Essays on the Evolution, Stability, and Heterogeneity of Social Preferences

Dissertation

Zur Erlangung der Würde des Doktors der
Wirtschafts- und Sozialwissenschaften
des Fachbereichs Volkswirtschaftslehre
der Universität Hamburg

vorgelegt von

Philipp Schliffke
aus Lingen/Ems

Hamburg, 31. Mai 2012

Vorsitzender:	Prof. Dr. Gerd Mühlheußer
Erstgutachterin:	Prof. Dr. Anke Gerber
Zweitgutachter:	Prof. Dr. Andreas Lange
Datum der Disputation:	10. Oktober 2012

Wasch alles weg, lass es ziehen Richtung Meer,
Und die Leute lieben scheitern und ich scheitere so sehr,
Zieh Dir etwas Hübsches an und halte meine Hand,
Heute Abend ist der letzte Abend in diesem Land.

Thees Uhlmann

I thank Manfred J. Holler for employing me at his institute and the freedom. I thank Gerd Mühlheußer for taking chair in my commission and, deeply, Anke Gerber and Andreas Lange for supervising my thesis. Many colleagues have contributed to this work both in academic and personal terms. I thank, among others and in alphabetical order, Menusch Kahdjavi, Nicola Maaser, Jakob Neitzel, Andreas Nicklisch, Andreas Nohn, Johannes Schwarze, Jens Tiedemann, and Wenke Wegner. Special thanks are attributed to Martin Leroch - it was not the same without you. I am indebted to my wife, kids and parents.

Contents

Part I

Introduction

Chapter 1

Evolution, Stability, and Heterogeneity of Social Preferences

Microeconomic theory can explain human cooperation via two main channels. Either it is driven by strategic concerns and the goal to obtain future benefits, or it is directly based on preferences and independent of potential future benefits. The decisive element to discriminate between both causes is the power to predict behavior in one-shot interactions with anonymous partners. As an example, consider the game of trust played sequentially by two players. The first mover has the choice of either trusting or not trusting the other player. If trust is not shown, the game ends immediately and both players receive some default payoff greater zero. If trust is shown, the second mover must decide whether to reward or exploit. If the second mover rewards, both players receive an equal payoff which is larger compared to the default payoff. If the second mover exploits, however, the first mover receives zero, i.e. less than the default payoff, but the second mover receives more compared to the payoff under reward. The setup sets clear incentives favoring exploitation but both players can gain if trust is shown and rewarded, i.e. if both players cooperate. Whether or not trust will be shown depends critically on the expectations of the first mover. If he expects exploitation, it is either an act of irrationality to nevertheless trust, or an act of altruism, i.e. the pure desire to make the other player better off. If, on the other hand, he can expect to be rewarded, it is in the material interest to trust. However, unless the first mover has absolutely perfect foresight with respect to the second movers' reaction, the final payoff from the interaction remains uncertain and the sender needs to *trust* that the second mover will indeed chose reward.

The crucial question is what the second mover should do and the answer

depends very critically on the number of interactions. If both players interact only once, do not know, or see each other, and never meet again, there is no obvious reason to return something. Not to return anything is the money maximizing choice in such a one-shot setting of the game. Obviously, a rational first mover will anticipate that reaction and accordingly, will not trust. Formally, $(not\ trust, exploit)$ is the unique subgame perfect Nash equilibrium of the game. Cooperation will not realize and while both players would be better off given that trust is shown and repaid, individual rationality and money maximization prohibit this mutually beneficial outcome. The situation is fundamentally different if the game is repeated and as long as the probability of continuation is not zero. Given that trust is shown, the payoff for the second mover is strictly higher compared to a situation where trust is not shown. Thus, if the second mover can induce trust in subsequent rounds by returning, he can generate a payoff stream which is superior to one where trust is not shown. This leads to a simple trade-off. Either the second mover increases his short run benefits by exploitation, which is expected to crowd out trust and thus lowers future payoffs, or he forgoes short run benefits in order to obtain a greater sum of payoffs over time. The actual decision then depends on how much the second mover discounts future payoffs. Given that future payoffs are sufficiently important, it is rational not to exploit and thus, under repeated interaction, cooperation can strive.[1]

Cooperation, in this case, is not at all based on, for example, empathy. Rather, it is the result of strategic considerations and money maximizing behavior over time. One consequence is that cooperation will break down once the strategic incentives are removed. Consider, for example, a company that expands its supplier base, subsequently interacts less frequently with its original suppliers, and can even do without a specific supplier. Such a setting results in a lowered probability of continuation which has the same effect as a higher discount of future benefits. Accordingly, the incentives, for example to pay bills on time, are lowered. In the most extreme case, the probability of continuation drops to zero, for example if the company already decided to rely on a different supplier in the future. Then the interaction turns into a one-shot setting and strategic considerations can never explain trustworthy behavior on that last interaction.[2]

[1] A more complete and formal presentation of the argument is provided classically by Axelrod (1984) for the related prisoners' dilemma game.

[2] The argument here relies on a probability of continuation which drops to zero based on exogenous factors which are unknown to the subjects on an a priori basis. If player know in advance that there is a final period, then arguments of backward induction and subgame perfection would prohibit the rise of cooperation in the first place. That is a different topic, however.

Thus, strategic considerations can be causal for cooperation but not in one-shot settings. Experimental tests, however, frequently reveal that trust is shown and repaid exactly under such one-shot conditions with anonymous partners, see e.g. Berg et al. (1995); Bolle (1998); Burks et al. (2003); McCabe et al. (2003); Ortmann et al. (2000). By far, the experimental evidence is not restricted to trust games. It extends to related games as the prisoners' dilemma, or gift-exchange game. It extends to the provision of public goods, or the use of common pool resources. It extents to rejecting low offers in bargaining situations or simply sharing money with strangers, and it extends to the punishment of players who behaved in a certain way either interacting with oneself or other individuals. In all those cases, money maximization in a one-shot setting calls for no returns, no effort, no contributions, no rejections, no sharing, or no punishment. Nevertheless, such behavior is present. The answer by microeconomic theory to explain such observations is a more pronounced separation of utility and payoffs. A *game* is generally described by the set of players, the available strategies and the utility that arises for each player under each combination of strategies. If utility is assumed to coincide with pecuniary payoffs, the dilemma of trust, for example, cannot be resolved rigorously for one-shot interactions. However, utility need not coincide with pecuniary payoffs. Utility as such is a pure representation of the underlying preferences by each individual but preferences need not be restricted to material gains either. They can entail empathy components and considerations of others' well-being in multiple forms. Given that such components exist, however, the payments that arise from a specific form of interaction are not equivalent to utility. Rather, utility arises by the transformation of payments given the intrinsic system of reference points, values, or norms by each individual, and dependent on the weighting of each factor within each individual. Games that are clearly defined in payoffs can then take multiple forms in terms of individual utility and what constitutes a dilemma situation in payoffs need not be a dilemma in terms of utility.

Preferences that take account of interaction partners are called either *other-regarding*, or *social preferences*. The most well established forms are preferences that imply an aversion against inequity between individuals, see e.g. Fehr and Schmidt (1999) and Bolton and Ockenfels (2000), or preferences that imply reciprocal behavior, see e.g. Charness and Rabin (2002); Cox et al. (2008); Dufwenberg and Kirchsteiger (2004); Falk and Fischbacher (2006); Rabin (1993). Social preferences are distinct from pure altruism. As pointed out, altruism is the desire to make the other player better off. In an extreme version, altruism implies, for example, that the first mover trusts even if it is known that the second mover will exploit. Social preferences differ from that in the sense that they take both the own and the others' expected payoff into

account and weight them against each other in the one or the other way. This gives rise, for example, to conditional cooperation where players are, again for example, willing to contribute to a public good but if and only if they expect others' to contribute as well, a clear distinction to altruism. It further explains why the existence of social preferences does not stand in contrast to competitive behavior on markets. For example, Fehr and Schmidt (1999) point out that if a market outcome will be characterized by inequity anyway, because, for example, only one seller will eventually sell the product, each attempt to reduce this inequity must fail and accordingly, considerations of inequality will not affect behavior (p. 834f.).

With respect to cooperation in general, the crucial aspect is that social preferences can explain cooperation in one-shot settings. Again referring to the trust game, a sufficient aversion against inequality can explain why the second mover will not exploit (since this would yield the most unequal outcome). Alternatively, preferences for reciprocity can imply that reward is chosen. Since the second mover faces a strictly higher payoff given that trust is shown, showing trust is kind and with a sufficiently reciprocal inclination, the second mover will react in kind by not choosing to exploit. If one leaves the domain of one-shot interactions, the predictions by models of social preferences and by models taking into account strategic considerations under repeated interaction can obviously coincide. Differences in predictions may then arise due to exact setup of a game or the assumption of one specific utility function, but both approaches yield convincing arguments in favor of cooperative behavior.

