferred from A to B, with probability p the taken amount x is not transferred and, on top of that, agent B has to pay a fixed fine f. We use a fixed fine f that is independent of x for $x > 0$ in order to keep the design as simple as possible. The structure of the game is summarized in Figure 1.1.

Since we focus on incentive effects on B's behavior, we vary the detection probability p and the fine f across different treatments and fix w_A and w_B at levels 90 and 50, respectively. Table 1.1 presents the treatments.

Treatment T1, our benchmark treatment, implements no deterrent incentives. It is simply the mirror image of a dictator game.[6] In all other treatments a strictly

[5] $w_A > w_B$ allows to distinguish between subjects who have a preference for fair (equal) outcomes and subjects who simply do not want to take anything in treatment T1.

[6] Here subjects can decide how much they take away from (instead of to give to) another agent in

Table 1.1: Treatments of the stealing game

Treatment	p	f	B's expected payoff given x = 0	B's expected payoff given x = 90	Level of incentives
T1	0.0	0	50	140	zero
T2	0.6	6	50	82.4	small
T3	0.5	25	50	82.5	small
T4	0.6	20	50	74	intermediate
T5	0.7	40	50	49	high
T6	0.8	40	50	36	very high

positive p and a strictly positive f is implemented. We categorize the intensity of these incentives according to agent B's expected payoff when taking agent A's whole initial endowment. As Table 1.1 shows, the level of incentives (weakly) increases in the order of the treatment. In treatments T2, T3, and T4 taking everything pays off in expectation. Treatment T5 is characterized by a combination of p and f such that taking the maximally possible amount generates about the same expected payoff as taking nothing. In treatment T6, however, the expected payoff from taking everything is substantially smaller than the one from taking nothing. Since in treatments T2 and T3 the same intensity of incentives is implemented by different p and f, we can analyze whether p and f are interchangeable instruments in deterring taking behavior, at least for this level of incentives.

Each experimental session consisted of three parts: two different treatments of the stealing game and a dictator game.[7] After these three parts participants filled out a questionnaire eliciting data on their age, sex and subject of studies. We used a paid Holt and Laury (2002) procedure to get an indication of subjects' risk preferences.[8] The conducted sessions are presented in Table 1.2.

At the beginning of each session participants were told that one randomly picked part out of the three would be paid for all of them. After each part only the instructions

a purely distributional context without any strategic considerations.

[7] In the dictator game the dictator could give any amount of his initial endowment of 90 to a passive agent with an initial endowment of 50. The chosen amount may indicate the dictator's aversion to advantageous inequity. Note, however, that the donated amount might be affected by the treatments played in part 1 and part 2.

[8] The translated table and a brief report on the observed levels of risk aversion can be found in the appendix.

Table 1.2: Session plan of the stealing game

Session	Part 1	Part 2	Part 3	Questionnaire*	Number of participants
T1T3	T1	T3	DG	Yes	38
T3T1	T3	T1	DG	Yes	38
T2T3	T2	T3	DG	Yes	20
T3T2	T3	T2	DG	Yes	18
T2T4	T2	T4	DG	Yes	38
T4T2	T4	T2	DG	Yes	36
T5T6	T5	T6	DG	Yes	32
T6T5	T6	T5	DG	Yes	38

* includes a Holt and Laury (2002) table
DG dictator game

for the following part were handed out. Subjects did not receive any feedback before the end of the experiment. They were matched according to a perfect stranger design, i.e. a couple matched once is never matched again in the following parts. Those subjects who were randomly chosen to be agents B in part 1 remained agents B in part 2 and were assigned the role of the dictator in part 3. Consequently, passive subjects remained passive throughout all three parts of the session.[9]

This design offers the possibility to analyze the observed behavior in two different ways. First, we can compare behavior in part 1 across different treatments. This is the cleanest comparison because individual behavior in part 1 is not influenced by any preplay. Second, we can analyze how agents B adapt their behavior to the change in incentives from part 1 to part 2. Since the structure of the game is very simple, we assume that a change in behavior from part 1 to part 2 is stimulated by the change of incentives rather than learning.

Our experimental sessions were run in November 2006 and March 2007 at the experimental laboratory of the SFB 504 in Mannheim, Germany. 258 students of the Universities of Mannheim and Heidelberg participated in the experiment. Subjects were randomly assigned to sessions and could take part only once. The experiment was programmed and conducted with the experimental software z-Tree (Fischbacher, 2007). The sessions were framed neutrally[10] and lasted about 40 minutes. Subjects

[9] To keep the passive subjects busy, we asked them how they would decide if they were agent B.

[10] Translated instructions for agents B and a more detailed description of the procedure of a session

did not receive a show-up fee[11] and earned 12.34 € on average.

1.3 Behavioral predictions and hypotheses

We focus on the question how the intensity of incentives affects B's decision. This depends on the specific form of B's utility function. Different theoretical approaches make different assumptions and, therefore, have varying behavioral predictions.

1.3.1 Behavioral predictions

Model 1: The self-interest model

The standard neoclassical approach assumes that all people are selfish, i.e. their utility function U depends on their own material payoff m only and increases in m.

With these assumptions, the deterrence hypothesis holds, namely the optimal taken amount $x^*(p, f)$ monotonically (weakly) decreases in p and f.

Due to the fixed fine f, agent B who maximizes his expected utility either takes as much as possible (w_A) or nothing. This depends on the relative sizes of p, f, w_A, w_B, and on the level of risk aversion. B's optimal taken amount is

$$x^*(p,f) \in \begin{cases} \{0\} & \text{if} \quad p > \frac{U(w_A+w_B)-U(w_B)}{U(w_A+w_B)-U(w_B-f)} \\ \{0, w_A\} & \text{if} \quad p = \frac{U(w_A+w_B)-U(w_B)}{U(w_A+w_B)-U(w_B-f)} \\ \{w_A\} & \text{if} \quad p < \frac{U(w_A+w_B)-U(w_B)}{U(w_A+w_B)-U(w_B-f)} \end{cases} .$$

The higher p or the higher f, the less attractive it is to take everything. For sufficiently high values of p and f, agent B does not take anything. This holds for any risk preferences, i.e. it is independent whether U is concave or convex in m. For risk averse agents, i.e. agents whose U is concave in m, the set of p, f, w_A, w_B combinations for

are provided in the appendix.

[11] Six subjects did not earn anything in the randomly selected part and in the Holt and Laury (2002) table.

which taking everything is optimal is smaller than the one for risk neutral agents, i.e. agents whose U is linear in m.[12]

Numerous empirical studies have shown that individual behavior may systematically deviate from predictions of the standard neoclassical approach. In these studies observed behavior is often consistent with predictions of models of social preferences.[13] Our two-agent setup with unequal initial endowments seems to be a context in which it is very plausible to consider models of fairness concerns.

Model 2: A model of fairness concerns regarding final outcomes

Models of fairness concerns regarding final outcomes (e.g. Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) assume that an agent's utility function $\widetilde{U}$ is not only dependent on m but also on the material payoff inequality $|m - y|$ with y as the material payoff of the other agent. $\widetilde{U}$ increases in m for any given $|m - y|$ and $\widetilde{U}$ decreases in $|m - y|$ for any given m.

Given these assumptions, the deterrence hypothesis still holds if there exists a unique optimal decision $x^*(p = 0, f = 0)$ that maximizes agent B's expected utility for $p = 0$ and $f = 0$.

Due to the fixed fine f, agent B who maximizes his expected utility either takes an amount which is optimal for no incentives ($x^*(p = 0, f = 0)$) or nothing. For relatively low values of p and f, agent B takes $x^*(p = 0, f = 0)$, which may be smaller than w_A. For relatively high values of p and f, agent B is deterred and takes nothing.

The reason is that agents cannot trade off payoffs from different states, in our context payoffs if B's taking is detected and payoffs if B's taking is not detected. Hence, $x^*(p \geq 0, f \geq 0)$ cannot be strictly larger than $x^*(p = 0, f = 0)$: If B's taking is not detected, $\widetilde{U}$ is maximized at $x^*(p = 0, f = 0)$; if B's taking is detected, $\widetilde{U}$ is the same for any $x > 0$ and largest for $x = 0$. Analogously, taking a strictly positive but smaller amount than $x^*(p = 0, f = 0)$ yields less expected utility than taking

[12] This is because concavity of U in m implies $\frac{w_A}{w_A+f} > \frac{U(w_A+w_B)-U(w_B)}{U(w_A+w_B)-U(w_B-f)}$.

[13] Fehr and Schmidt (2006) survey empirical foundations of social preferences.

$x^*(p = 0, f = 0)$ and, therefore, cannot be optimal.

Model 3: A model of fairness concerns regarding expected outcomes

Models of fairness concerns regarding expected outcomes (e.g. Trautmann, 2009) assume that an agent's utility function $\widehat{U}$ is not only dependent on m but also on the absolute difference between the own expected payoff m^e and the other agent's expected payoff y^e ($|m^e - y^e|$).[14] $\widehat{U}$ increases in m for any given $|m^e - y^e|$) and decreases in $|m^e - y^e|$) for any given m.

If $|m^e - y^e|$ directly enters the utility function, the deterrence hypothesis may not hold any more, i.e. there may exist values of p and f for which $x^*(p, f)$ *strictly* increases in p or in f.

The reason is that agents can trade off payoffs from different states, e.g. an advantageous inequity in material payoffs if B's taking is not detected can compensate a disadvantageous inequity in material payoffs if B's taking is detected. As an illustration, consider the following utility function $\widehat{U} = m - \beta * max\{m^e - y^e, 0\}$ with $m^e = (1-p)*(w_B+x)+p*(w_B-f)$ and $y^e = (1-p)*(w_A-x)+p*w_A$ for $x > 0$. If $\beta > \frac{1}{2}$, agent B who maximizes his expected utility tries to perfectly equate m^e and y^e by choosing x. Hence, agent B takes more the higher p or the higher f. Nevertheless, deterrence by strong incentives may still occur in this illustration as x is bounded from above by w_A.[15]

Model 4: A model of fairness concerns (regarding final outcomes) that are crowded out by extrinsic incentives

The literature on crowding out of intrinsic motivation by extrinsic incentives uses the term "intrinsic motivation" very broadly. It may apply to fairness concerns as well. In our context crowding out implies that agents' fairness concerns decrease in the intensity of deterrent incentives. Formally, this assumption can be captured by the

[14] Consequently, the evaluation of a state is not independent of another state.

[15] If p and f are so high such that taking everything would generate $m^e < y^e$ with $m^e < w_B$, taking nothing is optimal.

following utility function:

$$V = \lambda(p, f) * U(m) + [1 - \lambda(p, f)] * \widetilde{U}(m, |m - y|),$$

where, as above, $U(m)$ represents the utility of a selfish agent and $\widetilde{U}(m, |m - y|)$ the utility of an agent with fairness concerns regarding final outcomes. The core of the crowding out assumption is that $\lambda(p, f) \in [0, 1]$, the weight of $U(m)$, increases in p and in f.

With these assumptions, there may be ranges of p, f combinations such that the optimal amount $x^*(p, f)$ *strictly* increases in p and f. Therefore, the deterrence hypothesis does not necessarily hold.

The intuition is that for small values of p and f, agent B is relatively fair-minded and takes a small amount, while for higher values of p and f, he is rather selfish and takes a high amount. If the level of incentives is very high such that both selfish and fair-minded subjects are deterred, agent B does not take anything.

If we observe crowding out by deterrent incentives, our data may contribute to the following two aspects of crowding out on which the verdict is still out.

(i) Continuity of crowding out: $\lambda(p, f) \in [0, 1]$ may increase continuously or discontinuously in p and f. Even if it increases continuously, $x^*(p, f)$ may increase discontinuously in p and f for some $\widetilde{U}(m, |m - y|)$.

The empirical results of Gneezy and Rustichini (2000b) and Gneezy (2003) suggest discontinuous crowding out. Frey and Oberholzer-Gee (1997), however, explain their data by assuming continuous crowding out.

In our context subjects who increase their taken amount to a level strictly less than the maximal amount of w_A as a reaction to an introduction or an increase in incentives are evidence for continuous rather than discontinuous crowding out.

(ii) Hysteresis: Extrinsic incentives may crowd out fairness concerns lastingly. As a consequence, the crowding out effect of an increase in incentives is larger than the crowding in effect of the *subsequent* decrease in incentives that reverses the increase in incentives by size.

Some studies (e.g. Irlenbusch and Sliwka, 2005; Gneezy and Rustichini, 2000a; Gächter et al., 2006) find evidence for hysteresis, i.e. evidence that incentives crowd out fairness concerns lastingly.

To give an example, if subjects in our setting take more in a given treatment with small incentives when it is played in part 2 after a part 1 with relatively strong incentives than when it is played in part 1, this is evidence for hysteresis. If we find both backfiring of incentives *and* hysteresis, our data can only be explained by a model of lasting crowding out of fairness concerns and not by a model of fairness concerns regarding expected outcomes. Thus, hysteresis might be a mean to distinguish between these two models (models 3 and 4) that can explain backfiring of incentives.

1.3.2 Hypotheses

The behavioral predictions of the various models differ. However, all four models predict Hypothesis 1.1.

Hypothesis 1.1: Deterrence by strong incentives

Relatively high values of the detection probability p and the fixed fine f deter agents from taking. The range of these values is larger, if an agent is risk averse.

The threshold of strong incentives may vary by subject. A risk neutral or risk averse selfish agent abstains from taking in treatments T5 and T6. A risk neutral or risk averse agent with Fehr and Schmidt (1999) fairness preferences may even abstain from taking in treatment T4.

In contrast to Hypothesis 1.1, only the self-interest model and the model of fairness concerns regarding final outcomes necessarily predict Hypothesis 1.2.

Hypothesis 1.2: Deterrence hypothesis

The taken amount x monotonically (weakly) decreases in the detection probability p and the fixed fine f.

Hypothesis 1.2 implies that the average taken amount x (weakly) decreases from treatments T1 to T6.

1.4 Results

In a first step, we compare behavior in part 1 across subjects. This step has the advantage that the analyzed behavior is not influenced by any preplay. However, we cannot draw any conclusion how (the same) subjects react to a change of incentives, and whether hysteresis occurs. In a second step, we address these issues by comparing behavior in part 1 with behavior in part 2.

1.4.1 Comparison of treatments in part 1

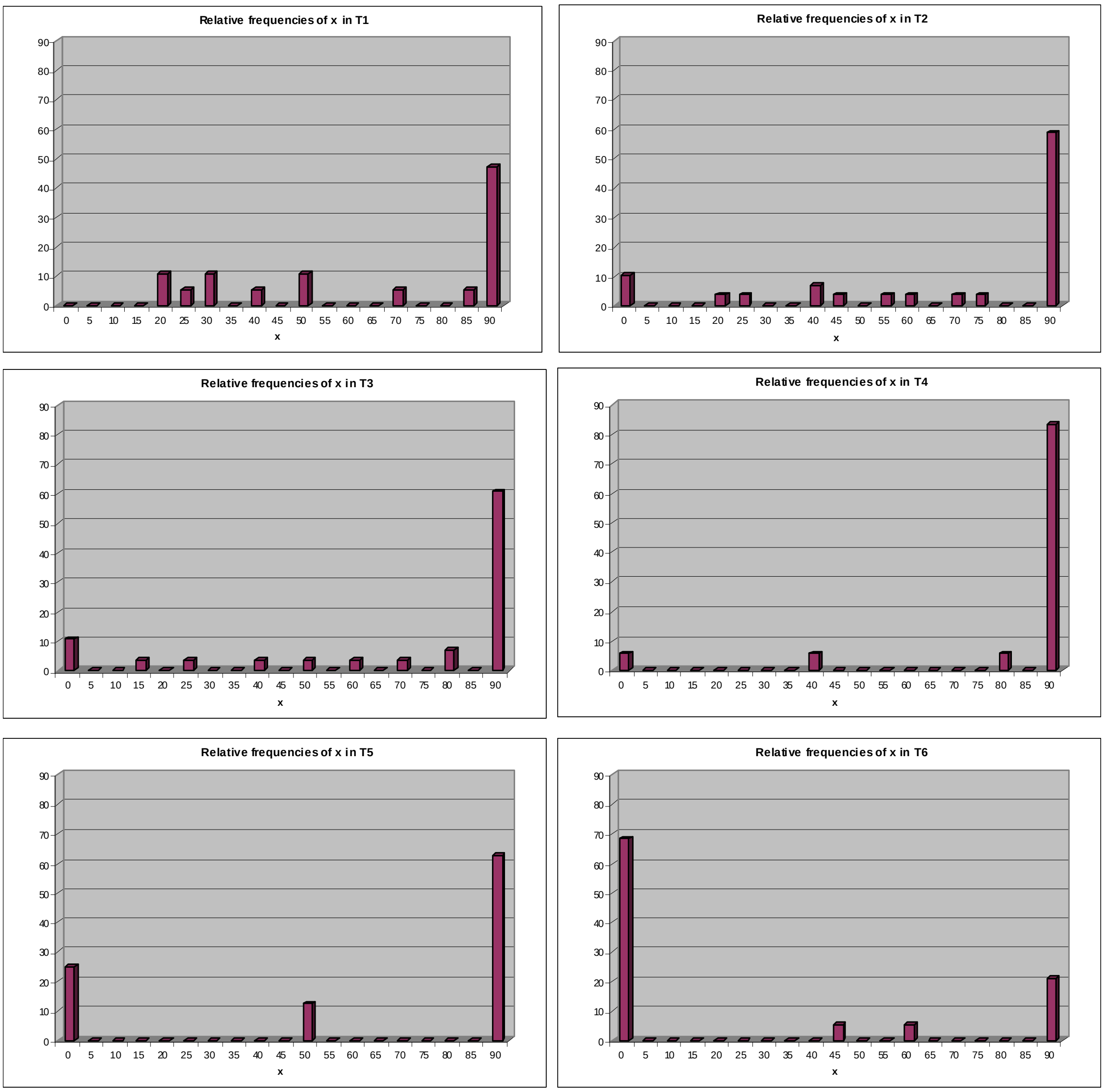
Summary statistics – Benchmark treatment

The experimental data of treatment T1 show how much people take in the absence of deterrent incentives. The upper left panel of Figure 1.2 summarizes the distribution of the taken amount x in the benchmark treatment.

As treatment T1 is the mirror image of a dictator game, we can compare behavior in T1 with standard results of dictator games as those from Forsythe et al. (1994). In line with their paper, we can identify two types of agents: selfish agents and fair-minded agents. In their benchmark treatment (the paid dictator game conducted in April with a pie of 5 \$) about 45 % of subjects are pure gamesmen who do not give anything, and the rest gives a strictly positive amount. These types of agents correspond remarkably well to the 47 % (52.5 %) of selfish subjects in treatment T1 who take everything (between 80 and 90), and the rest who takes a strictly positive amount below 90 (80).

To summarize, we have two types of agents: slightly less than 50 % of our subjects have selfish preferences while a bit more than 50 % have fairness concerns. As the model of fairness concerns regarding final outcomes shows, fairness concerns do not necessarily imply that the deterrence hypothesis fails. It may fail if fairness concerns are based on expected outcomes or if they are crowded out by deterrent incentives.

Figure 1.2: Distribution of the taken amount by treatment (in intervals of size 5)



Relative frequencies of x in T1

Relative frequencies of x in T2

Relative frequencies of x in T3

Relative frequencies of x in T4

Relative frequencies of x in T5

Relative frequencies of x in T6

Interval *i < 90* denoted on the horizontal axis is the union of all *x*∈ *[i,i+5)*.

Interval *i = 90* denoted on the horizontal axis considers *x = 90* only.

To figure out whether the deterrence hypothesis holds, we have a closer look at the treatments with deterrent incentives.

Summary statistics – Treatments with deterrent incentives

Figure 1.3 summarizes the average taken amount per treatment. Our treatments are ordered by the intensity of deterrent incentives, i.e. the combined effect of the detection probability p and the fine f (compare Table 1.1).

The average taken amount increases in the range of no, small, and intermediate incentives (from T1 to T4), while it decreases in the range of strong and very strong incentives (T5 and T6). Hence, the relationship between the average taken amount and the intensity of deterrent incentives is rather inverted-U shaped than monotonically decreasing.

Figure 1.2 shows that the fraction of subjects taking everything increases by treatment from T1 to T4. In treatment T4 it peaks at a value of more than 80 % which is considerably higher than the corresponding 47 % in the absence of any incentives as in treatment T1. From treatment T5 onwards, this fraction decreases.

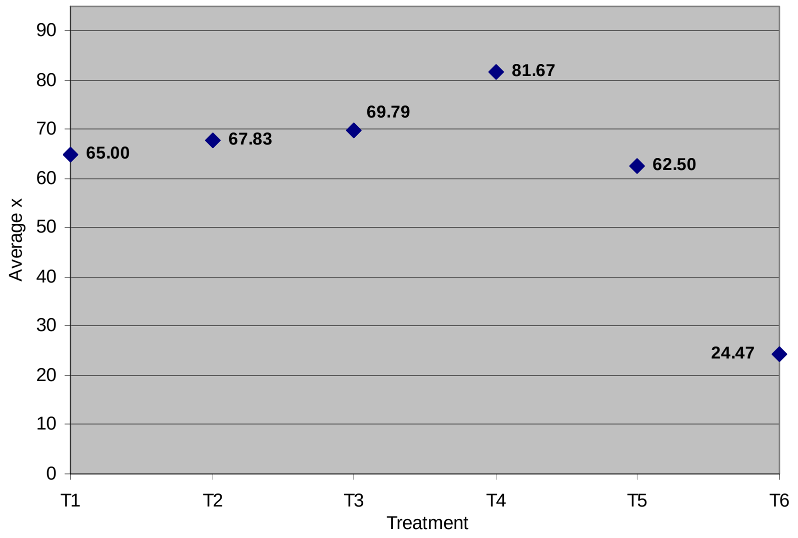
Still, the share of subjects not taking anything monotonically increases in the level of incentives. It is moderate with no, small and intermediate incentives ($\leq$ 10 %), quite substantial with strong incentives (about 25 %), and largest with very strong incentives (nearly 70 %).

Interestingly, there are always subjects taking interior values of their choice set, most so in the benchmark treatment. The share of these subjects decreases in the intensity of incentives. Moreover, the average of the chosen interior values increases in the order of the treatment from T1 to T4.

Compared to the benchmark treatment, deterrent incentives shift mass to the borders of the choice set. We observe both backfiring of small incentives and deterrence at the same time.[16] Small and intermediate incentives move mass predominately towards the upper border which stands in sharp contrast to the deterrence hypothesis but is

[16] In an experiment on corruption that uses probabilistic incentives as we do Schulze and Frank (2003) observe a similar pattern in their data.

Figure 1.3: Average taken amount by treatment



consistent with models 3 and 4. Strong and very strong incentives move mass exactly to the lower border which is consistent with Hypothesis 1.1.

Since the results of treatments T2 and T3 are very similar, the detection probability and the fine seem to be interchangeable instruments.

Analysis of hypotheses

A Kruskal-Wallis test on behavior in part 1 documents significant ($p < 0.01$) treatment effects. In order to identify and characterize the significant differences, we run pairwise Mann-Whitney-U tests. The one-sided p-values are recorded in Table 1.3.

Table 1.3: One-sided p-values of pairwise Mann-Whitney-U tests on taking

	T2	T3	T4	T5	T6
T1	0.287	0.234	0.015	0.400	< 0.001
T2		0.408	0.040	0.447	< 0.001
T3			0.058	0.390	< 0.001
T4				0.071	< 0.001
T5					0.005

In treatment T6 agents take significantly ($p < 0.01$) less than in any other treatment. This is consistent with Hypotheses 1.1 and 1.2. However, contradictory to Hypothesis

1.2 – the deterrence hypothesis – agents take significantly more in treatment T4 than in treatments T1 ($p < 0.05$), T2 ($p < 0.05$) and T3 ($p = 0.058$).[17] There is no significant difference in behavior between treatments T2 and T3.

In order to account for individual characteristics when comparing treatments, we estimate two specifications whose results are presented in Table 1.4.

Table 1.4: Regression results on taking

Dependent variable: x	OLS-r	Tobit
Intercept	+057.15***	+094.56***
Sex (1 if male, 0 else)	+011.51*	+025.70
Risk aversion (1 if risk averse, 0 else)	- 013.73**	- 054.78**
Inconsistent risk pref. (1 if inconsistent, 0 else)	- 027.29	- 105.99*
Economist (1 if economist, 0 else)	+010.12	+031.33
DG (donated amount in part 3)	- 000.12	- 000.51
T2	+010.79	+035.97
T3	+009.25	+028.02
T4	+019.80**	+096.04**
T5	- 007.22	- 020.41
T6	- 042.84***	- 132.32***
Number of observations	129	129
(Pseudo) R-squared	0.3098	0.0771

Ti: 1 if treatment = Ti, 0 else
Inconsistent risk pref.: several switching points in the Holt and Laury (2002) table
*, **, *** significant at 10, 5, 1 percent significance level
-r with robust standard errors

First, we regress the taken amount x on individual characteristics and treatment dummies using an OLS estimation with robust standard errors. Second, we address the fact that the taken amount x is truncated and estimate a Tobit specification with the same regressors. In both estimations the treatment dummy for T4 is significantly positive ($p < 0.05$), the treatment dummy for T6 is significantly negative ($p < 0.05$), and the treatment dummies for T2 and T3 are not significantly different from each other. Hence, these results are robust. Risk aversion has a significantly negative effect ($p < 0.05$) on the taken amount in both specifications (as risk averse subjects are more likely to be deterred).[18]

[17] One-sided Kolmogorov-Smirnov tests and χ^2-tests based on a grouping of subjects according to whether they are deterred, try to roughly equate payoffs (take between 15 and 29 points), have some fairness concerns (take between 30 and 79 points), or are selfish (take between 80 and 90 points) largely confirm the results of the Mann-Whitney-U tests presented here. In particular, subjects always take significantly more in treatment T4 than in T1.

[18] 5 subjects in the role of agent B indicated several switching points in the Holt and Laury (2002)

Given the results of the Mann-Whitney-U tests and the regressions, we do not reject Hypotheses 1.1 but we reject Hypothesis 1.2, the deterrence hypothesis.

Result 1.1: Deterrence by strong incentives

Very strong incentives as in treatment T6 significantly reduce the taken amount. On average, risk averse agents take significantly less.

Result 1.2: Backfiring of small incentives

The average taken amount does not monotonically (weakly) decrease in the level of deterrent incentives. Intermediate incentives as in treatment T4 significantly increase the average taken amount.

Result 1.3: Interchangeability of detection probability and fine

We do not find any significant differences between treatments T2 and T3. In that sense, the detection probability p and the fine f seem to be interchangeable policy instruments.

In sum, these results are consistent with the predictions of models 3 and 4.

1.4.2 Comparison of behavior in part 1 with behavior in part 2

Up to now, we have compared our treatments across *different subjects* in part 1. In contrast to the deterrence hypothesis, our results so far show that intermediate incentives backfire. A model of fairness concerns regarding expected outcomes or a model of fairness concerns that are crowded out by incentives can explain this phenomenon. Since each subject sequentially participated in two different treatments, we can further analyze how *the same subjects* react to a change of deterrent incentives.[19] Sessions in which we increase the intensity of incentives from part 1 to part 2 allow analyzing (i)

table. Since we could not classify them as risk averse, risk neutral, or risk loving, we report them as individuals with inconsistent risk preferences.

[19] Since subjects do not get any feedback after part 1, behavioral effects cannot be triggered by the realization of punishment.

whether backfiring of small and intermediate incentives is observed on an individual level and (ii) whether backfiring is a continuous or discontinuous process. Sessions in which we decrease the intensity of incentives from part 1 to part 2 enable us to check whether we observe hysteresis. Hysteresis can be explained by lasting crowding out of fairness concerns, but it is inconsistent with the model of fairness concerns regarding expected outcomes. Sessions with incentives of the same intensity in both parts indicate whether p and f are interchangeable instruments on an individual level.

Backfiring of incentives on an individual level

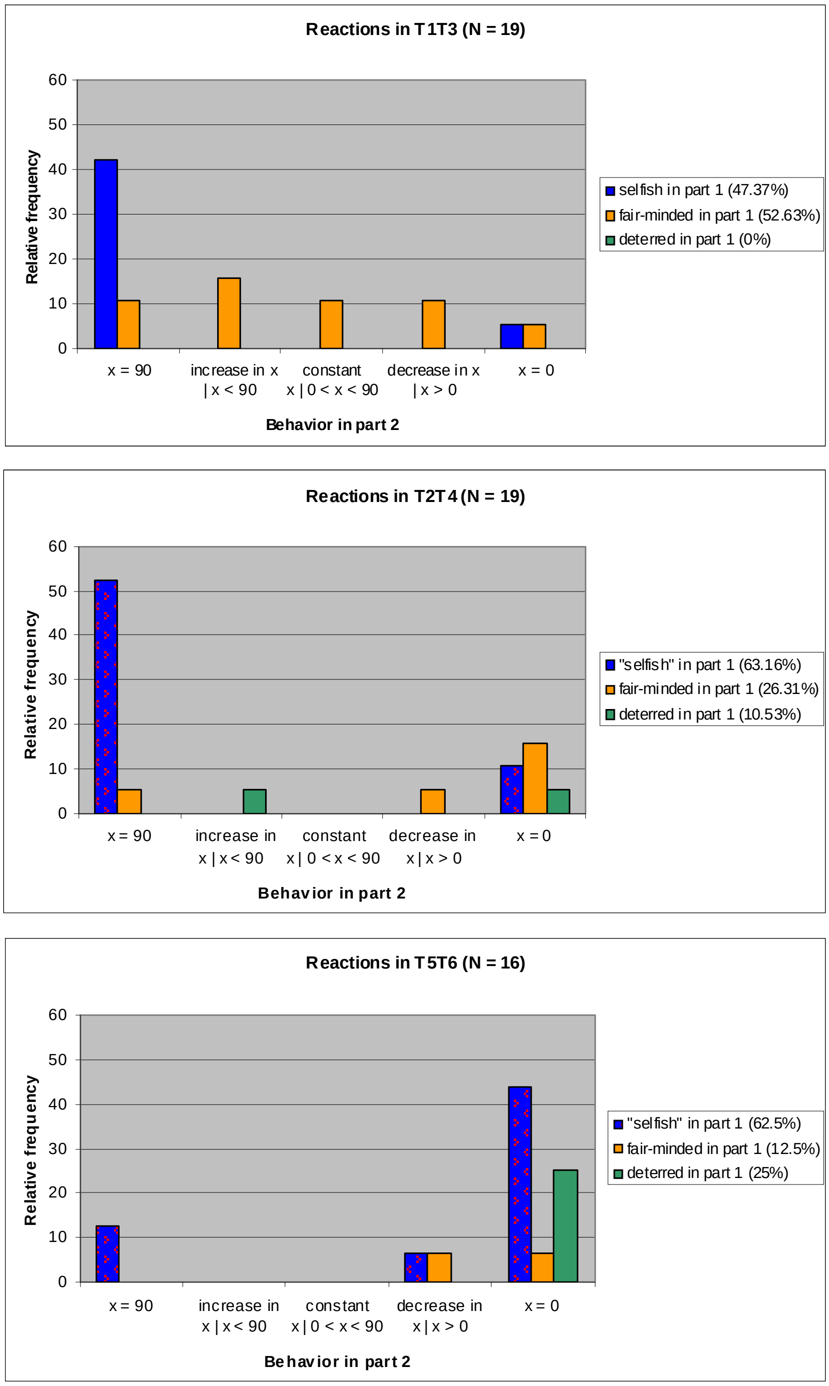
In three different sessions we increase the intensity of incentives from part 1 to part 2: in session T1T3 from no to small incentives, in session T2T4 from small to intermediate incentives, in session T5T6 from strong to very strong incentives. Figure 1.4 summarizes how subjects behave in part 2 conditional on whether they acted selfishly ($x = 90$), acted fair-mindedly ($0 < x < 90$) or were deterred ($x = 0$) in part 1.

Since the benchmark treatment was played in the first part of session T1T3, we can identify about 47 % of subjects with selfish preferences. All except one take everything in part 2 again. About 53 % of all subjects take a positive amount strictly less than everything in part 1. About a third of them increase the taken amount x to a level smaller than 90, a fifth switch to taking everything in part 2, and another fifth keep x constant. Hence, for 50 % of fair-minded subjects small incentives seem to strictly backfire. Only one selfish and one fair-minded subject are deterred by small incentives.

In session T2T4 about 63 % of subjects take everything already in part 1. We cannot distinguish whether they have selfish preferences or fairness concerns which are completely crowded out by the small incentives present in part 1. Again, the majority of these subjects is not deterred and keeps taking everything in part 2. The share of subjects taking intermediate amounts in part 1 is considerably smaller than in session T1T3. For 20 % of these subjects the increase of incentives completely backfires. The majority, however, is deterred. Note that a moderate fraction of deterrence can already be found in part 1.[20]

[20] None of the proposed models can explain the behavior of subjects in sessions T1T3 and T2T4

Figure 1.4: Reactions to an increase in the intensity of incentives



In session T5T6 62.5 % of subjects still take everything in part 1. More than two thirds of them are deterred by the increase of incentives, though. 25 % of all subjects are deterred in part 1 and stay deterred in part 2. Only 12.5 % of subjects take a strictly positive amount below 90 in part 1. Half of them are deterred in the second part.

These observations can be summarized by the following two results:

Result 1.4: Backfiring of small incentives on an individual level

Subjects seem to be heterogeneous. There are selfish agents for which the deterrence hypothesis holds. However, there are also fair-minded agents for which small and intermediate incentives backfire. Independent of the type of agent, strong incentives deter.

Result 1.5: Continuous and discontinuous backfiring of incentives

We find evidence for both continuous and discontinuous backfiring of incentives.

Hysteresis

Whether hysteresis (lasting crowding out of fairness concerns) is present in our data can be seen by comparing behavior of a given treatment played in part 1 with behavior of the same treatment played in part 2 after a part 1 with stronger incentives. Hysteresis implies that we observe sequence effects for these treatments. Table 1.5 records two-sided p-values of pairwise Mann-Whitney-U tests that compare the *same* treatment played in *different* parts of a session.[21]

As Table 1.5 indicates, we observe sequence effects in treatments T1, T2, and T5. Subjects in T1 take significantly ($p < 0.05$) more when it is played after T3 (81.3 instead of 65.0 points on average). Preplay in T3 with small incentives increases the average taken amount in treatment T1. Similarly, the average taken amount in T2 is

who react to increased incentives by decreasing the taken amount to a level strictly larger than 0 or increasing it from 0 to a strictly positive amount.

[21] Treatments T2 and T3 are played second in two different sessions. Since the observations from the second parts are not significantly different (p=0.71 and p=0.34, respectively according to two-sided Mann-Whitney-U tests) for different sessions, we do not report each session comparison separately.

Table 1.5: Non-parametric comparisons of different sequences (Mann-Whitney-U tests)

Treatment	played first in	played second in	p-value (two sided)
T1	T1T3	T3T1	0.082
T2	T2T3 T2T4	T3T2 T4T2	0.099
T3	T3T1 T3T2	T1T3 T2T3	0.676
T4	T4T2	T2T4	0.061
T5	T5T6	T6T5	0.014
T6	T6T5	T5T6	0.617

significantly ($p < 0.05$) higher when it is played second (after a harsher or a constant intensity of incentives) than first. Both results are consistent with a model of lasting crowding out of fairness concerns but cannot be reconciled with the predictions of a model of fairness concerns regarding expected outcomes. In contrast, subjects in T5 take significantly ($p < 0.05$) less when it is played after T6. Preplay in T6 with very strong incentives seems to increase deterrence in treatment T5. This is inconsistent with a model of fairness concerns regarding expected outcomes and with lasting crowding out of fairness concerns (if fairness concerns imply less taking in T5 than selfish concerns imply).

Result 1.6: Hysteresis

Small and intermediate incentives have a lasting effect. They still backfire when incentives are decreased or even removed in the following period.

Since we observe hysteresis, a model of fairness concerns that are crowded out by incentives explains our data better than a model of fairness concerns regarding expected outcomes. Hysteresis also underlines how costly extrinsic incentives are. In addition to the effect incentives have in the current period, they may also influence behavior in future periods. From this perspective, also strong and very strong incentives could backfire by crowding out fairness concerns in future periods in which incentives are smaller.

In treatments with an increase in incentives there are no significant sequence effects for treatments T3 and T6. Subjects in treatment T4, however, take significantly ($p < 0.05$) less when it is played in part 2 after treatment T2 than when it is played in part 1.

Substitutability of detection probability and fine

Since treatments T2 and T3 have the same intensity of deterrent incentives implemented by different values of the detection probability p and the fine f, we can test – at least for this specific level of incentives – whether these two instruments are interchangeable. We have already observed that treatments T2 and T3 do not differ significantly across subjects in part 1 (Result 1.3). Our within subject analysis in Figure 1.5 also finds this result on an individual level.