Given the above background, this thesis is located in the domain of social preferences. It takes a journey from the theoretical analysis of the evolution of preferences for reciprocity within the trust and the related gift-exchange game towards experimental works on the consistency and potential specificity of social preferences. The motivation to study the evolution of preferences is based on the idea that evolutionary success can add a more ultimate cause to assume the existence of social preferences in general, as well as to assume the existence of specific preferences. Evolutionary success in this regard needs to be based on payoff superiority of specific behavior patterns compared to others. If it would not be possible to establish evolutionary stability of cooperative preferences in general, it would be quite surprising to observe them. In addition, an evolutionary analysis can help to sort out carefully the conditions that need to be satisfied in order for cooperative behavior to evolve. Probably the most fundamental finding in this regard is that players must have the ability to identify individual types and accordingly condition their behavior on the type of the other player, see e.g. Güth and Kliemt (1994, 1998) and Herold and Kuzmics (2009). To pick up the trust game example,

first movers with perfect type information will trust those players of whom they know that they will reward, but they will not trust others. This excludes exploitive individuals from the interaction but since inclusion is superior to exclusion in material terms, it also yields an explanation why trust and trustworthiness can be evolutionarily stable. Beyond the establishment of such general conditions for the evolutionary success of cooperative preferences, an evolutionary analysis can also help to understand specific preferences much better.

The first work of this thesis, *Would you Trust Yourself? On the Long Run Stability of Reciprocal Trust*, is an attempt into the direction of understanding specific forms of social preferences better. It studies the evolution of Falk and Fischbacher (2006) preferences for reciprocity in the game of trust. As pointed out, reciprocity can explain trustworthiness in the dilemma of trust. The specific feature of the Falk and Fischbacher (2006) model is that it combines an outcome based approach with the intentions behind actions. Similar to the purely outcome based model by Fehr and Schmidt (1999), the utility function contains material reward and the payoff difference between both players (which is used to measure the kindness of the other player in this case). It additionally contains a term measuring the impact of a players' action on the other players' outcome relative to expected payoffs, i.e. a term measuring the reciprocal response. More crucially, however, the model is based on psychological game theory, see Geanakoplos et al. (1989), and expected payoffs are calculated based on a second-order belief structure. The second mover, for example, will take into account whether or not, and to which degree the first mover expects him to be trustworthy. Ceteris paribus, the first mover is less kind the more he is expected to expect the second mover to be trustworthy. The reason is that if trustworthiness can be anticipated, showing trust is in the material interest of the first mover and the intention behind the action is not necessarily restricted to just being kind. The assumption is then that a model that takes account of such more psychological motives is more realistic compared to models which leave such motives aside.

With respect to evolution, a medium reciprocal inclination turns out to be evolutionarily stable. Second movers will reward with a probability that is just sufficient in order for trust to be shown with probability one. However, this does not imply that the reward probability is one as well. This in turn implies that the evolutionary prediction is neither characterized by efficiency, nor by equity. Second movers obtain a larger share of the pie and, in an additional analysis, the advantage is shown to increase if one assumes that first and second movers form two separate populations. I argue that second movers evolve to be *constrained dictators*. They need to ensure that trust is

shown with probability one but beyond that, no generosity as such evolves. Given that the equilibrium is symmetric and unique, each player evolves to one just sufficiently reciprocal to ensure a trust probability of one. Given that players interact in both roles, it also implies that each player will just trust each other player of exactly that type. In a way, each player ends up being just sufficiently reciprocal such that he would just trust himself, which inspired the title.

The second work, *The Co-Evolution of Reciprocity-Based Wage Offers and Effort Choices*, expands the analysis.[3] It studies the evolution of Falk and Fischbacher (2006) preferences in the gift-exchange game. In principle, the gift-exchange game has the same structure as the trust game but here, it is not the first movers' choice which is efficiency enhancing. The main difference is that the gift-exchange game is typically framed in terms of wage contracts. The first mover can pay a wage to the second mover who then decides upon how much effort to invest. The money maximizing prediction is that the worker shows the lowest possible effort, irrespective of the wage, and accordingly, the employer should pay the lowest possible wage. It is a dilemma game again since individual rationality prohibits a mutually beneficial outcome which could be reached given that a high wage is paid and rewarded by efficiency enhancing high effort. With respect to the evolutionary analysis, the difference is that employers and workers are not assumed to frequently switch positions. While the analysis of separate first and second mover population is a subchapter of the evolution of trust paper, it is the only approach taken in the evolution of wages paper.

The evolutionary prediction is that a medium reciprocal inclination is stable on the side of the workers but reciprocity must vanish on the side of the employers. Among workers, those who require higher wages to show full effort obtain higher payoffs and the respective preferences thus receive evolutionary support. Among employers, it is always beneficial if the wage ensures maximal effort. However, employers with a positive reciprocal inclination ceteris paribus pay smaller wages. They dislike if the workers obtain a higher share of the total pie for themselves. But since inducing the maximal effort is superior in evolutionary terms, a positive reciprocal inclination is not stable. The prediction is similar to the one obtained in the work on the evolution if trust. Especially, the evolutionary equilibrium is characterized by strong inequity in favor of the second movers, i.e. the workers in this case. It is different because the equilibrium is efficient, but the main difference is the different interpretation. The reciprocity based solution to the gift-exchange game, paying high wages to induce higher than the minimal effort, can be

[3]This work is forthcoming in *Economics Letters*.

regarded as a form of efficiency wages in the spirit of Akerlof (1982). The strong inequity in favor of the workers suggests, however, that employers might prefer to search for other solutions to secure effort, such as classical contract theory solutions including fines. While such solutions are typically inefficient, they might nevertheless ensure a higher share of the sum of payoffs for the employers. Overall, the evolutionary analysis thus suggests that while efficiency wage may well work in the short run, they may be unstable over time.

The evolutionary prediction obtained for the trust game, which is qualitatively confirmed by the prediction for the gift-exchange game, inspired the second part of this thesis. The second part applies experimental methods to study the consistency and specificity of social preferences. A crucial aspect of the evolutionary predictions obtained for the trust game is that it stands in sharp contrast to predictions obtained for the ultimatum and dictator game reached under the application of the same utility concept and the same evolutionary assumptions. For the ultimatum game, an infinitely reciprocal inclination and equal splits are predicted while for the dictator game, behavior not different from money maximizations is predicted. Thus, the evolutionary predictions for those three games are clearly distinct. The divergent prediction for the dictator game can be resolved since *acceptance* of the second mover in the dictator game is clearly unintentional (the second mover has no choice) and since the Falk and Fischbacher (2006) model contains an additional parameter which can discount for example giving if the other player does not have reasonable alternative actions, or none. However, even this additional parameter cannot explain differences across the trust and ultimatum game. If one assumes that the evolutionary predictions carry over to short run behavior by individuals as well, because the material forces behind the predictions are constantly at work, the divergent predictions imply the question of how consistent behavior can be. Note that the default assumption of the Falk and Fischbacher (2006) model is one parameter for reciprocity which would then predict behavior in both the trust and ultimatum game. Similar assumptions are made in other models of social preferences as well. Typically one or two parameters scale the impact of whatever norm is modeled but the parameters are not assumed to systematically vary across games.[4] Accordingly, choices across games should not vary systematically either.

The third work, *Inconsistent People? An Experiment on the Impact of Social Preferences Across Games*, seeks to find whether or not behavior is

[4] An exception is the work by Dufwenberg and Kirchsteiger (2004) who at least mention the possibility of game specific parameters in a footnote.

indeed consistent. It follows a very general idea of consistency, namely that choices across many different games should reveal a similar deviation from money maximization, respectively, a similar impact of social preferences. Players are assumed to have one stable character trait, but occasional deviations from that trait are accepted. The approach differs from one-to-one comparisons of choice behavior in one situation with choice behavior in other situations which is otherwise the typical standard to test for consistency, see especially Blanco et al. (2011). Subjects in the experiment play a total of six games and face each position in each game once. Choices are then categorized as being made either under a low, medium, or high impact of other-regarding motives. Then I count how often a particular impact occurs within an individual and define behavior as consistent whenever a clear majority of choices fall into one category. The results do not support the consistency of behavior in general. The rate of consistent profiles is below 50% and it remains below 50% if the analysis is restricted to subclasses of games where games are more similar. Correlations across games are positive and highly significant in general, but they remain of medium strength at most. Survival analyses additionally reveal that the likelihood that a player shows a different impact of social preferences in one other randomly selected choice is above 50%, and that the inconsistency expands into the domain of conditional cooperation and unconditional defection.