In session T2T3 7 out of 10 subjects do not change their behavior. In session T3T2 only a single subject is apart from the 45° line. 6 subjects keep taking everything, 2 keep taking the same intermediate amount.

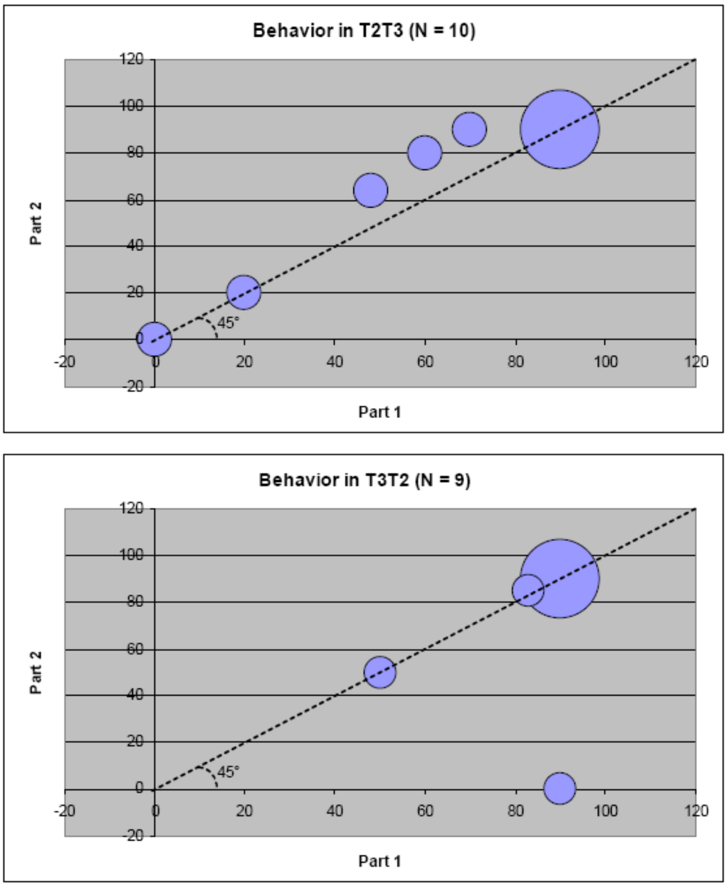
Result 1.7: Interchangeability of detection probability and fine on an individual level

Our within subjects comparison confirms Result 1.3 that p and f seem to be interchangeable instruments.

1.5 Robustness check - Framing

So far, we have presented results from neutrally framed experiments. This is a valid approach to test the deterrence hypothesis which relies on incentive effects that are independent of all other factors that may influence crime as e.g. the frame. While a non-neutral frame may ceteris paribus affect the taken amount (e.g. due to additional moral costs), comparative statics should remain unchanged. For any given (neutral or non-neutral) frame, the deterrence hypothesis predicts the taken amount to monotonically decrease in the detection probability and the fine. However, it is not clear whether

Figure 1.5: Reactions to a change of incentives keeping their intensity constant



Behavior in T2T3 (N = 10)

Behavior in T3T2 (N = 9)

a non-neutral frame interacts with incentives in a model of fairness concerns regarding final outcomes that are crowded out by incentives which fits our data best. While neutrally framed incentives crowd out fairness concerns, this may not necessarily be the case for incentives that are combined with a strong moral connotation.

In real life deterrent incentives often have a moral connotation and policy makers may try to make use of that. This is why we run two additional, morally framed sessions and check whether a non-neutral, moral frame changes our results. In the morally framed sessions B's decision was labeled as "stealing" if $x > 0$ and the fixed fine f was called "penalty" instead of "minus points". Apart from these two different labels, the neutrally and morally framed sessions were conducted completely identically. In order to check whether framing affects behavior in the absence of incentives, we run a morally framed version of treatment T1 (T1f). To analyze whether framing and incentives interact, we run a morally framed version of treatment T4 (T4f).[22] 38 subjects participated in session T1fT4f, 32 subjects in session T4fT1f.

The results in the morally and neutrally framed treatments are similar. There is no significant framing effect in part 1 in the absence of incentives, i.e. between T1 and T1f (two-sided Mann-Whitney-U test: $p > 0.5$). In contrast, subjects take more in part 1 in treatment T4 than in treatment T4f (two-sided Mann-Whitney-U test: $p = 0.075$). There is no significant difference in parts 1 between treatments T1f and T4f. However, the within subjects analysis documents crowding out: when incentives are introduced in part 2 of session T1fT4f more than 30 % of subjects flip from taking intermediate amounts to taking everything. This parallels the results obtained in the neutrally framed sessions T1T3 and T2T4. We conclude that also with moral framing backfiring of intermediate incentives is a non-negligible phenomenon.

[22] We choose treatment T4 since the intensity of deterrent incentives in this treatment is (i) low enough to not deter the majority of subjects and (ii) high enough to potentially crowd out fairness concerns significantly.

1.6 Conclusion

We have presented an experimental test of the deterrence hypothesis applied to the context of stealing. Our across subjects analysis of part 1 rejects the hypothesis that the average taken amount monotonically (weakly) decreases in deterrent incentives. On average, subjects take most when intermediate incentives are present. Only very strong incentives deter.

Both our across subjects comparison of behavior in part 1 and our within subjects comparison of behavior in part 1 with behavior in part 2 reflect two different types of subjects. We identify about 50 % selfish subjects whose behavior is consistent with the deterrence hypothesis and about 50 % fair-minded subjects for which intermediate incentives backfire. Since we observe hysteresis, a model of lasting crowding out of fairness concerns explains our data best.

We have contributed to the empirical literature on crowding out in various ways. First, we observe crowding out of fairness concerns in a very simple setting. Second, we have established the existence of crowding out as a reaction to probabilistic incentives[23] and in a new domain, namely when incentives are set to deter criminal activities. Third, our comparison of behavior in part 1 with behavior in part 2 provides further evidence for lasting crowding out as it is observed by Irlenbusch and Sliwka (2005), Gneezy and Rustichini (2000a), and Gächter et al. (2006). While it exists for many subjects, we have also observed some subjects whose fairness concerns are – at least partially – reestablished when incentives are reduced or removed completely. Fourth, our study has explicitly focused on the domain of small and intermediate incentives that are especially important in real life[24]: We have run four out of six treatments with small and intermediate incentives which according to standard neoclassical theory do not deter risk neutral subjects. Thus, we have several treatments to analyze whether

[23] To our knowledge the only other paper that documents the existence of crowding out of intrinsic motivation due to probabilistic incentives is Schulze and Frank (2003).

[24] In Germany the clearance rate for thefts with (without) aggravating circumstances was 14 % (44 %) in 2005 (Polizeiliche Kriminalstatistik, 2005, Table 23). Andreoni et al. (1998) present figures for tax evasion in the US: In 1995 the audit rate for individual tax return was only 1.7 %, the penalty for underpayment of taxes usually 20 % of the underpayment. Polinsky and Shavell (2000b) point out that, in general, the severity of punishment is quite low in relation what potential offenders are capable to pay.

crowding out is a continuous or discontinuous process. Our within subject analysis finds evidence for both.

Interestingly, incentives – even in this very simple and plain context – backfire. Tversky and Kahneman's (1986) argument that extrinsic incentives shift the context from an ethical and other-regarding to an instrumental and self-regarding one seems to be adequate for our results. Similarly, the findings confirm those of Houser et al. (2008) who show that crowding out of intrinsic motivation is not only caused by the intentions that incentives signal but also by incentives *per se*.

What are the policy implications from our experimental study? Taking our data literally would imply to punish criminal behavior either hard or not at all in order to avoid backfiring of incentives. Of course, the laboratory may abstract from social norms and stigmata that could be the driving forces behind punishment in reducing criminal behavior. Thus, we do not conclude that punishment does not work outside the laboratory. However, our data directly reject the deterrence hypothesis that relies on punishment whose effectiveness is independent of all other factors that may influence crime. Our results show that if crime was a gamble – as economists generally argue and as we have modeled it in the laboratory – incentives may not work: Especially small and intermediate incentives backfire and may crowd out fairness concerns lastingly. Thus, to convincingly contribute to the discussion on how to efficiently deter crime, economists should go beyond the standard deterrence hypothesis.

1.7 Appendix

1.7.1 Experimental sessions

The order of events during each experimental session was the following: Subjects were welcomed and randomly assigned a cubicle in the laboratory where they took their decisions in complete anonymity from the other participants. The random allocation to a cubicle also determined a subject's role in all three parts. Subjects were handed out the general instructions for the experiment as well as the instructions for part

1. After all subjects had read both instructions carefully and all remaining questions had been answered, we proceeded to the decision stage of the first part. Part 2 and 3 were conducted in an analogous way. We finished each experimental session by letting subjects answer a questionnaire that asked for demographic characteristics and included a paid Holt and Laury (2002) table. This table was explained in detail in the questionnaire and it was highlighted that one randomly drawn decision from the table was paid out in addition to the earnings in the previous parts.

Instructions, the program, and the questionnaire were originally written in German. The translated general instructions, the translated instructions of the neutrally framed treatment T4 in part 1 for agent B, and the translated Holt and Laury (2002) table can be found in the following. Instructions for part 2 and part 3 are as similar to part 1 as possible. For the framed treatments, we used the expressions "steal any integer amount between 0 and 90 from participant A" instead of "choose any integer amount between 0 and 90 that shall be transferred from participant A to you", and the term "minus a penalty of x points" instead of "minus an amount of x points".

1.7.2 Translated general instructions

General explanations concerning the experiment

Welcome to this experiment. You and the other participants are asked to make decisions. Your decisions as well as the decisions of the other participants will determine the result of the experiment. At the end of the experiment you will be paid **in cash** according to the **actual** result. So please read the instructions thoroughly and think about your decision carefully.

During the experiment you are not allowed to talk to the other participants, to use cell phones or to start any other programs on the computer. The neglect of these rules will lead to the immediate exclusion from the experiment and all payments. If you have any questions, please raise your hand. An experimenter will then come to your seat to answer your questions.

During the experiment we will talk about points instead of Euros. Your total income will therefore be calculated in points first. At the end of the experiment the total amount of points will be converted into Euros according to the following exchange rate:

1 point = 15 Cents.

The experiment consists of three **independent** parts in which you can accumulate points. Before each part only the instructions of this part will be handed out.

During the experiment neither you nor the other participants will receive any information on the course of the experiment (e.g. decisions of other participants or results of a particular part).

The results of each single part will be calculated only after all three parts will be finished. **Then, one of these three parts will be chosen randomly. At the end of the whole experiment only this part will be paid out in cash according to your decisions.**

1.7.3 Translated instructions of the neutrally framed treatment T4 in part 1

Part 1

In this part there are **participants in role A and participants in role B. You have been randomly assigned role B for this part. You will be randomly and anonymously matched to another participant in role A.** This random matching lasts only for this part. The matched participant will not be matched to you in the following two parts again. Neither before nor after the experiment will you receive any information about the identity of your matched participant. Likewise, your matched participant will not receive any information about your identity.

As participant B you have an initial endowment of 50 points. Participant A has an initial endowment of 90 points.

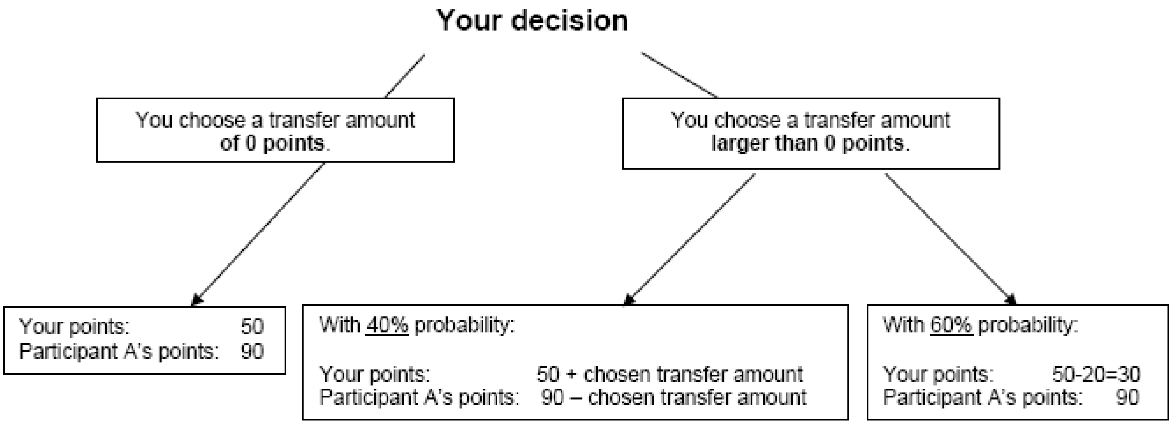
As a participant in role B you can choose any **integer amount** between 0 and 90 points (including 0 and 90) **which shall be transferred from participant A to you**. Participant A does not make any decision. In order to make your decision, please enter your chosen amount on the corresponding computer screen and push the OK button.

- If you choose a transfer amount of 0 points, you will receive your initial endowment of 50 points, and participant A will receive his initial endowment of 90 points.
- If you choose a transfer amount larger than 0 points,
  - **with *40 %* probability you will receive your initial endowment of 50 points plus your chosen transfer amount and participant A will receive his initial endowment of 90 points minus your chosen transfer amount.**
  - **with *60 %* probability you will receive your initial endowment of 50 points minus an amount of 20 points, i.e. 30 points, and participant A will receive his initial endowment of 90 points.**

Example 1: You choose a transfer amount of 22 points. With *40 %* probability you will receive 50 + 22 points = 72 points, and with *60 %* probability you will receive 50 – 20 points = 30 points. Participant A will receive 90 – 22 points = 68 points with *40 %* probability and his initial endowment of 90 points with *60 %* probability.

Example 2: You choose a transfer amount of 0 points. You will receive 50 points. Participant A will receive 90 points.

The course of action of part 1 is illustrated by the following figure:


Your decision
You choose a transfer amount **of 0 points**.
You choose a transfer amount **larger than 0 points**.
Your points: 50
Participant A's points: 90
With 40% probability:
Your points: 50 + chosen transfer amount
Participant A's points: 90 – chosen transfer amount
With 60% probability:
Your points: 50-20=30
Participant A's points: 90

If you have any questions, please raise your hand. An experimenter will come to your seat to answer your questions.

1.7.4 Translated Holt and Laury (2002) table

Decision	Option A	Option B
Decision 1	10 points	25 points with a probability of 10 % 0 points with a probability of 90 %
Decision 2	10 points	25 points with a probability of 20 % 0 points with a probability of 80 %
Decision 3	10 points	25 points with a probability of 30 % 0 points with a probability of 70 %
Decision 4	10 points	25 points with a probability of 40 % 0 points with a probability of 60 %
Decision 5	10 points	25 points with a probability of 50 % 0 points with a probability of 50 %
Decision 6	10 points	25 points with a probability of 60 % 0 points with a probability of 40 %
Decision 7	10 points	25 points with a probability of 70 % 0 points with a probability of 30 %
Decision 8	10 points	25 points with a probability of 80 % 0 points with a probability of 20 %
Decision 9	10 points	25 points with a probability of 90 % 0 points with a probability of 10 %
Decision 10	10 points	25 points with a probability of 100 % 0 points with a probability of 0 %

Participants made 10 separate decisions whether they preferred option A to option B. While option A was the same in all 10 decisions, option B varied in the associated probabilities as displayed above. At the end of the experiment one decision was chosen randomly (all with equal probability) and paid.

In the following we have a look at the decisions of agents B in the neutrally and morally framed treatments. We classify the observed 51 subjects who prefer option A to option B in decisions 1 to 4 and option B to option A otherwise as risk-neutral. The observed 16 subjects preferring option A in decisions 1 to k, with $k < 4$, and option

B else are categorized as risk-seeking. We observe 88 risk-averse subjects indicating option A in decisions 1 to k, with $k > 4$, and option B else. Three subjects behave irrationally in the sense that they always prefer option A to option B, even in decision 10.

Chapter 2

An experimental study on information flows in teams*

2.1 Introduction

Team members often work inefficiently little as their rewards are related to the team output but the costs of contributing to the team output accrue to the individual member only. In this study we address the question whether team members are stimulated to work more when one member receives a signal on his colleague's performance prior to his own decision. Furthermore, we analyze how the signal's type affects behavior.

In many situations agents of a team act sequentially and the second mover observes some signal on the first mover's performance. The signal may affect the second mover's incentives and, thereby, provide the first mover with commitment power. Consider, for example, the case where individual contributions to the team output are substitutes (i.e. the more one agent contributes, the lower are the incentives for the other agent to work hard) and the second mover can perfectly observe the first mover's contribution. If all agents are self-interested, it is optimal for the first mover to work very little in order to induce the second mover to work hard. However, if agents are concerned about fairness or reciprocity, it may be optimal for the first mover to work very hard for the project in order to induce his fellow worker to work very hard as well. An

*This chapter is joint work with Sandra Ludwig from the University of Munich.

important role may be also played by the signal's type: The second mover may observe how hard the first mover worked for the team project or how much value he added to the team output. Take, for example, the development of a new product consisting of a design and a construction phase. The constructor may observe the effort the designer devoted to designing the product (e.g. if their offices are situated closely together), or the quality of the design (e.g. results of a customer test series), or both. If the constructor's and the designer's wages depend only on the quality of the product, a selfish constructor focuses on the information on the quality of the design, whereas a fair-minded constructor takes also the designer's effort into account. This raises the question how the structure of information flows in teams affects behavior.

In order to put this question to a test, we conduct a series of laboratory experiments in which two team members work sequentially. We observe that on average both agents work more when the second mover receives an informative signal on the first mover's performance. Especially the first mover's effort turns out to be a stimulating signal.

In our experiment a team consists of two agents who sequentially decide on their effort to contribute to the team output. Prior to the second mover's effort choice he observes a signal on the first mover's performance. The higher an agent's effort, the higher is the probability that his contribution to the team output is high rather than low. An agent's payoff equals his wage minus his effort costs. Both agents receive the same wage that positively depends on the team output. The wage scheme is given exogenously, e.g. by a principal or a market that is not modeled in the game.

As we focus on the pure incentive effect of different information flows, we vary the signal across treatments, while we keep everything else constant. Our treatments are defined according to the degree of information the signal provides about the first mover's contribution – the only performance variable of the first mover that is directly payoff relevant for the second mover. Given the first mover's contribution, the second mover's payoff does not depend on the first mover's effort. In our benchmark treatment there is no signal at all, while in all other treatments the signal is either the first mover's effort (i.e. an imperfect signal), or the first mover's contribution (i.e. a perfect signal), or both. Observing both does, however, not provide more information on the first mover's contribution than observing his contribution only. Still, the effort provides additional information about the first mover's payoff.

We implement a wage scheme such that agents' contributions are substitutes. In this case standard theory predicts for all treatments with an informative signal that the second mover's effort is negatively related to the signal[25] and the first mover works less than in the benchmark treatment. We observe, however, that on average the first mover works more when informative signals are available and the second mover positively reacts to the signals. Even if the second mover observes the first mover's contribution and effort, the second mover's effort increases in the first mover's effort. This contradicts standard theory since in this case the first mover's effort is payoff irrelevant for the second mover. Overall, both agents work more on average when informative signals are available.

In an extension we test whether these results are robust to the strategic context. By varying the wage scheme we implement a setting where agents' contributions are complements, i.e. the more one agent contributes to the team output, the higher are the incentives for the other agent to work hard. In this case standard theory predicts for all treatments with an informative signal that the second mover's effort is positively related to the signal and the first mover works more than in the benchmark treatment. Similar to before – but more consistent with standard theory – we find that the first mover works more in the presence of informative signals and the second mover's effort increases in the first mover's effort. The positive relation with the first mover's effort is again inconsistent with standard theory when both the first mover's effort and contribution are observable. As for substitutes, both agents work more when informative signals are available.

Overall, our results indicate that standard theory fails to explain how team members behave when informative signals are available. A simple model of inequity aversion, however, can largely explain our results. Taking into account that agents care about payoff inequalities, the information on the first mover's effort becomes directly relevant for the second mover's utility: The first mover's effort determines the first mover's

[25] If the first mover's effort and contribution are observable, the second mover only reacts to the first mover's contribution.

effort costs and, thereby, payoff inequalities.

The issue of team production when agents can observe signals on team mates' performance has been addressed in theoretical papers relying on standard preferences by Goldfayn (2006), Winter (2006a,b), and Ludwig (2008). They find mixed results on whether sequentially working teams outperform simultaneously working teams in which no signals are observable. Their results crucially depend on the team's production function and the strategic nature of the game. Moreover, the theoretical paper by Huck and Rey-Biel (2006) shows that sequentially working teams outperform simultaneously working teams when agents dislike effort inequalities between team mates. This indicates that also social comparison is crucial in team settings where signals are available. In our empirical analysis we can address the role of the strategic nature of the game, different signals, and social preferences.

Some empirical studies investigate the effect of signals in settings where the agents' payments are completely independent of those agents' actions on whom they receive the signal: Falk and Ichino (2006) study the performance in a real-effort task when individuals work either separately or in a shared room. Sausgruber (2009) investigates behavior in a public good experiment when individuals receive information on the overall contribution of *another* group of individuals. Both studies find that individuals work or contribute more when signals are available.

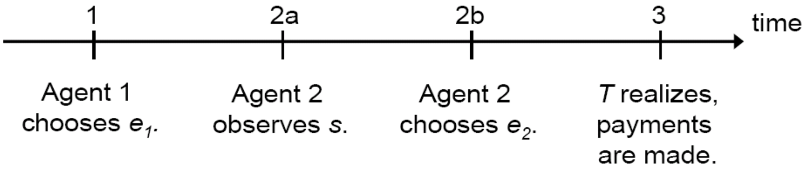
Mohnen et al. (2008), in contrast, consider signals in dynamic teams where the agents' payments directly depend on those agents' actions on whom they receive the signal. Nevertheless, agents' actions in their setting are strategically independent according to standard theory since agents have a dominant strategy. Hence, standard theory predicts that signals on the other agent's action have no effect. They conduct a real-effort experiment in which two agents work simultaneously in two stages successively. After the first stage agents either mutually observe their performance or observe nothing. They find that agents work harder in the first stage and also achieve a better aggregate team performance when agents receive interim information. While in their study both agents send and receive a signal, in our study one agent is the sender and the other one the receiver. Therefore, full commitment of the sender and a reaction without strategic uncertainty of the receiver are only possible in our study.

Similar to Mohnen et al. (2008), the literature on sequential contributions to public goods considers signals within groups. The signal in these settings is equal to one or more predecessors' contributions. In the standard (linear) public good game agents' actions are again strategically independent. Consequently, standard theory predicts no difference between sequential and simultaneous contributions. Standard public good experiments find that a sequential structure alone does not or only slightly increase the contributions to the public good, e.g. Güth et al. (2007), Levati et al. (2007), Potters et al. (2007), Gächter et al. (2008), and Rivas and Sutter (2008). These studies observe, however, that many individuals are conditional cooperators (Fischbacher et al., 2001), i.e. they contribute if others do so but they do not if others do not contribute. In contrast to the aforementioned studies, Gächter et al. (2009) implement a non-standard public good experiment in which actions are strategic substitutes. In line with their theoretical predictions, sequential contributions result in lower overall provision than simultaneous contributions. Nevertheless, there is evidence for conditional cooperation. In contrast to our setup, an agent's investment is deterministically related to his contribution and, in addition, the signal's type does not vary in all these public good experiments.

Huck and Müller (2000) vary the noisiness of the signal a second mover receives on the first mover's action. Their experimental setup considerably differs from a voluntary contribution game: Its payoffs resemble those of a market with quantity competition and are designed such that fairness considerations do not play a major role. They find that behavior converges to the subgame perfect outcome, irrespective of the signal's noise. While their signals differ in the degree of information, our signals differ in the type of information: The first mover's contribution directly affects both agents' wages, the first mover's effort directly affects his effort costs.

This chapter proceeds as follows. In Section 2.2 we present the experimental design and procedure, and in Section 2.3 the behavioral predictions and hypotheses. We summarize our results in Section 2.4. In Section 2.5 we extend our analysis to the case of complements. We discuss our results in Section 2.6 and conclude in Section 2.7.

Figure 2.1: Timeline of the team production setting



2.2 Experimental design and procedure

We consider a team of two agents, agent 1 and agent 2, that generates a team output T. Agents sequentially exert effort: First, agent 1 exerts effort $e_1 \geq 0$. Second, agent 2 observes a signal s and, then, exerts effort $e_2 \geq 0$. Agent i's effort e_i, with $i \in \{1, 2\}$, does not directly influence the team output T but the probability that agent i's contribution to T is high ($b_i = 1$) rather than low ($b_i = 0$), $Pr\,(b_i = 1|e_i) = p\,(e_i)$ with $p'\,(e_i) > 0$. The probability that agent i's contribution to T is low equals $Pr\,(b_i = 0|e_i) = 1 - p\,(e_i)$. $p\,(e_i)$ is independent of e_j. The sum of both agents' contributions generates the team output $T := b_1 + b_2$. Each agent receives a high wage w_H if $T = 2$, i.e. if the contributions of both agents are high, an intermediate wage w_M if $T = 1$, i.e. if the contribution of only one agent is high, and a low wage w_L if $T = 0$, i.e. if the contributions of both agents are low, with $w_H > w_M > w_L = 0$. If agent i exerts effort e_i, he incurs private costs $c(e_i)$ with $c(0) = 0$, $c'(e_i) \geq 0$, and $c''(e_i) > 0$. Agent i's expected payoff is his expected wage minus his effort costs. Figure 2.1 illustrates the timing of the game.

For our experimental design we choose the following parameters. Agent $i \in \{1, 2\}$ could choose effort $e_i \in \{0, 1, ..., 80\}$ with corresponding effort costs $c(e_i) = \frac{e_i^2}{80}$. An effort of e_i yields a high contribution ($b_i = 1$) with probability $p\,(e_i) = 0.1 + 0.01 \cdot e_i$. The agents' wages conditional on the team output are $w_H = 280$, $w_M = 172$, and $w_L = 0$.

In this experimental study we focus on the question how the availability of different signals affects team mates' behavior. Hence, we vary s across treatments and keep everything else constant. Table 2.1 presents our four treatments.

Table 2.1: Treatments of the team production setting

Treatment	Signal	Number of participants
T-NO	no signal	18
T-IMP	e$_1$	34
T-P	b$_1$	36
T-PPlus	e$_1$ and b$_1$	30

As agent 1's contribution – in contrast to his effort – directly affects agent 2's payoff, we define our treatments according to the degree of information the signal provides about agent 1's contribution. In treatment T-NO, our benchmark treatment, there is **no** signal at all, i.e. $s = \{\}$. Hence, agent 1's and agent 2's decision problems are theoretically identical and so agents behave as if they had to decide simultaneously.[26] In treatment T-IMP $s = e_1$ which is an imperfect signal on b_1. In treatment T-P $s = b_1$, i.e. agent 2 receives a perfect signal on b_1. In T-PPlus agent 2 receives $s = (e_1, b_1)$ and is again perfectly informed on b_1. Observing e_1 in addition to b_1 does, however, not provide more information on agent 1's contribution.

In each of our experimental sessions the team production setting was played for 14 rounds. Individuals' roles as well as the treatment conditions were fixed for all rounds. The matching of agents, in contrast, was random such that individuals were not able to distinguish whether their current team mate had already been matched with them in a previous round or not. After each round individuals received feedback concerning $e_{1,}$ e_2, b_1, b_2, and both agents' payoffs of the past round. Therefore, treatment differences cannot be explained by shame. Only one randomly determined round was paid for all participants of a session. At the end of the 14th round individuals were informed which round is paid. All this was told the participants in the instructions that were handed out at the beginning of the experiment and were framed as a team work setting.[27] In a post-experimental questionnaire we elicited data on the participants' sex, subject of

[26] Empirically, the mere knowledge about the physical timing of moves may affect behavior even if the informational condition is equivalent to a simultaneous setting: Duffy et al. (2007), for instance, find that the timing itself matters in a dynamic public good experiment, whereas Masclet et al. (2007) find that it has no effect in a standard public good experiment.

[27] See the appendix for a more detailed description of the procedure of a session and translated instructions. Original instructions are written in German and are available upon request.

studies and risk attitude[28].

Our experimental sessions were run at the Munich Experimental Laboratory for Economic and Social Sciences (MELESSA) in Germany in 2008. 118 individuals participated in the experiment. They were randomly assigned to sessions and could take part in one session only. For the recruitment we used the software ORSEE by Greiner (2004). The experiment was programmed and conducted with the experimental software z-Tree by Fischbacher (2007). The sessions lasted about 1 hour. Individuals earned on average 12.6 € (at the time of the experiment 1 € $\approx$ 1.57 USD), including a show-up fee of 7 €[29].

2.3 Behavioral predictions and hypotheses

In this section we derive behavioral predictions on how agents in a team react when one member receives a signal on his colleague's performance. As a benchmark, we consider the standard neoclassical approach.

2.3.1 Behavioral predictions of a self-interest model

The standard neoclassical approach assumes that all individuals are selfish, i.e. their utility depends only on their own material payoff and increases in this payoff. Furthermore, we assume that all individuals maximize their expected utility and are risk-neutral.

T-NO

If there is no signal at all, we solve for the Nash equilibrium of the simultaneous move game. Agent i's expected payoff given agent j chooses e_j, with $i, j \in \{1, 2\}$ and $j \neq i$,

[28] Individuals indicated on a scale ranging from 0 to 10 whether they are willing to take risks (or try to avoid risks). 0 represented a very weak willingness to take risks, while 10 represented a strong willingness to take risks. Dohmen et al. (2005) show that this general risk question is a good predictor of actual risk-taking behavior.

[29] We chose a relatively high show-up fee as losses were deducted from the show-up fee. This was applied for about 10 % of the participants.

and $w_L = 0$ is

$$U_i\left(e_i, e_j\right) = w_H \cdot p\left(e_i\right) \cdot p\left(e_j\right) + w_M \cdot \left[p\left(e_i\right) \cdot \left(1 - p\left(e_j\right)\right) + \left(1 - p\left(e_i\right)\right) \cdot p\left(e_j\right)\right] - \tfrac{e_i^2}{80}.$$

Maximizing $U_i\left(e_i, e_j\right)$ with respect to e_i, agent i's reaction function is

$$e_i^*(e_j) = 0.4 \cdot [w_M + (0.1 + 0.01 \cdot e_j) \cdot (w_H - 2 \cdot w_M)].$$

$e_i^*(e_j)$ strictly increases in e_j if $w_H > 2 \cdot w_M$, while it strictly decreases if $w_H < 2 \cdot w_M$. Hence, efforts are strategic complements if $w_H > 2 \cdot w_M$, while they are strategic substitutes in our setting as $w_H < 2 \cdot w_M$. Solving for the intersection of both agents' reaction functions and substituting $w_H = 280$ and $w_M = 172$, we derive the unique Nash equilibrium in which $e_1^* = e_2^* = \frac{165.6}{3.14} \approx 52.74$.

T-IMP

In T-IMP agent 2 receives the signal $s = e_1$ and we solve for the subgame perfect Nash equilibrium given $s = e_1$. Agent 2's reaction to $s = e_1$ is equivalent to his reaction function in T-NO, i.e.

$$e_2^*(s) = 0.4 \cdot [w_M + (0.1 + 0.01 \cdot s) \cdot (w_H - 2 \cdot w_M)].$$

Equivalently to above, $e_2^*(s)$ strictly increases (decreases) in $s = e_1$ if $w_H >(<) 2 \cdot w_M$. $e_2^*(s) = e_2^*(e_1)$ maximizes agent 2's expected payoff $U_2\left(e_2, s\right)$ given $s = e_1$. Agent 1 anticipates agent 2's reaction to s and maximizes his own expected payoff

$$\begin{gathered} U_1\left(e_1, e_2^*(e_1)\right) = \\ w_H \cdot p\left(e_1\right) \cdot p\left(e_2^*(e_1)\right) + w_M \cdot \left[p\left(e_1\right) \cdot \left(1 - p\left(e_2^*(e_1)\right)\right) + \left(1 - p\left(e_1\right)\right) \cdot p\left(e_2^*(e_1)\right)\right] - \tfrac{e_1^2}{80} \end{gathered}$$

with respect to e_1. Agent 1's expected payoff is maximized at $e_1 = \frac{1.616256}{0.0434464} \approx 37.20$. Consequently, in the unique subgame perfect Nash equilibrium $e_1 = s \approx 37.20$ and $e_2 = \frac{2.464128}{0.0434464} \approx 56.72$ according to agent 2's reaction ($e_2^*(e_1) = 66.24 - 0.256 \cdot e_1$).

T-P

In T-P we solve for the unique Nash equilibrium given $s = b_1$. Agent 2's expected payoff given agent 1's contribution is high, i.e. $s = b_1 = 1$, is

$$U_2\left(e_2, b_1 = 1\right) = w_H \cdot p\left(e_2\right) + w_M \cdot \left(1 - p\left(e_2\right)\right) - \frac{e_2^2}{80}. \tag{2.1}$$

$U_2\left(e_2, b_1 = 1\right)$ is maximized at $e_2^*(b_1 = 1) = 0.4 \cdot (w_H - w_M) = 43.20$. Agent 2's expected payoff given agent 1's contribution is low, i.e. $s = b_1 = 0$, is

$$U_2\left(e_2, b_1 = 0\right) = w_M \cdot p\left(e_2\right) - \frac{e_2^2}{80}. \tag{2.2}$$

$U_2\left(e_2, b_1 = 0\right)$ is maximized at $e_2^*(b_1 = 0) = 0.4 \cdot w_M = 68.80$. Note that $e_2^*(b_1 = 1) >(<) e_2^*(b_1 = 0)$ if $w_H >(<) 2 \cdot w_M$. Agent 1 anticipates agent 2's reaction to s and maximizes his own expected payoff

$$U_1\left(e_1, e_2^*(b_1)\right) = w_H \cdot p\left(e_1\right) \cdot p\left(e_2^*(b_1 = 1)\right) + w_M \cdot \\ \left[p\left(e_1\right) \cdot \left(1 - p\left(e_2^*(b_1 = 1)\right)\right) + \left(1 - p\left(e_1\right)\right) \cdot p\left(e_2^*(b_1 = 0)\right)\right] - \tfrac{e_1^2}{80}$$

with respect to e_1. $e_1 = 37.568$ maximizes $U_1\left(e_1, e_2^*(b_1)\right)$. Hence, in the unique Nash equilibrium $e_1 = 37.568$, $e_2^*(b_1 = 1) = 43.20$, and $e_2^*(b_1 = 0) = 68.80$.

T-PPlus

In T-PPlus agent 2 also observes agent 1's effort (in addition to his contribution). This additional piece of information does, however, neither influence agent 2's expected payoff nor his reaction to s in comparison to T-P since agent 2's expected payoff is independent of e_1 given b_1 (cf. Equations 2.1 and 2.2). Consequently, agent 1's decision problem is the same as in T-P. Thus, in the unique Nash equilibrium $e_1 = 37.568$, $e_2^*(b_1 = 1) = 43.20$, and $e_2^*(b_1 = 0) = 68.80$.

Table 2.2 summarizes the predictions of the standard self-interest model. In all treatments agents work inefficiently little since an effort of 80 always maximizes the sum of payoffs.

Table 2.2: Behavioral predictions of e_1 and e_2

Treatment	e_1	e_2
T-NO	52.74	52.74
T-IMP	37.20	66.24 - 0.256 * e_1 = 56.72
T-P	37.57	43.20 if b_1 = 1 68.80 if b_1 = 0
T-PPlus	37.57	43.20 if b_1 = 1 68.80 if b_1 = 0

The predictions are rounded to two digits after the decimal point.

2.3.2 Hypotheses

Based on the predictions of the self-interest model, we formulate the following hypotheses.

Hypothesis 2.1: *Decreasing reaction of the second mover to informative signals:*

(i) *In T-IMP the second mover exerts less effort the higher the first mover's effort.*

(ii) *In T-P and T-PPlus the second mover exerts less effort when the first mover's contribution is high than when it is low.*

If $w_H < 2 \cdot w_M$, the second mover faces higher incentives to work when either the first mover works less in T-IMP, or when the first mover's contribution is low rather than high in T-P and T-PPlus.

Hypothesis 2.2: *Less effort of the first mover when signals are informative: The first mover's effort is smaller in T-IMP, T-P, and T-PPlus than in T-NO.*

If $w_H < 2 \cdot w_M$, the first mover leans back when signals are informative. This is due to the first mover's anticipation that the second mover works more the smaller the informative signal. Hence, by working less he can shift work load to the second mover and save effort costs.

Hypothesis 2.3: *No additional information by the first mover's effort when the first mover's contribution is known: Efforts in T-P and T-PPlus are the same.*

As individuals only care about their own expected payoff, the information on the first mover's effort in addition to the information on the first mover's contribution does not alter the second mover's maximization problem. Therefore, both agents' decisions are identical in T-P and T-PPlus.