As an explanation for the low rates of observed consistency, I offer a combination of, on the one hand, multiple behavioral forces that drive behavior (which is also the explanation by Blanco et al. (2011) for their similar findings) combined with stochastic elements of behavior, but, on the other hand, also weaknesses of the experimental method with respect to measuring the consistency of preferences. That behavior is driven by multiple forces, for example concerns for equity, reciprocity, efficiency, etc., can clearly cause behavior to be different comparing two choices. However, given that the motives may also overlap, it does not follow on an a priori ground that behavior as categorized in the approach here, and simultaneously measured across multiple choices, is inconsistent. Ex post, it is though. Experimental effects such as being unfamiliar with the decision situations, scrutiny and demand effects, or simply framing, can add to the observed inconsistency. I conclude by arguing that experiments with respect to consistency should be based on repeated observations because this can eliminate some of those effects, as least partially.

The fourth work, *Game Specific Social Preferences: Different Types and a Canceling-Out Effect*, applies such a repeated observations approach. At first, however, it takes a step back. While it is true that the question of consistency follows from the divergent evolutionary predictions, it follows only

indirectly. Inconsistent behavior can follow if either the analysis is restricted to the ultimatum and trust game, or if it is assumed that the evolutionary analysis of additional games will yield yet different predictions. One may further argue that the evolutionary prediction does not necessarily predict inconsistency, if inconsistency is understood as arising from more or less random choice behavior. Rather, it predicts game specific social preferences. Yet further, one may argue that specific preferences for the trust and ultimatum game are not necessarily surprising given that the two games represent fundamentally different decision problems. However, the reason to study the general consistency of behavior first was that if it would have been possible to establish consistency, the question of specificity could be argued to be one of minor importance. Since it was not possible to establish consistency, however, the last work of this thesis tries to establish specificity as a potential explanation for the observed inconsistencies.

The experimental setup is inspired by the assumptions of the indirect evolutionary approach which was applied to gain the evolutionary predictions. Subjects play a dictator, ultimatum and trust game in both positions each. Each game appears in random order and is played multiple times. Crucially, players receive information regarding the average play of their current matching partner in previous rounds. This allows them to discriminate between different types and prohibits that cooperation breaks downs as often observed, for example, in public good games. Players cannot, however, identify each other which rules out the application of repeated game strategies. The results are threefold. At first, average play in the dictator game is clearly distinct from average play in the trust and ultimatum game both at the aggregate and at the individual level. This is well in line with the theoretical predictions given that the Falk and Fischbacher (2006) model can capture situations where actions are unintentional (see above). Secondly, aggregate play in the trust and ultimatum game is not significantly different both in terms of choices, but also in terms of parameters estimated for the Falk and Fischbacher (2006) model. Thirdly, however, the aggregated non-specificity is not due to consistent behavior at the individual level but due to a canceling out effect. On the individual level, two clearly distinct types can be identified. While both types show similar behavior in the ultimatum game, one type is characterized by trust game returns just at the threshold where showing trust becomes an investment with negative return, but the other type is characterized by returns close to the equity implying action. In addition, while individual trust and ultimatum game behavior operates on different levels, the correlation between both choices is greater than .5 and highly significant. Within both types, the correlations are even greater than .7. Together with additional results showing that the impact of signaling,

and thus strategic behavior which is certainly possible given the repeated interaction setup with information transmission, is probably low, I take the large correlations as evidence in favor of a preference based explanation. At least to some degree, the observed inconsistencies can thus be traced back to game specific preferences.

The work on game specific preferences ends the thesis. Given than almost all models of social preferences are formalized in a way which suggest that individuals can be characterized by one or two preference parameters, the question regarding the consistency of behavior can be asked without the specific background of the divergent evolutionary predictions. Nevertheless, the evolutionary predictions suggest that the use of context free parameters might be misleading, at least if the goal is to explain individual behavior. My own findings with respect to consistency highlight the dimension of the problem and add to the literature by obtaining the results via a different approach which does not rely on one-to-one comparisons, but is based on the very general idea that social preferences explain deviations from money maximization. Finally, the work on game specific social preferences provides an explanation for low rates of consistency. Crucially, the explanation differs from previous explanations. The assignment of a multiplicity of social norms which trigger inconsistent behavior goes back to Blanco et al. (2011). The authors argue however, that the different norms may be uncorrelated within each individual. The specificity explanation is different because while choice behavior operates on very different levels, it is nevertheless highly correlated.

Part II

Theory: Evolution of Reciprocity

Chapter 2

Would You Trust Yourself? - On the Long Run Stability of Reciprocal Trust

2.1 Introduction

It is widely acknowledged that trust is one of the fundamentals of successful economic activity, see e.g. Arrow (1974) or Ostrom (2010). From a strategic perspective, however, trust remains to be a dilemma game. The efficient solution requires that trust is shown and rewarded but if the second mover can individually gain by taking advantage of shown trust, the first mover should never trust. In ignorance of this subgame perfect, money maximizing prediction, people in experiments frequently reveal to be trustful as well as being trustworthy, see e.g. Berg et al. (1995), Bolle (1998), McCabe et al. (2003), or Schotter and Sopher (2006).

Deviations from the money maximizing predictions are not restricted to trust games or two player games but are a rather frequent result in experimental economics. Today, the main approaches to explain such observations are *social* or *other-regarding preferences*. Well known are outcome based models of inequity aversion as for instance introduced by Fehr and Schmidt (1999) or Bolton and Ockenfels (2000). A different strand of literature arises from psychological games introduced by Geanakoplos et al. (1989). In psychological games, not only outcomes but also beliefs held by players about others' actions and possible intentions behind actions affect utility. Rabin (1993) explicitly considers reciprocity, i.e. the desire to reward kind actions and punish unkind ones. Falk and Fischbacher (2006) combine an equitable reference standard with the desire to reciprocate and the impact of intentions.

Their model does quite well in explaining stylized facts about behavior in experimental settings for example with respect to dictator, ultimatum, and gift-exchange games. By capturing the impact of intentions, their model can also explain differences in experimental findings regarding reduced ultimatum and best-shot games that cannot be explained with pure outcome based models.[1]

Preferences of the Falk-Fischbacher-type can also explain trust and trustworthiness. Second movers regard shown trust as kind and with a sufficiently high reciprocal inclination, they reward the kindness by being trustworthy. Unlike in a pure outcome orientated model which would directly contrast the own material gain for example with the payoff difference between players, the background that yields the prediction is more complex, however. Kindness, for example, is evaluated relative to equitable expected payoffs. But the calculation of expected payoffs is not only based on the first-order belief regarding the other players' action, but also on the second-order belief regarding the belief by the other player about the own action. Via the second-order belief structure, the believed intentions behind the other player' action enter utility directly. Applied to the game of trust and from the perspective of the second mover, it makes a great difference whether the first mover expects to be rewarded or not. If the returnee believes that the first mover expects to be rewarded, then the attributed kindness is lower compared to a situation in which the returnee believes that the first mover does not expect to be rewarded with high probability. In the first case, and with correct beliefs, the decision to trust is not really risky and potentially beneficial in material terms for the first mover. Hence, it is not particularly kind. In the latter case, however, the first mover deliberately puts his own payoff at stake by showing trust. This is indeed kind.

The fact that the model is able to capture such differences in underlying psychological motives has a clear disadvantage in terms of model complexity. However, it also has a clear advantage in terms of plausibility. This is not to say that players should be expected to evaluate the situation exactly in the way as they do under the assumption of Falk and Fischbacher (2006) preferences. But that intentions indeed make a difference is very plausible and the model provides one particular way to formalize the idea. The specific predictions can then be judged with regard to their plausibility ex post. For the present case, it indeed appears plausible that the kindness of trust depends on the expected risk taken by the trustee. Another example for an intuitive prediction by the model is that first movers with a high reciprocal

[1]For general overviews on social preferences, see for example Camerer (2003) or Fehr and Schmidt (2006).

inclination ceteris paribus trust less. In addition to the material loss in case of exploitation, reciprocal players suffer from being exploited as such.