2.4 Results

First, we focus on the second mover's behavior and analyze how it is affected by different signals. Then, we consider the first mover's behavior.

2.4.1 The second mover's behavior

Table 2.3 reports for each treatment the mean and the standard deviation of the second mover's effort across individuals and all rounds.

Table 2.3: Mean and standard deviation of e_2

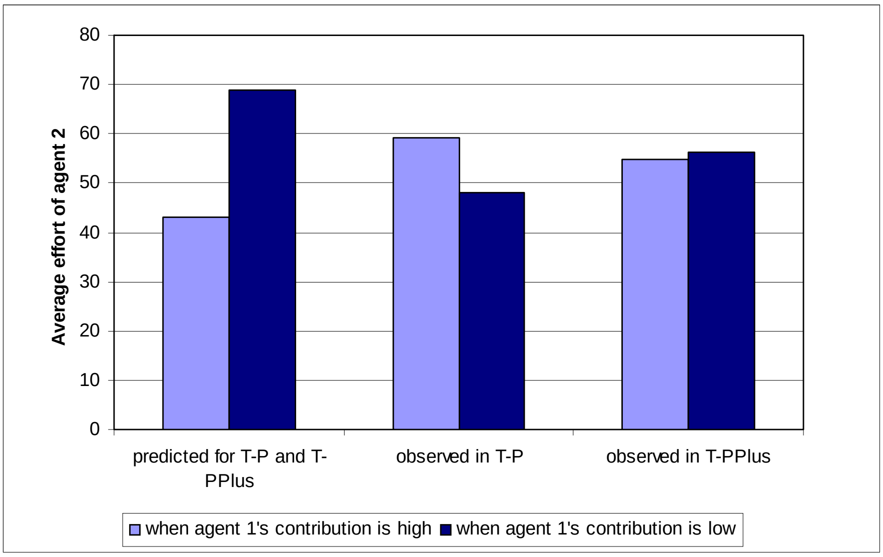
Treatment	Mean	Standard deviation	Number of observations
T-NO	50.29	13.23	126
T-NO*	49.07	14.21	252
T-IMP	53.49	23.98	238
T-P	54.25	23.72	252
T-PPlus	55.26	18.34	210

*: e_1 and e_2 are considered.

In T-NO the second mover's average effort is around 50 (with or without the decisions of the first movers)[30] which is close to our prediction of about 52.74. If, in contrast, an informative signal is available for the second mover, the second mover works more on average. This may be caused by the signal's realized value the second

[30] As in T-NO the second mover does not receive a signal prior to his decision, the first mover's decision problem is theoretically identical to the second mover's decision problem. Therefore, both agents' decisions of T-NO may be considered in the analysis of the second mover's behavior.

Figure 2.2: The second mover's average reaction in T-P and T-PPlus



mover observes in these treatments. In the following we, therefore, illustrate how the second mover reacts to the signals. The second mover's average reaction to the first mover's contribution is considerably different than predicted: In T-P the second mover works more when the first mover's contribution is high rather than low, and in T-PPlus the second mover's average effort does not vary in the first mover's contribution. Figure 2.2 reports the second mover's average effort conditional on the first mover's contribution in T-P and T-PPlus.

If we consider the second mover's reaction to the first mover's effort in T-IMP and T-PPlus, we find that the second mover's effort is positively correlated with the first mover's effort: In T-IMP and in T-PPlus the correlation coefficient is equal to 0.24 and 0.12, respectively. The self-interest model, in contrast, predicts a negative correlation in T-IMP and no correlation in T-PPlus.

In order to analyze the second mover's behavior more closely and to test our hypotheses regarding the second mover, we regress the second mover's effort[31] on treatment dummies, the interactions of treatments dummies with available signals, and control variables such as sex and risk aversion[32]. In the first two columns of Table

[31] In our regressions we also consider the first mover's decisions in T-NO. Our results do not change considerably when we only consider the second mover's decisions.

[32] We create the variable for risk aversion from our measure of risk attitude that we elicited in the post-experimental questionnaire.

2.4 we report the results of two random effects panel regressions where one (Tobit) captures that our dependent variable is (weakly) between 0 and 80.

Table 2.4: Regression results on e_2 and e_1

Dependent variable:	e_2 Panel (re)	e_2 Panel (re) (Tobit)	e_1 Panel (re)	e_1 Panel (re) (Tobit)
Intercept	+47.30	+46.73	+45.58	+44.65
	(0.000)	(0.000)	(0.000)	(0.000)
T-IMP	- 09.72	- 13.17	+07.49	+07.72)
	(0.066)	(0.045)	(0.092)	(0.124)
T-P	- 01.44	- 01.73	+02.54	+01.88
	(0.737)	(0.752)	(0.568)	(0.707)
T-PPlus	- 05.46	- 05.65	+10.66	+12.17
	(0.432)	(0.496)	(0.020)	(0.019)
T-IMP * e_1	+00.24	+00.34		
	(0.000)	(0.000)		
T-PPlus * e_1	+00.25	+00.27		
	(0.008)	(0.014)		
T-P * b_1	+12.56	+16.01		
	(0.000)	(0.000)		
T-PPlus * b_1	- 02.88	- 03.03		
	(0.275)	(0.313)		
Sex	+05.65	+08.05	+06.85	+08.54
(1 if male, 0 else)	(0.059)	(0.035)	(0.040)	(0.023)
Risk aversion	- 01.65	- 02.88	- 02.53	- 01.92
(1 if risk averse, 0 else)	(0.582)	(0.453)	(0.454)	(0.617)
Number of observations	952	952	952	952
(Pseudo) R-squared	0.069		0.089	

(re): random effects
In all four regressions e_1 and e_2 in T-NO are considered.
Numbers in brackets represent the p-values of the coefficients.

When only the first mover's effort is observable (T-IMP), the second mover's effort significantly increases in the first mover's effort. When only the first mover's contribution is observable (T-P), the second mover's effort significantly increases in the first mover's contribution. When both the first mover's effort and contribution are observable (T-PPlus), the second mover's effort significantly increases in the first mover's effort. We do, however, not find a significant effect of the first mover's contribution on the second mover's effort in T-PPlus. These results confirm our previous observations from Figure 2.2 and from the correlation coefficients in T-IMP and T-PPlus: In contrast to our predictions, the second mover positively reacts to informative signals

– in T-PPlus at least to the first mover's effort. The first mover's effort seems to be an especially stimulating signal as it even affects the second mover's effort in T-PPlus. Consequently, we reject Hypotheses 2.1(i) and 2.1(ii). We also reject Hypothesis 2.3 since the second mover's behavior in T-PPlus varies with the first mover's effort and, therefore, does not equal the second mover's behavior in T-P.

We summarize our findings on the second mover's behavior in the following three results:

Result 2.1: *Higher average effort of the second mover in T-IMP, T-P, and T-PPlus than in T-NO: On average the second mover exerts more effort when informative signals are available for the second mover.*

Result 2.2: *Increasing reaction of the second mover in T-IMP, T-P, and T-PPlus: In contrast to Hypotheses 2.1(i) and 2.1(ii), the second mover exerts more effort, the higher the first mover's effort in T-IMP and T-PPlus, or the higher the first mover's contribution in T-P.*

Result 2.3: *Effect of the first mover's effort on the second mover's behavior when the first mover's contribution is observable: In contrast to Hypothesis 2.3, the second mover's effort increases in the first mover's effort in T-PPlus.*

2.4.2 The first mover's behavior

We now turn to the first mover. Our theoretical predictions of the first mover's behavior are based on the fact that the first mover expects the second mover to react as predicted by the self-interest model. Yet, the second mover's behavior systematically deviates from the predictions. The first mover may anticipate the second mover's actual behavior and, therefore, behave differently than hypothesized even if he maximizes his own material payoff. Table 2.5 reports for each treatment the mean and the standard deviation of the first mover's effort across individuals and rounds. Similar to the second mover's behavior in T-NO, the first mover's average effort in T-NO is about 50 which is relatively close to the prediction of about 52.74. In all treatments with

Table 2.5: Mean and standard deviation of e_1

Treatment	Mean	Standard deviation	Number of observations
T-NO	47.84	15.08	126
T-NO*	49.07	14.21	252
T-IMP	55.26	17.74	238
T-P	49.28	24.15	252
T-PPlus	58.43	13.34	210

*: e_1 and e_2 are considered.

an informative signal the first mover works more than in T-NO and considerably more than predicted. Thus, the first mover does on average not lean back when the second mover receives an informative signal. One can show that – given the actual behavior of the second movers – it paid off for first movers to work more when informative signals are available.

In order to test our hypotheses regarding the first mover, we regress the first mover's effort on treatment dummies and control variables such as sex and risk aversion.[33] In the third and fourth column of Table 2.4 we report the results of two random effects panel regression where one (Tobit) captures that our dependent variable is (weakly) between 0 and 80.

When the second mover only observes the first mover's effort (T-IMP), the first mover works (marginally) significantly more compared to T-NO. When the second mover observes only the first mover's contribution (T-P), the first mover neither works significantly more nor less compared to T-NO. When the second mover observes both the first mover's effort and contribution (T-PPlus), the first mover works significantly more than in T-NO. Thus, the availability of signals (weakly) increases the first mover's effort. Especially the observability of his effort stimulates the first mover to work more. The reason may be that the second mover positively reacts to the first mover's effort and the first mover anticipates this. Although the second mover positively reacts to the first mover's contribution in T-P, the first mover does not work significantly more in T-P than in T-NO. The reason may be that the first mover's effort does not directly

[33] In our regressions we also consider the second mover's decisions in T-NO. Our results do not change considerably when we only consider the first mover's decisions.

map into his contribution. On the basis of these results, we reject Hypothesis 2.2. We also reject Hypothesis 2.3 since the coefficient of T-PPlus is significantly larger than the one of T-P in both regressions. These findings are summarized in the following two results:

Result 2.4: *No decreasing effect of informative signals on the first mover's effort: On average, the first mover works more in T-IMP, T-P, and T-PPlus than in T-NO. In T-IMP and T-PPlus, in which the first mover's effort is observable, the first mover works (marginally) significantly more than in T-NO.*

Result 2.5: *Effect of the first mover's effort on his behavior when the first mover's contribution is observable: In contrast to Hypothesis 2.3, the first mover exerts more effort when the second mover can observe the first mover's effort in addition to his contribution.*

Our results show that both the first and the second mover's behavior systematically differ from the predictions of the self-interest model when informative signals on the first mover's performance are available. The second mover positively reacts to informative signals, especially to the first mover's effort, and on average both agents work more. Finally, we test whether the sum of the first and the second mover's effort is higher when informative signals are available. Table 2.6 presents the results of two random effects panel regressions where we regress the sum of the first and the second mover's effort on treatment dummies. In one of these regressions (Tobit) we capture that the sum of the first mover's and the second mover's effort is (weakly) between 0 and 160.

When informative signals are available, the sum of efforts is at least as high as in T-NO. When the first mover's effort is available for the second mover (T-IMP or T-PPlus), the sum of efforts is even significantly larger.

Table 2.6: Regression results on the sum of e_1 and e_2

Dependent variable:	$e_1 + e_2$ Panel (re)	$e_1 + e_2$ Panel (re) (Tobit)
Intercept	+98.13	+98.13
	(0.000)	(0.000)
T-IMP	+10.61	+10.81
	(0.082)	(0.070)
T-P	+05.39	+05.49
	(0.372)	(0.353)
T-PPlus	+15.56	+15.53
	(0.013)	(0.011)
Number of observations	826	826
(Pseudo) R-squared	0.031	

(re): random effects
Numbers in brackets represent the p-values of the coefficients.

2.5 Extension

So far, we have focused on the case of substitutes in which $w_H < 2 \cdot w_M$. In this section we vary the strategic setting and consider the opposite case, i.e. $w_H > 2 \cdot w_M$, to check whether our results still hold when agents' contributions are complements. We invited 130 additional individuals in order to run our four treatments with $w_M = 110$ instead of $w_M = 172$. All other parameters of the game remained the same. Table 2.7 presents our additional treatments that are indicated by ^{X}.

Table 2.7: Additional treatments of the team production setting

Treatment	Signal	Number of participants
T-NO^X	no signal	18
T-IMP^X	e_1	36
T-P^X	b_1	38
T-PPlus^X	e_1 and b_1	38

As described in Section 2.3, efforts are strategic complements in T-NO^X and the second mover's reaction is predicted to increase in the signals in T-IMP^X, T-P^X, and

T-PPlus^X.[34] Furthermore, the self-interest model predicts the first mover to not work less but more when informative signals are available. Similar to the case of substitutes, no treatment differences are predicted between T-P^X and T-PPlus^X. Table 2.8 summarizes the predictions of the self-interest model for our additional treatments.[35]

Table 2.8: Behavioral predictions for the additional treatments

Treatment	e_1	e_2
T-NO^X	61.05	61.05
T-IMP^X	77.61	$46.40 + 0.240 * e_1 = 65.02$
T-P^X	73.28	68.00 if $b_1 = 1$ 44.00 if $b_1 = 0$
T-PPlus^X	73.28	68.00 if $b_1 = 1$ 44.00 if $b_1 = 0$

The predictions are rounded to two digits after the decimal point.

Based on these predictions we formulate the following hypotheses.

Hypothesis 2.4: *Increasing reaction of the second mover when signals are informative and* $w_M = 110$*:*

(i) *In T-IMP*^{X} *the second mover exerts more effort the higher the first mover's effort.*

(ii) *In T-P*^{X} *and T-PPlus*^{X} *the second mover exerts more effort when the first mover's contribution is high than when it is low.*

If $w_H > 2 \cdot w_M$, the second mover faces higher incentives to work when the first mover works more in T-IMP^X, or when the first mover's contribution is high rather than low in T-P^X and T-PPlus^X.

Hypothesis 2.5: *Higher effort of the first mover when signals are informative and* $w_M = 110$*: The first mover's effort is higher in T-IMP*^{X}*, T-P*^{X}*, and T-PPlus*^{X} *than in T-NO*^{X}*.*

[34] In T-PPlus^X the increasing reaction only refers to the first mover's contribution and not to his effort. This is analogous to T-PPlus.

[35] The derivation of these predictions is analogue to the one in Section 2.3.

The first mover anticipates that the second mover positively reacts to the informative signals and that he can, therefore, induce the second mover to exert a high effort by working a lot in T-IMP^X, T-P^X, and T-PPlus^X.

Hypothesis 2.6: *No effect of the additional information on the first mover's effort when the first mover's contribution is known and $w_M = 110$: Efforts in T-P^X and T-PPlus^X are the same.*

Table 2.9 reports the average behavior of both agents in our additional treatments.

Table 2.9: Mean of e_1 and e_2 in the additional treatments

Treatment	Mean of e_1	Mean of e_2	Number of observations
T-NO^X	45.37	46.53	126
T-NO^X*	45.95	45.95	252
T-IMP^X	62.27	59.67	252
T-P^X	61.05	61.88	266
if $b_1 = 1$		61.60	199
if $b_1 = 0$		62.70	67
T-PPlus^X	55.54	55.82	266
if $b_1 = 1$		62.94	155
if $b_1 = 0$		45.88	111

*: e_1 and e_2 are considered.

In T-NO^X agents work on average about 46 which is lower than the prediction of about 61.05. As w_M is smaller than in the case of substitutes (and w_H and w_L remained the same), agents may be less motivated to exert effort. Similar to the previous section, both agents work more on average when the second mover receives an informative signal. The second mover's reaction to the first mover's contribution differs, however, compared to what we observed for substitutes: In T-P^X the second mover's average effort does not vary with the first mover's contribution, while in T-PPlus^X the second mover on average works more when the first mover's contribution is high than when it is low.

In order to analyze behavior more closely and to test our hypotheses for the additional treatments, we regress the second mover's effort on treatment dummies, interactions of treatments dummies with available signals, and control variables and we regress the first mover's effort on treatment dummies, and control variables.[36] Table 2.10 reports our regression results.

Table 2.10: Regression results on e_2 and e_1 in the additional treatments

Dependent variable:	e_2 Panel (re)	e_2 Panel (re) (Tobit)	e_1 Panel (re)	e_1 Panel (re) (Tobit)
Intercept	+46.08	+44.79	+43.52	+42.56
	(0.000)	(0.000)	(0.000)	(0.000)
T-IMP^X	- 09.85	- 13.09	+17.23	+20.58
	(0.083)	(0.076)	(0.001)	(0.001)
T-P^X	+14.96	+20.67	+14.08	+19.09
	(0.000)	(0.000)	(0.005)	(0.003)
T-PPlus^X	- 12.64	- 14.38	+09.15	+11.75
	(0.003)	(0.010)	(0.066)	(0.064)
T-IMP^X * e_1	+00.37	+00.48		
	(0.000)	(0.000)		
T-PPlus^X * e_1	+00.30	+00.34		
	(0.000)	(0.000)		
T-P^X * b_1	+00.41	+00.72		
	(0.863)	(0.814)		
T-PPlus^X * b_1	+09.22	+11.08		
	(0.000)	(0.000)		
Sex	+01.02	+02.73	+05.46	+06.48
(1 if male, 0 else)	(0.680)	(0.421)	(0.133)	(0.162)
Number of observations	1036	1036	1036	1036
(Pseudo) R-squared	0.173		0.106	

(re): random effects
In all four regressions e_1 and e_2 in T-NO^X are considered.
Numbers in brackets represent the p-values of the coefficients.

Similar to the results for substitutes, the second mover's effort significantly increases in the first mover's effort in T-IMP^X. The coefficient of the interaction of T-IMP^X with the first mover's effort is even significantly larger than the predicted 0.24. In contrast to Hypothesis 2.4(ii) and the results for substitutes, the second mover's effort does not increase in the first mover's contribution in T-P^X. In T-PPlus^X, however, the second mover's effort significantly increases in both the first mover's effort and his contribution. Hence, the first mover's effort, again, is a very stimulating signal for the second mover, while the first mover's contribution significantly affects the second

[36] In our regressions we again consider the first and the second mover's decisions in T-NO^X. Our results do not change considerably if we separately consider the first mover's and the second mover's decisions in T-NO^X.

mover's effort only if also the first mover's effort is observable. With respect to the first mover's behavior we find that in all treatments with informative signals the first mover works significantly more than in our benchmark treatment. On the basis of these results, we reject Hypothesis 2.4(ii) with respect to T-P^X. We also reject Hypothesis 2.6 regarding the second mover's behavior as his effort in T-PPlus^X varies with the first mover's effort and contribution, while it is independent of both in T-P^X. We summarize our findings in the following results.

Result 2.6: *Higher average effort in T-IMP^X, T-P^X, and T-PPlus^X than in T-NO^X: On average, the first and the second mover exert more effort when informative signals are available.*

Result 2.7: *Increasing reaction of the second mover in T-IMP^X and T-PPlus^X: In line with Hypotheses 2.4(i) and 2.4(ii), the second mover works more the higher the first mover's effort in T-IMP^X, and the higher the first mover's contribution in T-PPlus^X. In T-P^X the first mover's contribution does not significantly affect the second mover's effort.*

Result 2.8: *Increasing effect of informative signals on the first mover's effort when $w_M = 110$: The first mover works significantly more in T-IMP^X, T-P^X, and T-PPlus^X than in T-NO^X.*

Result 2.9: *Effect of the additional information on the first mover's effort on the second mover's behavior when $w_M = 110$: In contrast to Hypothesis 2.6, the second mover's effort increases in the first mover's effort in T-PPlus^X. The first mover's behavior does, however, not differ between T-P^X and T-PPlus^X.*

2.6 Discussion

Our results of both the case of substitutes and complements show that (i) the second mover positively reacts to the first mover's effort when the first mover's effort is the only available signal, (ii) the second mover (weakly) positively reacts to the first mover's contribution when the first mover's contribution is the only available signal, (iii) the second mover positively reacts to the first mover's effort when the first mover's effort and contribution are observable, and (iv) the first mover does not lean back but (weakly) forward when informative signals are available. These observations are inconsistent with the predictions of the self-interest model.

We try to explain our data with an alternative model that considers elements of social comparison. If individuals compare their payoff with their team mate's payoff, the information on the team mate's effort becomes crucial – even if the information on the team mate's contribution is given. This is because the team mate receives the same wage and his payoff may only differ due to a different effort level. In order to make a converse assumption to the self-interest model, we make the extreme assumption that all individuals are inequity averse. We choose the inequity aversion model by Fehr and Schmidt (1999) as it has proved to explain human behavior in series of different games and is comparably simple.[37] Inequity averse agents dislike payoff inequalities which implies for our team setting that they dislike effort inequalities (as they receive the same wage).

We summarize the main predictions of our alternative model as follows:[38] If only the first mover's effort is observable (in T-IMP and T-IMP^X), the second mover's effort increases in the first mover's effort[39], and the first mover works at least as much than if no informative signal is available (in T-NO and T-NO^X). If only the first mover's contribution is observable (in T-P and T-P^X), the second mover does not react to the first mover's contribution, and the range of the first mover's equilibrium effort is about the same as if no informative signal is available. If the first mover's effort and contribution are observable (in T-PPlus and T-PPlus^X), the second mover's effort increases in the first mover's effort and is constant in the first mover's contribution[40],

[37] Bolton and Ockenfels (2000) also model inequity aversion. For the sake of simplicity, however, we focus on the model of Fehr and Schmidt (1999). In particular, we assume that the parameter of aversion to disadvantageous inequity is $\alpha = 2$ and the parameter of aversion to advantageous inequity is $\beta = 0.6$. These values were also assumed in Fehr et al. (2007, 2008) for inequity averse individuals in order to explain observed behavior.

[38] The detailed predictions and their derivations are in the appendix.

[39] In T-IMP the second mover's reaction decreases for $e_1 < \frac{66.24}{3.256} \approx 20.34$. Yet, more than 95 % of our observations of e_1 are strictly larger than 20.34.

[40] In T-PPlus (T-PPlus^X) this holds for $e_1 > \frac{68.8}{3} \approx 22.93$ ($\frac{68}{3} \approx$22.67). Yet, more than 99 % (84 %) of our observations of e_1 are strictly larger than 22.93 (22.67).

and the first mover works at least as much than if no informative signal is available.

In T-IMP and T-IMP^X our observations are in line with the predictions of the model of inequity aversion for both agents: The corresponding coefficients are significantly positive (cf. Tables 2.4 and 2.10).

For T-P and T-P^X the model of inequity aversion predicts that the second mover chooses the same effort independent of the first mover's contribution. In equilibrium the second mover knows the first mover's effort. Therefore, the signal b_1 does not give any additional information on the first mover's effort. If the second mover is inequity averse, he tries to match the first mover's effort independent of b_1. In T-P we observe that the second mover exerts more effort if the first mover's contribution is high (cf. Table 2.4), which is inconsistent with the predictions of the model of inequity aversion and also the self-interest model. In line with the model of inequity aversion, we observe in T-P^X that the second mover does not react to the first mover's contribution (cf. Table 2.10). As the model of inequity aversion predicts multiple equilibria for T-P and T-P^X, it is not clear for the second mover which effort the first mover has chosen. Thus, one may argue that the signal b_1 provides some information to the second mover whether the first mover's effort was rather high or low. As a low contribution rather indicates a low effort, the second mover's behavior in T-P could be interpreted as a punishment of the first mover for choosing a low effort. Why does this argument not hold for T-P^X? Figure 2.2 and Table 2.9 show that the second mover's effort after a high contribution is similar in T-P as in T-P^X, but after a low contribution the second mover's effort is lower in T-P than in T-P^X. Hence, there seems to be no such punishment in T-P^X. This observation could be related to the fact that in T-P^X the second mover ex ante has a second mover advantage in terms of material payoffs, while in T-P there is a first mover advantage. If an agent faces an ex ante advantage, he may hesitate to punish his team mate in an environment where effort is unobservable and he cannot be sure whether his team mate, indeed, exerted a low effort. Regarding the first mover's behavior in T-P and T-P^X, we observe that he works at least as much as if no informative signal is available. This does not contradict the prediction of the inequity aversion model.

In T-PPlus and T-PPlus^X the second mover positively reacts to the first mover's effort, and the first mover works at least as much as if no informative signal is available (cf. Tables 2.4 and 2.10). Furthermore, the second mover does not react to the first mover's contribution in T-PPlus. These results are in line with the model of inequity aversion. We observe that the second mover positively reacts to the first mover's contribution in T-PPlus^X, which is in line with the self-interest model but only in line with the inequity aversion model for a small range of e_1 (cf. Footnote 40).

Overall, our model of inequity aversion with the extreme assumption that all individuals are inequity averse largely explains our results. Nevertheless, our simple parameterization, in particular that all individuals are inequity averse with the corresponding parameters, does not comply with all of our observations. This concerns especially the exact point predictions. Anyhow, our results suggest that social comparison plays a crucial role in teams and shapes behavior: The reaction to signals may be contrary to the predictions of the self-interest model and signals that are predicted to have no effect may, in fact, be influential.

2.7 Conclusion

We have presented an experimental study on teams in which one team member receives a signal on his colleague's performance prior to his own decision. We observe that both the first and the second mover tend to work more on average when informative signals are available. This holds when agents' contributions are substitutes as well as when they are complements. Especially the first mover's effort seems to be a stimulating signal: If the first mover's effort is observable, the second mover works more the higher the first mover's effort, and the first mover chooses a higher effort. This observation is independent of whether the first mover's contribution is observable or not. If the first mover's contribution is the only observable signal, the second mover reacts (weakly) positively and the first mover works (weakly) more. If the first mover's contribution is observable in addition to his effort, the second mover does not react to the contribution in the case of substitutes and positively reacts to it in the case of complements.

Overall, the self-interest model fails to capture our results, while a simple model of inequity aversion largely explains behavior. Therefore, we conclude that social comparison plays a crucial role in teams in which informative signals are available. Agents' behavior may not only be contrary to the predictions of the self-interest model but also signals that are predicted to have no effect may, in fact, be influential and signals that are predicted to have an effect may be redundant.

Our results suggest that signals in teams may reduce free-riding irrespective of the strategic context. Providing information on effort seems to be an effective way to achieve more efficient outcomes. With regard to the organization of teams it seems particularly promising if agents work closely together: In this case they are not only able to observe their colleagues' efforts but also social comparison may be very pronounced.

2.8 Appendix

2.8.1 Experimental sessions

The order of events during each experimental session was the following: Individuals were welcomed and randomly assigned a cubicle in the laboratory where they took their decisions in complete anonymity from the other participants. The random allocation to a cubicle also determined an individual's role. Individuals were handed out the instructions for the experiment and the experimenter read them aloud. Then, individuals had time to go through the instructions on their own and ask questions. After all individuals had finished going through the instructions and all remaining questions had been answered, we proceeded to the decision stages. After the 14^{th} round individuals were informed about which of the 14 rounds is paid for all participants of the session, and their payment. We finished each experimental session by letting individuals answer a questionnaire that asked for demographic characteristics such as sex, subject of studies, and risk attitude. We also asked individuals to describe how they came to their decisions.

Instructions, the program, and the questionnaire were originally written in German.

In the following we give a translation of the instructions for T-NO. Instructions for the other treatments are as similar to T-NO as possible, with the difference being the description of what worker 2 knows when choosing his effort.

2.8.2 Translated instructions for T-NO

Instructions for the experiment

Welcome to this experiment. You and the other participants are asked to make decisions. **At the end of the experiment you will be paid in cash according to your decisions and the decisions of the other participants.** In addition, you receive a payment of 7 Euro.

During the whole experiment you are not allowed to speak to other participants, to use cell phones, or to start any other program on the computer. If you have questions, please raise your hand. An instructor of the experiment will then come to your seat to answer your questions.

During the experiment we do not speak of Euros but of points. Your payment will initially be calculated in points. At the end of the experiment your actual amount of total points will be converted into Euro according to the following exchange rate:

1 point = 4 Eurocents.

In this experiment there are participants in the role of **worker 1** and participants in the role of **worker 2**. One worker 1 and one worker 2 form a work team. When the experiment starts, you will be informed whether you are in the role of worker 1 or worker 2. The roles are assigned randomly. During the whole experiment that consists of 14 rounds you keep your assigned role. At the beginning of each round it will be determined anew which workers 1 and workers 2 form work groups. This assignment is random and anonymous. No participant receives any information about the identity of his matched participants.

The procedure of a round

One worker 1 and one worker 2 form a work group that has to complete two tasks. Each task can be completed with either high or low quality. Whether a task is completed with high or low quality, depends on a worker's effort and on chance.

At the beginning of each round worker 1 chooses his effort X1. X1 is an integer between 0 and 80 (including 0 and 80). The effort chosen by worker 1 influences the probability that the first task is completed with high quality and, therefore, also the probability that the first task is completed with low quality. The probability that the first task is completed with high quality is (10 + effort of worker 1) percent. Accordingly, the probability that the first task is completed with low quality is (90 – effort of worker 1) percent.

Probability that the first task is completed with high quality $= (10 + X1)\ \%$

Probability that the first task is completed with low quality $= (90 - X1)\ \%$

Examples: If worker 1 chooses an effort of 17, the first task will be completed with high quality with a probability of 27 % and with low quality with a probability of 73 %. If worker 1 chooses an effort of 72, the first task is completed with high quality with a probability of 82 % and with low quality with a probability of 18 %.

After worker 1 has chosen his effort X1, worker 2 chooses his effort X2. At this point in time worker 2 is neither informed about the effort of worker 1 nor about the quality with which first task is completed. Worker 2 chooses an effort that is an integer between 0 and 80 (including 0 and 80). The chosen effort of worker 2 influences the probability that the second task is completed with high quality and, therefore, also the probability that the second task is completed with low quality. The probability that the second task is completed with high quality is (10 + effort of worker 2) percent. Accordingly, the probability that the second task is completed with low quality is (90 - effort of worker 2) percent.

Probability that the second task is completed with high quality $= (10 + X2)\ \%$

Probability that the second task is completed with low quality $= (90 - X2)\ \%$

Examples: If worker 2 chooses an effort of 17, the second task will be completed with high quality with a probability of 27 % and with low quality with a probability of 73 %. If worker 2 chooses an effort of 72, the second task is completed with high quality with a probability of 82 % and with low quality with a probability of 18 %.

The **wage** of the workers in a work group depends on the quality with which both tasks are completed:

- If both tasks are completed with high quality, **both** workers receive a wage of **280 points** each.
- If one task is completed with high and the other task with low quality, **both** workers receive a wage of **172 points** each.
- If both tasks are completed with low quality, **both** workers receive a wage of **0 points** each.

Each worker bears the **effort costs** for **his** effort. The effort costs depend on the chosen effort of a worker and are the following:

$$\boxed{\text{effort costs} = \frac{(\text{chosen effort of a worker})^2}{80}}$$

The table at the end of the instructions indicates the effort costs for all possible effort levels of a worker.

A worker's payoff for a round is his wage minus his effort costs:

$$\boxed{\text{profit of a round} = \text{wage - effort costs}}$$

After worker 1 and worker 2 have chosen their effort, both workers are informed about the effort of both workers in their work group, the quality with which the first and the second task are completed, and each worker's profits in this round.

Number of rounds

The experiment consists of 14 repetitions of the procedure described above. This results in 14 constituent rounds. Each participant keeps his role as worker 1 or worker 2 throughout all 14 rounds. At the beginning of each round the work groups are randomly formed anew. No worker is able to distinguish whether his matched worker has already been assigned to him in one of the preceding rounds or not.

Payment of the experiment

In this experiment not all 14 rounds are paid. **One** randomly determined round is paid for all participants. Which of the 14 rounds is paid out, will be told all participants after the 14^{th} round. All participants are paid according to their profit in the randomly chosen round. If your profit in this round is negative, this amount is deducted from the 7 Euro mentioned at the beginning of the instructions.

Table of costs

Effort	Effort costs
0	0.00
1	0.01
2	0.05
3	0.11
4	0.20
5	0.31
6	0.45
7	0.61
8	0.80
9	1.01
10	1.25
11	1.51
12	1.80
13	2.11
14	2.45
15	2.81
16	3.20
17	3.61
18	4.05
19	4.51
20	5.00
21	5.51
22	6.05
23	6.61
24	7.20
25	7.81
26	8.45
27	9.11
28	9.80
29	10.51
30	11.25
31	12.01
32	12.80
33	13.61
34	14.45
35	15.31
36	16.20
37	17.11
38	18.05
39	19.01
40	20.00
41	21.01
42	22.05
43	23.11
44	24.20
45	25.31
46	26.45
47	27.61
48	28.80
49	30.01
50	31.25
51	32.51
52	33.80
53	35.11
54	36.45
55	37.81
56	39.20
57	40.61
58	42.05
59	43.51
60	45.00
61	46.51
62	48.05
63	49.61
64	51.20
65	52.81
66	54.45
67	56.11
68	57.80
69	59.51
70	61.25
71	63.01
72	64.80
73	66.61
74	68.45
75	70.31
76	72.20
77	74.11
78	76.05
79	78.01
80	80.00

Numbers are rounded to two digits after the decimal point.

2.8.3 Behavioral predictions of a model of inequity aversion

In this subsection we derive behavioral predictions for our treatments from a model of inequity aversion. In contrast to the self-interest model, an individual's utility does no longer only depend on his own material payoff but also on other individuals' payoffs. We use the assumptions from the model by Fehr and Schmidt (1999) in order to capture the notion of inequity aversion. For the case of two players, individual i's utility function is

$$u_i\left(e_i, e_j\right) =$$
$$\pi_i\left(e_i, e_j\right) - \alpha \cdot \max\left\{\pi_j\left(e_i, e_j\right) - \pi_i\left(e_i, e_j\right), 0\right\} - \beta \cdot \max\left\{\pi_i\left(e_i, e_j\right) - \pi_j\left(e_i, e_j\right), 0\right\},$$

where $\pi_j\left(e_i, e_j\right)$ denotes individual j's payoff given he chooses e_j and individual $i \neq j$ chooses e_i. α measures the aversion to disadvantageous inequity, and β the aversion to advantageous inequity.

In stark contrast to the self-interest model, in which $\alpha = \beta = 0$ for all individuals, we now assume $\alpha = 2$ and $\beta = 0.6$ for all individuals. Furthermore, we assume that individuals maximize their expected utility.

The case of substitutes

T-NO: If there is no signal at all, we solve for the Nash equilibrium of the simultaneous move game. In a first step, we derive agent i's reaction to any possible strategy of agent $j \neq i$, with $i, j \in \{1, 2\}$.

Given agent j chooses $e_j = \frac{165.6}{3.14} \approx 52.74$, the unique prediction of the self-interest model, agent i chooses $e_i = 66.24 - 0.256 \cdot e_j = \frac{165.6}{3.14}$ as a best response. This maximizes agent i's expected payoff, which is shown in the derivation of the predictions of the self-interest model, and minimizes the inequity of payoffs as $e_i = e_j$. Thus, it maximizes agent i's expected utility.

Given agent j chooses $e_j > \frac{165.6}{3.14}$, $e_i = 66.24 - 0.256 \cdot e_j$ maximizes agent i's expected payoff but not his utility as it generates advantageous inequity since $66.24 - 0.256 \cdot e_j < \frac{165.6}{3.14}$ for $e_j > \frac{165.6}{3.14}$. Therefore, agent i's optimal response is larger than $66.24 - 0.256 \cdot e_j$ (but smaller than e_j). It equals $\min\left\{\frac{66.24 - 0.256 \cdot e_j}{1-\beta}, e_j\right\}$, which is derived by maximizing agent i's expected utility for the case of advantageous inequity with respect to e_i. For $\beta = 0.6$ and $e_j \leq 80$ the minimum is equal to e_j.

Given agent j chooses $e_j < \frac{165.6}{3.14}$, again $e_i = 66.24 - 0.256 \cdot e_j$ maximizes agent i's expected payoff but not his expected utility as it generates disadvantageous inequity since $66.24 - 0.256 \cdot e_j > \frac{165.6}{3.14}$ for $e_j < \frac{165.6}{3.14}$. Therefore, agent i's optimal response is smaller (but larger than e_j). It equals $\max\left\{\frac{66.24-0.256\cdot e_j}{1+\alpha}, e_j\right\}$, which is derived analogously to above. For $\alpha = 2$ the maximum is equal to e_j if and only if $e_j \geq \frac{66.24}{3.256}$.

We summarize agent i's reaction function for $\alpha = 2$, $\beta = 0.6$ and $e_j \in [0, 80]$ as follows

$$e_i^*(e_j) = \begin{cases} e_j & \text{if} \quad e_j \geq \frac{66.24}{3.256} \\ \frac{66.24-0.256\cdot e_j}{3} & \text{if} \quad e_j < \frac{66.24}{3.256} \end{cases}.$$

In a next step, we solve for the intersections of both agents' reaction functions.

If $e_j \in \left[\frac{66.24}{3.256}, 80\right]$, all strategy combinations with $e_j = e_i \in \left[\frac{66.24}{3.256}, 80\right]$ are intersections.

If $e_j \in \left[0, \frac{66.24}{3.256}\right)$, there is no intersection. To see this, suppose, to the contrary, there was at least one intersection. Then $e_i = \frac{66.24-0.256\cdot e_j}{3}$ and $e_i < \frac{66.24}{3.256}$, as otherwise $e_j = e_i \geq \frac{66.24}{3.256}$ which was a contradiction. As $e_i < \frac{66.24}{3.256}$, $e_j = \frac{66.24-0.256\cdot e_i}{3}$. Yet, the two equalities can only hold if $e_j = e_i = \frac{66.24}{3.256}$. Since $\frac{66.24}{3.256} \notin \left[0, \frac{66.24}{3.256}\right)$, there cannot be an intersection.

Thus, the set of Nash equilibria is equal to the set of all strategy combinations with $e_j = e_i \in \left[\frac{66.24}{3.256}, 80\right]$.

T-IMP: In T-IMP agent 2 receives the signal $s = e_1$. We solve for the subgame perfect Nash equilibrium given $s = e_1$. Agent 2's reaction to $s = e_1$ is equivalent to his reaction function in T-NO as derived before. Agent 1 anticipates agent 2's reaction to s and maximizes his own expected utility[41]

$$U_1\left(e_1, e_2^*(e_1)\right) = w_H \cdot p\left(e_1\right) \cdot p\left(e_2^*(e_1)\right) + w_M \cdot \left[p\left(e_1\right) \cdot \left(1 - p\left(e_2^*(e_1)\right)\right) + \left(1 - p\left(e_1\right)\right) \cdot p\left(e_2^*(e_1)\right)\right] - \frac{e_1^2}{80} - \beta \cdot \left(\frac{e_2^*(e_1)^2}{80} - \frac{e_1^2}{80}\right).$$

with respect to e_1. For $e_1 \in \left[\frac{66.24}{3.256}, 80\right]$, $e_2^*(e_1) = e_1$ and agent 1's expected utility is maximized at $e_1 = 80$.[42] For $e_1 \in \left[0, \frac{66.24}{3.256}\right]$, $e_2^*(e_1) = \frac{66.24-0.256\cdot e_1}{3} \geq e_1$ and agent 1's

[41] Here we use the fact that agent 1 either faces no payoff inequality because agent 2 perfectly matches his effort $\left(\text{for } e_1 \geq \frac{66.24}{3.256}\right)$ or advantageous inequality because agent 2 chooses more effort $\left(\text{for } e_1 < \frac{66.24}{3.256}\right)$.

[42] This function's maximum is attained at $e_1 > 80$. As the function is concave, it increases in e_1 for the whole range of $e_1 \in \left[\frac{66.24}{3.256}, 80\right]$. Hence, $e_1 = 80$ is the maximum for $e_1 \in \left[\frac{66.24}{3.256}, 80\right]$.

expected utility is maximized at $e_1 = \frac{66.24}{3.256}$.[43] As $e_1 = \frac{66.24}{3.256}$ is agent 1's best choice out of the range $e_1 \in \left[0, \frac{66.24}{3.256}\right]$, and $e_1 = 80$ is agent 1's best choice out of the range $e_1 \in \left[\frac{66.24}{3.256}, 80\right]$, which includes $e_1 = \frac{66.24}{3.256}$, agent 1's best choice for the whole range of $e_1 \in [0, 80]$ is $e_1 = 80$. Consequently, in the unique subgame perfect Nash equilibrium $e_1 = s = 80$ and $e_2 = 80$ according to agent 2's reaction $e_2^*(e_1) = e_1$ for $e_1 \geq \frac{66.24}{3.256}$.

T-P: In T-P we solve for Nash equilibria given $s = b_1$. In a first step, we derive agent 2's reaction to any combination of e_1 and b_1. Consider $b_1 = 1$. What is agent 2's best response in this case?

Given agent 1 chooses $e_1 = 43.20$, which is agent 2's equilibrium effort in the self-interest model as derived in Section 2.3, agent 2 chooses $e_2 = 43.20$ as a best response. This maximizes agent 2's expected payoff, which is shown in the derivation of the predictions of the self-interest model, and minimizes the inequity of payoffs as $e_2 = e_1$.

Given agent 1 chooses $e_1 > 43.20$, $e_2 = 43.20$ maximizes agent 2's expected payoff (as the payoff maximizing effort is independent of e_1). This choice, however, generates advantageous inequity. Therefore, agent 2's utility maximizing response is larger than 43.20 (but smaller than e_1). It equals $\min\left\{\frac{43.20}{1-\beta}, e_1\right\}$, which is derived by maximizing agent 2's utility for the case of advantageous inequity with respect to e_2. For $\beta = 0.6$ and $e_1 \leq 80$ the minimum equals e_1.

Given agent 1 chooses $e_1 < 43.20$, $e_2 = 43.20$ again maximizes agent 2's expected payoff. This time, however, it generates disadvantageous inequity. Therefore, agent 2's optimal response is smaller (but larger than e_1). It equals $\max\left\{\frac{43.20}{1+\alpha}, e_1\right\}$, which is derived analogously to above. For $\alpha = 2$, the maximum equals e_1 if and only if $e_1 \geq 14.4$.

We summarize agent 2's reaction for $b_1 = 1$, $\alpha = 2$, $\beta = 0.6$ and $e_1 \in [0, 80]$ as follows

$$e_2^*(e_1|b_1 = 1) = \begin{cases} e_1 & \text{if} \quad e_1 \geq 14.4 \\ 14.4 & \text{if} \quad e_1 < 14.4 \end{cases}.$$

Equivalently, we can derive agent 2's reaction for $b_1 = 0$, $\alpha = 2$, $\beta = 0.6$ and $e_1 \in [0, 80]$

[43] Out of the range $e_1 \in \left[0, \frac{66.24}{3.256}\right]$, $e_1 = \frac{66.24}{3.256}$ yields the highest material payoff for agent 1 and avoids payoff inequalities since $e_2^*(e_1) = \frac{66.24 - 0.256 \cdot e_1}{3} = e_1$ for $e_1 = \frac{66.24}{3.256}$.

that is the following

$$e_2^*(e_1|b_1=0)=\begin{cases} e_1 & \text{if} \quad e_1 \geq \frac{68.80}{3} \\ \frac{68.80}{3} & \text{if} \quad e_1 < \frac{68.80}{3} \end{cases}.$$

Note that for $e_1 \geq \frac{68.80}{3}$ agent 2's reaction is $e_2 = e_1$ independent of b_1.

In a next step, we solve for the intersections of both agents' reaction functions, i.e. we search for strategy combinations $(e_1, (e_2(b_1=1), e_2(b_1=0)))$ in which e_1 is a best response to $(e_2(b_1=1), e_2(b_1=0))$ and $(e_2(b_1=1), e_2(b_1=0))$ is a best response to e_1.

Given $e_1 = \frac{165.6}{3.14}$, the equilibrium effort of T-NO in the self-interest model, agent 2's best response is $\left(\frac{165.6}{3.14}, \frac{165.6}{3.14}\right)$ since $e_1 \geq \frac{68.80}{3}$. Given agent 2 chooses $\frac{165.6}{3.14}$ independent of b_1, agent 1's best response is to choose $e_1 = \frac{165.6}{3.14}$. This maximizes agent 1's expected payoff, which is shown in the derivation of the predictions of the self-interest model, and minimizes the inequity of payoffs as $e_1 = e_2(b_1=1) = e_2(b_1=0)$. Hence, this strategy combination is a Nash equilibrium.

Given $e_1 = x \in \left(\frac{165.6}{3.14}, 80\right]$, agent 2's best response is (x, x) as $e_1 \geq \frac{68.80}{3}$. Given agent 2 chooses $x \in \left(\frac{165.6}{3.14}, 80\right]$ independent of b_1, agent 1's best response is to choose x as it is shown by the reaction function derived in T-NO for the model of inequity aversion. Thus, all strategy combinations with $e_1 = e_2(b_1=1) = e_2(b_1=0) \in \left(\frac{165.6}{3.14}, 80\right]$ are Nash equilibria.

Given $e_1 = x \in \left[\frac{68.80}{3}, \frac{165.6}{3.14}\right)$, agent 2's best response is (x, x). Given agent 2 chooses $x \in \left[\frac{68.80}{3}, \frac{165.6}{3.14}\right)$ independent of b_1, agent 1's best response is to choose x, as it is shown by the reaction function derived in T-NO for the model of inequity aversion. Consequently, all strategy combinations with $e_1 = e_2(b_1=1) = e_2(b_1=0) \in \left[\frac{68.80}{3}, \frac{165.6}{3.14}\right)$ are Nash equilibria.

Given $e_1 = x \in \left[14.4, \frac{68.80}{3}\right)$, agent 2's best response is $\left(x, \frac{68.80}{3}\right)$. Given agent 2 chooses x if $b_1 = 1$ and $\frac{68.80}{3}$ if $b_1 = 0$, agent 1's best response is not x. It can be shown that $e_1 = \frac{68.80}{3}$, for instance, yields a strictly higher expected utility than $e_1 = x$ $\in \left[14.4, \frac{68.80}{3}\right)$. Consequently, there is no Nash equilibrium with $e_1 \in \left[14.4, \frac{68.80}{3}\right)$.

Given $e_1 = x \in [0, 14.4)$, agent 2's best response is $\left(14.4, \frac{68.80}{3}\right)$. Given agent 2 chooses 14.4 if $b_1 = 1$ and $\frac{68.80}{3}$ if $b_1 = 0$, agent 1's best response is not x. Choosing

$e_1 = 14.4$, for instance, yields a strictly higher expected payoff and causes less payoff inequalities than $e_1 = x \in [0, 14.4)$. Hence, there is no Nash equilibrium with $e_1 \in [0, 14.4)$.

Thus, the set of Nash equilibria is equal to the set of all strategy combinations with $e_1 = e_2(b_1 = 1) = e_2(b_1 = 0) \in \left[\frac{68.80}{3}, 80\right]$.

T-PPlus: In T-PPlus agent 2 observes agent 1's effort in addition to his contribution. We solve for the subgame perfect Nash equilibrium given $s = (e_1, b_1)$. Agent 2's reaction to $s = (e_1, b_1)$ is equivalent to his best response in T-P as derived before. Agent 1 anticipates agent 2's reaction to s and maximizes his own expected utility[44]

$$U_1\left(e_1, e_2^*(e_1|b_1 = 1), e_2^*(e_1|b_1 = 0)\right) = w_H \cdot p\left(e_1\right) \cdot p\left(e_2^*(e_1|b_1 = 1)\right) + w_M \cdot \\ \left(p\left(e_1\right) \cdot \left(1 - p\left(e_2^*(e_1|b_1 = 1)\right)\right) + \left(1 - p\left(e_1\right)\right) \cdot p\left(e_2^*(e_1|b_1 = 0)\right)\right) - \frac{e_1^2}{80} - \beta \cdot p\left(e_1\right) \cdot \\ \left(\frac{e_2^*(e_1|b_1=1)^2}{80} - \frac{e_1^2}{80}\right) - \beta \cdot \left(1 - p\left(e_1\right)\right) \cdot \left(\frac{e_2^*(e_1|b_1=0)^2}{80} - \frac{e_1^2}{80}\right).$$

with respect to e_1. It can be shown that $e_1 = 14.4$ is agent 1's best choice out of the range $e_1 \in [0, 14.4]$, $e_1 = \frac{68.8}{3}$ out of the range $e_1 \in \left[14.4, \frac{68.8}{3}\right]$, and $e_1 = 80$ out of the range $e_1 \in \left[\frac{68.8}{3}, 80\right]$. Hence, agent 1's expected utility is maximized at $e_1 = 80$. Therefore, in the unique subgame perfect Nash equilibrium $e_1 = 80$ and $e_2 = 80$, according to agent 2's reaction $e_2^*(e_1|b_1 = 1) = e_2^*(e_1|b_1 = 0) = e_1$ for $e_1 \geq \frac{68.80}{3}$.

The case of complements

T-NO^X: If there is no signal at all, we solve for the Nash equilibrium of the simultaneous move game. In a first step, we derive agent i's reaction function to any possible strategy of agent $j \neq i$, with $i, j \in \{1, 2\}$. We proceed analogously to T-NO but start with the unique Nash equilibrium prediction of the self-interest model that is $e_j = \frac{116}{1.9} \approx 61.05$.

For $\alpha = 2$, $\beta = 0.6$ and $e_j \in [0, 80]$ agent i's reaction function is

$$e_i^*(e_j) = \begin{cases} e_j & \text{if} \quad e_j \geq \frac{46.40}{2.760} \\ \frac{46.40 + 0.240 \cdot e_j}{3} & \text{if} \quad e_j < \frac{46.40}{2.760} \end{cases}.$$

[44] Note that agent 2's reaction implies that agent 2's effort is at least as high as e_1. Thus, agent 1 faces either no or advantageous payoff inequality.

In a next step, we solve for the intersections of both agents' reaction functions. It can be shown that the set of Nash equilibria is equal to the set of all strategy combinations with $e_j = e_i \in \left[\frac{46.40}{2.760}, 80\right]$.

T-IMP^X: In T-IMP^X we solve for the subgame perfect Nash equilibrium given $s = e_1$ using agent 2's reaction function of T-NO^X. Agent 1 anticipates agent 2's reaction to s and maximizes his own expected utility with respect to e_1. Analogously to T-IMP, it can be shown that agent 1's expected utility is maximized at $e_1 = 80$. Consequently, in the unique subgame perfect Nash equilibrium $e_1 = s = 80$ and $e_2 = 80$, according to agent 2's reaction $e_2^*(e_1) = e_1$ for $e_1 \geq \frac{46.40}{2.760}$.

T-P^X: In T-P^X we solve for the Nash equilibrium given $s = b_1$. To derive agent 2's best response to e_1 if $b_1 = 1$, we proceed analogously to T-P and start with agent 2's equilibrium effort of the self-interest model $e_1 = 68$. Agent 2's reaction for $b_1 = 1$, $\alpha = 2$, $\beta = 0.6$ and $e_1 \in [0, 80]$ is as follows

$$e_2^*(e_1|b_1 = 1) = \begin{cases} e_1 & \text{if} \quad e_1 \geq \frac{68}{3} \\ \frac{68}{3} & \text{if} \quad e_1 < \frac{68}{3} \end{cases}.$$

Equivalently, we can derive agent 2's best response for $b_1 = 0$ that is the following

$$e_2^*(e_1|b_1 = 0) = \begin{cases} e_1 & \text{if} \quad e_1 \geq \frac{44}{3} \\ \frac{44}{3} & \text{if} \quad e_1 < \frac{44}{3} \end{cases}.$$

By solving for the intersections of both agents' reaction functions we derive all Nash equilibria of T-P^X. Proceeding in an equivalent way to T-P and starting with $e_1 = \frac{116}{1.9}$, the equilibrium effort in T-NO^X in the self-interest model, we can show that the set of Nash equilibria is equal to the set of all strategy combinations with $e_1 = e_2(b_1 = 1) = e_2(b_1 = 0) \in \left[\frac{68}{3}, 80\right]$.

T-PPlus^X: In T-PPlus^X we solve for the subgame perfect Nash equilibrium given $s = (e_1, b_1)$. Agent 2's reaction to $s = (e_1, b_1)$ is equivalent to his best response in T-P^X. Agent 1 anticipates agent 2's reaction to s and maximizes his own expected utility with respect to e_1. It can be shown that agent 1's expected utility is maximized at $e_1 = 80$. Consequently, in the unique subgame perfect Nash equilibrium $e_1 = 80$ and $e_2 = 80$, according to agent 2's reaction $e_2^*(e_1|b_1 = 1) = e_2^*(e_1|b_1 = 0) = e_1$ for $e_1 \geq \frac{68}{3}$.

Chapter 3

Can intentions spoil the kindness of a gift? An experimental study

3.1 Introduction

Consider a situation where person A undertakes a costly action that benefits person B. This behavior seems altruistic. However, if person A expects a reward in return, e.g. from person B, then person A's action may be motivated by the expected rewards rather than by pure altruism. If the expected rewards are sufficiently high, even selfish individuals have an incentive to behave in this way. The question we address in this study is how person B reacts to the intentions of person A. Does person B perceive person A's action as less kind if he expects person A to expect rewards, and – if person B can reciprocate – does he return less?

There are many situations where behavior seems altruistic but is obviously strategic. Companies, for example, give Christmas gifts to their business partners in order to improve the business relationship, hoping that this pays off in future transactions. Their business partners may well understand that the given Christmas gifts are part of the company's profit maximizing investment strategy. The question, however, is whether this knowledge spoils the perceived kindness of the gifts and makes them less effective.

We address this question experimentally in a series of modified trust games. In these games we vary the probability that the second mover can reciprocate and analyze effects on second mover behavior. Our results suggest that neither the perceived kindness of the first mover's action nor the rewards given by the second mover are spoiled by expected future rewards.

In our modified trust game agent A, the first mover, decides how much of his initial endowment he transfers to agent B, the second mover. Agent B receives the tripled amount of agent A's transfer. Then, a lottery determines whether agent B can decide on his return transfer to agent A or not. In the latter case nothing is returned to agent A. We conduct two treatments of this modified trust game that differ in the probability that agent B can decide on his return transfer: In treatment T-HIGH this probability is 80 % and in treatment T-LOW it is 50 %. Agent A behaves in a way that seems altruistic when he transfers a strictly positive amount to agent B. This is true in both treatments. Our treatment variation, however, changes the possibility for agent B to make a return transfer to agent A and, thereby, varies the chance for agent A to receive a return. Agent A's expected returns for his transfer are smaller in T-LOW if agent A has the same belief about agent B's reaction in both treatments. Consequently, agent B may perceive agent A as more kind in T-LOW than in T-HIGH and, therefore, may return more in T-LOW than in T-HIGH when he is asked to decide. Agent A's beliefs about agent B's reaction, however, may differ in both treatments. Models of intention-based reciprocity predict that agent B returns more in T-LOW than in T-HIGH. Nevertheless, agent A expects smaller future rewards for a given transfer in T-LOW than in T-HIGH. This is because the difference in the probability that agent B can decide on his return transfer dominates the difference in agent B's return transfer.

Our results suggest that expected future returns do not affect the perceived kindness of an action and the action's rewards. Agent B's return transfer for a given transfer of agent A does not differ across treatments. This is not because agent B does not care about agent A's action at all. Actually, we observe a lot of agents B who return strictly positive amounts and, in addition, agent B's average return transfer increases in agent A's transfer. This suggests that individuals reward actions that seem altruistic, irrespective of the actor's expectation of future rewards or of the actor's specific kind of intention. Consequently, we conclude that individuals in our setting condition their behavior on outcomes rather than on intentions or higher order beliefs.

We try to explain our findings by analyzing data from our questionnaire that participants filled out after they had made their decisions. First, our regressions of agent B's return transfer on agent B's elicited second order belief give no indication that the given expected future returns spoil the kindness of an action and the action's rewards. Second, we analyze whether agent B's perception of agent A's action is affected by the treatment. We do not find a significant effect. Third, we test treatment effects on agent B's stated emotions. For some interior values of agent A's transfer, negative emotions like anger and contempt are experienced significantly more intense in T-HIGH than in T-LOW, while appreciation is significantly more pronounced in T-LOW than in T-HIGH. Even though intentions may affect individuals' emotions, these effects do not seem to carry over to the perception of an action and the reaction to it.

Intentions have been modeled in a number of theoretical papers. Rabin (1993), Dufwenberg and Kirchsteiger (2004), and Falk and Fischbacher (2006) introduce theoretical models of intention-based reciprocity. In contrast to models of social preferences that are based on outcomes only (e.g. Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), they (also) take into account that intentions affect the perception of others' actions and, thereby, behavior.

Various experimental papers have focused on the empirical relevance of intentions. A couple of them (e.g. Blount, 1995; Offerman, 2002; McCabe et al., 2003; Charness, 2004; Cox, 2004; Falk et al., 2008) study the effect of intentionality, i.e. whether the second mover's reaction to the first mover's action is different when the first mover's action is chosen by the first mover himself (out of a non-singleton action set) than when it is exogenously determined by the experimenter or a lottery. These studies typically find that the second mover returns the favor or the disfavor of the first mover's action in a more pronounced way when it was intentionally chosen by the first mover himself. Hence, intentionality seems to matter. These studies, however, do not provide evidence for the effect of *different* intentions behind the *same intentional* action.

Charness and Levine (2007) go further in this direction. They present experimental results of a gift exchange game in which the principal determines an initial wage that is, then, hit by a random shock. Agents work less for the *same* final wage when it was brought about by a lousy offer of the principal and a positive shock than by a generous offer and a negative shock. While this study compares two different intentional actions that lead to the same outcome, Bolton et al. (1998) and Falk et al. (2003) compare the reaction to the same intentional action when different (non-singleton) action sets are available for the first mover. Typically, the same action is either the most or the least generous action of the first mover's action set. While Bolton et al. (1998) do not observe any significant effects in their experimental setting, Falk et al. (2003) find in an experimental ultimatum game that responders reject the same offer less often when it is the most generous offer of the first mover's action set than when it is the least generous. Hence, there is evidence that the relative position of an action in the first mover's choice set seems to matter. In our study, in contrast, we focus on gifts, i.e. on intentional actions that always seem to be altruistic or generous. In all of our treatments the first mover's action set is the same and so is the ranking of the actions' generosity. We only vary the second mover's possibility to reciprocate and, thereby, can study different intentions behind the same gift. In particular, we ask whether expected rewards spoil the kindness of a gift.

Stanca et al. (forthcoming) analyze in their experimental study whether the second mover's reaction is different when the first mover's action is extrinsically motivated rather than intrinsically. They compare the second mover's reaction in a standard trust game with the corresponding reaction in a trust game in which first movers are not informed that second movers can react to their transfer until they have made their decision.[45] They hypothesize and also find that the slope of the second mover's reaction function is larger when the first mover is intrinsically motivated. In our experimental study, in contrast, we do not distinguish between extrinsic and intrinsic motivation since the first mover may expect a strictly positive return in both treatments and, therefore, may be extrinsically motivated in both treatments. Furthermore, we test the hypothesis that the second mover returns more *for a given transfer* in T-LOW

[45] Hence, they implement an asymmetry of information conditions, which is not present in our experiment. In our experiment all participants (in all treatments) receive all relevant information at the beginning of the experiment.

than in T-HIGH,[46] which is supported by models of intention-based reciprocity.

This chapter proceeds as follows. Section 3.2 presents the experimental design and procedure, Section 3.3 the behavioral predictions and hypotheses. Our results are summarized and discussed in Section 3.4. Section 3.5 concludes.

3.2 Experimental design and procedure

We consider a modified trust game with two agents, A and B. Agent A, the trustor, is initially endowed with $w_A = 20$ and can transfer an amount $x \in \{0, 5, 10, 15, 20\}$ to agent B, the trustee, who is initially endowed with $w_B = 0$. Agent B receives the tripled amount of agent A's transfer, $3{*}x$. After agent A's decision a lottery determines whether the game stops at this point in time or continues. With probability $1 - q$ the game stops and agent A earns his initial endowment minus his transfer, $20 - x$, while agent B earns agent A's tripled transfer, $3 * x$. With probability q, though, the game continues and agent B can transfer an amount $y(x) \in [0, 3 * x]$ to agent A. In the latter case agent A earns his initial endowment minus his transfer plus agent B's return transfer, $20 - x + y(x)$, and agent B earns agent A's tripled transfer minus his return transfer, $3 * x - y(x)$.[47] The structure of this game is summarized in Figure 3.1.

The modification of the trust game consists in the random move of nature after agent A's decision on x. If $q = 1$, the game resembles the standard trust game introduced by Berg et al. (1995).[48] In contrast, if $q = 0$, the game boils down to a dictator game in which agent B can never return anything to agent A.[49] The higher $q \in (0, 1)$, the higher the chance that agent B can make a return transfer (given $x > 0$) and the more similar the game is to the standard trust game. The smaller $q \in (0, 1)$, the smaller the chance that agent B can make a return transfer (given $x > 0$) and the more similar the game is to a dictator game.

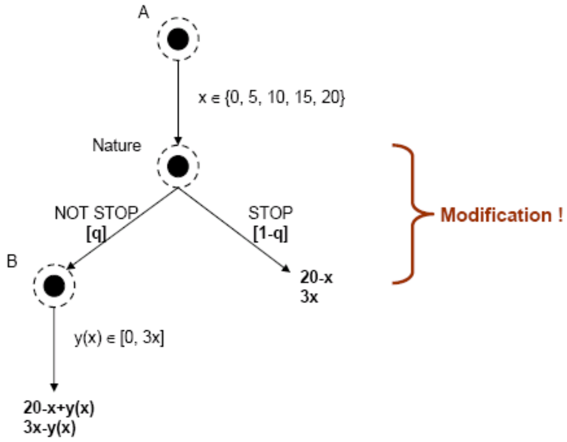
[46] This hypothesis does not necessarily imply that the slope of the second mover's reaction function is larger in T-LOW than in T-HIGH.

[47] Note that agent A does not receive the tripled amount of agent B's return transfer.

[48] One major difference to the game introduced by Berg et al. (1995) is that in their version agent B also has a strictly positive initial endowment (which equals w_A).

[49] One major difference to standard dictator games is that in the typical versions the dictator's transfer is not tripled.

Figure 3.1: Structure of the modified trust game



As we address the question whether the perceived kindness of an action that seems altruistic and the action's rewards are reduced by the actor's expectation of future rewards, we vary q, the probability that agent B can return a positive amount (given $x > 0$), across treatments and keep everything else constant. Table 3.1 presents our treatments.

Table 3.1: Treatments of the modified trust game

Treatment	q	Number of participants
T-HIGH	0.8	40
T-LOW	0.5	60

In treatment T-HIGH q is higher than in treatment T-LOW. We do not implement probabilities close or equal to 0 and 1 since we would like to avoid the effect that a certain event is going to happen almost for sure. Furthermore, we are restrained to take higher values of q since agents B are only asked to decide on $y(x)$ when the game, indeed, continues.[50] If q was small, we expected very few observations of $y(x)$ for a given number of participants.[51]

[50] We could have asked all agents B to decide on $y(x)$ *given the game continues.* Then, however, treatment effects may also have been caused by social preferences based on expected outcomes and it would have been difficult to disentangle the source of observed treatment effects.

[51] For instance, if $q = 0.2$ and 100 individuals participated in this treatment, 50 individuals were

In each experimental session one treatment of the modified trust game is conducted. The implemented treatment of a session is played once. At the beginning of each session the roles of the game are assigned randomly. Participants are informed about their assigned roles after they have correctly answered a set of control questions. Agents A are always asked to decide on x, while agents B are only asked to decide on $y(x)$ when the game continues after agent A's decision. Given agents B are asked to decide, we elicit $y(x)$ by the strategy method, i.e. agents B are informed that the game continues but are not informed about x and decide on their return transfer for each possible x.[52] After participants have made their decisions, they fill out a questionnaire concerning their emotions, beliefs, perception of the other player's action, and individual data such as sex, age, and subject of studies.

Our experimental sessions were run at the Center for Experimental Economics of the University of Innsbruck, Austria, in April 2008. 100 individuals participated in the experiment, which was conducted with the software z-Tree by Fischbacher (2007). Individuals were randomly assigned to sessions and could take part only once. The sessions were framed neutrally and lasted about an hour.[53] Individuals earned on average 10.34 € including a show-up fee of 5 €.[54]

3.3 Behavioral predictions and hypotheses

We address the question whether the perceived kindness of an action that seems altruistic, i.e. a costly action that benefits others, and the action's rewards are reduced by the actor's expectation to receive future rewards. This may be the case since future rewards can partially cover the actor's initial costs and reduce the others' net benefit. In the presented modified trust game agent A behaves in a way that seems altruistic when he transfers a strictly positive amount to agent B. First, this action is costly

agents B out of which we expect 10 to be asked to decide on $y(x)$.

[52] We apply the strategy method here in order to get agent B's reaction function. We are aware that this elicitation method may affect $y(x)$. However, we expect this effect to be orthogonal to our treatment variation. Furthermore, Stanca et al. (forthcoming) argue that the strategy method applied in their trust games does not significantly affect decisions.

[53] Translated instructions and a more detailed description of the procedure of a session are provided in the appendix.

[54] The maximum payoff was 23 € and the minimum 5 €.

because the amount is deducted from his initial endowment. Second, it benefits agent B since the tripled transfer is assigned to agent B. This is true for both treatments. Agent A's expectation to receive a transfer from agent B in return may reduce the perceived kindness of agent A's action. In particular, the more agent A expects in return for his (given) transfer, the more of agent A's initial costs are covered in expectation, the less expected payoff is assigned to agent B, and, therefore, the less kind agent B may perceive agent A and the less agent B may, in fact, return when he is asked to decide. Our treatment variation changes the possibility for agent B to make a return transfer to agent A and varies the chance for agent A to receive a return. Hence, agent A's expected returns for a given transfer are smaller in T-LOW when agent A has the *same* belief about agent B's reaction in both treatments. Consequently, agent B may perceive agent A as more kind in T-LOW and, therefore, may return more in T-LOW when he is asked to decide. If this, indeed, is the case and agent A's belief about agent B's reaction is correct and agent B's belief about agent A's belief is correct, then agent A expects agent B to transfer more in T-LOW when agent B is asked to decide. Nevertheless, agent A faces less expected future returns in T-LOW since the difference in q compensates the difference in agent B's reaction. If it did not and agent A expected higher future returns in T-LOW, agent B perceived agent A as less kind in T-LOW and, therefore, he transferred less in T-LOW. Consequently, agent A's expectation were incorrect.

In the following we present standard models of social preferences that differ in their assumptions on the individuals' utility function and, consequently, in their behavioral predictions. Some of them explicitly model how the perceived kindness is reduced by the actor's expectation to receive future rewards and predict that agent B returns more in T-LOW for a given transfer.

3.3.1 Behavioral predictions

Model 1: The self-interest model

The standard neoclassical model assumes that all individuals are selfish, i.e. their utility function U depends on their own material payoff m only and increases in m.

Given these assumptions, agent B's decision does not vary in $q \in (0,1)$.

As agent B maximizes his own material payoff only, he transfers $y(x) = 0 \ \forall \ x$ in the unique subgame perfect Nash equilibrium. This is true for all $q \in (0,1)$.

Model 2: A model of social preferences based on final outcomes

Models of social preferences based on final outcomes (e.g. Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) assume that an individual's utility function $\widetilde{U}$ does not only depend on m but also on another individual's material payoff r. This does not necessarily imply that an individual is altruistic. Individuals with $\widetilde{U}$ may also be spiteful, envious, inequity averse or inequity loving.

Given these assumptions, agent B's decision does not vary in $q \in (0,1)$.

As agent B's decision is affected by final outcomes only (and not by how these outcomes came about),[55] agent B faces the same decision problem at his decision node independent of $q \in (0,1)$. Hence, his optimal decision does not vary across treatments.

Model 3: Models of intention-based reciprocity

Models of intention-based reciprocity (e.g. Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006) assume that an individual's utility function V is not only dependent on outcomes but also on how these outcomes came about, e.g. whether the underlying decision problem was determined randomly or whether the underlying decision problem was intentionally brought about by another individual.

[55] Models of social preferences based on expected outcomes (e.g. Trautmann, 2009) predict the same, as long as agent B's decision is based on his expectations formed at the moment of his decision making.

Crucial roles are played by the perceived kindness of an individual's own action and the perceived kindness of other individuals' actions. Typically, the kinder an individual perceives the action of another individual, the kinder the individual treats this other individual. The perceived kindness of an action is shaped by the actor's intentions. How kindness is defined exactly and how intentions concretely enter the utility varies across models. In the following we present the predictions of a modified version of the model by Dufwenberg and Kirchsteiger (2004) and a similar model that implements central elements from the model by Falk and Fischbacher (2006).[56]

A modified version of the model by Dufwenberg and Kirchsteiger (2004): As the model of Dufwenberg and Kirchsteiger (2004) is intended for finite multi-stage games *without* nature, we modify it in a simple and straight-forward way that accounts for random moves of nature. Our modification consists in the way how agent B perceives the kindness of agent A's strategy in the course of the game, in particular, after the lottery determined that the game continues.

Given the assumptions of this model, $y(x)$ is (weakly)[57] higher in T-LOW than in T-HIGH for $x = 20$ in any sequential reciprocity equilibrium (SRE) in which agent B chooses a pure strategy.

A modified version of the model by Dufwenberg and Kirchsteiger (2004) with central elements of the model by Falk and Fischbacher (2006): We take our modified version of the model by Dufwenberg and Kirchsteiger (2004) and implement central elements of the model by Falk and Fischbacher (2006) that concern how kindness is defined.

Given the assumptions of this model, $y(x)$ is (weakly)[58] higher in T-LOW than in T-HIGH $\forall\ x > 0$ in any sequential reciprocity equilibrium (SRE) in which agent B chooses a pure strategy.

[56] In the appendix we present these models and derive their predictions.

[57] No treatment differences are predicted if either agent B is hardly sensitive to reciprocity concerns such that he chooses $y(x) = 0$ in both treatments, or agent B is extremely sensitive to reciprocity concerns such that he chooses $y(x) = 3 * x$ in both treatments.

[58] No treatment differences are predicted if either agent B is hardly sensitive to reciprocity concerns such that he chooses $y(x) = 0$ in both treatments, or agent B is extremely sensitive to reciprocity concerns such that he chooses $y(x) = 3 * x$ in both treatments.

3.3.2 Hypotheses

The various theoretical models predict different behavioral patterns of agent B. We focus on the predicted equilibria in which agent B chooses a pure strategy and summarize these predictions in the following three hypotheses.

Hypothesis 3.1: No returns in all treatments

Agent B returns nothing to agent A in T-HIGH and in T-LOW.

This hypothesis is supported by the self-interest model and implies that $y(x) = 0$ for all x and all treatments. Actions that seem altruistic are never rewarded.

Hypothesis 3.2: The same returns in all treatments

Agent B returns a weakly positive amount to agent A. Agent B's return transfer for a given x is the same in all treatments.

Models of social preferences based on final outcomes support this hypothesis. Similar to the self-interest model, there are no treatment effects *with respect to agent B's behavior*. In contrast to the self-interest model, agent B returns a weakly positive amount to agent A. Actions that seem altruistic are rewarded, irrespective of the actor's intentions.

Hypothesis 3.3: Higher returns in T-LOW

Agent B returns a weakly positive amount to agent A. $y(x)$ is higher in T-LOW than in T-HIGH for $x > 0$.[59]

Models of intention-based reciprocity take into consideration how a decision problem came about and, therefore, capture the effect of intentions. They predict that the perceived kindness of an action that seems altruistic and the action's rewards are reduced by the expectation of future returns.

[59] This is supported by the modified version of the model by Dufwenberg and Kirchsteiger (2004) with central elements of the model by Falk and Fischbacher (2006). The modified version of the model by Dufwenberg and Kirchsteiger (2004) predicts $y(x)$ to be higher in T-LOW than in T-HIGH for $x = 20$, but not necessarily for all $x > 0$. The reason for the possibly different predictions is that the two models use different definitions of kindness.

3.4 Results

First, we summarize the descriptive results of our experiment and compare them with standard results from trust games and dictator games. In a next step, we test our hypotheses and analyze whether the perceived kindness of an action that seems altruistic and the action's rewards are reduced by the expectation of future rewards. Finally, we try to explain our findings with questionnaire data.

3.4.1 Summary statistics

Behavior of agent A

Table 3.2 presents the mean and the standard deviation of agent A's transfer in T-HIGH, T-LOW and both treatments together.

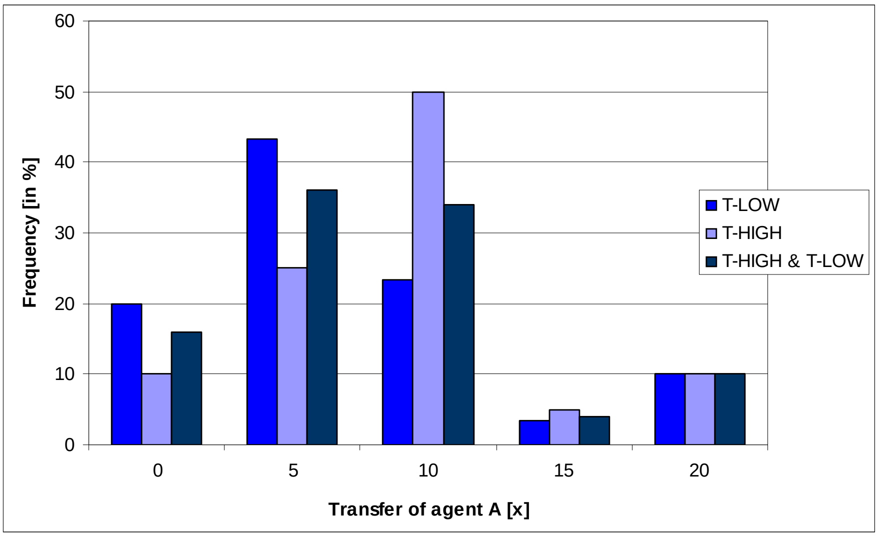
Table 3.2: Mean and standard deviation of A's transfer

Treatment	Mean	Standard deviation	Number of observations
T-HIGH	9.00	5.28	20
T-LOW	7.00	5.81	30
T-HIGH + T-LOW	7.80	5.64	50

On average, agent A transfers 7.8 points (out of 20 available points) to agent B. This is considerably larger than 0. Camerer (2003, p. 86), however, reports that in standard trust games, in which $q = 1$ and agent B often has the same initial endowment as agent A, agent A transfers on average half of his initial endowment. This is more than in our experiment, in particular in T-LOW, what suggests that agent A transfers less when the probability that agent B can reciprocate is low. Figure 3.2 illustrates the aggregate distribution of x in T-HIGH, T-LOW and both treatments together.