Besides the success of the Falk-Fischbacher model, or models of social preferences in general, in explaining experimental observations, it is a different question why other regarding preferences develop in the first place and whether one can expect them to be stable in the long run. One method that may deepen our understanding with respect to both issues is an evolutionary analysis. If it is possible to establish the evolutionary stability or reciprocal preferences, then the theory does not just explain observed behavior, but one additionally gains a more ultimate reason to assume their existence. Accordingly, such an evolutionary analysis is carried out in this paper. More specifically, an indirect evolutionary approach in the spirit of Güth and Kliemt (1994) or Güth and Napel (2006) is adopted. There, and in contrast to a direct evolutionary approach, the material gains associated with end nodes of a specific game are evaluated in a dual fashion. On the one hand, they enter a subjective evaluation by players who may or may not take other than pecuniary motives into account, for example considerations of reciprocity. On the other hand, the actions that arise based on the subjective evaluation have material consequences which enter an evolutionary evaluation. Specific preferences then spread or decline dependent on their relative success. The approach captures both, forward looking behavior driven by expectations and subjective payoffs, as well as evolutionary path dependence.

In a similar work, Berninghaus et al. (2007) study Falk-Fischbacher preferences in ultimatum and dictator games. They find that reciprocity cannot induce any behavior that is different from money maximizing in the dictator game but that the inclination to reciprocate approaches infinity in the ultimatum game. Contrary to those findings, it is shown that a medium level of reciprocal inclination is stable in the game of trust. In a way, second movers develop to *constrained dictators*. As in the dictator game, they have complete discretion over allocating a certain amount of money but unlike in the dictator game, the pie to be divided is not exogenous to the individuals' choices. Thus, second movers need to behave in such a manner that first movers will show trust and the pie evolves and indeed, second movers should reward with a probability which guarantees that first movers trust them for sure. Nevertheless, any level of trustworthiness above the critical level is suboptimal and no generosity as such develops. Since each player is assumed to play in first and second mover position, the result also implies that every individual is just sufficiently reciprocal to trust a player of the exact same type, i.e. he would just trust himself.

The paper is organized as follows. Section 2.2 presents and discusses

the reciprocity equilibrium for a game of trust. In Section 2.3, a preference game is constructed and by showing that the preference game has a unique, symmetric, and strict Nash equilibrium, asymptotic stability of reciprocal preferences is established. Section 2.4 introduces different parameters concerning positive and negative reciprocity. The reciprocity equilibrium as such is unaffected by the change. However, while a medium level of positive reciprocity remains stable, negative reciprocity must vanish in the long run. Section 2.5 concludes.

2.2 Reciprocity Equilibrium in the Game of Trust

The analysis is based on a standard *game of trust*, as illustrated in Figure 2.1 below. There are two players 1 and 2. Player 1 moves first and decides whether to *trust* (T) or *not to trust* (N). Whenever the choice is not to trust, the game ends immediately and both players receive a material payoff $s > 0$. Whenever the choice is to trust, then player 2 has the opportunity to either *reward* (R) or to *exploit* (E). In case player 2 rewards, both players receive material payoff $r > s$. The second mover receives $1 > r$ when exploiting, in which case the first mover is left with 0. It is assumed that mutual cooperation is the efficient outcome, i.e. $2r > 1$.

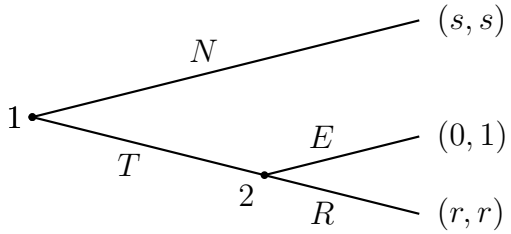


Figure 2.1: The game of trust

Obviously, if material interest is the only factor that determines players decisions and if the game is one-shot, then the unique subgame perfect Nash equilibrium of the above game is (N, E), i.e. the second mover exploits and the first mover, anticipating the exploitation, does not trust. The game is a dilemma game since individual rationality leads to equilibrium payoffs (s, s) which are lower as in the case of mutual cooperation, (r, r).

Now it is assumed that utility is not solely based on material interest. Rather, players care about *being nice* as well as about *being treated nicely*. More precisely, it is assumed that players have reciprocal preferences in the

sense of Falk and Fischbacher (2006). Applied to the above two player game with each player having exactly one node to decide in, utility for player i is defined as

$$u_i(f, s_i'', s_i') \equiv \pi_i(f) + \rho_i \; \varphi_j(n, s_i'', s_i') \; \sigma_i(n, f, s_i'', s_i')$$

Basically, utility is the sum of material reward in terminal node f, i.e. $\pi_i(f)$, and a component referring to reciprocity, $\rho_i \varphi_j(\cdot)\sigma_i(\cdot)$ called *reciprocity utility*. The *reciprocity parameter* $\rho_i \in \mathbb{R}_+$ scales the impact of reciprocity utility on overall utility. If $\rho_i = 0$, then the model is identical to the standard money maximizing approach. If $\rho_i > 0$, then the kindness by player j, measured via the *kindness term* $\varphi_j(\cdot)$, and the possibility for i to reciprocate, measured via the *reciprocation term* $\sigma_i(\cdot)$, affects overall utility as well.

Whether player j's behavior in node n is perceived as kind or unkind by i depends on the expected payoffs for i and j. Expected payoffs are calculated with the first-order belief by player i upon the strategy applied by j, which is s_i', and the second-order belief by i upon the belief held by j upon his strategy, which is s_i''. As a reference, equity between i and j is used such that i perceives j as kind whenever his expected payoff for given beliefs is larger than j's and as unkind if i expects j to get more. Formally, $\varphi_j(n, s_i'', s_i') \equiv \pi_i(n, s_i'', s_i') - \pi_j(n, s_i'', s_i')$.[2] Note that this approach is based on a direct comparison of payoffs between i and j and not on a comparison of one's own payoff relative to some *fair* payoff a player expects for himself like in e.g. Rabin (1993).

The third component of reciprocity utility, besides ρ_i and $\varphi_j(\cdot)$, is the reciprocation term $\sigma_i(\cdot)$. It captures the reaction by i toward the kindness by j by measuring how much i's choice affects the expected payoff by j. Formally, $\sigma_i(n, f, s_i'', s_i') \equiv \pi_j(\nu(n, f), s_i'', s_i') - \pi_j(n, s_i'', s_i')$ with $\pi_j(n, s_i'', s_i')$ the ex ante expected payoff for j in n, from the perspective of i, and $\pi_j(\nu(n, f), s_i'', s_i')$ the ex post expected payoff for j from the perspective of i if i chooses s_i'' ($\nu(n, f)$ indicates a decision in n on the path towards f). Note that if j is unkind such that $\varphi_j(\cdot) < 0$ and i can reciprocate by lowering j's payoff such that $\sigma_i(\cdot) < 0$ as well, then reciprocity utility derived from the punishing action is positive. Further, the higher ρ_i the more likely it is in this case that i chooses the punishing action even if it is costly to him. On the other

[2] In fact, the original kindness term also contains a term measuring the role of intentions. Kindness might be discounted if it appears to be unintentional. However, in the game of trust it turns out that all actions are fully intentional and the respective term is equal to one. In order to avoid the great complexity that comes along with intentions and since they do not have an impact here, their inclusion is set aside. Nevertheless, the proof of Proposition 1 contains the needed formal arguments. For the possible impact of intention in trust games, see McCabe et al. (2003).

hand, if j is kind such that $\varphi_j(\cdot) > 0$ and i can increase j's payoff such that $\sigma_i(\cdot) > 0$, then a positively reciprocal reaction becomes increasingly likely as ρ_i increases.

Finally, it is assumed that information is perfect and that beliefs are correct in equilibrium. Let p be the probability with which player 1 chooses to trust and let q be the probability with which player 2 rewards trust.

Proposition 1 *In the game of trust as defined above and given that players are endowed with Falk-Fischbacher preferences with $\rho_1 > 0, \rho_2 \geq 0$, the reciprocity equilibrium is given by*

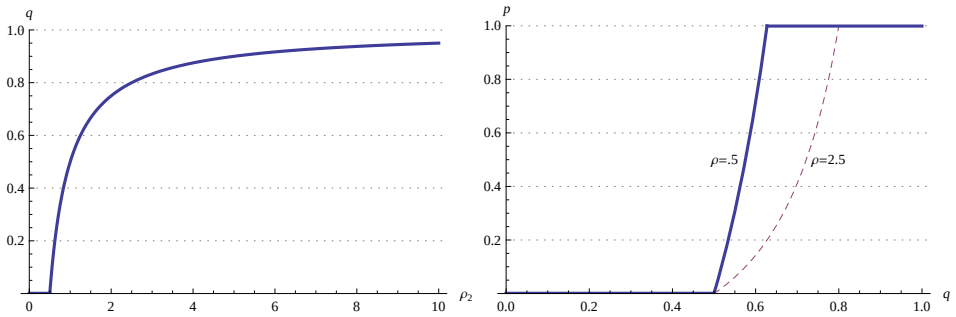
$$q^* = \begin{cases} 0 & if \quad \rho_2 \leq \frac{1-r}{r} \\ 1 - \frac{1-r}{\rho_2 r} & if \quad \rho_2 > \frac{1-r}{r} \end{cases} \tag{2.1}$$

$$p^* = \begin{cases} 0 & if \quad q \leq \frac{s}{r} \\ \min\left\{1, \frac{qr-s}{\rho_1(1-q)[1-s-q(1-r)]}\right\} & if \quad q > \frac{s}{r} \end{cases} \tag{2.2}$$

If $\rho_1 = 0$, p^* turns into a step function from 0 to 1 at $q = \frac{s}{r}$.

Proof 1 *See the Appendix.*

In order to illustrate equilibrium behavior, Figure 2.2 plots optimal behavior for players 2 (depending on ρ_2) and 1 (depending on q and ρ_1). Payoff parameters are set to $r = \frac{2}{3}$ and $s = \frac{1}{3}$. The thick and dashed lines illustrate behaviors for reciprocity parameters $\rho_1 = .5$ and $\rho_1 = 2.5$ respectively.



(a) Reward probability q^*. (b) Trust probability p^*.

Figure 2.2: First and second mover equilibrium behavior

Player 2 will always regard shown trust as fully intentional since player 1 could be less kind by choosing not to trust. Furthermore, for any $p > 0$, player 2's expected payoff is strictly higher compared to the expected payoff by 1 and thus the decision to trust is perceived as kind. Reciprocity utility is

positive if 2 reacts to the kindness by putting positive weight on rewarding this behavior. If reciprocal inclination is sufficiently high, player 2 will reward with strictly positive probability. However, player 2 will never reward for sure. The more player 2 believes that player 1 anticipates a high probability to reward, i.e. the higher player 2's second-order belief, the less 2 regards 1's trust as particularly kind. In addition, the possibility to surprise the first mover by a higher than expected return is decreasing in player 2's second-order belief as well. Hence, the probability to reward strictly increases in reciprocal inclination but at a diminishing rate.

As pointed out, the expected payoff for player 2 is greater than the expected payoff for player 1 given the choice of trust and as long as $q \neq 1$. This implies that 2 regards 1 as kind but it also implies that 1 regards 2 as unkind. Since player 2 has the option to reward for sure, and since that would not put 2 in a situation where he has less than 1 (both receive r), player 1 regards the unkindness as fully intentional. Besides full intentionality, a positive probability to trust would imply that the unkindness by 2 is rewarded such that reciprocal utility is subtracted from material reward in case player 1 chooses to trust (a positive reciprocation term multiplied with a negative kindness term). On the other hand, if player 1 chooses not to trust, the expected payoff for player 2 is smaller compared to the case in which the decision is to trust. Player 1's behavior may be interpreted as punishment in advance such that reciprocal utility is actually added in case of not showing trust (a negative reciprocation term multiplied with a negative kindness term). As a first consequence, any reciprocal inclination can never induce player 1 to trust as long as doing so is not favorable in material terms. In the absence of reciprocal preferences, trusting behavior is favorable whenever $q > \frac{s}{r}$.

The special payoff structure induces that player 1 regards 2's behavior as unkind *per se*. Indeed, the more player 1 believes that 2 expects him to trust (high second-order belief), the more unkind is player 2 from the perspective of 1. This effect is, however, dampened by two other effects. First, the higher the probability to reward (high first-order belief), the higher is the expected material payoff for player 1. Second, the higher the probability to reward, the less unkind is player 2's behavior from the perspective of 1. Beyond the threshold $q = \frac{s}{r}$, both effects start to outweigh the impact of reciprocal utility until player 1 finally shows trust for sure. Note, however, that the higher the reciprocal inclination, the more disutility 1 obtains from exploitation after he has shown trust such that players with a higher reciprocal inclination, ceteris paribus, trust less. Compared to players who do not care, or care little about reciprocity, those who do care face not only the material loss of being exploited but also the pain of being cheated.

2.3 Evolution of Reciprocity

As already pointed out, the evolution of reciprocal preferences will be studied in an indirect evolutionary approach similar to the one used in e.g. Güth and Kliemt (1994), Güth and Napel (2006), or Berninghaus et al. (2007). The idea behind this approach is that while players behave based on a subjective evaluation of payoffs in each stage game, every stage game in embedded in an evolutionary process solely driven by material success.

In order to derive the evolutionarily equilibrium, a preference game $\hat{\Gamma}$ is constructed and analyzed. The intuition is that players are bound to their reciprocity parameter ρ_i just like they are bound to certain strategies in direct evolutionary games. The difference is that strategies in typical direct evolutionary games are identical to certain actions while in this approach, the reciprocity parameter determines behavior according to Proposition 1.

Consider the following setup. In each point in (continuous) time $t \in [0, \infty)$, nature first matches each player in a (random) pair and randomly assigns player positions a or b in the preference game. Irrespective of player position a, b in the preference game $\hat{\Gamma}$, nature also determines first or second mover position in trust game Γ. For both assignments, the probability for each player to be in the one or the other role is 0.5. Each player's strategy consists of some $\delta_i \equiv \rho_i$ from a finite, but possibly arbitrarily fine grid $\Delta_i \equiv \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \overline{P}\right\}$ with $n \in \mathbb{N}$ and an upper bound $\bar{P} > 0$.[3] Any strategy profile $\delta \in \Delta^2$ implies a specific equilibrium behavior (p_i^*, q_i^*) in trust game Γ. To be more precise, equilibrium behavior in game Γ becomes a function of the own strategy δ_i as well as of the strategy of the other player δ_j in preference game $\hat{\Gamma}$, i.e. $p_i^* = p_i^*(\delta_i, \delta_j)$ and $q_i^* = q_i^*(\delta_i, \delta_j)$. Payoffs for preference game $\hat{\Gamma}$ are thus indirectly determined via (p_i^*, q_i^*). It is assumed that players have perfect information. It is well known that the ability to discriminate between different player types, which requires information, is crucial for the evolutionary success of social preferences, see e.g. Güth and Kliemt (1994) or more recently Herold and Kuzmics (2009).

Preference game $\hat{\Gamma}$ is thus a $k \times k$ simultaneous move game with payoffs dependent on underlying trust games played by subjects with Falk/Fischbacher preferences. Since it is true for both players that they are in first or second mover position with probability 0.5 in trust game Γ, the 0.5 is ignored. In addition, notation is simplified by writing p_i for $p_i^*(\delta_i, \delta_j)$ and q_i for $q_i^*(\delta_i, \delta_j)$.

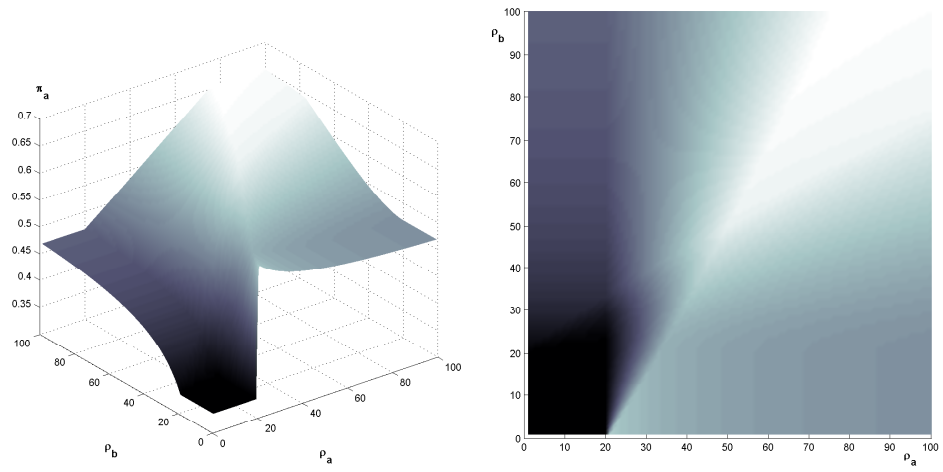
[3] The construction of the preference game is similar to Berninghaus et al. (2007). The restriction to a finite strategy space is due to a lack of general sufficient conditions for stability in evolutionary games with infinite strategy spaces, see e.g. Oechssler and Riedel (2001). Note, however, that all results will hold for the limit case of an infinitely fine grid.