In both treatments together 84 % of agents A transfer strictly positive amounts, more than 40 % half of their initial endowment or more, and more than 10 % even more than 60 % of their initial endowment (some even their whole initial endowment). This is considerably different to results from standard dictator games, in which $q = 0$

Figure 3.2: Distribution of A's transfer



and transfers are not tripled. For instance, in the benchmark treatment by Forsythe et al. (1994) (the paid dictator game conducted in April with a pie of 5 \$) about 55 % of dictators transfer a strictly positive amount, less than 20 % half of their endowment or more, and no dictator transfers more than 60 % of his endowment. This suggests that the distribution of x shifts towards higher values when $q > 0$ compared to $q = 0$.[60] When we consider the distributions of x in T-HIGH and in T-LOW, we observe that that the distribution is more to the right in T-HIGH than in T-LOW. Hence, it seems that agents A react to differences in q. They tend to send more the higher the probability that agent B can reciprocate.

Behavior of agent B

Table 3.3 presents the mean and the standard deviation of agent B's return transfer per x in T-HIGH, T-LOW and both treatments together.

In line with the results from the standard trust game by Berg et al. (1995), the more agent A transfers, the more agent B returns on average. The observed average

[60] This shift could also be caused by the fact that in standard dictator games agent A's transfer is not tripled. Cox (2004), though, observes that the distribution of transfers in a standard trust game is centered on higher values than the distribution of transfers in the corresponding trust game with $q = 0$.

Table 3.3: Mean and standard deviation of B's return transfer

Treatment	x	Mean	Standard deviation	y(x)/x	Number of observations
T-HIGH	5	01.75	02.59	0.35	16
	10	06.75	06.14	0.68	16
	15	11.06	08.83	0.74	16
	20	17.31	13.68	0.87	16
T-LOW	5	01.93	02.76	0.39	15
	10	04.93	04.35	0.49	15
	15	10.13	08.19	0.68	15
	20	16.47	12.00	0.82	15
T-HIGH + T-LOW	5	01.84	02.63	0.37	31
	10	05.87	05.34	0.59	31
	15	10.61	08.40	0.71	31
	20	16.90	12.69	0.85	31

return transfers, however, seem to be lower than the ones by Berg et al. (1995).[61] Table 3.3 also reports the rate of the average return transfer, i.e. the average return transfer divided by the transfer. Independent of $x > 0$ the rate of average return transfer is below 1. Hence, a strictly positive transfer does not pay for agent A on average, even if agent A knew beforehand that the game is not stopped. Nevertheless, the rate of the average return transfer increases in x and peaks at a value more than 0.8 at $x = 20$.

If we separately examine agent B's return transfers in the two treatments, we observe that on average agent B returns more in T-HIGH than in T-LOW for all $x \in \{10, 15, 20\}$.

3.4.2 Analysis of hypotheses

Hypothesis 3.1: No returns in all treatments

Table 3.3 shows that agent B's average return transfers are considerably higher than 0 for $x > 0$. P-values of one sample median tests on $y(x) = 0$ per treatment and per x are reported in Table 3.4.

On the basis of these tests, we reject Hypothesis 3.1 for all $x > 0$ and all treatments.

[61] One explanation for this difference could be that in the experiment by Berg et al. (1995) agents B have the same initial endowment as agents A.

Table 3.4: P-values of one sample median tests on Hypothesis 3.1

Treatment	x	Percentage of agents B with y(x) = 0	Number of observations	p-value
T-HIGH	5	62.50	16	0.015
	10	31.25	16	0.001
	15	25.00	16	0.001
	20	25.00	16	0.001
T-LOW	5	60.00	15	0.015
	10	33.33	15	0.002
	15	20.00	15	0.001
	20	20.00	15	0.001

Nevertheless, Table 3.4 shows that there are some agents B that return nothing given $x > 0$. The percentage of these observations decreases in x. Still, 25 % of agents B in T-HIGH and 20 % of agents B in T-LOW return nothing given $x = 20$.

Hypothesis 3.2: The same returns in all treatments

Table 3.3 indicates that agent B's average return transfer for a given $x > 0$ does not considerably vary across treatments. Table 3.5 reports per x the two-sided p-values of Mann-Whitney-U tests on whether $y(x)$ differs across treatments.

Table 3.5: Two-sided p-values of pairwise Mann-Whitney-U tests on Hypothesis 3.2

x	Number of observations in T-HIGH	Number of observations in T-LOW	p-value
5	16	15	0.856
10	16	15	0.405
15	16	15	0.873
20	16	15	0.873

On the basis of these tests, we are far from rejecting Hypothesis 3.2. Agent B's return transfer does not seem to differ across treatments.

Hypothesis 3.3: Higher returns in T-LOW

From the results presented in Table 3.5 we conclude that $y(x)$ is not significantly smaller in T-HIGH than in T-LOW, neither for $x = 20$ nor for any other $x > 0$. If anything differed between T-HIGH and T-LOW regarding $y(x)$, then $y(x)$ was larger in T-HIGH than in T-LOW, at least for $x \in \{10, 15, 20\}$. Hence, our data seem to be inconsistent with Hypothesis 3.3. One may, however, argue that the introduced models of intention-based reciprocity predict no treatment difference either when agent B is hardly sensitive to reciprocity concerns such that he chooses $y(x) = 0$ in both treatments, or when agent B is extremely sensitive to reciprocity concerns such that he chooses $y(x) = 3{*}x$ in both treatments. Table 3.4 reports that in both treatments a fraction of agents B return nothing, even if $x = 20$. This suggests that a fraction of agents B are, indeed, hardly sensitive to reciprocity concerns. We, however, do not observe a single individual with $y(x) = 3 * x$ in any of our treatments. This rules out the possibility that a fraction of agents B are extremely sensitive to reciprocity concerns such that no treatment effects are predicted. Thus, either agents B are, in general, not sensitive to reciprocity concerns, or too few of agents B are sufficiently sensitive to reciprocity concerns.

We summarize our findings in the following two results:

Result 3.1: Rewards for actions that seem altruistic

As in previous studies on trust games (see Camerer, 2003, p. 86), we observe that agent B returns significantly positive amounts. On average, these amounts increase in agent A's transfer.

Result 3.2: No effect of the intention of a gift

Agent B's return transfer (given $x > 0$) does not vary across treatments.

These results are consistent with the predictions of models of social preferences based on final outcomes but inconsistent with the predictions of the self-interest model. The introduced models of intention-based reciprocity may predict no treatment differences, but only for individuals that are sufficiently insensitive to reciprocity concerns or for individuals that are extremely sensitive to reciprocity concerns. In our data, however, there is no evidence for individuals that are extremely sensitive to reciprocity concerns. There may be individuals that are sufficiently insensitive to reciprocity concerns and return nothing. On average, though, agents B return strictly positive amounts. Consequently, the predictions of the introduced models of intention-based reciprocity are inconsistent with our aggregate results. We conclude that the kindness of an action and the action's rewards are not spoiled by the actor's expectation to receive future rewards. On average, actions that seem altruistic are rewarded by others. The rewards vary in the action: The more altruistic they seem, the higher are the average rewards. The average rewards for a given action, though, are independent of the actor's expectation to receive future rewards.

3.4.3 Possible explanations for our findings

In this section we try to find explanations for our findings by analyzing data from our questionnaire.

Incorrect higher order beliefs of agent B

The perceived kindness of an action that seems altruistic can only be spoiled by the actor's expectation of future rewards if individuals expect the actor to expect future rewards. From other experimental studies we know that individuals have difficulties to draw inferences from other individual's actions and correctly form beliefs.[62] In the following we analyze whether the given elicited second order beliefs of agent B directly affect his behavior.[63] We regress agent B's return transfer for a given x on x and on the product of agent B's second order belief with q for a given x, i.e. agent B's expectation of agent A's expected future returns for a given x. First, we estimate an OLS regression. Second, we run a two-stage least squares instrumental variable regression in which we

[62] Prominent examples are experimental studies on information cascades, e.g. by Anderson and Holt (1997), Hung and Plott (2001), Kariv (2005), Nöth and Weber (2003), and Goeree et al. (2007).

[63] Agent B's second order belief was elicited in a non-incentivized way after agent B has made his decision. We are aware that these second order beliefs may be affected by agent B's own decision. Therefore, we checked whether agent B's elicited second order belief significantly differs from those elicited by agents B who have not decided upon $y(x)$ because the lottery stopped the game after agent A's decision. We run pairwise Mann-Whitney-U tests and do not find a significant difference. Hence, we assume that an agent's own action does not influence his second order beliefs to a large extent.

instrument for the product of agent B's second order belief with q for a given x. The instrument we use is q itself as it is exogenous and, by definition, correlated with the instrumented variable. We run this additional regression since agent B's second order belief for x could be endogenous and, therefore, our estimated OLS coefficient could be biased and inconsistent. Table 3.6 presents the results of our regressions for $x > 0$.[64]

Table 3.6: Regressions of the return transfer for x > 0

Dependent variable: y(x)	OLS-c	2SLS-IV-c
Intercept	- 03.05***	- 01.67
x	+00.79***	+00.33
Agent B's second order belief * q	+00.24	+00.77
Number of observations	124	124
R-squared	0.3384	0.2700

*, **, *** significant at 10, 5, 1 percent significance level
-c with individual clusters

In all of our regressions agent B's belief about agent A's expected return does not significantly affect agent B's return for a given x. In OLS-c the only significant regressor is agent A's transfer: The higher agent A's transfer, the more agent B returns.[65]

Result 3.3: No effect of agent B's belief about agent A's expected returns

Agent B's elicited beliefs about agent A's expected returns do not affect agent B's returns.

Consequently, we conclude that incorrect higher order beliefs of agent B are not the explanation for why the kindness of an action that seems altruistic and the action's future rewards are not spoiled by the actor's expectation to receive future rewards.

Effect on the perception or on emotions

The perceived kindness of an action may be spoiled by the actor's expectation to receive future rewards without affecting the reaction to that action. Similarly, the reactor's emotions may be affected in the sense that he experiences more negative emotions and

[64] In all regressions we consider $x > 0$ since the restriction on $x = 20$ would considerably reduce our data set.

[65] These results do not change considerably if we control for sex, age and subject of studies.

less positive emotions in T-HIGH than in T-LOW. Table 3.7 reports one-sided p-values of Mann-Whitney-U tests on whether the perceived kindness of agent A's action differs across treatments.[66]

Table 3.7: One-sided p-values of Mann-Whitney-U tests on perceived kindness across treatments

x	Number of observations in T-HIGH	Number of max. int. in T-HIGH	Number of observations in T-LOW	Number of max. int. in T-LOW	p-value
0	16	0	15	0	0.1058
5	16	0	15	0	0.0964
10	16	0	15	1	0.0284
15	16	1	15	2	0.2567
20	16	10	15	12	0.2294

max. int. observations perceiving the kindness with maximal intensity

For any $x \in \{0, 15, 20\}$ we do not identify any significant differences in the perceived kindness across treatments.[67] $x = 5$ and $x = 10$ are perceived as less kind in T-HIGH at a significance level of 10 %. We take this as weak evidence that agent B's perception of agent A's action is affected by the treatment variation.

Result 3.4: Hardly no effect of agent A's intention on agent B's perception of agent A's action

Agent B's perceived kindness of agent A's action does not significantly vary across treatments. This is true for $x = 20$ and for the most other values of x.

Table 3.8 reports one-sided p-values of Mann-Whitney-U tests on whether the intensity of hypothetically sensed emotions is higher in one of the treatments.[68]

For $x = 20$, stated anger is sensed significantly more strongly in T-HIGH than in T-LOW at a significance level of 10 %. There is no significant difference in contempt, gladness, and appreciation for $x = 20$.

[66] In our questionnaire agents B had to indicate on a scale ranging from 1 to 7 how kind they perceive a given transfer by agent A. 1 represented "very unkind", while 7 represented "very kind".

[67] For $x = 20$ this could be due to the fact that the majority of agents B choose the maximal intensity.

[68] In our questionnaire individuals had to indicate on a scale ranging from 1 to 7 with which intensity they hypothetically sensed an emotion for each x. If they did not sense an emotion at all, they were asked to indicate 1 for this emotion and the given x.

Table 3.8: One-sided p-values of Mann-Whitney-U tests on emotions across treatments

Emotion	x	Number of observations in T-HIGH	Number of max. int. in T-HIGH	Number of observations in T-LOW	Number of max. int. in T-LOW	p-value
Anger	0	16	3	15	1	0.1449
	5	16	0	15	0	0.0669
	10	16	0	15	0	0.0154
	15	16	0	15	0	0.0023
	20	16	0	15	0	0.0820
Contempt	0	16	4	15	0	0.0358
	5	16	0	15	0	0.0384
	10	16	0	15	0	0.0079
	15	16	0	15	0	0.0097
	20	16	1	15	0	0.4185
Gladness	0	16	0	15	0	0.2011
	5	16	1	15	0	0.1242
	10	16	1	15	0	0.0989
	15	16	1	15	3	0.3273
	20	16	12	15	13	0.1791
Appreciation	0	16	0	15	0	0.1439
	5	16	0	15	0	0.0054
	10	16	0	15	1	0.0060
	15	16	0	15	3	0.0579
	20	16	9	15	9	0.3954

max. int. observations sensing an emotion with maximal intensity

For interior values of x, anger is significantly more strongly pronounced in T-HIGH than in T-LOW. The same holds for contempt. Only for $x = 10$ gladness is significantly less pronounced in T-HIGH than in T-LOW at a significance level of 10 %. Appreciation is significantly less pronounced in T-HIGH than in T-LOW for interior values of x.

For $x = 0$, we detect a significant treatment difference in contempt only.[69]

Result 3.5: Effect of A's intentions on anger, contempt, and appreciation for interior values of x

Negative emotions such as anger and contempt are significantly more pronounced in T-HIGH than in T-LOW for interior values of x. Furthermore, appreciation is significantly less strongly pronounced in T-HIGH than in T-LOW for interior values of

[69] For $x = 20$, one may, however, argue that regarding the positive (negative) emotions a large number of observations indicated the maximal (minimal) intensity and, therefore, no treatment differences are identified. For $x = 0$, a large number of observations indicated the minimal intensity for gladness and appreciation.

x. Gladness seems to be unaffected by the treatment variation for the most values of x.

Consequently, we conclude that agent B's emotions may be affected by agent A's intentions. This effect, however, does not seem to carry over to agent B's perception of agent A's action and to agent B's reaction.

Other explanations

There are other potential reasons for why the perceived kindness of an action that seems altruistic and the action's rewards are not spoiled by the actor's expectation of future rewards in our setting. One reason may be that agent B can voluntarily decide on his return transfer and is not forced to return a certain amount. Expecting a return that is voluntarily given may not spoil the kindness of an action. This may be different for expecting a return that is involuntarily given. The introduced models of intention-based reciprocity do not take this into account.

Another reason may be that kindness is not an absolute measure but a relative one that captures the ranking of actions for a *given* action set. $x = 20$, for instance, may be perceived as the kindest action of agent A and, therefore, would be evaluated as equally kind in both treatments.

3.5 Conclusion

We have presented an experimental study on whether the perceived kindness of an action that seems altruistic, i.e. a costly action that benefits others, and the action's rewards are reduced by the actor's expectation to receive future rewards.

In our experimental study second movers in a modified trust game return significantly positive rewards to first movers. On average, these rewards increase in the first mover's transfer. However, they do not significantly vary in the probability that the second mover can reciprocate. The second mover's return transfer is even slightly higher when the probability that the second mover can reciprocate is 0.8 rather than

0.5 for some values of x. On the basis of data from our questionnaire, we test whether this is due to incorrect higher order beliefs of second movers. Our regression results suggest that this, however, does not seem to be the case. Furthermore, we test whether the second mover's perception or emotions are affected by the probability that the second mover can reciprocate. We find significant effects on some of the second mover's emotions, at least for some values of x.

Our results suggest that behavior that seems altruistic is rewarded. The more altruistic it seems, the higher is the reward in return. The reward for a given action, however, does not vary in the actor's expectation to receive future rewards. This is consistent with the predictions of models of social preferences based on final outcomes but inconsistent with the predictions of the self-interest model and the introduced models of intention-based reciprocity. Hence, individuals in this setting seem to condition their behavior on outcomes rather than on intentions or higher order beliefs.

These results seem to be relevant for different kinds of contexts. Political as well as commercial campaigns often try to gain the support of a large group of individuals by behaving in a way that seems altruistic, e.g. by distributing small gifts. Individuals may well anticipate that these gifts are intended to gain their support. In the light of our results, however, we would conclude that this does not diminish the effectiveness of the small gifts. Similarly, in some organizations workers are financially incentivized to help their colleagues.[70] Workers, therefore, may anticipate that the help of a colleague is motivated by receiving financial rewards. We would conclude that this does not diminish the perceived kindness of help and does not harm the willingness to reward this action.

This experimental study contributes to the discussion of higher order beliefs and of intentions. Our results suggest that higher order beliefs and specific kinds of intentions do not significantly influence the reaction to an action that seems altruistic. It may well be that higher order beliefs and intentions are crucial for other sorts of behavior, though, e.g. for the reaction to socially undesired behavior. Criminal law often conditions penalties on the criminal's intentions. Hence, the effect of intentions may depend

[70] A worker's wage may, for instance, depend on the performance of his colleagues.

on the specific context.

3.6 Appendix

3.6.1 Experimental sessions and instructions

Experimental sessions

The order of events during each experimental session was the following: Individuals were welcomed and randomly assigned a cubicle in the laboratory where they took their decisions in complete anonymity from the other individuals. The random allocation to a cubicle also determined an individual's role. The instructions for the experiment, which each individual found in his cubicle, were read aloud. Then, individuals could go through the instructions on their own and ask questions. After all remaining questions had been answered and no individual needed more time to go through the instructions, they had to answer a set of control questions concerning the procedure of the experiment. After each individual had answered all control questions correctly, participants were informed about their role in the experiment and we proceeded to the decision stages. First, agents A decided upon x. Second, a computer program determined randomly which games of a session were stopped. Each game of a session had the same probability that it is stopped, which corresponded to q of the implemented treatment of the session. Third, agents B were informed about whether the game was stopped or not. In case the game was not stopped agents B decided upon the return transfer for each x. In case the game was stopped agents B were asked what they would have transferred in return for each x if the game had continued. After participants had made their decisions, they were asked questions whose answers were not related to any payments, e.g. agents A were asked how many points they believe agent B transfers in return for each x given the game is not stopped, and agents B were asked which intensities of certain emotions they would experience for each x. After all participants had answered the questions posed to them, all agents were informed about the outcome of the game, i.e. agent A's decision, nature's random move on whether the game stops right after agent A's decision, and - in case the game was not stopped - agent B's decision for the corresponding x. Finally, we elicited demographic variables such as sex, age and subject of studies. At the end of the session individuals were paid in cash according to their earned amount in the modified trust game plus a show-up fee of 5 Euro.

The instructions, the control questions, the program, and the questionnaire were originally written in German. The translated instructions for T-HIGH can be found in the following. The instructions for T-LOW are similar except that the probability that the game is not stopped right after agent A's decision is 50 % (instead of 80 %).

Translated instructions of T-HIGH

Instructions for the experiment

Welcome to this experiment. You and the other participants are asked to make decisions. Your decisions as well as the decisions of the other participants determine the result of the experiment. At the end of the experiment you will be paid **in cash** according to the **actual** result of the experiment. So please read the instructions attentively and think about your decisions carefully. In addition, you receive – independent of the result of the experiment - a show up fee of 5 Euro.

During the whole experiment it is not allowed to talk with other participants, to use mobile phones, or to start other programs on the computer. The contempt of these rules immediately leads to the exclusion of the experiment and of all payments. If you have any questions, please raise your hand. An instructor of the experiment will then come to your seat in order to answer your questions.

During the experiment we talk about points rather than about Euros. Your whole income is initially calculated in points. At the end of the experiment your actual amount of total points is converted into Euros according to the following rate:

1 point = 30 Cents.

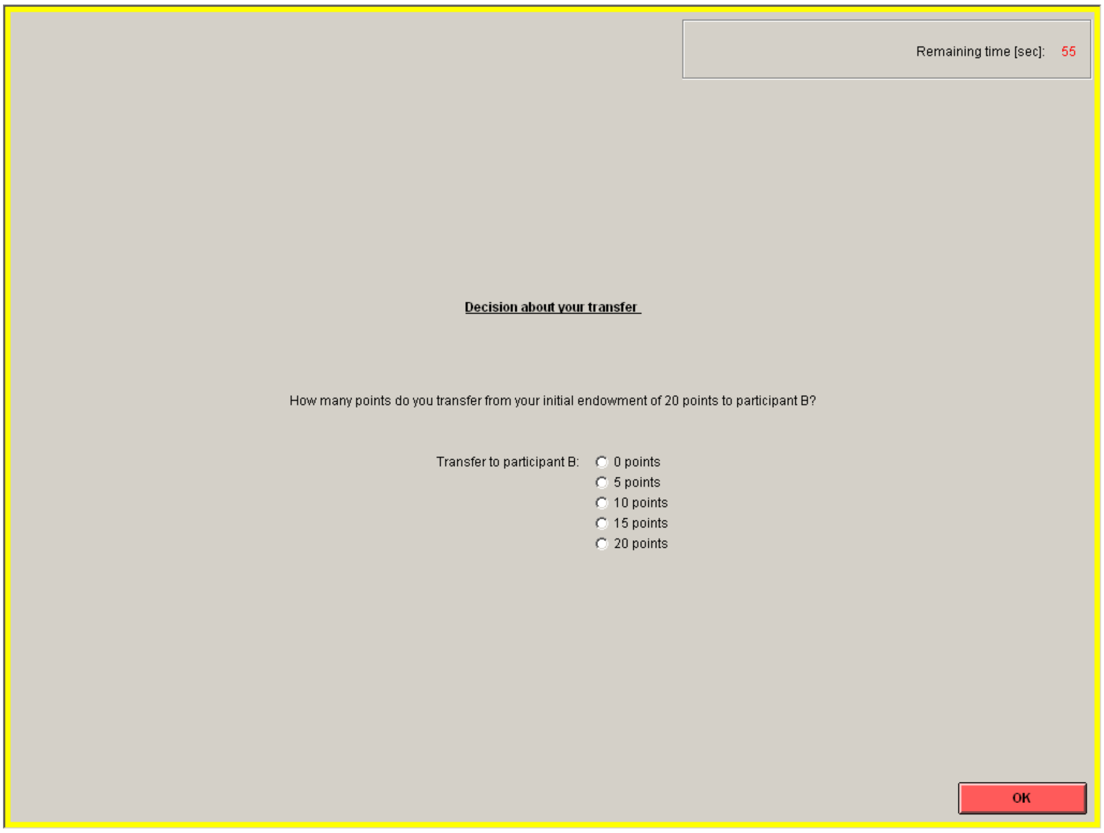
In this experiment there are **participants A** and **participants B**. Before the experiment starts, you are informed whether you are a participant A or a participant B. While entering the room this was randomly determined. If you are participant A, you are randomly and anonymously matched to a participant B. If you are participant B, you are randomly and anonymously matched to a participant A. Neither during nor after the experiment you receive any information about the identity of your matched participant. Likewise, your matched participant does not receive any information about your identity.

The procedure

Participant A has an initial endowment of 20 points. Participant B has an initial endowment of 0 points.

Participant A can decide how much of his initial endowment he transfers to participant B. **Participant A can either choose 0, 5, 10, 15 or 20 points**.

In order to make this decision, participant A selects one amount on the following computer screen and presses the OK-button.



Remaining time [sec]: 55

Decision about your transfer

How many points do you transfer from your initial endowment of 20 points to participant B?

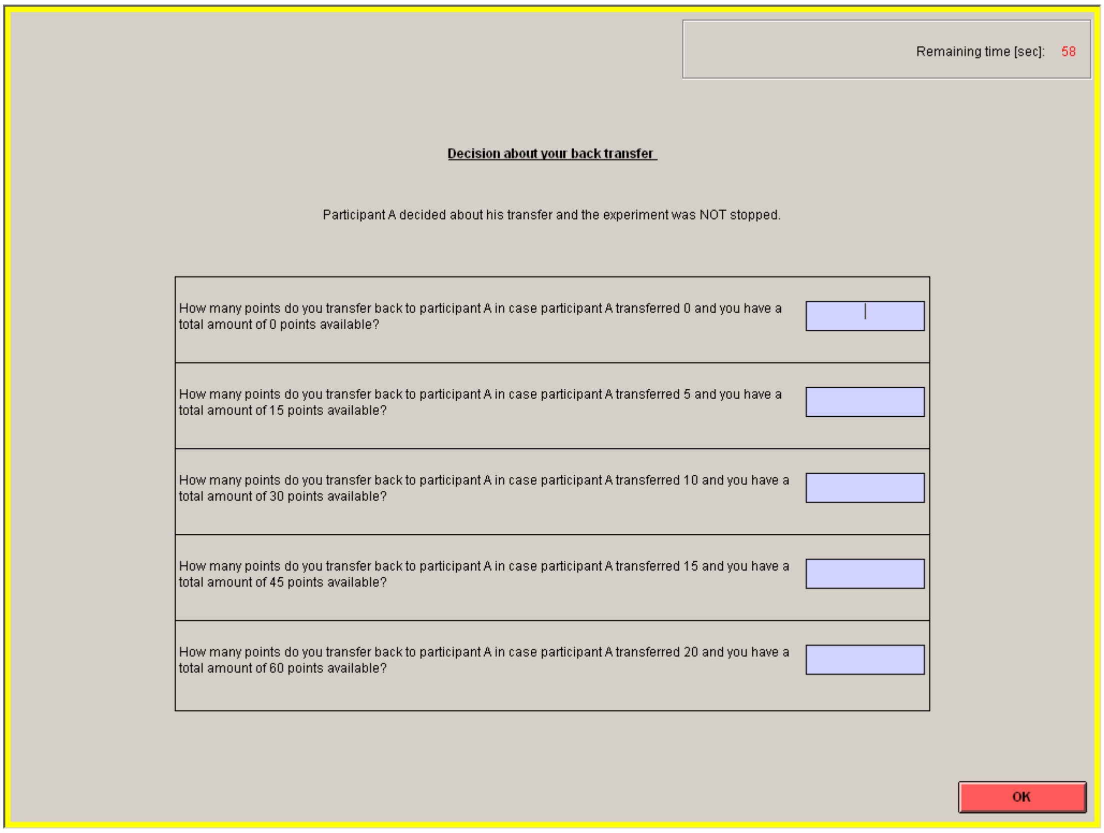
Transfer to participant B:
○ 0 points
○ 5 points
○ 10 points
○ 15 points
○ 20 points

Participant A's transfer is then **tripled** and sent to participant B.

After participant A chose his transfer and participant A's tripled transfer was sent to participant B, it is randomly determined whether the experiment is stopped at this point in time.

- With the probability of 20% the experiment is stopped at this point in time. **In this case participant A receives his initial endowment minus his transfer, and participant B receives participant A's tripled transfer**.
- With the probability of 80% the experiment is not stopped at this point in time and participant B decides which integer between 0 and participant A's tripled transfer (including 0 and participant A's tripled transfer) he transfers back to participant A. **In this case participant A receives his initial endowment minus his transfer plus participant B's back transfer, and participant B receives participant A's tripled transfer minus his back transfer**.

In case the experiment is not stopped right after participant A's decision, participant B makes the decision about the back transfer. In order to do that, participant B indicates for each possible transfer of participant A his selected amount on the following computer screen and presses the OK-button. Depending on what participant A transferred, participant B's corresponding entry is transferred back to participant A.



Remaining time [sec]: 58

Decision about your back transfer

Participant A decided about his transfer and the experiment was NOT stopped.

How many points do you transfer back to participant A in case participant A transferred 0 and you have a total amount of 0 points available?

How many points do you transfer back to participant A in case participant A transferred 5 and you have a total amount of 15 points available?

How many points do you transfer back to participant A in case participant A transferred 10 and you have a total amount of 30 points available?

How many points do you transfer back to participant A in case participant A transferred 15 and you have a total amount of 45 points available?

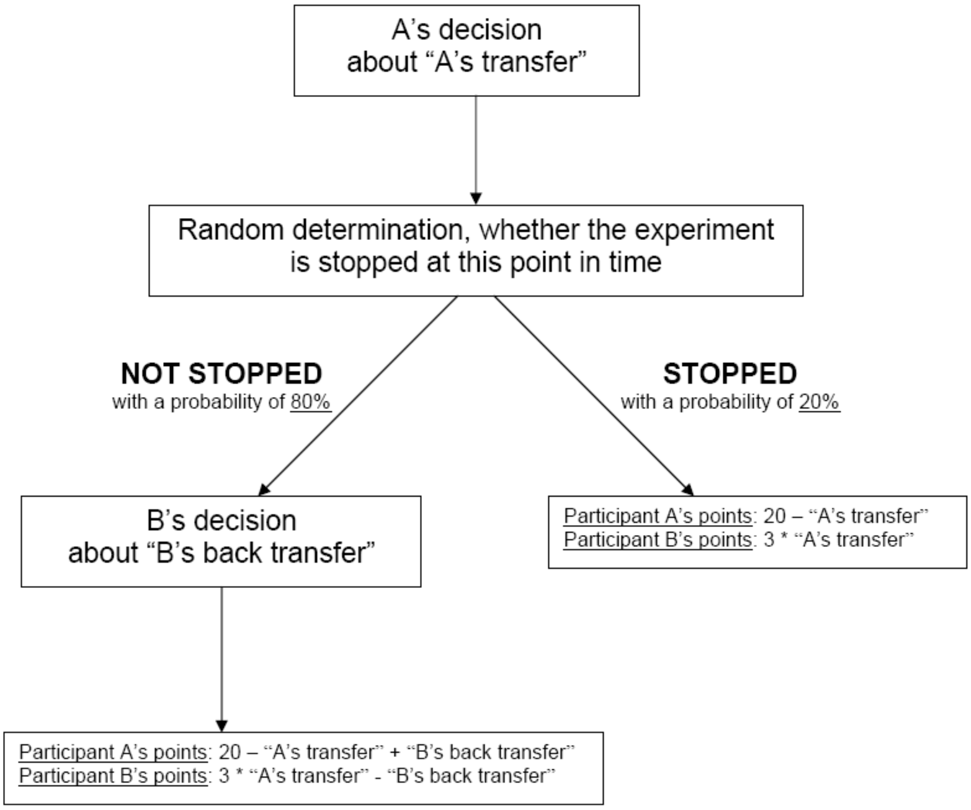
How many points do you transfer back to participant A in case participant A transferred 20 and you have a total amount of 60 points available?

Participant B makes this decision only if the experiment was not stopped right after participant A's decision.

Example 1: Participant A chooses a transfer of 15 points. Then, it is randomly determined that the experiment is stopped right after participant A's decision. Participant A receives 20 – 15 points = 5 points. Participant B receives 3 * 15 points = 45 points.

Example 2: Participant A chooses a transfer of 15 points. Then, it is randomly determined that the experiment is not stopped right after participant A's decision. Participant B chooses a back transfer of 39 points if participant A transferred 15 points. Participant A receives 20 – 15 + 39 points = 44 points. Participant B receives 3 * 15 – 39 points = 6 points.

The procedure is illustrated by the following graph:


A's decision
about "A's transfer"
Random determination, whether the experiment is stopped at this point in time
NOT STOPPED
with a probability of 80%
STOPPED
with a probability of 20%
B's decision
about "B's back transfer"
Participant A's points: 20 – "A's transfer"
Participant B's points: 3 * "A's transfer"
Participant A's points: 20 – "A's transfer" + "B's back transfer"
Participant B's points: 3 * "A's transfer" - "B's back transfer"

After this procedure participant A and participant B are both informed about participant A's transfer, about whether the experiment was stopped right after participant A's decision, and - in case the experiment was not stopped right after participant A's decision - about participant B's back transfer. Then, the experiment ends. The procedure is not repeated.

During the course of the experiment you might be asked to answer questions. The answers to these questions do not affect the payments and the procedure of the experiment. They are treated anonymously and are not sent to your matched participant or any other participant.

Before you are informed whether you are participant A or participant B and the experiment starts, you are asked to answer several control questions concerning the procedure of the experiment.

If you have any questions, please raise your hand. An instructor of the experiment will then come to your seat in order to answer your questions.

3.6.2 Behavioral predictions of the modified version of the model by Dufwenberg and Kirchsteiger (2004)

The basic model by Dufwenberg and Kirchsteiger (2004)

In Dufwenberg and Kirchsteiger (2004) individual i's utility function in a 2-player game with individual j is defined in the following way:

$$U_i = \pi_i + Y_i * \kappa_i * \lambda_i,$$

where π_i represents individual i's expectation of his own material payoff that depends on his strategy and his belief about individual j's strategy, $Y_i \geqslant 0$ individual i's parameter of sensitivity to reciprocity concerns, κ_i individual i's perception of the kindness of his own strategy, and λ_i individual i's perception of the kindness of individual j's strategy. Y_i is a parameter that is exogenously given, whereas π_i, κ_i, and λ_i depend on individual i's strategy, individual i's belief about individual j's strategy, and individual i's belief about individual j's belief about individual i's strategy.

Dufwenberg and Kirchsteiger (2004) define κ_i as individual i's expectation of individual j's material payoff minus a reference payoff which is the mean of the maximum and the minimum expected material payoff individual i beliefs he could assign to individual j by varying his strategy.[71] λ_i is defined as individual i's belief about individual j's expectation of individual i's material payoff minus a reference payoff which is the mean of the maximum and the minimum expected material payoff individual i beliefs that individual j beliefs he could assign to individual i by varying his strategy.

Note that an individual's beliefs are updated in the course of the game and, therefore, may differ after different histories of play. Updated beliefs after a given history equal initial beliefs, except for the choices that were already made and lead to the given history. Updated beliefs assign a probability of 1 to already made choices. Consider, for example, individual i that initially believes individual j to play action a with probability p and action b with probability $1 - p$ (which may, indeed, be correct). After

[71] Dufwenberg and Kirchsteiger (2004) define the reference payoff in a more general way that is equivalent to our notion in our setup.

individual j's action a has realized, individual i believes that individual j has chosen a with probability 1 (and not p). As beliefs are updated, also an individual's perception of the kindness of his own strategy and of the other individual's strategy are updated in the course of the game and may differ after different histories of play.

Dufwenberg and Kirchsteiger (2004) introduce the sequential reciprocity equilibrium (SRE) in which each player in each of his decision nodes makes choices that maximize his utility for the given history, given his updated first and second order beliefs, and given that he follows his equilibrium strategy at other histories. Furthermore, all players' initial first and second order beliefs are correct.

Our modification

Dufwenberg and Kirchsteiger (2004) restrict attention to finite multi-stage games without nature. For our context, we could simply use their framework and consider nature as a third player who always chooses to stop the game with probability $1 - q$ and to continue the game with probability q, and to whom agent A and agent B are insensitive to reciprocity concerns. This, however, leads to an unintuitive way of evaluating agent A's kindness in the course of the game: At the beginning of the game agent B has some initial belief about agent A's strategy, nature's strategy, and agent A's belief about nature's strategy. After agent A's chosen amount is transferred and the lottery has chosen to continue the game, agent B's updated beliefs are that agent A has chosen the given transfer (with probability 1), that nature has chosen to continue the game (with probability 1), and that agent A believes that nature has chosen to continue the game (with probability 1). If agent B evaluates the kindness of agent A's strategy given his updated beliefs, he takes into consideration that agent A believes that nature has chosen to continue the game with probability 1. However, agent A's belief about nature's strategy was different at agent A's decision node and, therefore, agent A's intentions were different.

In order to avoid that, we undertake a small and natural modification of the basic model by Dufwenberg and Kirchsteiger (2004). Our modification consists in the way how agent B perceives the kindness of agent A's strategy in the course of the game. At agent B's decision node we let him evaluate the kindness of agent A's strategy on the basis of his belief that agent A believes that nature has chosen to continue the game with probability q rather than with probability 1.