The payoff for players $i \in \{a, b\}$ in preference game $\hat{\Gamma}$ is determined as:

$$\hat{\pi}_i = p_i(q_j r - s) + p_j(q_i r - s + 1 - q_i) + 2s \tag{2.3}$$

where $p_i(q_j r - s) + s$ is the payoff as a first mover and $p_j(q_i r - s + 1 - q_i) + s$ is the payoff as a second mover with respect to Γ. One important observation is that the above specified preference game $\hat{\Gamma}$ is symmetric, i.e. it will suffice to focus on one player for equilibrium derivation.

Figure 2.3 illustrates the payoffs for player a dependent on ρ_a and ρ_b. Payoff parameters are set to $r = \frac{2}{3}$ and $s = \frac{1}{3}$. Dark shading reflects comparably low payoffs and Figure 2.3(b) represents the view on Figure 2.3(a) *from the top*. The bright *line* ranging from medium left bottom to medium right top in (b) indicates the respective best response by a for given ρ_b.



(a) 3-D-perspective (b) 2-D-perspective

Figure 2.3: Payoffs in $\hat{\Gamma}$ for player a.

It is well known that there is a close connection between Nash equilibria and stable states in evolutionary games, see e.g. Weibull (1996). Therefore, I start by searching for Nash equilibria of preference game $\hat{\Gamma}$. Recall from Proposition 1 that both p_i, q_i are monotone in ρ_i such that q_i monotonically increases in ρ_i beyond the threshold $\rho_i = \frac{1-r}{r}$ and p_i strictly decreases in ρ_i but strictly increases in q_j both if $p_i \neq \{0, 1\}$. Besides the monotonicity, p_i, q_i are continuous but not everywhere differentiable functions. Throughout the analysis it is assumed that the initial state is truly mixed which implies that there are at least some players with a ρ_i such that $p_j > 0$ and the system is not initially at rest.

Now, let $\rho_j = \bar{\rho}_j$ be given such that $q_j = \bar{q}_j$ and $p_j = p_j(\rho_i)$. Define $\{\rho_i | p_j = 1\}$ as the set of ρ_i-values such that the other player trusts with probability 1. It is straightforward to see that whenever $\bar{q}_j$ is such that player i either never trusts, $p_i = 0$, or always trusts, $p_i = 1$, then the best response by i must satisfy $\min\{\rho_i | p_j = 1\}$. In both cases the first mover payoff is independent of ρ_i. Second mover payoff, however, is positive if $p_j > 0$ (since this requires $q_i > \frac{s}{r}$) and strictly increasing in ρ_i as long as $p_j \leq 1$. If player j trusts for sure, then the payoff of i decreases in ρ_i because player j would still trust for sure but i exploits less frequently.

There is a slight difference if $\bar{q}_j$ is such that $0 < p_i < 1$. In this case, i's payoff is decreasing in ρ_i given that ρ_i is sufficiently low in order to induce j to never trust i (and second mover payoff drops out). Indeed, given that case, player i should lower ρ_i until $p_i = 1$ such that the payoff to i becomes $\bar{q}_j r - s$. However, player i might as well choose some ρ_i such that $0 < p_j < 1$ and $0 < p_i < 1$. For this case, substituting the respective equations for p^* and q^* according to Proposition 1 into equation (2.3) and taking the derivative of (2.3) with respect to ρ_i yields:

$$\partial_{\rho_i} \hat{\pi}_i = -\frac{(\bar{q}_j r - s)^2}{\rho_i^2 (1 - \bar{q}_j)(\bar{q}_j r - s + 1 - \bar{q}_j)} + \frac{r(r-s)}{\rho_j (1-r)}$$

The parameter restrictions are not sufficient to directly follow whether the derivative is positive or negative. However, observe that for $p_i > 0$ it must hold that $q_j = s/r + \varepsilon$ with $\varepsilon > 0$. Taking the limit $\varepsilon \to 0$, the negative term of the derivative vanishes as $(\bar{q}_j r - s)^2$ approaches 0 and the second term is strictly positive. In addition, the second order derivative

$$\partial^2_{\rho_i} \hat{\pi}_i = \frac{2(\bar{q}_j r - s)^2}{\rho_i^3 (1 - \bar{q}_j)(\bar{q}_j r - s + 1 - \bar{q}_j)} > 0$$

is strictly positive. Thus, the derivative is positive at the lower bound and then increases. Player i should thus increase ρ_i up to the boundary where $p_j = 1$. If $p_j = 1$, second mover payoff alone is $q_i r - s + 1 - q_i$ and overall utility is certainly greater than $\bar{q}_j r - s$ such that the alternative with $p_j = 1$ is superior to the one with $p_j = 0$. Again, it cannot be optimal for i to choose an even higher ρ_i since his payoff starts to decrease beyond $\min\{\rho_i | p_j = 1\}$.

Since $\hat{\Gamma}$ is symmetric, the following Lemma can be stated.

Lemma 1 *In a Nash equilibrium of preference game $\hat{\Gamma}$, players must choose*

$$\rho_i^* = \min\{\rho_i | p_j = 1\}$$

Given Lemma 1, both players will choose the same reciprocity parameter in equilibrium. Using that information, the ρ satisfying Lemma 1 is given by:

$$\rho^* = \frac{1 - 3r + 4r^2 - 2r^3}{r(2r-1)(r-s)} \tag{2.4}$$

Recall that it was assumed that mutual cooperation is the efficient outcome of trust game Γ, i.e. $2r > 1$. This guarantees that $\rho^* > 0$. For example, with $r = \frac{2}{3}$ and $s = \frac{1}{3}$, one obtains $\rho^* = 2.5$ and $q^* = 0.8$. To complete the evolutionary analysis, observe that the equilibrium implied by Lemma 1 is unique, symmetric, and strict. The following Proposition 2 immediately follows.

Proposition 2 *Any payoff monotone selection dynamic on $\hat{\Gamma}$, as well as fictitious play and the best response dynamic, all starting from a truly mixed environment that is not initially at rest, must converge toward ρ^* specified in equation (2.4). The associated strategy ρ_i^* is the unique evolutionarily stable strategy (ESS) of $\hat{\Gamma}$.*

Proof 2 *For an arbitrary two player symmetric game with $k \times k$ payoff matrix, a symmetric strict Nash equilibrium is asymptotically stable for payoff-monotone selection dynamics as well as for fictitious play and the best-response dynamics (see Theorem 2.5.3 in Cressman (2003) for a proof). Since ρ^* yields a symmetric and strict equilibrium, it must be asymptotically stable. Since ρ^* is also the unique Nash equilibrium, it must be globally asymptotically stable and any dynamic must converge given that the process is not initially at rest. The fact that ρ^* is an unique ESS follows from uniqueness of the equilibrium and from the fact that a strategy that yields a symmetric and strict Nash equilibrium is an ESS, see Definition 2.5.1 in Cressman (2003).*

Note that the class of payoff-monotone selection dynamics includes the well known replicator dynamics. As pointed out in the introduction, this result contrasts to the findings by Berninghaus et al. (2007) who have shown that reciprocity cannot induce behavior any different from purely selfish behavior in the dictator game but that the reciprocity parameter approaches infinity in the ultimatum game yielding fair split offers.[4] Here, a medium level

[4]Actually, players in the dictator game can have a strong concern for reciprocity. Since the second mover cannot do anything else than accepting, his action is perceived as unintentional which extends the strategy space by players because they now have free parameters ρ_i and the intention factor. A strong concern for reciprocity is evolutionarily stable if it is combined with a sufficiently low intention factor such that the resulting action is an offer of zero. Thus, reciprocity might be there but it cannot, from the evolutionary perspective, induce any behavior that is different from money maximizing behavior.

of reciprocal inclination is stable. Like in the ultimatum game, reciprocity is not successful because players care for equity or reciprocity as such. Rather, second movers behave reciprocal for strategic reasons, namely to induce trust by first movers. The result also allows for a neat interpretation already mentioned above. It is optimal for every player to be reciprocal to such a degree that the others will just trust with probability one. Due to the symmetry of the solution, every player, in equilibrium, will be reciprocal exactly up to that degree and hence each player will trust each other player just with probability one. With each player of a unique type, this is equivalent to every player being exactly so reciprocal that he would just trust himself.

2.4 Discrimination Between Positive and Negative Reciprocity

So far it was assumed that each player has exactly one reciprocity parameter ρ_i. Consequently, the reciprocal inclination that drives the desire to reward kind actions (positive reciprocity) and punish unkind ones (negative reciprocity) was taken to be the same. However, it might be that this assumption does not hold. In fact, there is some empirical evidence that supports the view that positive and negative reciprocity are not two sides of the same token. For example, Dohmen et al. (2009) present evidence from a representative survey of more than 20.000 German citizens which reveals a literally tiny .01 albeit 5%-level significant correlation between individual statements regarding positive and negative reciprocity. In addition, the idea of distinct traits regarding social preferences with respect to positive and negative deviations from a norm is also common in the theoretical literature. The inequity aversion model by Fehr and Schmidt (1999), for example, includes distinct parameters regarding aversion against advantageous and disadvantageous inequality.

The specific game of trust used in this paper easily allows studying the impact or different reciprocal inclinations regarding positive and negative reciprocity. In fact, players who are assigned the role of the second mover in trust game Γ are asked to be positively reciprocal since shown trust must be seen as a kind action. On the other hand, for almost all beliefs that a first mover can hold, player 1 will perceive player 2 as unkind and may possibly punish the second mover by not trusting, i.e. by being negatively reciprocal. Hence, to study the impact of different parameters for positive and negative reciprocity, one can simply substitute ρ_{i1} for ρ_i in p^* and ρ_{i2} for ρ_i in q^* with both p^*, q^* according to Proposition 1.

The introduction of a separate concern for positive and negative reciprocity does not affect equilibrium behavior in the one shot version of Γ as Proposition 1 is unaffected by this change. However, the discrimination makes a difference in the evolutionary game. At first, the strategy space for each player is now extended such that $\delta_i \equiv (\rho_{i1}, \rho_{i2}) \in \Delta_i \equiv \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \overline{P}\right\} \times \left\{0, \frac{1}{k}, \frac{2}{k}, \ldots, \overline{Q}\right\}$ with $n, k \in \mathbb{N}$ and upper bounds $\bar{P}, \bar{Q} > 0$. Now recall that the payoff for player i in $\hat{\Gamma}$ is given by

$$\hat{\pi}_i = p_i(q_j r - s) + p_j(q_i r - s + 1 - q_i) + 2s$$

Player i can now choose (ρ_{i1}, ρ_{i2}) such that he optimizes first and second mover payoff independent of each other. The result concerning second mover behavior is unaffected by the change, i.e. the best response in second mover position is still $\rho_{i2} = \min\{\rho_{i2} | p_j = 1\}$. For a given $\bar{q}_j > \frac{s}{r}$, however, first mover payoff is always maximal whenever $p_i = 1$, i.e. if $\rho_{i1} \in \{\rho_{i1} | p_i = 1\}$. Recall that p_i decreases in ρ_{i1} such that if some ρ_{i1} yields $p_i = 1$, then any other ρ'_{i1} with $\rho'_{i1} < \rho_{i1}$ will also yield $p_i = 1$ and the payoff for i is unaffected. However, with a lower ρ_{i1}, player j would want to decrease his reciprocity parameter such that in Nash equilibrium, the maximal value of $\{\rho_{i1} | p_i = 1\}$ must be chosen in order to satisfy the condition of mutual best responses.

Lemma 2 *In preference game $\hat{\Gamma}$ with players allowed to have different reciprocal inclinations regarding positive and negative reciprocity, any Nash equilibrium must be characterized by*

$$\rho^*_{i1} = \max\{\rho_{i1} | p_i = 1\}, \text{ and } \rho^*_{i2} = \min\{\rho_{i2} | p_j = 1\}$$

The crucial aspect is that any constellation satisfying Lemma 2 will yield a Nash equilibrium of preference game $\hat{\Gamma}$. But as long as $\{\rho_{i1} | p_i = 1\}$ is more than single valued, none of these equilibria with $\rho_{i1} > 0$ will be strict. Player i could always choose a lower parameter value related to negative reciprocity and would still earn the same payoff. In fact, strategies with lower ρ_{i1}-values weakly dominate those with higher ones since while they never do worse than strategies with a high ρ_{i1}-value, they will do better against lower ρ_{i2}-values.

The impact of this result is straightforward. Given some distribution of reciprocity parameters in the population, first movers with a comparably low negative reciprocity parameter will either earn an equal or higher payoff than those with a comparably high negative reciprocal inclination. Further, second movers with a sufficiently high parameter for positive reciprocity will always be trusted. However, whenever interacting with first movers who fully trust but would already do so at lower levels of trustworthiness, those second movers who indeed exploit more often will fare better. Thus, the

average rate of trustworthiness faces downward pressure. Given that, some low, but within this class comparably large parameter values for negative reciprocity will no longer satisfy $p_i = 1$ which is suboptimal and implies downward pressure on negative reciprocal inclination. This process will not stop besides in the limit where negative reciprocal inclination vanishes.

To conclude, note that with $\rho_{i1} = 0$, reciprocity payoff in the utility function of a first mover drops out and the player simply compares the expected payoff from trusting (for given belief) with the payoff from not trusting. The threshold value for q_j is $q_j = \frac{s}{r}$ and I assume that players who are indifferent choose to trust, i.e. p_i becomes a step function from zero to one such that at the threshold, $p_i = 1$. Proposition 3 summarizes the results.

Proposition 3 *The unique evolutionarily stable strategy in preference game $\hat{\Gamma}$ with players allowed to exhibit distinct reciprocal inclination regarding positive and negative reciprocity is $\delta_i^* = (\rho_{i1}^*, \rho_{i2}^*) = (0, \frac{1-r}{r-s})$.*

Proof 3 *Since δ_i^* yields a symmetric and strict Nash equilibrium, it is an ESS, see the proof of Proposition 2. Following Weibull (1996, Proposition 2.3), a strategy that is weakly dominated cannot be an ESS. Since all other strategies $\delta = (\rho_{i1}, \cdot)$ are weakly dominated by $\delta_i^* = (0, \cdot)$, δ_i^* is the unique ESS.*

The value $\rho_{i2}^* = \frac{1-r}{r-s}$ is obtained by solving $\frac{s}{r} = 1 - \frac{1-r}{\rho_{i2} r}$ with respect to ρ_{i2}. With the example payoff parameters $r = \frac{2}{3}$ and $s = \frac{1}{3}$, one obtains $\rho_{i2}^* = 1$ and $q^* = 0.5$. For positive reciprocity, again an intermediate level or reciprocal inclination is evolutionarily stable. Also, it is still true that players should be exactly so reciprocal that others just trust them for sure. On the other hand, negative reciprocity must vanish in the long run. A more stylized prediction would be that negative reciprocal inclination should typically be lower than positive reciprocal inclination since the first faces permanent downward pressure while the latter receives support at least for intermediate values. This is in fact in line with the findings by Dohmen et al. (2009) who report a rather strong concern for positive reciprocity but obtain weaker (and more dispersed) support for negative reciprocity.

2.5 Conclusion

In this paper, a formal proof for the reciprocity equilibrium in the game of trust was provided. It should be pointed out that the intuition behind the equilibrium has been present since the working paper version of the Falk-Fischbacher model, see Falk and Fischbacher (1998), and was already applied

in e.g. Altmann et al. (2008). Nevertheless, to the best of my knowledge, a formal proof was lacking until today. The main aspect here is that players in second mover positions will apply a mixed strategy and only in the limit of an infinite reciprocal inclination, the probability to reward approaches one.

Whereas equilibrium derivation for the one-shot game is a preliminary exercise, the focus of the paper is the evolutionary analysis. As pointed out, the evolutionary prediction differs structurally from previous findings because a medium level of reciprocity and a medium impact on behavior is found to be evolutionarily stable. This contrasts the extreme predictions of either no impact on behavior as for the dictator game, or the implication of fair split offers in the ultimatum game associated with an infinitely strong reciprocal inclination.

One observation is that although players use equity as a reference standard to evaluate kindness, and although they have a perfect ability to discriminate between different types, the solutions are not characterized by equity between the first and second movers. If players have a joint reciprocity parameter for positive and negative reciprocity, and with the parameter specifications used in the examples, the first mover ends up with 42% of the total payoff and the second mover gets 58%. If players discriminate between positive and negative reciprocity, then the first mover is left with 29% of the realized total payoff and the second mover gets 71%.

Another interesting aspect is that the evolutionary predictions are not characterized by efficiency. The efficient solution would call for a reward probability of one but in the first treatment with a joint reciprocity parameter, total realized payoffs remains 5% short of the efficient solution and in the second treatment, overall payoff is 12.5% less than possible payoff. Hence, the dilemma character of the game is not removed entirely.