Agent B's utility function when he is asked to decide

Consider agent A has chosen on x and the lottery has determined to continue the game. Agent B, then, decides on $y(x) \in [0, 3 * x]$.[72] At his decision node he believes that agent A has chosen x (with probability 1), that nature has chosen to continue the game (with probability 1), and that agent A believes that agent B returns $\widetilde{y}(0) = 0$, $\widetilde{y}(5) \in [0, 15]$, $\widetilde{y}(10) \in [0, 30]$, $\widetilde{y}(15) \in [0, 45]$, $\widetilde{y}(20) \in [0, 60]$, where $\widetilde{y}(x)$ represents agent B's second order belief for x.

Then, agent B's expectation of his own material payoff is equal to

$$\pi_B\left(y(x), x\right) = 3 * x - y(x),$$

and agent B's perception of the kindness of his own strategy, $y(x)$, is equal to

$$\kappa_B\left(y(x), x\right) = 20 - x + y(x) - ref_{\kappa_B}\left(x\right), \text{ with } ref_{\kappa_B}\left(x\right) = \tfrac{(20-x+0)+(20-x+3*x)}{2}.$$

The first term of $\kappa_B\left(y(x), x\right)$, $20 - x + y(x)$, refers to agent B's expectation of agent A's material payoff, while the second, $ref_{\kappa_B}\left(x\right)$, to the corresponding reference payoff that is the mean of the minimum he (believes he) can assign to agent A with $y(x) \in [0, 3 * x]$, $20 - x + 0$, and its maximum, $20 - x + 3 * x$.

Furthermore, agent B's perception of the kindness of agent A's strategy, x, is equal to

$$\lambda_B\left(\widetilde{y}(\cdot), x\right) = 3 * x - q * \widetilde{y}(x) - ref_{\lambda_B}\left(\widetilde{y}(\cdot)\right).$$

The first term of $\lambda_B\left(\widetilde{y}(\cdot), x\right)$, $3 * x - q * \widetilde{y}(x)$, represents agent B's belief about agent A's expectation of agent B's material payoff, which depends on agent B's belief about

[72] Here and in the following we focus on agent B's pure strategies only.

agent A's action, x, as well as agent B's belief about agent A's belief about agent B's strategy, $\widetilde{y}(x)$, and nature's move. The second term of λ_B, $ref_{\lambda_B}(\widetilde{y}(\cdot))$, represents the corresponding reference payoff that is the mean of the minimum agent B believes agent A believes he (agent A) can assign to agent B with $x \in \{0, 5, 10, 15, 20\}$ and its maximum. As $\widetilde{y}(x) \in [0, 3 * x]$, the minimum of $3 * x - q * \widetilde{y}(x)$ is equal to 0 which is attained at $x = 0$. The maximum of $3 * x - q * \widetilde{y}(x)$ depends on agent B's second order beliefs $\widetilde{y}(x)$ for all x. Note that it is not necessarily equal to $3 * 20 - q * \widetilde{y}(20)$ which is attained at $x = 20$.

Hence, agent B's utility function is the following

$$U_B(y(x), x, \widetilde{y}(\cdot)) = \pi_B(y(x), x) + Y_B * \kappa_B(y(x), x) * \lambda_B(\widetilde{y}(\cdot), x) =$$
$$3 * x - y(x) + Y_B * \left(y(x) - \tfrac{3*x}{2}\right) * (3 * x - q * \widetilde{y}(x) - ref_{\lambda_B}(\widetilde{y}(\cdot))).$$

Equilibrium predictions

In this subsection we derive some statements that hold in any SRE in which agent B chooses a pure strategy $y(x) \in [0, 3 * x]$ for all $x \in \{0, 5, 10, 15, 20\}$.

1. $y(x)$ **(weakly) increases in** x $\forall$ $q \in (0, 1)$.

Suppose not. Then there exist $x', x \in \{0, 5, 10, 15, 20\}$ with $x' > x$ but $y(x') < y(x)$. As in any SRE initial beliefs about strategies are correct, e.g. $y(x) = \widetilde{y}(x)$ and $y(x') = \widetilde{y}(x')$, agent B believes that agent A intends to assign him more expected payoff with x' than with x since $3 * x' - y(x') * q > 3 * x - y(x) * q$. As $ref_{\lambda_B}(\widetilde{y}(\cdot))$ is the same under x' and x, we have $\lambda_B(\widetilde{y}(\cdot), x') > \lambda_B(\widetilde{y}(\cdot), x)$, i.e. agent B perceives strategy x' as kinder than strategy x. Nevertheless, agent B returns less when he receives $3 * x'$ than when he receives $3 * x$. From revealed preferences it must be the case that:

$$U_B(y(x'), x', \widetilde{y}(\cdot)) \geq U_B(y(x), x', \widetilde{y}(\cdot))$$

and

$$U_B(y(x), x, \widetilde{y}(\cdot)) \geq U_B(y(x'), x, \widetilde{y}(\cdot))$$

because $y(x)$ is available given x' (since $y(x) \leq 3 * x < 3 * x'$), and $y(x')$ is available given x (since $y(x') < y(x) \leq 3 * x$). The two inequalities can be written as

$$3 * x' - y(x') + Y_B * \left(y(x') - \tfrac{3}{2} * x'\right) * \lambda_B\left(\widetilde{y}(\cdot), x'\right) \geq$$
$$3 * x' - y(x) + Y_B * \left(y(x) - \tfrac{3}{2} * x'\right) * \lambda_B\left(\widetilde{y}(\cdot), x'\right)$$

and

$$3 * x - y(x) + Y_B * \left(y(x) - \tfrac{3}{2} * x\right) * \lambda_B\left(\widetilde{y}(\cdot), x\right) \geq$$
$$3 * x - y(x') + Y_B * \left(y(x') - \tfrac{3}{2} * x\right) * \lambda_B\left(\widetilde{y}(\cdot), x\right)$$

which can be rewritten as

$$\tfrac{1}{Y_B} \geq \lambda_B\left(\widetilde{y}(\cdot), x'\right) \text{ and } \tfrac{1}{Y_B} \leq \lambda_B\left(\widetilde{y}(\cdot), x\right).$$

This implies $\lambda_B\left(\widetilde{y}(\cdot), x\right) \geq \lambda_B\left(\widetilde{y}(\cdot), x'\right)$ which is a contradiction.

2. $(3 * x - y(x) * q)$ **(weakly) increases in** x $\forall\ q \in (0, 1)$**.**

Suppose not. Then, there exist $x', x \in \{0, 5, 10, 15, 20\}$ with $x' > x$ but $(3 * x' - y(x') * q) < (3 * x - y(x) * q)$. As in any SRE initial beliefs about strategies are correct, e.g. $y(x) = \widetilde{y}(x)$ and $y(x') = \widetilde{y}(x')$, agent B believes that agent A intends so assign him less payoff with x' than with x. As $ref_{\lambda_B}\left(\widetilde{y}(\cdot)\right)$ is the same under x' and x, we have $\lambda_B\left(\widetilde{y}(\cdot), x'\right) < \lambda_B\left(\widetilde{y}(\cdot), x\right)$, i.e. agent B perceives strategy x as kinder than strategy x'. From revealed preferences it must be the case that:

$$U_B\left(y(x'), x', \widetilde{y}(\cdot)\right) \geq U_B\left(y(x), x', \widetilde{y}(\cdot)\right)$$

and

$$U_B\left(y(x), x, \widetilde{y}(\cdot)\right) \geq U_B\left(\left(y(x') - 3 * (x' - x)\right), x, \widetilde{y}(\cdot)\right)$$

because $y(x)$ is available given x' (since $y(x) \leq 3 * x < 3 * x'$), and $y(x') - 3 * (x' - x)$ is available given x. The latter is true since $y(x') - 3 * (x' - x) \leq 3 * x$, which is equivalent to $y(x') \leq 3 * x'$, and $y(x') - 3 * (x' - x) \geq 0$ since the above assumed $(3 * x' - y(x') * q) < (3 * x - y(x) * q)$ implies $y(x') - 3 * (x' - x) > y\,(x) \geq 0$. The two inequalities can be written as

$$3 * x' - y(x') + Y_B * \left(y(x') - \tfrac{3}{2} * x'\right) * \lambda_B\left(\widetilde{y}(\cdot), x'\right) \geq$$
$$3 * x' - y(x) + Y_B * \left(y(x) - \tfrac{3}{2} * x'\right) * \lambda_B\left(\widetilde{y}(\cdot), x'\right)$$

and

$$3 * x - y(x) + Y_B * \left(y(x) - \tfrac{3}{2} * x\right) * \lambda_B\left(\widetilde{y}(\cdot), x\right) \geq$$
$$3 * x - (y(x') - 3 * (x' - x)) + Y_B * \left((y(x') - 3 * (x' - x)) - \tfrac{3}{2} * x\right) * \lambda_B\left(\widetilde{y}(\cdot), x\right).$$

$3 * x' - y(x') * q < 3 * x - y(x) * q$ is equivalent to $3 * (x' - x) < q * (y(x') - y(x))$ and implies (i) $y(x') > y(x)$ and (ii) $3 * (x' - x) < y(x') - y(x)$ which is equivalent to $y(x') - 3 * (x' - x) > y(x)$. Therefore, we can be rewrite the two inequalities as

$$\lambda_B\left(\widetilde{y}(\cdot), x'\right) \geq \tfrac{1}{Y_B} \text{ and } \lambda_B\left(\widetilde{y}(\cdot), x\right) \leq \tfrac{1}{Y_B}.$$

This implies $\lambda_B\left(\widetilde{y}(\cdot), x'\right) \geq \lambda_B\left(\widetilde{y}(\cdot), x\right)$ which is a contradiction.

3. $\lambda_B\left(\widetilde{y}(\cdot), x\right)$ **(weakly) increases in** x $\forall$ $q \in (0, 1)$**.**

As in any SRE initial beliefs about strategies are correct, e.g. $y(x) = \widetilde{y}(x)$ for all $x \in \{0, 5, 10, 15, 20\}$, and our second property holds, agent B believes that agent A intends to assign him (weakly) more expected material payoff the higher x. As $ref_{\lambda_B}\left(\widetilde{y}(\cdot)\right)$ is the same for any $x \in \{0, 5, 10, 15, 20\}$, $\lambda_B\left(\widetilde{y}(\cdot), x\right)$ (weakly) increases in x.

4. The higher q **the (weakly) smaller** $y(x)$ $\forall$ $q \in (0, 1)$ **and** $x = 20$**.**

Suppose not. Then, there exist $q', q \in (0, 1)$ with $q' > q$ and an SRE for q' with $y(20)_{q'}$ and an SRE for q with $y(20)_q$ such that $y(20)_{q'} > y(20)_q$. As in any SRE initial beliefs about strategies are correct, agent B believes that agent A intends to assign him more expected payoff with $x = 20$ when the probability that the game is not stopped is q rather than q' because $3 * 20 - y(20)_q * q > 3 * 20 - y(20)_{q'} * q'$. $ref_{\lambda_B}(\widetilde{y}(\cdot))$ may be different for q' and q. Due to our second property and correct initial beliefs about strategies, we can simply calculate $ref_{\lambda_B}(\widetilde{y}(\cdot))$ as the mean of the minimum agent B believes agent A believes he (agent A) can assign to agent B with $x \in \{0, 5, 10, 15, 20\}$, which is attained at $x = 0$, and its maximum, which is attained at $x = 20$. Hence, $ref_{\lambda_B}\left(\widetilde{y}(\cdot)_q\right)_q = (3 * 20 - y(20)_q * q) * \frac{1}{2}$ and $ref_{\lambda_B}\left(\widetilde{y}(\cdot)_{q'}\right)_{q'} = (3 * 20 - y(20)_{q'} * q') * \frac{1}{2}$. As a consequence, agent B perceives $x = 20$ as kinder in the SRE with q than in the one with q', i.e. $\lambda_B\left(\widetilde{y}(\cdot)_q, 20\right)_q > \lambda_B\left(\widetilde{y}(\cdot)_{q'}, 20\right)_{q'}$.[73] This is because $(3 * 20 - y(20)_q * q) * \frac{1}{2} > (3 * 20 - y(20)_{q'} * q') * \frac{1}{2}$. Nevertheless, agent B returns more in the SRE with q' than in the one with q. From revealed preferences it must be the case that:

$$U_B\left(y(20)_{q'}, 20, \widetilde{y}(\cdot)_{q'}\right)_{q'} \geq U_B\left(y(20)_q, 20, \widetilde{y}(\cdot)_{q'}\right)_{q'}$$

and

$$U_B\left(y(20)_q, 20, \widetilde{y}(\cdot)_q\right)_q \geq U_B\left(y(20)_{q'}, 20, \widetilde{y}(\cdot)_q\right)_q$$

because $y(20)_q$ and $y(20)_{q'}$ are both available given $x = 20$. The two inequalities can be written as

$$\begin{gathered} 3 * 20 - y(20)_{q'} + Y_B * \left(y(20)_{q'} - \tfrac{3}{2} * 20\right) * \lambda_B\left(\widetilde{y}(\cdot)_{q'}, 20\right)_{q'} \geq \\ 3 * 20 - y(20)_q + Y_B * \left(y(20)_q - \tfrac{3}{2} * 20\right) * \lambda_B\left(\widetilde{y}(\cdot)_{q'}, 20\right)_{q'} \end{gathered}$$

and

$$\begin{gathered} 3 * 20 - y(20)_q + Y_B * \left(y(20)_q - \tfrac{3}{2} * 20\right) * \lambda_B\left(\widetilde{y}(\cdot)_q, 20\right)_q \geq \\ 3 * 20 - y(20)_{q'} + Y_B * \left(y(20)_{q'} - \tfrac{3}{2} * 20\right) * \lambda_B\left(\widetilde{y}(\cdot)_q, 20\right)_q \end{gathered}$$

which can be rewritten as

$$\lambda_B\left(\widetilde{y}(\cdot)_{q'}, 20\right)_{q'} \geq \tfrac{1}{Y_B} \text{ and } \lambda_B\left(\widetilde{y}(\cdot)_q, 20\right)_q \leq \tfrac{1}{Y_B}.$$

[73] This may not be the case for $x < 20$.

This implies $\lambda_B\left(\widetilde{y}(\cdot)_{q'},20\right)_{q'} \geq \lambda_B\left(\widetilde{y}(\cdot)_q,20\right)_q$ which is a contradiction.

5. If $y(20)_q = y(20)_{q'}$ for $q',q \in (0,1)$ with $q' > q$, then either $y(20)_q = y(20)_{q'} = 60$ or $y(20)_q = y(20)_{q'} = 0$.

Suppose not. Then, there exist $q',q \in (0,1)$ with $q' > q$ and an SRE for q' with $y(20)_{q'}$ and an SRE for q with $y(20)_q$ such that $0 < y(20)_{q'} = y(20)_q < 60$. As in any SRE initial beliefs about strategies are correct, agent B believes that agent A intends to assign him more expected payoff with $x = 20$ when the probability that the game is not stopped is q rather than q' because $3*20-y(20)_q*q > 3*20-y(20)_{q'}*q'$ (with $y(20)_{q'} = y(20)_q > 0$). Due to our second property and correct initial beliefs about strategies, we can simply calculate $ref_{\lambda_B}\left(\widetilde{y}(\cdot)_q\right)_q = (3*20-y(20)_q*q)*\frac{1}{2}$ and $ref_{\lambda_B}\left(\widetilde{y}(\cdot)_{q'}\right)_{q'} = (3*20-y(20)_{q'}*q')*\frac{1}{2}$. As a consequence, agent B perceives $x = 20$ as kinder in the SRE with q than in the one with q', i.e. $\lambda_B\left(\widetilde{y}(\cdot)_q,20\right)_q > \lambda_B\left(\widetilde{y}(\cdot)_{q'},20\right)_{q'}$.[74] Nevertheless, agent B returns the same in the SRE with q' as in the one with q. From revealed preferences it must be the case that:

$$U_B\left(y(20)_{q'},20,\widetilde{y}(\cdot)_{q'}\right)_{q'} \geq U_B\left(0,20,\widetilde{y}(\cdot)_{q'}\right)_{q'}$$

and

$$U_B\left(y(20)_q,20,\widetilde{y}(\cdot)_q\right)_q \geq U_B\left(60,20,\widetilde{y}(\cdot)_q\right)_q$$

because 0 and 60 are available given $x = 20$. The two inequalities can be written as

$$\begin{gathered} 3*20-y(20)_{q'}+Y_B*\left(y(20)_{q'}-\tfrac{3}{2}*20\right)*\lambda_B\left(\widetilde{y}(\cdot)_{q'},20\right)_{q'} \geq \\ 3*20-0+Y_B*\left(0-\tfrac{3}{2}*20\right)*\lambda_B\left(\widetilde{y}(\cdot)_{q'},20\right)_{q'} \end{gathered}$$

and

$$\begin{gathered} 3*20-y(20)_q+Y_B*\left(y(20)_q-\tfrac{3}{2}*20\right)*\lambda_B\left(\widetilde{y}(\cdot)_q,20\right)_q \geq \\ 3*20-60+Y_B*\left(60-\tfrac{3}{2}*20\right)*\lambda_B\left(\widetilde{y}(\cdot)_q,20\right)_q \end{gathered}$$

[74] This may not be the case for $x < 20$.

which can be rewritten as

$$\lambda_B\left(\widetilde{y}(\cdot)_{q'}, 20\right)_{q'} \geq \tfrac{1}{Y_B} \text{ and } \lambda_B\left(\widetilde{y}(\cdot)_{q}, 20\right)_{q} \leq \tfrac{1}{Y_B}.$$

This implies $\lambda_B\left(\widetilde{y}(\cdot)_{q'}, 20\right)_{q'} \geq \lambda_B\left(\widetilde{y}(\cdot)_{q}, 20\right)_{q}$ which is a contradiction.

6. $\lambda_B\left(\widetilde{y}(\cdot), 20\right)$ **is (weakly) larger when** $q = 0.5$ **than when** $q = 0.8$**.**

Suppose not. Then, there exist an SRE for $q = 0.5$ with $y(20)_{0.5}$ and an SRE for $q = 0.8$ with $y(20)_{0.8}$ such that $\lambda_B\left(\widetilde{y}(\cdot)_{0.5}, 20\right)_{0.5} < \lambda_B\left(\widetilde{y}(\cdot)_{0.8}, 20\right)_{0.8}$. Due to our second property and correct initial beliefs about strategies, this implies that $(60 - 0.5 * y(20)_{0.5}) * \frac{1}{2} < (60 - 0.8 * y(20)_{0.8}) * \frac{1}{2}$ and, therefore, $y(20)_{0.5} > y(20)_{0.8}$. From revealed preferences it must be the case that:

$$U_B\left(y(20)_{0.5}, 20, \widetilde{y}(\cdot)_{0.5}\right)_{0.5} \geq U_B\left(y(20)_{0.8}, 20, \widetilde{y}(\cdot)_{0.5}\right)_{0.5}$$

and

$$U_B\left(y(20)_{0.8}, 20, \widetilde{y}(\cdot)_{0.8}\right)_{0.8} \geq U_B\left(y(20)_{0.5}, 20, \widetilde{y}(\cdot)_{0.8}\right)_{0.8}$$

because $y(20)_{0.5}$ and $y(20)_{0.8}$ are available given $x = 20$. The two inequalities can be written as

$$\begin{gathered} 3 * 20 - y(20)_{0.5} + Y_B * \left(y(20)_{0.5} - \tfrac{3}{2} * 20\right) * \lambda_B\left(\widetilde{y}(\cdot)_{0.5}, 20\right)_{0.5} \geq \\ 3 * 20 - y(20)_{0.8} + Y_B * \left(y(20)_{0.8} - \tfrac{3}{2} * 20\right) * \lambda_B\left(\widetilde{y}(\cdot)_{0.5}, 20\right)_{0.5} \end{gathered}$$

and

$$\begin{gathered} 3 * 20 - y(20)_{0.8} + Y_B * \left(y(20)_{0.8} - \tfrac{3}{2} * 20\right) * \lambda_B\left(\widetilde{y}(\cdot)_{0.8}, 20\right)_{0.8} \geq \\ 3 * 20 - y(20)_{0.5} + Y_B * \left(y(20)_{0.5} - \tfrac{3}{2} * 20\right) * \lambda_B\left(\widetilde{y}(\cdot)_{0.8}, 20\right)_{0.8} \end{gathered}$$

which can be rewritten as

$$\lambda_B\left(\widetilde{y}(\cdot)_{0.5}, 20\right)_{0.5} \geq \tfrac{1}{Y_B} \text{ and } \lambda_B\left(\widetilde{y}(\cdot)_{0.8}, 20\right)_{0.8} \leq \tfrac{1}{Y_B}.$$

This implies $\lambda_B\left(\widetilde{y}(\cdot)_{0.5}, 20\right)_{0.5} \geq \lambda_B\left(\widetilde{y}(\cdot)_{0.8}, 20\right)_{0.8}$ which is a contradiction.

7. Agent A's expected return from $x = 20$, $q * y(20)$, is (weakly) smaller when $q = 0.5$ than when $q = 0.8$.

Our sixth property states $\lambda_B\left(\widetilde{y}(\cdot)_{0.5}, 20\right)_{0.5} \geq \lambda_B\left(\widetilde{y}(\cdot)_{0.8}, 20\right)_{0.8}$. Due to our second property and correct initial beliefs about strategies, this implies $(60 - 0.5 * y(20)_{0.5}) * \frac{1}{2} \geq (60 - 0.8 * y(20)_{0.8}) * \frac{1}{2}$ which is equivalent to $0.8 * y(20)_{0.8} \geq 0.5 * y(20)_{0.5}$.

Existence of an SRE

So far, we have developed a couple of statements that hold in any SRE in which agent B chooses a pure strategy $y(x) \in [0, 3 * x]$ for all $x \in \{0, 5, 10, 15, 20\}$. In the following we show that at least one such SRE exists for each of our treatments.

Lemma 1: $\forall\ x \in \{0, 5, 10, 15, 20\}$ **and** $q \in (0, 1)$ **there exists an optimal pure action for agent** B, $y(x) \in [0, 3 * x]$, **such that agent** B**'s initial beliefs about agent** A**'s beliefs about agent** B**'s actions are all correct, i.e.** $y(x) = \widetilde{y}(x)$ **for all** $x \in \{0, 5, 10, 15, 20\}$**.**

Take an $x \in \{0, 5, 10, 15\}$, a $\widetilde{y}(20) \in [0, 60]$, and the fact that $ref_{\lambda_B}\left(\widetilde{y}(\cdot)\right) = \frac{60 - q * \widetilde{y}(20) + 0}{2}$ (as it is the case in all SRE in which agent B chooses a pure strategy due to our second property). Then, agent B's utility function is $U_B\left(y(x), x, \widetilde{y}(\cdot)\right) = 3 * x - y(x) + Y_B * \left(y(x) - \frac{3 * x}{2}\right) * \left(3 * x - q * \widetilde{y}(x) - \frac{60 - q * \widetilde{y}(20) + 0}{2}\right)$. As $U_B\left(y(x), x, \widetilde{y}(\cdot)\right)$ does not depend on $\widetilde{y}(x')$ with $x' \in \{0, 5, 10, 15\} \setminus x$ and x and $\widetilde{y}(20)$ are fixed, we rewrite agent B's utility function as $U_B\left(y(x), \widetilde{y}(x)\right)$. $U_B\left(y(x), \widetilde{y}(x)\right)$ is continuous in $y(x)$ and $\widetilde{y}(x)$, and $U_B\left(\cdot, \widetilde{y}(x)\right)$ is quasi-concave in $\widetilde{y}(x)$. By choosing a $y(x) \in G\left(\widetilde{y}(x)\right) = [0, 3 * x]$ agent B can maximize his utility. The correspondence $G\left(\widetilde{y}(x)\right)$ is constant and continuous in $\widetilde{y}(x)$. Furthermore, for any $\widetilde{y}(x)$ $G\left(\widetilde{y}(x)\right)$ is non-empty, compact, and convex-valued. Consequently, we can apply Berge's Maximum Theorem and conclude that for any $\widetilde{y}(x) \in [0, 3 * x]$ there exists at least one $y(x) \in [0, 3 * x]$ that maximizes $U_B(y(x), \widetilde{y}(x))$ and the correspondence $Y^*(\widetilde{y}(x)) : [0, 3 * x] \to [0, 3 * x]$ that maps $\widetilde{y}(x) \in [0, 3 * x]$ into the set of $y(x) \in [0, 3 * x]$ which maximize $U_B(y(x), \widetilde{y}(x))$

is non-empty, compact-valued, upper-hemicontinuous, and convex-valued. It remains to show that $Y^*(\widetilde{y}(x))$ has a fixed point $\widetilde{y}(x) \in Y^*(\widetilde{y}(x))$, i.e. agent B's initial beliefs about agent A's beliefs about agent B's actions for x are correct. We apply Kakutani's Fixed Point Theorem and conclude that at least one fixed point exists.

Now, take $x = 20$, and the fact that $ref_{\lambda_B}\left(\widetilde{y}(\cdot)\right) = \frac{60-q*\widetilde{y}(20)+0}{2}$. Then, agent B's utility function is $U_B\left(y(20), 20, \widetilde{y}(\cdot)\right) = 3*20 - y(20) + Y_B * \left(y(20) - \frac{3*20}{2}\right) * \left(3*20 - q*\widetilde{y}(20) - \frac{60-q*\widetilde{y}(20)+0}{2}\right)$. As $U_B\left(y(20), 20, \widetilde{y}(\cdot)\right)$ does not depend on $\widetilde{y}(x')$ with $x' \in \{0, 5, 10, 15\}$ and x is fixed at 20, we rewrite agent B's utility function as $U_B\left(y(20), \widetilde{y}(20)\right)$. Again, $U_B\left(y(20), \widetilde{y}(20)\right)$ is continuous in $y(20)$ and $\widetilde{y}(20)$, $U_B(\cdot, \widetilde{y}(20))$ is quasi-concave, and $[0, 60]$, the attainable set of pure actions, is continuous in $\widetilde{y}(20)$, non-empty, compact, and convex-valued. As above, we can conclude that for any $\widetilde{y}(20) \in [0, 60]$ there exists at least one $y(20) \in [0, 60]$ that maximizes $U_B(y(20), \widetilde{y}(20))$ and that there exist at least one $\widetilde{y}(20)$ that is correct.

From our second property we know that if (i) agent B has some $ref_{\lambda_B}\left(\widetilde{y}(\cdot)\right)$, which is the same under all $x \in \{0, 5, 10, 15, 20\}$, and (ii) his initial beliefs are correct, e.g. $y(x) = \widetilde{y}(x)$ for all $x \in \{0, 5, 10, 15, 20\}$, and (iii) he behaves rational in the sense that he chooses an action when its derived utility is (weakly) highest, then $3*x - \widetilde{y}(x)*q$ increases in x and $ref_{\lambda_B}\left(\widetilde{y}(\cdot)\right) = \frac{60-q*\widetilde{y}(20)+0}{2}$.

Proposition 1: For any $q \in (0, 1)$ there exists an SRE, in which agent B chooses a pure strategy.

Due to Lemma 1, it remains to show that given agent B's pure optimal strategy agent A has an optimal (possibly randomized) strategy a that is correctly expected by initial beliefs, i.e. $a = \widetilde{a}$ with $\widetilde{a}$ as agent A's initial second order belief on a.

Take any optimal pure strategy of agent B $y(x)$ for all $x \in \{0, 5, 10, 15, 20\}$ which is correctly expected by agent A. Then, agent A's utility function is $U_A\left(a, y(\cdot), \widetilde{a}\right) = \pi_A\left(a, y(\cdot)\right) + Y_A * \kappa_A\left(a, y(\cdot)\right) * \lambda_A\left(y(\cdot), \widetilde{a}\right)$. Let us define $E\left(x\right)$ and $\widetilde{E}\left(x\right)$ as the mean of x resulting with strategy a and $\widetilde{a}$, respectively, and $E\left(y\left(x\right)\right)$ and $\widetilde{E}\left(y\left(x\right)\right)$ as the mean of $y\left(x\right)$ resulting with strategy a and $\widetilde{a}$, respectively. Then, $\pi_A\left(a, y(\cdot)\right) = 20 - E\left(x\right) - q*E\left(y\left(x\right)\right)$, $\kappa_A\left(a, y(\cdot)\right) = 3*E\left(x\right) - q*E\left(y\left(x\right)\right) - \frac{0+3*20-q*y(20)}{2}$, and $\lambda_A(y(\cdot),\widetilde{a}) = 20 - \widetilde{E}(x) + q * \widetilde{E}(y(x)) - \frac{20-\widetilde{E}(x)+q*0+20-\widetilde{E}(x)+q*3*\widetilde{E}(x)}{2}$. Hence, $U_A(a,y(\cdot),\widetilde{a}) = 20 - E(x) - q*E(y(x)) + Y_A * \left(3 * E(x) - q * E(y(x)) - \frac{3*20-q*y(20)}{2}\right) * \left(q * \widetilde{E}(y(x)) - \frac{q*3*\widetilde{E}(x)}{2}\right)$. As $y(\cdot)$ is fixed, we can rewrite agent A's utility function as $U_A(a,\widetilde{a})$. $U_A(a,\widetilde{a})$ is continuous in a and $\widetilde{a}$, $U_A(\cdot,\widetilde{a})$ is quasi-concave, and agent A's set of possibly randomized strategies X is continuous in $\widetilde{a}$, non-empty, compact and convex-valued. Hence, we can apply Berge's Maximum Theorem and conclude that for any $\widetilde{a}$ there exists a set of strategies $X^*(\widetilde{a})$ out of which each strategy is part of the set X and maximizes agent A's utility given $\widetilde{a}$. Furthermore, $X^*(\widetilde{a}) : X \rightarrow X$ is a non-empty, compact, convex-valued, and upper-hemicontinuous correspondence. Consequently, we can apply Kakutani's Fixed Point Theorem and conclude that $X^*(\widetilde{a})$ has at least one fixed point.

3.6.3 Behavioral predictions of the modified version of the model by Dufwenberg and Kirchsteiger (2004) with central elements of the model by Falk and Fischbacher (2006)

The model

We consider the same utility function of individual i as in the modified version of the model by Dufwenberg and Kirchsteiger (2004) but define κ_i and λ_i differently. The interpretation of these terms, though, remains the same. The reference payoff used for κ_i is equal to individual i's expectation of his own material payoff, π_i, while the reference payoff used for λ_i is equal to individual i's belief about individual j's expectation of individual j's material payoff. Everything else remains the same.

Agent B's utility function when he is asked to decide

In comparison to agent B's corresponding utility function in the modified version of the model by Dufwenberg and Kirchsteiger (2004), $\kappa_B(y(x),x)$ and $\lambda_B(\widetilde{y}(x),x)$ change. Now,

$$\kappa_B\left(y(x), x\right) = 20 - x + y\left(x\right) - \left(3 * x - y\left(x\right)\right)$$

and

$$\lambda_B\left(\widetilde{y}(x), x\right) = 3 * x - q * \widetilde{y}(x) - \left(20 - x + q * \widetilde{y}(x)\right).$$

Hence, agent B's utility function is the following

$$U_B\left(y(x), x, \widetilde{y}(x)\right) = 3*x - y(x) + Y_B*\left(20 - 4 * x + 2 * y\left(x\right)\right)*\left(4 * x - 2 * q * \widetilde{y}(x) - 20\right).$$

Equilibrium predictions

In this subsection we derive some statements that hold in any SRE in which agent B chooses a pure strategy $y(x) \in [0, 3 * x]$ for all $x \in \{0, 5, 10, 15, 20\}$.

1. $y(x)$ **(weakly) increases in** x $\forall$ $q \in (0, 1)$**.**

Suppose not. Then there exist $x', x \in \{0, 5, 10, 15, 20\}$ with $x' > x$ but $y(x') < y(x)$. As in any SRE initial beliefs about strategies are correct, e.g. $y(x) = \widetilde{y}(x)$ and $y(x') = \widetilde{y}(x')$, $\lambda_B\left(\widetilde{y}(x'), x'\right) > \lambda_B\left(\widetilde{y}(x), x\right)$ since $4*x' - 2*q*y(x') - 20 > 4*x - 2*q*y(x) - 20$. Nevertheless, agent B returns less when he receives $3 * x'$ than when he receives $3 * x$. From revealed preferences it must be the case that:

$$U_B\left(y(x'), x', \widetilde{y}(x')\right) \geq U_B\left(y(x), x', \widetilde{y}(x')\right)$$

and

$$U_B\left(y(x), x, \widetilde{y}(x)\right) \geq U_B\left(y(x'), x, \widetilde{y}(x)\right)$$

because $y(x)$ is available given x' (since $y(x) \leq 3 * x < 3 * x'$), and $y(x')$ is available given x (since $y(x') < y(x) \leq 3 * x$). The two inequalities can be written as

$$\begin{gathered} 3 * x' - y(x') + Y_B * \left(20 - 4 * x' + 2 * y\left(x'\right)\right) * \lambda_B\left(\widetilde{y}(x'), x'\right) \geq \\ 3 * x' - y(x) + Y_B * \left(20 - 4 * x' + 2 * y\left(x\right)\right) * \lambda_B\left(\widetilde{y}(x'), x'\right) \end{gathered}$$

and

$$3 * x - y(x) + Y_B * (20 - 4 * x + 2 * y(x)) * \lambda_B(\widetilde{y}(x), x) \geq$$
$$3 * x - y(x') + Y_B * (20 - 4 * x + 2 * y(x')) * \lambda_B(\widetilde{y}(x), x)$$

which can be rewritten as

$$\tfrac{1}{2*Y_B} \geq \lambda_B(\widetilde{y}(x'), x') \text{ and } \tfrac{1}{2*Y_B} \leq \lambda_B(\widetilde{y}(x), x).$$

This implies $\lambda_B(\widetilde{y}(x), x) \geq \lambda_B(\widetilde{y}(x'), x')$ which is a contradiction.

2. $\lambda_B(\widetilde{y}(x), x)$ (weakly) increases in x $\forall$ $q \in \{0.5, 0.8\}$.

Suppose not. Then, there exist $x', x \in \{0, 5, 10, 15, 20\}$ with $x' > x$ but $\lambda_B(\widetilde{y}(x'), x') < \lambda_B(\widetilde{y}(x), x)$. As in any SRE initial beliefs about strategies are correct, e.g. $y(x) = \widetilde{y}(x)$ and $y(x') = \widetilde{y}(x')$, this implies $4 * x' - 2 * q * y(x') - 20 < 4 * x - 2 * q * y(x) - 20$ which is equivalent to $4 * (x' - x) < 2 * q * (y(x') - y(x))$ and implies $y(x') > y(x)$. Furthermore, for $q \in \{0.5, 0.8\}$ $y(x) < 3 * x$ in SRE. If not and $y(x) = \widetilde{y}(x) = 3 * x$, $\lambda_B(\widetilde{y}(x), x) = 4 * x - 2 * 3 * x * q - 20$, which is equal or smaller than 0 for $q \in \{0.5, 0.8\}$. Given $\lambda_B(\widetilde{y}(x), x) \leq 0$, agent B preferred to return nothing instead of $y(x) = 3 * x$.

From revealed preferences it must be the case that:

$$U_B(y(x'), x', \widetilde{y}(x')) \geq U_B(y(x), x', \widetilde{y}(x'))$$

and

$$U_B(y(x), x, \widetilde{y}(x)) \geq U_B(3 * x, x, \widetilde{y}(x))$$

because $y(x)$ is available given x' (since $y(x) < 3 * x < 3 * x'$), and $3 * x$ is available given x. The two inequalities can be written as

$$3 * x' - y(x') + Y_B * (20 - 4 * x' + 2 * y(x')) * \lambda_B(\widetilde{y}(x'), x') \geq$$
$$3 * x' - y(x) + Y_B * (20 - 4 * x' + 2 * y(x)) * \lambda_B(\widetilde{y}(x'), x')$$

and

$$3 * x - y(x) + Y_B * (20 - 4 * x + 2 * y\,(x)) * \lambda_B\,(\widetilde{y}(x), x) \geq$$
$$3 * x - 3 * x + Y_B * (20 - 4 * x + 2 * 3 * x) * \lambda_B\,(\widetilde{y}(x), x).$$

As $y(x') > y(x)$ and $y(x) < 3 * x$, the two inequalities can be rewritten as

$$\lambda_B\,(\widetilde{y}(x'), x') \geq \tfrac{1}{2*Y_B} \text{ and } \lambda_B\,(\widetilde{y}(x), x) \leq \tfrac{1}{2*Y_B}.$$

This implies $\lambda_B\,(\widetilde{y}(x'), x') \geq \lambda_B\,(\widetilde{y}(x), x)$ which is a contradiction.

3. The higher q the (weakly) smaller $y(x)$ $\forall\ q \in (0, 1)$ and $x \in \{0, 5, 10, 15, 20\}$.

Suppose not. Then, there exist $q', q \in (0, 1)$ with $q' > q$ and an SRE for q' with $y(x)_{q'}$ and an SRE for q with $y(x)_q$ such that $y(x)_{q'} > y(x)_q$. As in any SRE initial beliefs about strategies are correct, e.g. $y(x)_{q'} = \widetilde{y}(x)_{q'}$ and $y(x)_q = \widetilde{y}(x)_q$, $\lambda_B\,(\widetilde{y}(x)_q, x)_q > \lambda_B\,(\widetilde{y}(x)_{q'}, x)_{q'}$ since $4 * x - 2 * q * y(x)_q - 20 > 4 * x - 2 * q' * y(x)_{q'} - 20$. Nevertheless, agent B returns more in the SRE with q' than in the one with q for x. From revealed preferences it must be the case that:

$$U_B\,(y(x)_{q'}, x, \widetilde{y}(x)_{q'})_{q'} \geq U_B\,(y(x)_q, x, \widetilde{y}(x)_{q'})_{q'}$$

and

$$U_B\,(y(x)_q, x, \widetilde{y}(x)_q)_q \geq U_B\,(y(x)_{q'}, x, \widetilde{y}(x)_q)_q$$

because $y(x)_q$ and $y(x)_{q'}$ are both available given x. The two inequalities can be written as

$$3 * x - y(x)_{q'} + Y_B * (20 - 4 * x + 2 * y(x)_{q'}) * \lambda_B\,(\widetilde{y}(x)_{q'}, x)_{q'} \geq$$
$$3 * x - y(x)_q + Y_B * (20 - 4 * x + 2 * y(x)_q) * \lambda_B\,(\widetilde{y}(x)_{q'}, x)_{q'}$$

and

$$3 * x - y(x)_q + Y_B * (20 - 4 * x + 2 * y(x)_q) * \lambda_B\left(\widetilde{y}(x)_q, x\right)_q \geq$$
$$3 * x - y(x)_{q'} + Y_B * (20 - 4 * x + 2 * y(x)_{q'}) * \lambda_B\left(\widetilde{y}(x)_q, x\right)_q$$

which can be rewritten as

$$\lambda_B\left(\widetilde{y}(x)_{q'}, x\right)_{q'} \geq \tfrac{1}{2*Y_B} \text{ and } \lambda_B\left(\widetilde{y}(x)_q, x\right)_q \leq \tfrac{1}{2*Y_B}.$$

This implies $\lambda_B\left(\widetilde{y}(x)_{q'}, x\right)_{q'} \geq \lambda_B\left(\widetilde{y}(x)_q, x\right)_q$ which is a contradiction.

4. For $x \in \{5, 10, 15, 20\}$ it holds that if $y(x)_q = y(x)_{q'}$ for $q', q \in (0,1)$ with $q' > q$, then either $y(x)_q = y(x)_{q'} = 3 * x$ or $y(x)_q = y(x)_{q'} = 0$.

Suppose not. Then, there exist $q', q \in (0,1)$ with $q' > q$ and an SRE for q' with $y(x)_{q'}$ and an SRE for q with $y(x)_q$ such that $0 < y(x)_{q'} = y(x)_q < 3 * x$. As in any SRE initial beliefs about strategies are correct, $\lambda_B\left(\widetilde{y}(x)_q, x\right)_q > \lambda_B\left(\widetilde{y}(x)_{q'}, x\right)_{q'}$ since $4 * x - 2 * q * y(x)_q - 20 > 4 * x - 2 * q' * y(x)_{q'} - 20$. Nevertheless, agent B returns the same in the SRE with q' as in the one with q. From revealed preferences it must be the case that:

$$U_B\left(y(x)_{q'}, x, \widetilde{y}(x)_{q'}\right)_{q'} \geq U_B\left(0, x, \widetilde{y}(x)_{q'}\right)_{q'}$$

and

$$U_B\left(y(x)_q, x, \widetilde{y}(x)_q\right)_q \geq U_B\left(3 * x, x, \widetilde{y}(x)_q\right)_q,$$

because 0 and $3 * x$ are available given x. The two inequalities can be written as

$$3 * x - y(x)_{q'} + Y_B * (20 - 4 * x + 2 * y(x)_{q'}) * \lambda_B\left(\widetilde{y}(x)_{q'}, x\right)_{q'} \geq$$
$$3 * x - 0 + Y_B * (20 - 4 * x + 2 * 0) * \lambda_B\left(\widetilde{y}(x)_{q'}, x\right)_{q'}$$

and

$$3 * x - y(x)_q + Y_B * (20 - 4 * x + 2 * y(x)_q) * \lambda_B\left(\widetilde{y}(x)_q, x\right)_q \geq$$
$$3 * x - 3 * x + Y_B * (20 - 4 * x + 2 * 3 * x) * \lambda_B\left(\widetilde{y}(x)_q, 20\right)_q$$

which can be rewritten as

$$\lambda_B\left(\widetilde{y}(x)_{q'},x\right)_{q'} \geq \tfrac{1}{2{*}Y_B} \text{ and } \lambda_B\left(\widetilde{y}(x)_{q},x\right)_{q} \leq \tfrac{1}{2{*}Y_B}.$$

This implies $\lambda_B\left(\widetilde{y}(x)_{q'},x\right)_{q'} \geq \lambda_B\left(\widetilde{y}(x)_{q},x\right)_{q}$ which is a contradiction.

5. $\lambda_B\left(\widetilde{y}(x),x\right)$ **is (weakly) larger when** $q=0.5$ **than when** $q=0.8$.

Suppose not. Then, there exist an SRE for $q=0.5$ with $y(x)_{0.5}$ and an SRE for $q=0.8$ with $y(x)_{0.8}$ such that $\lambda_B\left(\widetilde{y}(x)_{0.5},x\right)_{0.5} < \lambda_B\left(\widetilde{y}(x)_{0.8},x\right)_{0.8}$. Due to correct initial beliefs about strategies, this implies that $4*x-2*0.5*y(x)_{0.5}-20 < 4*x-2*0.8*y(x)_{0.8}-20$ and, therefore, $y(x)_{0.5} > y(x)_{0.8}$. From revealed preferences it must be the case that:

$$U_B\left(y(x)_{0.5},x,\widetilde{y}(x)_{0.5}\right)_{0.5} \geq U_B\left(y(x)_{0.8},x,\widetilde{y}(x)_{0.5}\right)_{0.5}$$

and

$$U_B\left(y(x)_{0.8},x,\widetilde{y}(x)_{0.8}\right)_{0.8} \geq U_B\left(y(x)_{0.5},x,\widetilde{y}(x)_{0.8}\right)_{0.8},$$

because $y(x)_{0.5}$ and $y(x)_{0.8}$ are available given x. The two inequalities can be written as

$$\begin{gathered}3*x-y(x)_{0.5}+Y_B*(20-4*x+2*y(x)_{0.5})*\lambda_B\left(\widetilde{y}(x)_{0.5},x\right)_{0.5} \geq \\ 3*x-y(x)_{0.8}+Y_B*(20-4*x+2*y(x)_{0.8})*\lambda_B\left(\widetilde{y}(x)_{0.5},x\right)_{0.5}\end{gathered}$$

and

$$\begin{gathered}3*x-y(x)_{0.8}+Y_B*(20-4*x+2*y(x)_{0.8})*\lambda_B\left(\widetilde{y}(x)_{0.8},x\right)_{0.8} \geq \\ 3*x-y(x)_{0.5}+Y_B*(20-4*x+2*y(x)_{0.5})*\lambda_B\left(\widetilde{y}(x)_{0.8},x\right)_{0.8}\end{gathered}$$

which can be rewritten as

$$\lambda_B\left(\widetilde{y}(x)_{0.5},x\right)_{0.5} \geq \tfrac{1}{2{*}Y_B} \text{ and } \lambda_B\left(\widetilde{y}(x)_{0.8},x\right)_{0.8} \leq \tfrac{1}{2{*}Y_B}.$$

This implies $\lambda_B\left(\widetilde{y}(x)_{0.5}, x\right)_{0.5} \geq \lambda_B\left(\widetilde{y}(x)_{0.8}, x\right)_{0.8}$ which is a contradiction.

6. Agent A's expected return from x, $q * y(x)$, is (weakly) smaller when $q = 0.5$ than when $q = 0.8$.

Our fifth property states $\lambda_B\left(\widetilde{y}(x)_{0.5}, x\right)_{0.5} \geq \lambda_B\left(\widetilde{y}(x)_{0.8}, x\right)_{0.8}$. Due to correct initial beliefs about strategies, this implies $4*x-2*0.5*y(x)_{0.5}-20 \geq 4*x-2*0.8*y(x)_{0.8}-20$ which is equivalent to $0.8 * y(x)_{0.8} \geq 0.5 * y(x)_{0.5}$.

Existence of an SRE

So far, we have developed a couple of statements that hold in any SRE in which agent B chooses a pure strategy $y(x) \in [0, 3*x]$ for all $x \in \{0, 5, 10, 15, 20\}$. In the following we show that at least one such SRE exists for each of our treatments.

Lemma 1': $\forall\, x \in \{0, 5, 10, 15, 20\}$ **and** $q \in (0,1)$ **there exists an optimal pure action for agent** B, $y(x) \in [0, 3*x]$, **such that agent** B**'s initial beliefs about agent** A**'s beliefs about agent** B**'s actions are all correct, i.e.** $y(x) = \widetilde{y}(x)$ **for all** $x \in \{0, 5, 10, 15, 20\}$**.**

Take an $x \in \{0, 5, 10, 15, 20\}$. Then, agent B's utility function is $U_B\left(y(x), x, \widetilde{y}(x)\right) = 3*x - y(x) + Y_B*(20 - 4*x + 2*y(x))*(4*x - 2*q*\widetilde{y}(x) - 20)$. As x is fixed, we rewrite agent B's utility function as $U_B\left(y(x), \widetilde{y}(x)\right)$. $U_B\left(y(x), \widetilde{y}(x)\right)$ is continuous in $y(x)$ and $\widetilde{y}(x)$, and $U_B\left(\cdot, \widetilde{y}(x)\right)$ is quasi-concave in $\widetilde{y}(x)$. By choosing a $y(x) \in G\left(\widetilde{y}(x)\right) = [0, 3*x]$ agent B can maximize his utility. The correspondence $G\left(\widetilde{y}(x)\right)$ is constant and continuous in $\widetilde{y}(x)$. Furthermore, for any $\widetilde{y}(x)$ $G\left(\widetilde{y}(x)\right)$ is non-empty, compact and convex-valued. Consequently, we can apply Berge's Maximum Theorem and conclude that for any $\widetilde{y}(x) \in [0, 3*x]$ there exists at least one $y(x) \in [0, 3*x]$ that maximizes $U_B(y(x), \widetilde{y}(x))$ and the correspondence $Y^*(\widetilde{y}(x)) : [0, 3*x] \rightarrow [0, 3*x]$ that maps $\widetilde{y}(x) \in [0, 3*x]$ into the set of $y(x) \in [0, 3*x]$ which maximize $U_B(y(x), \widetilde{y}(x))$ is non-empty, compact-valued, upper-hemicontinuous, and convex-valued. It remains to show that $Y^*(\widetilde{y}(x))$ has a fixed point $\widetilde{y}(x) \in Y^*(\widetilde{y}(x))$, i.e. agent B's initial beliefs about agent A's beliefs about agent

B's actions for x are correct. We apply Kakutani's Fixed Point Theorem and conclude that at least one fixed point exists.

Proposition 1': For any $q \in (0,1)$ there exists an SRE, in which agent B chooses a pure strategy.

Due to Lemma 1', it remains to show that given agent B's pure optimal strategy agent A has an optimal (possibly randomized) strategy a that is correctly expected by initial beliefs, i.e. $a = \widetilde{a}$ with $\widetilde{a}$ as agent A's initial second order belief on a.

Take any optimal pure strategy of agent B $y(x)$ for all $x \in \{0, 5, 10, 15, 20\}$ which is correctly expected by agent A. Then, agent A's utility function is $U_A\left(a, y(\cdot), \widetilde{a}\right) = \pi_A\left(a, y(\cdot)\right) + Y_A * \kappa_A\left(a, y(\cdot)\right) * \lambda_A\left(y(\cdot), \widetilde{a}\right)$. Let us define $E\left(x\right)$ and $\widetilde{E}\left(x\right)$ as the mean of x resulting with strategy a and $\widetilde{a}$, respectively, and $E\left(y\left(x\right)\right)$ and $\widetilde{E}\left(y\left(x\right)\right)$ as the mean of $y\left(x\right)$ resulting with strategy a and $\widetilde{a}$, respectively. Then, $\pi_A\left(a, y(\cdot)\right) = 20 - E\left(x\right) - q * E\left(y\left(x\right)\right)$, $\kappa_A\left(a, y(\cdot)\right) = 3 * E\left(x\right) - q * E\left(y\left(x\right)\right) - \left(20 - E\left(x\right) + q * E\left(y\left(x\right)\right)\right)$, and $\lambda_A\left(y(\cdot), \widetilde{a}\right) = 20 - \widetilde{E}\left(x\right) + q * \widetilde{E}\left(y\left(x\right)\right) - \left(3 * \widetilde{E}\left(x\right) - q * \widetilde{E}\left(y\left(x\right)\right)\right)$. Hence, $U_A\left(a, y(\cdot), \widetilde{a}\right) = 20 - E\left(x\right) - q * E\left(y\left(x\right)\right) + Y_A * \left(4 * E\left(x\right) - 2 * q * E\left(y\left(x\right)\right) - 20\right) * \left(20 - 4 * \widetilde{E}\left(x\right) + 2 * q * \widetilde{E}\left(y\left(x\right)\right)\right)$. As $y(\cdot)$ is fixed, we can rewrite agent A's utility function as $U_A\left(a, \widetilde{a}\right)$. $U_A\left(a, \widetilde{a}\right)$ is continuous in a and $\widetilde{a}$, $U_A\left(\cdot, \widetilde{a}\right)$ is quasi-concave, and agent A's set of possibly randomized strategies X is continuous in $\widetilde{a}$, non-empty, compact, and convex-valued. Hence, we can apply Berge's Maximum Theorem and conclude that for any $\widetilde{a}$ there exists a set of strategies $X^*\left(\widetilde{a}\right)$ out of which each strategy is part of the set X and maximizes agent A's utility given $\widetilde{a}$. Furthermore, $X^*\left(\widetilde{a}\right) : X \rightarrow X$ is a non-empty, compact, convex-valued, and upper-hemicontinuous correspondence. Consequently, we can apply Kakutani's Fixed Point Theorem and conclude that $X^*\left(\widetilde{a}\right)$ has at least one fixed point.

Chapter 4

The curse of wealth and power in an experimental jungle-game*

4.1 Introduction

The jungle – as introduced by Piccione and Rubinstein (2007) – is an economy where power relationships govern the involuntary exchange of assets. Contrary to the voluntary and mutually beneficial trade of goods in the textbook exchange economy, individuals in the jungle can use their power to seize control of assets that are held by weaker individuals. Examples for the jungle in real life are vertical hierarchies. In many organizations hierarchical power often enables stronger individuals to exploit weaker ones. For example, it is often the case that upper-level managers are authorized to decide on the allocation of time of subordinated lower-level managers and employees to certain tasks. This means that upper-level managers have the power to consume the manpower of subordinates for their own interests and careers (see, for instance, Bowles and Gintis, 1992).

An intuitive conjecture when thinking about the jungle would be that being more powerful is always an advantage in the sense that more powerful individuals will never end up less wealthy than less powerful ones. In fact, many organizations rely on (tournament) promotions to upper hierarchies with more power as instruments to incentivize

*This chapter is joint work with Matthias Sutter from the University of Innsbruck.

less powerful individuals. Piccione and Rubinstein (2004) have shown that this conjecture may be misleading, though. Their so-called "curse of wealth and power" states that if the order of power is perfectly correlated with the order of initial wealth and there are at least 3 agents, then there is always an agent (who is not the most powerful one) that finally ends up less wealthy than a less powerful agent. Hence, less power may be a blessing in the jungle. The intuition is that initial wealth and power, which enables an agent to seize the wealth of others, attracts attacks.

In this study we use the model of Piccione and Rubinstein (2004) for an experimental test of behavior in the jungle, i.e. in an economy where the involuntary exchange of wealth is determined by a hierarchical structure of power relations. We focus our attention on the following questions: (i) Do individuals exploit their hierarchical power in order to seize the wealth of less powerful individuals, as predicted by Piccione and Rubinstein (2004)? (ii) How frequently does the curse of wealth and power occur when it is expected under standard assumptions of selfishness? (iii) Does the involuntary exchange of assets create more inequality in final wealth compared to the distribution of initial wealth in the jungle?

Our results show that experimental participants try to seize the wealth of less powerful individuals *more* (instead of less) often than selfish individuals are expected to do under standard assumptions of rationality and payoff maximization. This pattern of behavior is robust to the correlation of the order of power and of initial wealth, and the costs of seizing another individual's wealth. The curse of wealth and power occurs almost always when it is predicted to occur (in the treatments with perfect correlation of the order of power and of initial wealth), but it also occurs quite frequently when it should not occur according to standard analysis (when the order of power and of initial wealth are not perfectly correlated). We also find that the behavior of individuals in the laboratory *increases* (rather than decreases) the inequality in final wealth in comparison to the initial wealth distribution.

These results are unexpected since more powerful individuals do not hesitate to exploit less powerful ones, even if that generates extremely unequal distributions of final wealth. This seems to be at odds with the abundant literature on the importance of social preferences, like fairness and inequality aversion. Recent surveys of Camerer (2003) or Fehr and Schmidt (2006), for instance, report that a considerable fraction of individuals care for a fair distribution of wealth in simple distributional games. In the penultimate section of this chapter we show that our findings could be reconciled with the presence of social preferences and that social preferences need not preclude the curse of wealth and power in our setting, though.

A few earlier experimental papers have investigated behavior in the jungle, albeit in a context that is different from Piccione and Rubinstein (2004, 2007). Durham et al. (1998) have examined the so-called paradox of power (Hirshleifer, 1991) which states that in the course of a bilateral battle an initially poorer side will gain in relative position in comparison with an initially richer side. In their experiment two parties have to allocate their exogenously given resources between a productive and an appropriative effort. The latter determines endogenously a party's strength. A higher investment in appropriation yields ceteris paribus a larger share of the common output from the productive effort. Durham et al. (1998) have found that individuals do not abstain from investing in appropriation. Furthermore, their results on the occurrence of the paradox of power are broadly in line with the theoretical predictions. Similar to Durham et al. (1998), Duffy and Kim (2005) or Powell and Wilson (2008) have experimentally analyzed situations where parties can invest into productive, appropriative, or defensive activities. Duffy and Kim (2005) have observed parties engaging in appropriative activities as predicted in the Nash equilibrium of an anarchy model. Installing a dictator (i.e., the state) has led to Pareto superior outcomes. Powell and Wilson (2008) have reported a high degree of inefficiency in their Hobbesian jungle, with parties being unwilling to sign (and keep) a constitutional contract that requests all parties to abstain from appropriative activities. In line with our results, Durham et al. (1998), Duffy and Kim (2005), and Powell and Wilson (2008) have found that individuals do not shy away from seizing the wealth of others and behaving relatively aggressively. However, in these three studies there is no clear order of power relations and each agent can attack each other agent. In our experiment, in contrast, there is a clear hierarchical structure and agents can only attack less powerful agents, i.e. an agent can be sure that he is not attacked by a less powerful agent.

The rest of this chapter is organized as follows. Section 4.2 introduces the jungle-game and our treatments. The experimental results are presented in Section 4.3. In Section 4.4 we discuss how the existence of social preferences might fit with our results. Section 4.5 concludes.

4.2 The jungle-game

4.2.1 General structure of the jungle-game

The jungle-game includes a set of N agents $\{A_1, ..., A_N\}$ that are strictly ordered by their power such that agent A_i is more powerful than agent A_j if $i < j$. Each agent A_i is initially endowed with wealth w_i according to a vector of initial endowments $\mathbf{W} = (w_1, ..., w_N)$. All agents decide simultaneously whether they want to seize the wealth of a less powerful agent. Each agent may – but need not – attack at most one less powerful agent at positive costs c. If an agent A_k is attacked simultaneously by two (or more) more powerful agents A_i and A_j, then only the most powerful agent A_i (with $i < j < k$) acquires the wealth of agent A_k. Nevertheless, all attackers have to bear the costs c. Note that acquiring the wealth of a less powerful agent may also work indirectly through a chain of attacks. If agent A_i attacks agent A_j, and agent A_j attacks agent A_k, then agent A_i receives the wealth of agents A_j and A_k, whereas agent A_j loses his initial endowment and has to bear costs c.

4.2.2 Experimental treatments

We consider the simplest possible version of Piccione and Rubinstein's (2004) jungle that consists of a set of 3 agents $\{A_1, A_2, A_3\}$ with initial endowments $\mathbf{W} = (w_1, w_2, w_3)$ and costs of c. Our treatments differ with respect how the order of power and of initial wealth are correlated since this is crucial for the existence of the curse of wealth and power. In treatments with $\mathbf{W}_{Mon}$ power and initial endowments are monotonically related, while in treatments with $\mathbf{W}_{NonMon}$ the relation between initial endowments and power is non-monotonic. Furthermore, we vary c, the costs of attacking, across treatments. Table 4.1 presents our four treatments.

Table 4.1: Treatments of the jungle-game

	W_{Mon} = (320, 200, 80)	W_{NonMon} = (320, 80, 200)
c = 20	Mon20	NonMon20
c = 60	Mon60	NonMon60

4.2.3 Experimental procedure

In each session one treatment of the jungle-game was played for five rounds in a perfect stranger matching, meaning that no participant interacted with any other participant more than once. Participants were randomly assigned the role of agent A_1, A_2, or A_3 at the beginning of each round. After each round all individuals received feedback about the actions in their group of three agents and their individual payoffs.

The experiment was conducted at the Max Planck Institute of Economics in Jena, Germany, where individuals decided at visually separated computer terminals. A total of 120 undergraduate students from the University of Jena participated in the computerized experiment (using zTree by Fischbacher, 2007, for programming and ORSEE by Greiner, 2004, for recruiting). No individual was allowed to participate in more than one session. In each treatment we had 30 participants. Individuals had to answer control questions before the decision stages. At the end of the experiment all five rounds were paid. The sessions were framed neutrally[75] and lasted less than 45 minutes. Individuals received 2.5 € for showing up on time, plus the amount of money earned in the experiment (where 100 points of final wealth corresponded to 1 €).

4.2.4 Behavioral predictions

Treatments $Mon20$ and $Mon60$ have a unique Nash equilibrium where agent A_1 attacks agent A_2, while agent A_2 does not attack agent A_3. This yields a final distribution of

[75] The experimental instructions for $Mon20$ are included in the appendix. The instructions for the other treatments are as similar as possible.

wealth $W_1 > W_3 > W_2$, where W_i denotes the final wealth of agent A_i. Hence, the curse of wealth and power is predicted to occur.

Treatments $NonMon20$ and $NonMon60$ have also a unique Nash equilibrium where agent A_1 attacks agent A_3, while agent A_2 does not attack agent A_3. As a consequence, the distribution of final wealth satisfies $W_1 > W_2 > W_3$, meaning that the order of power and of final wealth are perfectly correlated and the curse of wealth and power does not occur in equilibrium.

4.3 Experimental results

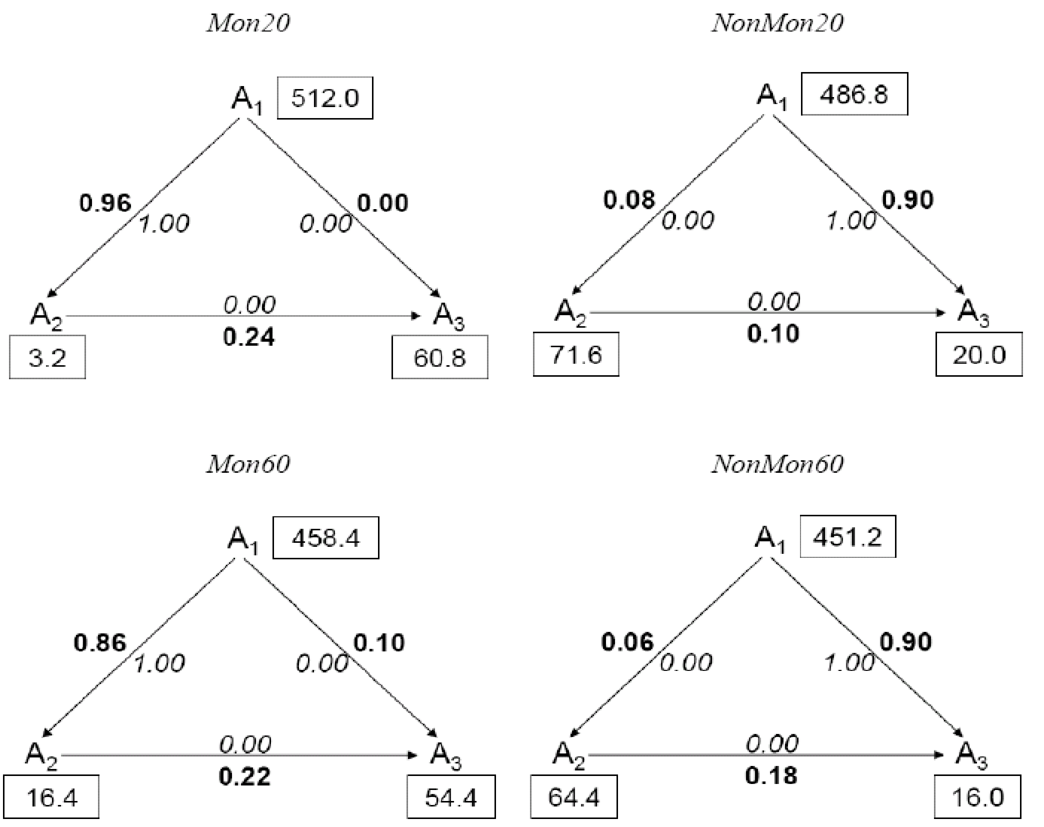
In Figure 4.1 we summarize the relative frequencies with which agents A_1 and A_2 attack less powerful agents in each treatment.

In treatment $Mon20$ agents A_1 attack the second wealthiest agent A_2 in more than 95 % of cases, which is almost completely in line with our prediction. Only in 4 % of cases agents A_1 do not attack a less powerful agent. A noteworthy finding is also that nearly one fourth of agents A_2 attack agent A_3. When the costs of attacking are higher, like in $Mon60$, the results are similar. In sum, the evidence from treatments $Mon20$ and $Mon60$ suggests that agents behave in the jungle even more aggressively than predicted under standard assumptions.

The curse of wealth and power occurs in 48 out of 50 groups (96 %) in $Mon20$, respectively in 43 out of 50 groups (86 %) in $Mon60$. In 36 (33) groups the curse is caused by equilibrium play in $Mon20$ ($Mon60$). Applying a Wilcoxon signed rank test, we can confirm that in a group of 3 members agent A_2's final wealth is significantly smaller than agent A_3's final wealth (in nearly each separate round: one-sided $p < 0.05$) in $Mon20$ as well as in $Mon60$.

Recall that in treatment $NonMon20$ and $NonMon60$ the relation between power and initial endowments is non-monotonic, which changes the equilibrium prediction in comparison to treatment $Mon20$ and $Mon60$. We observe an attack of agent A_1 on agent A_3 in 90 % of cases, both in $NonMon20$ and $NonMon60$. Contrary to the prediction, agents A_1 attack agent A_2 in some cases, perhaps due to expecting an attack

Figure 4.1: Frequencies of decisions and average final wealth in the jungle-game



Bold figures outside the triangles indicate observed frequencies of A_i attacking A_j (with $i < j$).
Italic figures inside the triangles indicate predicted frequencies of A_i attacking A_j (with $i < j$).
Figures in squared boxes report the average final wealth of a round.

of agent A_2 on agent A_3. The latter happens in 10 % (18 %) of cases in $NonMon20$ ($NonMon60$), showing again, like in $Mon20$ and $Mon60$, that individuals behave more aggressively in the jungle than predicted under standard assumptions.

The curse of wealth and power occurs in 10 % of cases both in $NonMon20$ and $NonMon60$, although it shouldn't occur in equilibrium. Overall, a Wilcoxon signed rank test confirms that in a group of 3 members $W_2 > W_3$ (in nearly each separate round: one-sided $p < 0.05$) in both treatments.

Across all treatments we observe a considerably larger degree of inequality in final wealth in comparison to the distribution of initial endowments. Overall, agents A_1 receive on average 86 % of the total final wealth, while their initial endowments amount to only 53 % of the sum of initial endowments.

We summarize the findings of this section in the following: (i) On average, agents attack other agents more often than predicted, (ii) the curse of wealth and power occurs in 91 % of cases when the order of power is perfectly correlated with the order of initial endowment (where the curse is predicted to occur), and still in 10 % of cases when the order of power is not perfectly correlated with the order of initial endowment (where the curse should not occur), (iii) the distribution of final wealth is more unequal than the distribution of initial endowments.

4.4 Reconciling the results with the existence of social preferences?

Our results have shown that individuals seize the assets of less powerful individuals in the jungle, even if that increases the inequality in final wealth substantially. At first sight, these findings seem to be at odds with the abundant literature on the importance of social preferences. In this section we would like to show that some of our results can be reconciled with the existence of social preferences. Note that this section is not intended to provide a test of social preference models, but rather to demonstrate that social preferences may not necessarily prevent individuals from seizing other individuals' assets in the jungle.

We consider a model in which 60 % of individuals are selfish and maximize their own expected material payoff and 40 % of individuals are inequity averse according to the model by Fehr and Schmidt (1999). We have chosen this model since it has been found to explain the general patterns of human behavior in series of different games quite well and because it is comparatively simple. For the sake of simplicity we assume that inequity averse individuals have a parameter of aversion to disadvantageous inequity of 2 and a parameter of aversion to advantageous inequity of 0.6.[76] Given these assumptions, the behavior in the unique Bayesian Nash equilibria is the following.[77]

In $Mon20$ all types of agents A_1 – the selfish one and the inequity averse one – have a strictly dominant action, namely the attack of agent A_2. This is because the material benefit from attacking agent A_2 is larger than the utility loss from increasing inequity by attacking agent A_2. Anticipating agent A_1's attack, agent A_2 does not attack agent A_3, irrespective of his type. In $Mon60$ only selfish agents A_1 have a strictly dominant action, namely the attack of agent A_2. As the costs of attacking are larger in $Mon60$, inequity averse agents A_1 are not willing to attack agent A_2 if he does not attack agent A_3 (and accumulate more wealth). Anticipating that at least selfish agents A_1 attack agent A_2, neither type of agent A_2 attacks agent A_3. Consequently, inequity averse agents A_1 do not attack.

Hence, a simple model of social preferences can explain why so many agents A_1 attack agent A_2 in $Mon20$. However, this model cannot predict that so many agents A_1 attack agent A_2 in $Mon60$, and that agent A_2 attacks agent A_3.

In $NonMon20$ the attack of nobody is strictly dominated by the attack of agent A_3 for both selfish and inequity averse agents A_1 because the material benefit from an attack is larger than the utility loss from an increase in inequity of final wealth. Anticipating agent A_1 attacking either agent A_2 or A_3, neither type of agent A_2 attacks agent A_3. Consequently, selfish and inequity averse agents A_1 attack agent A_3. In

[76] In the notation of Fehr and Schmidt (1999) the sensitivity to disadvantageous inequity is captured by the coefficient α, and the sensitivity to advantageous inequity by β. The distribution of types with respect to α and β is a simplification of the distribution presented in Fehr and Schmidt (1999). It has also been used in the experimental studies of Fehr et al. (2007, 2008).

[77] The derivation of the unique Bayesian Nash equilibria is included in the appendix.

$NonMon60$, in which the costs of attacking are higher, the attack of nobody is strictly dominated by the attack of agent A_3 for selfish agents A_1 only. Inequity averse agents A_1 are only willing to attack agent A_2 if he attacks agent A_3 (and accumulates more wealth) with at least some strictly positive probability. In equilibrium of $NonMon60$ the selfish agents A_1 attack agent A_2, while the inequity averse agents A_1 attack agent A_2 only with a probability of $\frac{32}{257} \approx 0.12$, but do not attack with probability $\frac{225}{257} \approx 0.88$. As a consequence, the selfish types of agent A_2 attack agent A_3, and the inequity averse agents A_2 attack agent A_3 with probability $\frac{1}{2}$.

This model of social preferences can explain why agents A_1 attack agent A_3 in $NonMon20$ and why agents A_2 attack agent A_3 in $NonMon60$. However, it does not explain why agents A_2 attack agent A_3 in $NonMon20$ and why agents A_1 attack agent A_3 in $NonMon60$.

In sum, a simple model of social preferences is compatible with several features of observed behavior that seems to be unsocial. From this it should be clear that social preferences may not necessarily make life in the jungle peaceful. Why? In this simple model of social preferences there is still a fraction of individuals that are selfish. These individuals are not disciplined by the presence of inequity averse individuals to behave less aggressively or selfishly, as it might be the case e.g. in multistage games in which punishment of selfish behavior is possible. In $NonMon60$ selfish agents A_2 are predicted to behave even more aggressively than in the self-interest model. Furthermore, inequity averse individuals themselves often do not abstain from attacking others. This is because in the model of Fehr and Schmidt (1999) the payoff inequity in comparison to each matched agent is weighted with $\frac{1}{N-1}$, with N as the number of agents in the jungle. A successful attack increases the payoff inequity in comparison to the attacked agent to a large extent, but it only slightly increases (or even decreases) the payoff inequity in comparison to the other (not attacked) agent. Hence, the material benefit from an attack (weakly) exceeds its generated utility loss from increased inequity for inequity averse agents A_1 in $Mon20$, $NonMon20$, and $NonMon60$, and for inequity averse agents A_2 in $NonMon60$.

Of course, our simple model of social preferences leaves several features of the observed behavior unexplained. This might be driven by agents having incorrect beliefs about other agents' actions. It might also be the case that social preferences play at best a minor role in the jungle when power relationships – rather than mutually beneficial and voluntary trade – determine the exchange of assets.

4.5 Conclusion

Motivated by the thought-provoking work of Piccione and Rubinstein (2004, 2007) we have introduced an experimental jungle-game to test human behavior in situations where power relationships govern the involuntary exchange of wealth. In the jungle-game more powerful agents have been able to attack less powerful ones in order to seize their wealth. With our four simple versions of this game we have shown that individuals in the jungle behave largely as predicted from the standard analysis of Piccione and Rubinstein (2004). If anything differed, they are even more aggressively attacking others than expected. As a consequence of this kind of behavior, the resulting distribution of final wealth is very unequal. As predicted for the case of a perfect correlation of the order of power and of initial endowments, the curse of wealth and power occurs almost always, and it even occurs when the order of power and of initial endowment are not perfectly correlated, in which case the curse should not appear, though. The results observed are remarkably close to the predictions derived from the model introduced by Piccione and Rubinstein (2004). Of course, our results do not necessarily imply that agents in the jungle do not have social preferences. Our findings and a simple model with social preferences indicate that social preferences – if existent – may not prevent an overwhelming majority of agents from seizing the assets of less powerful agents.

We would like to conclude with some possible implications of our experimental study. Hierarchically structured organizations in which the exchange of resources is determined by power relations are part of our daily life. Our results suggest that (i) individuals often use their power to seize the assets of less powerful agents and increase inequality and that (ii) the curse of wealth and power can occur as a consequence of such behavior. Both aspects of our results may have undesirable consequences for organizations. The exploitation of weaker individuals by more powerful individuals could decrease the incentives of weaker individuals to invest *ex ante*, e.g. in firm-specific knowledge. It may even support notions of shirking when less powerful individuals expect their work to be exploited to the benefit of superiors. In addition to this disincentive effect, which works irrespective of how the order of power and of initial wealth are correlated, the curse of wealth and power may add another disincentive effect because many organizations rely on (tournament) promotions to upper hierarchies with more power as an instrument to incentivize less powerful individuals. Lower-level managers, for example, are not incentivized by promotion prospects when middle-level managers end up worse. Hence, alarm bells should be ringing in organizations if employees compare working conditions to life in the jungle.

4.6 Appendix

4.6.1 Experimental instructions of treatment $Mon20$

Welcome to the experiment!

In the following you find the rules for this experiment. At the end of this experiment you will receive 2.50 € for showing up on time plus the amount that you can earn in this experiment.

Number of rounds and group size

This experiment consists of **5 rounds** that proceed according to the same scheme. In each round **groups of 3 member**s are formed. The composition of a group changes each round such that you are matched with each person of this room at most once. This means that you are not matched with a person of this room more than once.

Within a group each member in each round is randomly assigned a number. I.e. you are either member 1, member 2, or member 3. This number can change in each round.

A member with a higher number will be called *inferior* in the following, whereas a member with a lower number will be referred to as a *superior* member.

Endowment of each member

Each member receives an endowment.

Member 1 receives an endowment of **320** points.

Member 2 receives an endowment of **200** points.

Member 3 receives an endowment of **80** points.

Your decision – Whether or not to form a direct link

Your only decision in this experiment is whether or not you want to form a direct link to another member of your group, and – if so – to which member. All group members have to decide simultaneously whether to form a direct link or not, and – if so – to which member.

Rules for forming a direct link

Each member can only form **one** direct link, and it is only possible to form a direct link to an *inferior* member (i.e. one with a higher member number).

More precisely, member 1 can form a direct link either to member 2 or to member 3. Member 2 can only form a link to member 3. And member 3 cannot form any link.

Please note the following with respect to the formation of direct links. If two members form a direct link to the same other member, then only the link of the superior member will be successfully established, whereas the link of the inferior member will be cancelled.

An example: If both member 1 and member 2 directly link to member 3, then only member 1 has a direct link to member 3, but member 2 does not.

Costs of forming a direct link

The costs of forming a direct link are **20** points. Note that these costs have to be borne whenever you decide to form a direct link, be it successful or cancelled later on. (That means that in the above example both member 1 *and* member 2 would have to bear costs of 20 points each.)

Indirect links

Through the formation of a direct link you may also form an **indirect** link. That means that if you form successfully a link to an inferior member, then you form an indirect link to that member to whom the inferior member has formed a direct link.

Another example: If member 1 directly links to member 2, and member 2 directly links to member 3, then member 1 has a direct link to member 2 *plus* an indirect link to member 3.

Note that an indirect link of a superior member dominates a direct link of an inferior member. "Domination" means that the link of the inferior member is not valid any longer for the inferior member himself. That means that in the example just given it is member 1 who has established links to both member 2 and member 3, but that member 2 himself does not have a direct link to member 3 any longer (even though member 2 has to bear the costs of forming a direct link).

Further note that the formation of an indirect link does not have any costs.

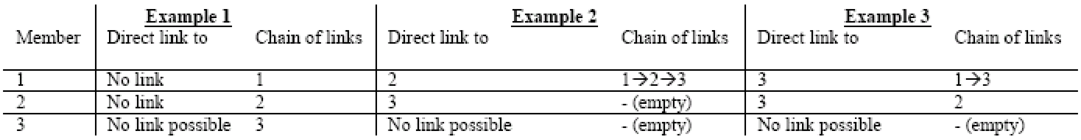
Chains of links (= sum of direct and indirect links)

Establishing direct and indirect links creates **chains of links**. A chain **starts** with the member to whom no superior member has established a link. Then, the direct link of this member and its indirect links follow.

It is possible that a member has an **empty chain of links**. This is the case if a superior member has established a link to this member.

It is possible that a member has a chain of links that consists of **his own member number only**. This is the case if (1) no superior member has established a link to this member, **and** if (2) either this member has not established a direct link or the direct link of this member is not valid due to the indirect link of a superior member.

Examples for chains of links: The following table presents some examples (Not



	Example 1		Example 2		Example 3	
Member	Direct link to	Chain of links	Direct link to	Chain of links	Direct link to	Chain of links
1	No link	1	2	1→2→3	3	1→3
2	No link	2	3	- (empty)	3	2
3	No link possible	3	No link possible	- (empty)	No link possible	- (empty)

all possible chains of links are listed.).

Your earnings of a round and the chain of links

Your earnings of a round is **the sum of the endowments of all those members** who are in **your chain of links**.

If a member formed a direct link (was it successful or cancelled later on), then the costs of forming a direct link (= 20 points) are deducted of the chain's sum of endowments.

Note that only the first member of a chain of links receives the endowments of all members of the chain. The other members of this link do no receive anything, but eventually have to bear costs of forming a direct link.

At the end of the experiment **100 points** earned will be worth **1 Euro** (= 100 Euro-cents).

4.6.2 Derivation of the Nash equilibria in our treatments

Independent of whether $c = 20$ or $c = 60$ and whether $\mathbf{W} = (w_1, w_2, w_3)$ is monotonically decreasing or not, the best response correspondences of agents A_1 and A_2 (on the other player's pure strategies) are single-valued.[78] In the following we illustrate for our four treatments the two-dimensional decision matrices and their Nash equilibria characterized by the (unique) intersection of both player's best response correspondences.[79]

[78] Agent A_3 – the least powerful agent - is not considered as an active player since his strategy set is a singleton.

[79] In the treatments with the monotonically decreasing endowment vector the Nash equilibrium strategy of agent A_1 is a strictly dominant strategy. In the treatments with the not monotonically decreasing endowment vector the Nash equilibrium equals the unique strategy combination that follows from the iterated elimination of strictly dominated actions. Hence, in all of our four treatments the presented Nash equilibria are unique.

Mon20 (c=20)

		A_2 attacks	
		nobody	A_3
	nobody	320 / 200	320 / $\boxed{260}$
A_1 attacks	A_2	$\boxed{500}$ / $\boxed{0}$	$\boxed{580}$ / -20
	A_3	380 / $\boxed{200}$	380 / 180

Mon60 (c=60)

		A_2 attacks	
		nobody	A_3
	nobody	320 / 200	320 / $\boxed{220}$
A_1 attacks	A_2	$\boxed{460}$ / $\boxed{0}$	$\boxed{540}$ / -60
	A_3	340 / $\boxed{200}$	340 / 140

NonMon20 (c=20)

		A_2 attacks	
		nobody	A_3
	nobody	320 / 80	320 / $\boxed{260}$
A_1 attacks	A_2	380 / $\boxed{0}$	$\boxed{580}$ / -20
	A_3	$\boxed{500}$ / $\boxed{80}$	500 / 60

NonMon60 (c=60)

		A_2 attacks	
		nobody	A_3
	nobody	320 / 80	320 / $\boxed{220}$
A_1 attacks	A_2	340 / $\boxed{0}$	$\boxed{540}$ / -60
	A_3	$\boxed{460}$ / $\boxed{80}$	460 / 20

4.6.3 Derivation of the Bayesian Nash equilibria in our treatments when agents are heterogeneous with respect to their utility function

We assume that 60 % of individuals are selfish and 40 % are inequity averse according to the model by Fehr and Schmidt (1999). Furthermore, we assume for inequity averse individuals that $\alpha = 2$, the parameter of aversion to disadvantageous inequality, and $\beta = 0.6$, the parameter of aversion of advantageous inequality. In our model each individual knows his own preferences but not the preferences of his matched individuals. However, each individual believes that each of the other individuals in his group is selfish with probability 0.6 and inequity averse with probability 0.4.

$Mon20$

An agent's strategy in this context specifies an action for each type of an agent. The selfish agent A_1 has a strictly dominant action, namely an attack of agent A_2. For the inequity averse agent A_1 it is also a strictly dominant action to attack agent A_2. This is due to the following two inequalities that hold for $\beta = 0.6$ and any $p \in [0, 1]$, the probability that agent A_2 attacks nobody:

$$p*\big(500-\tfrac{\beta}{2}*(500-0)-\tfrac{\beta}{2}*(500-80)\big)+(1-p)*$$
$$\big(580-\tfrac{\beta}{2}*(580+20)-\tfrac{\beta}{2}*(580-0)\big)>p*\big(320-\tfrac{\beta}{2}*(320-200)-\tfrac{\beta}{2}*(320-80)\big)+$$
$$(1-p)*\big(320-\tfrac{\beta}{2}*(320-260)-\tfrac{\beta}{2}*(320-0)\big), \text{ and}$$

$$p*\big(500-\tfrac{\beta}{2}*(500-0)-\tfrac{\beta}{2}*(500-80)\big)+(1-p)*$$
$$\big(580-\tfrac{\beta}{2}*(580+20)-\tfrac{\beta}{2}*(580-0)\big)>p*\big(380-\tfrac{\beta}{2}*(380-200)-\tfrac{\beta}{2}*(380-0)\big)+$$
$$(1-p)*\big(380-\tfrac{\beta}{2}*(380-180)-\tfrac{\beta}{2}*(380-0)\big)$$

Hence, in any Bayesian Nash equilibrium all types of agent A_1 attack agent A_2.

Given agent A_1 attacks agent A_2 with probability 1, the selfish agent A_2 attacks nobody and the inequity averse agent A_2 also attacks nobody. The latter is true since

$$0-\tfrac{\alpha}{2}*(500)-\tfrac{\alpha}{2}*(80)>-20-\tfrac{\alpha}{2}*(580+20)-\tfrac{\alpha}{2}*(20)$$

is satisfied for $\alpha=2$. Therefore, in the unique Bayesian Nash equilibrium all types of agent A_1 attack agent A_2 and all types of agent A_2 attack nobody.

Mon60

For a selfish agent A_1 it is a strictly dominant action to attack agent A_2. The inequity averse agent A_1 does not have a strictly dominant action.[80] However, the inequity averse agent A_1's action to attack agent A_3 is strictly dominated by the action to attack nobody. This is because the following inequality holds for $\beta=0.6$ and any $p\in[0,1]$, the probability that agent A_2 attacks nobody:

$$p*\big(320-\tfrac{\beta}{2}*(320-200)-\tfrac{\beta}{2}*(320-80)\big)+(1-p)*$$
$$\big(320-\tfrac{\beta}{2}*(320-220)-\tfrac{\beta}{2}*(320-0)\big)>$$
$$p*\big(340-\tfrac{\beta}{2}*(340-200)-\tfrac{\beta}{2}*(340-0)\big)+(1-p)*$$
$$\big(340-\tfrac{\beta}{2}*(340-140)-\tfrac{\beta}{2}*(340-0)\big)$$

Hence, in any Bayesian Nash equilibrium all types of agent A_1 do not attack agent A_3.

[80] If agent A_2 attacks nobody with probability 1, the inequity averse agent A_1 prefers to attack nobody. If agent A_2 attacks agent A_3 with probability 1, the inequity averse agent A_1 prefers attacking agent A_2.

Given agent A_1 attacks agent A_2 with a probability that is at least equal to 0.6, the selfish agent A_2 prefers to attack nobody since

$$s * 0 + (1 - s) * 200 > s * (-60) + (1 - s) * 220$$

for $s \in [0.6, 1]$, where s represents the probability that agent A_1 attacks agent A_2 and $1 - s$ the probability that agent A_1 attacks nobody. Similarly, the inequity averse agent A_2 prefers to attack nobody since for $\alpha = 2$ and $\beta = 0.6$ it holds that

$$\begin{gathered} s * \left(0 - \tfrac{\alpha}{2} * (460 - 0) - \tfrac{\alpha}{2} * (80 - 0)\right) + (1 - s) * \\ \left(200 - \tfrac{\alpha}{2} * (320 - 200) - \tfrac{\beta}{2} * (200 - 80)\right) > \\ s * \left(-60 - \tfrac{\alpha}{2} * (540 + 60) - \tfrac{\alpha}{2} * (60)\right) + (1 - s) * \left(220 - \tfrac{\alpha}{2} * (320 - 220) - \tfrac{\beta}{2} * (220 - 0)\right) \end{gathered}$$

for $s \in [0.6, 1]$. Hence, in any Bayesian Nash equilibrium all types of agent A_2 attack nobody.

Given agent A_2 attacks nobody with probability 1, the inequity averse agent A_1 attacks nobody since for $\beta = 0.6$

$$320 - \tfrac{\beta}{2} * (320 - 200) - \tfrac{\beta}{2} * (320 - 80) > 460 - \tfrac{\beta}{2} * (460 - 0) - \tfrac{\beta}{2} * (460 - 80) \, .$$

Therefore, in the unique Bayesian Nash equilibrium the selfish agent A_1 attacks agent A_2, the inequity averse agent A_1 attacks nobody, and all types of agent A_2 attack nobody.

$NonMon20$

The selfish agent A_1's action to attack nobody is strictly dominated by the action to attack agent A_3. Similarly, the inequity averse agent A_1's action to attack nobody is strictly dominated by the action to attack agent A_3. This is because the following inequality holds for $\beta = 0.6$ and any $p \in [0, 1]$, the probability that agent A_2 attacks nobody:

$$\begin{gathered}p * \left(500 - \tfrac{\beta}{2} * (500 - 80) - \tfrac{\beta}{2} * (500 - 0)\right) + (1 - p) * \\ \left(500 - \tfrac{\beta}{2} * (500 - 60) - \tfrac{\beta}{2} * (500 - 0)\right) > \\ p * \left(320 - \tfrac{\beta}{2} * (320 - 80) - \tfrac{\beta}{2} * (320 - 200)\right) + (1 - p) * \\ \left(320 - \tfrac{\beta}{2} * (320 - 260) - \tfrac{\beta}{2} * (320 - 0)\right)\end{gathered}$$

Hence, in any Bayesian Nash equilibrium all types of agent A_1 do not abstain from an attack.

Given that agent A_1 attacks either agent A_2 or agent A_3, the selfish agent A_2 attacks nobody. Similarly, the inequity averse agent A_2 attacks nobody since for $\alpha = 2$ and $\beta = 0.6$ it holds that

$$\begin{gathered}s * \left(0 - \tfrac{\alpha}{2} * (380 - 0) - \tfrac{\alpha}{2} * (200 - 0)\right) + (1 - s) * \left(80 - \tfrac{\alpha}{2} * (500 - 80) - \tfrac{\beta}{2} * (80 - 0)\right) > \\ s * \left(-20 - \tfrac{\alpha}{2} * (580 + 20) - \tfrac{\alpha}{2} * (20)\right) + (1 - s) * \left(60 - \tfrac{\alpha}{2} * (500 - 60) - \tfrac{\beta}{2} * (60 - 0)\right)\end{gathered}$$

for $s \in [0, 1]$, where x represents the probability that agent A_1 attacks agent A_2 and $1 - s$ the probability that agent A_1 attacks agent A_3. Hence, in any Bayesian Nash equilibrium all types of agent A_2 do not attack anybody.

Given agent A_2 attacks nobody with probability 1, a selfish agent A_1 attacks agent A_3 and an inequity averse agent A_1 attacks agent A_3. The latter is true since for $\beta = 0.6$

$$500 - \tfrac{\beta}{2} * (500 - 80) - \tfrac{\beta}{2} * (500 - 0) > 380 - \tfrac{\beta}{2} * (380 - 0) - \tfrac{\beta}{2} * (380 - 200)\,.$$

Therefore, in the unique Bayesian Nash equilibrium all types of agent A_1 attack agent A_3 and all types of agent A_2 attack nobody.

$NonMon60$

A selfish agent A_1's action to attack nobody is strictly dominated by the action to attack agent A_3. An inequity averse agent A_1's action to attack agent A_3 is strictly dominated by the action to attack nobody since for $\beta = 0.6$ and any $p \in [0, 1]$, the probability that agent A_2 attacks nobody

$$p*\big(320-\tfrac{\beta}{2}*(320-80)-\tfrac{\beta}{2}*(320-200)\big)+(1-p)*$$
$$\big(320-\tfrac{\beta}{2}*(320-220)-\tfrac{\beta}{2}*(320-0)\big)>p*\big(460-\tfrac{\beta}{2}*(460-80)-\tfrac{\beta}{2}*(460-0)\big)+$$
$$(1-p)*\big(460-\tfrac{\beta}{2}*(460-20)-\tfrac{\beta}{2}*(460-0)\big)$$

The following weak inequality holds if and only if $p \leq 0.4$:

$$p*340+(1-p)*540 \geq 460.$$

Hence, if $p < 0.4$, a selfish agent A_1 strictly prefers to attack agent A_2, if $p > 0.4$, a selfish agent A_1 strictly prefers to attack agent A_3, and if $p = 0.4$, a selfish agent A_1 is indifferent between attacking agent A_2 and agent A_3.

For $\beta = 0.6$ the following weak inequality holds if and only if $p \geq 0.2$:

$$p*\big(320-\tfrac{\beta}{2}*(320-80)-\tfrac{\beta}{2}*(320-200)\big)+(1-p)*$$
$$\big(320-\tfrac{\beta}{2}*(320-220)-\tfrac{\beta}{2}*(320-0)\big)\geq$$
$$p*\big(340-\tfrac{\beta}{2}*(340-0)-\tfrac{\beta}{2}*(340-200)\big)+(1-p)*$$
$$\big(540-\tfrac{\beta}{2}*(540+60)-\tfrac{\beta}{2}*(540-0)\big)$$

Hence, if $p > 0.2$, an inequity averse agent A_1 strictly prefers to attack nobody, if $p < 0.2$, an inequity averse agent A_1 strictly prefers to attack agent A_2, and if $p = 0.2$, an inequity averse agent A_1 is indifferent between attacking nobody and agent A_2.

The following weak inequality holds if and only if $k \leq 0.3$, the probability that agent A_1 attacks nobody:

$$k*80+s*0+(1-k-s)*80 \geq k*220+s*(-60)+(1-k-s)*20,$$

where s represents the probability that agent A_1 attacks A_2. Hence, if $k < 0.3$, the selfish agent A_2 strictly prefers to attack nobody, if $k > 0.3$, the selfish agent A_2 strictly prefers to attack agent A_3, and if $k = 0.3$, the selfish agent A_2 is indifferent between attacking nobody and agent A_3.

For $\alpha = 2$ and $\beta = 0.6$ the following weak inequality holds if and only if $k \leq \frac{102+78*s}{436}$:

$$k * \left(80 - \tfrac{\alpha}{2} * 240 - \tfrac{\alpha}{2} * 120\right) + s * \left(-\tfrac{\alpha}{2} * 340 - \tfrac{\alpha}{2} * 200\right) + (1 - k - s) *$$
$$\left(80 - \tfrac{\alpha}{2} * 380 - \tfrac{\beta}{2} * 80\right) \geq k * \left(220 - \tfrac{\alpha}{2} * 100 - \tfrac{\beta}{2} * 220\right) + s *$$
$$\left(-60 - \tfrac{\alpha}{2} * 600 - \tfrac{\alpha}{2} * 60\right) + (1 - k - s) * \left(20 - \tfrac{\alpha}{2} * 440 - \tfrac{\beta}{2} * 20\right)$$

Hence, if $k < \frac{102+78*s}{436}$, the inequity averse agent A_2 strictly prefers to attack nobody, if $k > \frac{102+78*s}{436}$, the inequity averse agent A_2 strictly prefers to attack agent A_3, and if $k = \frac{102+78*s}{436}$, the inequity averse agent A_2 is indifferent between attacking nobody and agent A_3.

Assume $p \in [0, 0.2)$. Then, a selfish agent A_1 prefers to attack agent A_2, and an inequity averse agent A_1 also prefers to attack agent A_2. Hence, $k = 0$ and $s = 1$. Consequently, all types of agent A_2 prefer to attack nobody, which means that $p = 1$. This is a contradiction. Therefore, in any Bayesian Nash equilibrium $p \notin [0, 0.2)$.

Assume $p = 0.2$. Then, a selfish agent A_1 prefers to attack agent A_2, and an inequity averse agent A_1 is indifferent between attacking agent A_2 and nobody. Consider the inequity averse agent A_1 to attack nobody with probability n and to attack agent A_2 with probability $1 - n$. Hence, $k = 0.4 * n$ and $s = 0.6 + 0.4 * (1 - n)$. If $n \in \left[0, \frac{3}{4}\right)$, all types of agent A_2 prefer to attack nobody and $p = 1$. This is a contradiction. If $n = \frac{3}{4}$, the selfish agent A_2 is indifferent between attacking nobody and agent A_3, and the inequity averse agent A_2 prefers to attack nobody. Hence, $p \in [0.4, 1]$. This is a contradiction. If $n \in \left(\frac{3}{4}, \frac{225}{257}\right)$, the selfish agent A_2 prefers to attack agent A_3, and the inequity averse agent A_2 prefers to attack nobody. Hence, $p = 0.4$. This is a contradiction. If $n = \frac{225}{257}$, the selfish agent A_2 prefers to attack agent A_3, and the inequity averse agent A_2 is indifferent between attacking nobody and agent A_3. Consider the inequity averse agent A_2 attacking nobody with probability l and agent A_3 with probability $1 - l$. Hence, $p = 0.4 * l$. If and only if $l = 0.5$, this is not a contradiction. If $n \in \left(\frac{225}{257}, 1\right]$, all types of agent A_2 prefer to attack agent A_3. Hence, $p = 0$. This is a contradiction.

Assume $p \in (0.2, 0.4)$. Then, a selfish agent A_1 prefers to attack agent A_2, and an inequity averse agent A_1 prefers to attack nobody. Hence, $k = 0.4$ and $s = 0.6$. Consequently, all types of agent A_2 prefer to attack agent A_3 and $p = 0$. This is a contradiction. Therefore, in any Bayesian Nash equilibrium $p \notin (0.2, 0.4)$.

Assume $p = 0.4$. Then, a selfish agent A_1 is indifferent between attacking agent A_2 and agent A_3, and an inequity averse agent A_1 prefers to attack nobody. Consider the selfish agent A_1 to attack agent A_2 with probability n and agent A_3 with probability $1 - n$. Hence, $k = 0.4$ and $s = 0.6 * n$. Consequently, all types of agent A_2 prefer to attack agent A_3 and it follows that $p = 0$. This is a contradiction. Therefore, in any Bayesian Nash equilibrium $p \neq 0.4$.

Assume $p \in (0.4, 1]$. Then, a selfish agent A_1 prefers to attack agent A_3, and an inequity averse agent A_1 prefers to attack nobody. Hence, $k = 0.4$ and $s = 0$. Consequently, all types of agent A_2 prefer to attack agent A_3 and it follows that $p = 0$. This is a contradiction. Therefore, in any Bayesian Nash equilibrium $p \notin (0.4, 1]$.

Therefore, in the unique Bayesian Nash equilibrium a selfish agent A_1 attacks agent A_2, an inequity averse agent A_1 attacks nobody with probability $\frac{225}{257}$ and agent A_2 with probability $1 - \frac{225}{257}$, a selfish agent A_2 attacks agent A_3, and an inequity averse agent A_2 attacks nobody with probability 0.5 and agent A_3 with probability 0.5.

Bibliography

[1] Abbink, K., Irlenbusch, B., Renner, E., 2002. An experimental bribery game. The Journal of Law, Economics, and Organization 18, 428-454.

[2] Anderson, L. R., Holt, C. A., 1997. Information cascades in the laboratory. American Economic Review 87, 847-862.

[3] Andreoni, J., Erard, B., Feinstein, J., 1998. Tax compliance. Journal of Economic Literature 36, 818-860.

[4] Becker, G. S., 1968. Crime and punishment: An economic approach. Journal of Political Economy 76, 169-217.

[5] Berg, J., Dickhaut, J., McCabe, K., 1995. Trust, reciprocity, and social history. Games and Economic Behavior 10, 122-142.

[6] Bénabou, R., Tirole, J., 2003. Intrinsic and extrinsic motivation. Review of Economic Studies 70, 489-520.

[7] Blount, S., 1995. When social outcomes aren't fair: The effect of causal attributions on preferences. Organizational Behavior and Human Decision Processes 63, 131-144.

[8] Bohnet, I., Cooter, R. D., 2001. Expressive law: Framing or equilibrium selection? Berkeley Program in Law & Economics, Working Paper Series Paper 31, University of California, Berkeley.

[9] Bolton, G. E., Brandts, J., Ockenfels, A., 1998. Measuring motivations for the reciprocal responses observed in a simple dilemma game. Experimental Economics 1, 207-219.

[10] Bolton, G. E., Ockenfels, A., 2000. ERC: A theory of equity, reciprocity, and competition. American Economic Review 90, 166-193.

[11] Bowles, S., 2007. Social preferences and public economics: Are good laws a substitute for good citizens? Working Paper No. 496, University of Siena.

[12] Bowles, S., Gintis, H., 1992. Power and wealth in a competitive capitalist economy. Philosophy and Public Affairs 21, 324-353.

[13] Bundeskriminalamt (Ed.), 2005. Polizeiliche Kriminalstatistik Bundesrepublik Deutschland, Berichtsjahr 2005, Wiesbaden.

[14] Camerer, C. F., 2003. Behavioral game theory: Experiments in strategic interaction. Princeton University Press, Princeton.

[15] Charness, G., 2004. Attribution and reciprocity in an experimental labor market. Journal of Labor Economics 22, 665-688.

[16] Charness, G., Levine, D. I., 2007. Intention and stochastic outcomes: An experimental study. The Economic Journal 117, 1051-1072.

[17] Cox, J. C., 2004. How to identify trust and reciprocity. Games and Economic Behavior 46, 260-281.

[18] Deci, E. L., 1971. Effects of externally mediated rewards on intrinsic motivation. Journal of Personality and Social Psychology 18, 105-115.

[19] Deci, E. L., Koestner, R., Ryan, R. M., 1999. A meta-analytic review of experiments examining the effects of extrinsic rewards on intrinsic motivation. Psychological Bulletin 125, 627-668.

[20] Dohmen, T., Falk, A., Huffman, D., Sunde, U., Schupp, J., Wagner, G. G., 2005. Individual risk attitudes: New evidence from a large, representative, experimentally-validated survey. Discussion Papers of DIW No. 511, German Institute for Economic Research, Berlin.

[21] Duffy, J., Kim, M., 2005. Anarchy in the laboratory (and the role of the state). Journal of Economic Behavior and Organization 56, 297-329.

[22] Duffy, J., Ochs, J., Vesterlund, L., 2007. Giving little by little: Dynamic voluntary contribution games. Journal of Public Economics 91, 1708–1730.

[23] Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games and Economic Behavior 47, 268-298.

[24] Durham, Y., Hirshleifer, J., Smith, V. L., 1998. Do the rich get richer and the poor poorer? Experimental tests of a model of power. American Economic Review 88, 970-983.

[25] Eide, E., 2000. Economics of criminal behavior. In: Bouckaert, B., De Geest, G. (Eds.). Encyclopedia of Law and Economics, Vol. V. Edward Elgar, Cheltenham, 345-389.

[26] Ellingsen, T., Johannesson, M., 2008. Pride and prejudice: The human side of incentive theory. American Economic Review 98, 990-1008.

[27] Falk, A., Fehr, E., Fischbacher, U., 2003. On the nature of fair behavior. Economic Inquiry 41, 20-26.

[28] Falk, A., Fehr, E., Fischbacher, U., 2008. Testing theories of fairness – Intentions matter. Games and Economic Behavior 62, 287-303.

[29] Falk, A., Fischbacher, U., 2002. "Crime" in the lab – Detecting social interaction. European Economic Review 46, 859-869.

[30] Falk, A., Fischbacher U., 2006. A theory of reciprocity. Games and Economic Behavior 54, 293-315.

[31] Falk, A., Ichino, A., 2006. Clean evidence on peer effects. Journal of Labor Economics 24, 39–58.

[32] Fehr, E., Falk, A., 2002. Psychological foundations of incentives. European Economic Review 46, 687-724.

[33] Fehr, E., Klein, A., Schmidt, K. M., 2007. Fairness and contract design. Econometrica 75, 121-154.

[34] Fehr, E., Kremhelmer, S., Schmidt, K. M., 2008. Fairness and the optimal allocation of ownership rights. The Economic Journal 118, 1262-1284.

[35] Fehr, E., Schmidt, K. M., 1999. A theory of fairness, competition, and cooperation. Quarterly Journal of Economics 114, 817-868.

[36] Fehr, E., Schmidt, K. M., 2006. The economics of fairness, reciprocity and altruism – Experimental evidence and new theories. In: Kolm, S.-C., Ythier, J. M. (Eds.). Handbook of the Economics of Giving, Altruism and Reciprocity, Vol. 1. Elsevier, Amsterdam, 615-691.

[37] Fischbacher, U., 2007. z-Tree: Zurich toolbox for ready-made economic experiments. Experimental Economics 10, 171–178.

[38] Fischbacher, U., Gächter, S., Fehr, E., 2001. Are people conditionally cooperative? Evidence from a public goods experiment. Economics Letters 71, 397–404.

[39] Forsythe, R., Horowitz, J. L., Savin, N. E., Sefton, M., 1994. Fairness in simple bargaining experiments. Games and Economic Behavior 6, 347-369.

[40] Frey, B. S., Jegen, R., 2001. Motivation crowding theory. Journal of Economic Surveys 15, 589-611.

[41] Frey, B. S., Oberholzer-Gee, F., 1997. The costs of price incentives: An empirical analysis of motivation crowding-out. American Economic Review 87, 746-755.

[42] Gächter, S., Kessler, E., Königstein, M., 2006. Performance incentives and the dynamics of voluntary cooperation. Mimeo.

[43] Gächter, S., Nosenzo, D., Renner, E., Sefton, M., 2008. Who makes a good leader? Social preferences and leading-by-example. IZA Discussion Paper No. 3914, Institute for the Study of Labor, Bonn.

[44] Gächter, S., Nosenzo, D., Renner, E., Sefton, M., 2009. Sequential versus simultaneous contributions to public goods: Experimental evidence. CESifo Working Paper Series No. 2602, Munich Society for the Promotion of Economic Research, Munich.

[45] Galbiati, R., Vertova, P., 2005. Law and behaviours in social dilemmas: Testing the effect of obligations on cooperation. CentER Discussion Paper No. 2005-56, University of Tilburg.

[46] Garoupa, N., 1997. The theory of optimal law enforcement. Journal of Economic Surveys 11, 267-295.

[47] Glaeser, E. L., 1999. An overview of crime and punishment. Mimeo.

[48] Gneezy, U., 2003. The w effect of incentives. Mimeo.

[49] Gneezy, U., Rustichini, A., 2000a. A fine is a price. Journal of Legal Studies 29, 1-17.

[50] Gneezy, U., Rustichini, A., 2000b. Pay enough or don't pay at all. Quarterly Journal of Economics 115, 791-810.

[51] Goeree, J. K., Palfrey, T. R., Rogers, B. W., McKelvey, R. D., 2007. Self-Correcting information cascades. Review of Economic Studies 74, 733-762.

[52] Goldfayn, E., 2006. Organization of R&D with two agents and a principal. Bonn Econ Discussion Papers 3/2006, University of Bonn.

[53] Greiner, B., 2004. An online recruitment system for economic experiments. In: Kremer, K., Macho, V. (Eds.). Forschung und wissenschaftliches Rechnen 2003, GWDG Bericht 63. Ges. für Wiss. Datenverarbeitung, Göttingen, 79–93.

[54] Güth, W., Levati, M. V., Sutter, M., van der Heijden, E., 2007. Leading by example with and without exclusion power in voluntary contribution experiments. Journal of Public Economics 91, 1023–1042.

[55] Hirshleifer, J., 1991. The paradox of power. Economics and Politics 3, 177-200.

[56] Holt, C. A., Laury, S. K., 2002. Risk aversion and incentive effects. American Economic Review 92, 1644-1655.

[57] Houser, D., Xiao, E., McCabe, K., Smith, V., 2008. When punishment fails: Research on sanctions, intentions and non-cooperation. Games and Economic Behavior 62, 509-532.

[58] Huck, S., Müller, W., 2000. Perfect versus imperfect observability – An experimental test of Bagwell's result. Games and Economic Behavior 31, 174–190.

[59] Huck, S., Rey-Biel, P., 2006. Endogenous leadership in teams. Journal of Institutional and Theoretical Economics 162, 253–261.

[60] Hung, A. A., Plott, C. R., 2001. Information cascades: Replication and an extension to majority rule and conformity-rewarding institutions. American Economic Review 91, 1508-1520.

[61] Irlenbusch, B., Sliwka, D., 2005. Incentives, decision frames, and motivation crowding out – An experimental investigation. IZA Discussion Paper No. 1758, Institute for the Study of Labor, Bonn.

[62] Kariv, S., 2005. Overconfidence and informational cascades. Mimeo.

[63] Lepper, M. R., Greene, D., Nisbett, R. E., 1973. Undermining children's intrinsic interest with extrinsic reward: A test of the "overjustification" hypothesis. Journal of Personality and Social Psychology 28, 129-137.

[64] Levati, M. V., Sutter, M., van der Heijden, E., 2007. Leading by example in a public goods experiment with heterogeneity and incomplete information. Journal of Conflict Resolution 51, 793–818.

[65] Levitt, S. D., 1997. Using electoral cycles in police hiring to estimate the effect of police on crime. American Economic Review 87, 270-290.

[66] Ludwig, S., 2008. Team production, incentives, and intermediate information. Mimeo.

[67] Masclet, D., Willinger, M., Figuières, C., 2007. The Economics of the telethon: Leadership, reciprocity and moral motivation. LAMETA Discussion Paper No. 2007-08, Laboratoire Montpelliérain D'Economie Théorique et Appliquée, Montpellier.

[68] McCabe, K. A., Rigdon, M. L., Smith, V. L., 2003. Positive reciprocity and intentions in trust games. Journal of Economic Behavior and Organization 52, 267-275.

[69] Mohnen, A., Pokorny, K., Sliwka, D., 2008. Transparency, inequity aversion, and the dynamics of peer pressure in teams: Theory and evidence. Journal of Labor Economics 26, 693–720.

[70] Nöth, M., Weber, M., 2003. Information aggregation with random ordering: Cascades and overconfidence. The Economic Journal 113, 166-189.

[71] Offerman, T., 2002. Hurting hurts more than helping helps. European Economic Review 46, 1423-1437.

[72] Piccione, M., Rubinstein, A., 2004. The curse of wealth and power. Journal of Economic Theory 117, 119-123.

[73] Piccione, M., Rubinstein, A., 2007. Equilibrium in the jungle. The Economic Journal 117, 883-896.

[74] Polinsky, A. M., Shavell, S., 2000a. The economic theory of public enforcement of law. Journal of Economic Literature 38, 45-76.

[75] Polinsky, A. M., Shavell, S., 2000b. The fairness of sanctions: Some implications for optimal enforcement theory. American Law and Economics Review 2, 223-237.

[76] Potters, J., Sefton, M., Vesterlund, L., 2007. Leading-by-example and signaling in voluntary contribution games: An experimental study. Economic Theory 33, 169–182.

[77] Powell, B., Wilson, B. J., 2008. An experimental investigation of hobbesian jungles. Journal of Economic Behavior and Organization 66, 669-686.

[78] Rabin, M., 1993. Incorporating fairness into game theory and economics. American Economic Review 83, 1281-1302.

[79] Rivas, M. F., Sutter, M., 2008. The dos and don'ts of leadership in sequential public goods experiments. Working papers in Economics and Statistics 2008-25, University of Innsbruck.

[80] Sausgruber, R., 2009. A note on peer effects between teams. Experimental Economics 12, 193-201.

[81] Schulze, G. G., Frank, B., 2003. Deterrence versus intrinsic motivation: Experimental evidence on the determinants of corruptibility. Economics of Governance 4, 143-160.

[82] Sliwka, D., 2007. Trust as a signal of a social norm and the hidden costs of incentive schemes. American Economic Review 97, 999-1012.

[83] Stanca, L., Bruni, L., Corazzini, L., forthcoming. Testing theories of reciprocity: Does motivation matter? Journal of Economic Behavior and Organization, forthcoming.

[84] Torgler, B., 2002. Speaking to theorists and searching for facts: Tax morale and tax compliance in experiments. Journal of Economic Surveys 16, 657-683.

[85] Trautmann, S. T., 2009. A tractable model of process fairness under risk. Mimeo.

[86] Tversky, A., Kahneman, D., 1986. Rational choice and the framing of decisions. Journal of Business 59, 251-278.

[87] Tyran, J.-R., Feld, L. P., 2006. Achieving compliance when legal sanctions are non-deterrent. Scandinavian Journal of Economics 108, 135-156.

[88] Winter, E., 2006a. Optimal incentives for sequential production processes. Rand Journal of Economics 37, 376–390.

[89] Winter, E., 2006b. Transparency among peers and incentives. Mimeo.

Eidesstattliche Versicherung

Ich versichere hiermit eidesstattlich, dass ich die vorliegende Arbeit selbständig und ohne fremde Hilfe verfasst habe. Die aus fremden Quellen direkt oder indirekt übernommenen Gedanken sowie mir gegebene Anregungen sind als solche kenntlich gemacht.

Die Arbeit wurde bisher keiner anderen Prüfungsbehörde vorgelegt und auch noch nicht veröffentlicht.

Datum: 3.7.2009

Christina Strassmair

Curriculum Vitae

October 2005 - July 2009	Research and Teaching Assistant Chair of Prof. Dr. Klaus M. Schmidt Ph.D. Program in Economics Munich Graduate School of Economics Ludwig-Maximilians-Universität, Munich
November 2007 - February 2008	Visiting Ph.D. Student in Economics University of Amsterdam
October 2001 - August 2005	Diplom in Economics Leopold-Franzens-Universität, Innsbruck
June 2001	Matura Höhere Lehranstalt für wirtschaftliche Berufe, Rankweil
July 31, 1982	Born in Feldkirch, Austria