Both observations are related to the weakness of negative reciprocity in the trust game. First movers are forgiving in the way that they might fully trust even though the reward probability is less than one. This gives rise to *constrained dictators* that ensure complete trust but are not at all generous beyond the needed level of cooperation. From the perspective of efficiency, it would be desirable to have players with a strong concern for negative reciprocity but evolution yields no support for negative reciprocity. Given that sure trust must be reached, second mover payoff is structurally higher than first mover payoff (in principle $qr - s$ vs. $qr - s + 1 - q$) so the dictator game aspect of the trust game sooner or later outweighs the effect of negative reciprocity. The effect of disjoining positive and negative reciprocity strongly illustrates the point.

The difference to the dictator game is obviously due to the fact that the dictator game leaves no room for discrimination between different types.

The difference to the ultimatum game is due to the much stronger impact of negative reciprocity. First movers in ultimatum games may base their decision on an intrinsic concern for reciprocity or on strategic concerns, i.e. to avoid rejection. However, in symmetric equilibria, the strategic aspect always outweighs the first movers concern for reciprocity and therefore, both first and second mover behavior is determined by the reciprocity parameter of the second mover (associated with negative reciprocity). In difference to that, both players concerns for reciprocity are decisive for decisions in the game of trust.

Using Falk-Fischbacher preferences makes the analysis more complicated compared to the approach taken by e.g. Güth and Kliemt (1994) who broadly separate trustworthy and non-trustworthy types. But it is worth the effort because the additional insight is a more differentiated picture, especially with respect to evolutionarily stable mixed strategies by second movers as well as the possibility to analyze positive and negative reciprocity separately. Besides this, the results are structurally identical as e.g. first movers should fully trust. This supports the result in the way that it is consistent with other approaches.

Chapter 3

The Co-Evolution of Reciprocity-Based Wage Offers and Effort Choices

This work is published in Economics Letters, Vol. 117(1), pp. 326-329

3.1 Introduction

Theories of social preferences based on reciprocity explain a positive wage-effort relation in the gift-exchange game frequently observed in experiments (e.g. Fehr et al., 1997, 2007). But when and why does reciprocal behavior evolve? And, is it persistent? A natural explanation would be that reciprocity yields superior payoffs in an evolutionary context. Accordingly, I study the evolution of reciprocity in a gift-exchange game.

Previous works have revealed that the ability to discriminate between different player types is crucial for the evolutionary success of other-regarding preferences (e.g. Güth and Kliemt, 1994; Herold and Kuzmics, 2009). Applying the reciprocity model introduced by Falk and Fischbacher (2006), Berninghaus et al. (2007) have shown that an infinitely large reciprocal inclination associated with fair-split offers is stable in the ultimatum game but behavior corresponding to money-maximization is successful in the dictator game.

In contrast to the ultimatum game, second mover choices in the gift-exchange game are associated with positive rather than negative reciprocity. Similar to the dictator game, workers make quasi-dictatorial decisions but they need the employers to trust them. Another characteristic is that there may be two equilibrium wages, either high ones inducing high effort or low

ones inducing low effort. On an a priori basis, it is unclear which kind of behavior will be evolutionary successful. Further, contrasting the standard one-population approach, the situation calls for a multi-population model since it appears unlikely that employers and workers frequently switch positions.

The paper proceeds as follows: Section 3.2 discusses the gift-exchange game and reviews Falk-Fischbacher preferences and the reciprocity equilibrium for the game. In sections 3.3 and 3.4, the evolution of reciprocity parameters is studied. Section 3.5 concludes.

3.2 Reciprocity Equilibrium in the Gift-Exchange Game

Using the specification by Falk and Fischbacher (2006), the gift-exchange game Γ is a two-player sequential game with an employer (E) who moves first offering a wage w to the worker (W). Given that the worker accepts the offer, the wage is paid and the worker chooses an effort level e. Pecuniary payoffs are given as $\pi_E = ve - w$ and $\pi_W = w - c(e)$. For simplicity, assume that $w \in [0,1], e \in [0,1]$, and $v = 1$. Further, let $c(e) = \alpha e^2$ with $\alpha \leq \frac{1}{4}$. Once the wage is paid, the worker has full discretion over the final outcome. If payoffs are equal to utility, $u(\pi) = \pi$, the unique subgame perfect equilibrium of the game is $e^* = 0, w^* = 0$.

Now assume that pecuniary payoffs do not equal utility. Rather, players hold Falk and Fischbacher (2006) preferences for reciprocity.[1] Agent i's utility is defined by:

$$u_i(f) \equiv \pi_i(f) + \rho_i \varphi_{ji}(n)\sigma_{ij}(n,f) \tag{3.1}$$

Utility is the sum of pecuniary reward at terminal node f, $\pi_i(f)$, and reciprocity utility $\varphi_{ji}(n)\sigma_{ij}(n,f)$ scaled with the individual reciprocity parameter $\rho_i \in \mathbb{R}_+$. The *kindness term* $\varphi_{ji}(n)$ evaluates the kindness by j toward i at non-terminal node n by comparing the expected payoffs for both players. Whenever i expects to get more (less) than j, player j's action is considered as kind (unkind). In addition, overall kindness depends on the intentions behind j's (un-)kindness. If, for example, player j is unkind but has no alternative to be less unkind, then the unkindness is considered as unintentional and the difference of expected payoffs is multiplied with the *outcome*

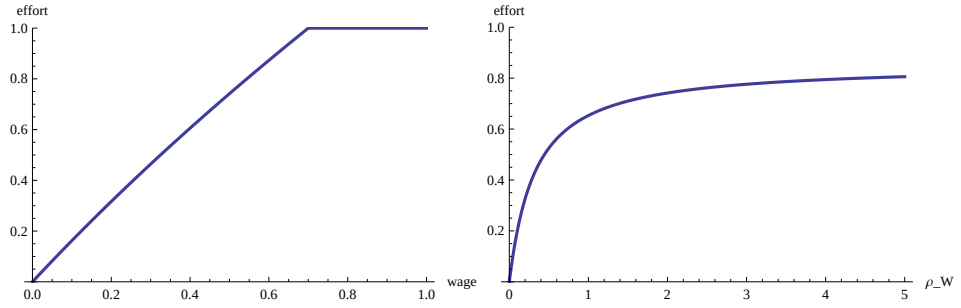
[1] The model is based on psychological game theory, see Geanakoplos et al. (1989), and combines outcome-based approaches to other-regarding preferences, like Fehr and Schmidt (1999), with intention-based models, like Rabin (1993).

concern parameter $\epsilon_i \in [0, 1]$.[2] The second component of reciprocal utility is the *reciprocation term* $\sigma_{ij}(n)$ capturing the impact of i's decision in n on j's final payoff.

The reciprocity equilibrium for the gift-exchange game is provided in Falk and Fischbacher (1998). Whenever the reciprocity parameter of the worker is zero, $\rho_W = 0$, then $e^* = 0, w^* = 0$ is the unique reciprocity equilibrium. Whenever $\rho_W > 0$, the optimal effort decision satisfies

$$e^* = \min \left[1, \frac{-2\alpha - \rho_W + \sqrt{(2\alpha + \rho_W)^2 + 8\alpha\rho_W^2 w}}{2\alpha\rho_W} \right] \tag{3.2}$$

Figure 3.1 illustrate the behavior of the workers ($\alpha = .2$).



(a) $e^*(w)$ with $\rho_W = 2$ (b) $e^*(\rho_W)$ with $w = .5$

Figure 3.1: Workers behavior

Since the worker can always assure an equal split but typically receives more than the employer, he judges the employer as kind and effort increases in w and ρ_W. Note that except for very low levels of ρ_W, effort will be equal to 1 for wages less than 1.

With regard to first-mover behavior, let $\bar{w}(\alpha, \rho_W) = \frac{1+\alpha}{2} + \frac{\alpha}{\rho_W}$ be the minimal wage that ensures an effort choice of 1. Moreover, let $\tilde{w}(\alpha, \rho_E, \rho_W)$ be the wage offer if w, e are not restricted to $w, e \leq 1$ but restrict $w^* \in [0, 1]$.[3] Then, there is always an equilibrium given by

$$w^* = \min \left[\bar{w}(\alpha, \rho_W), \tilde{w}(\alpha, \rho_E, \rho_W) \right] \tag{3.3}$$

Since the expected pecuniary payoff of the employer is smaller than the one of the worker, the employer judges the worker as unkind. Whenever the worker provides an effort of 1, however, the worker has no chance to be less

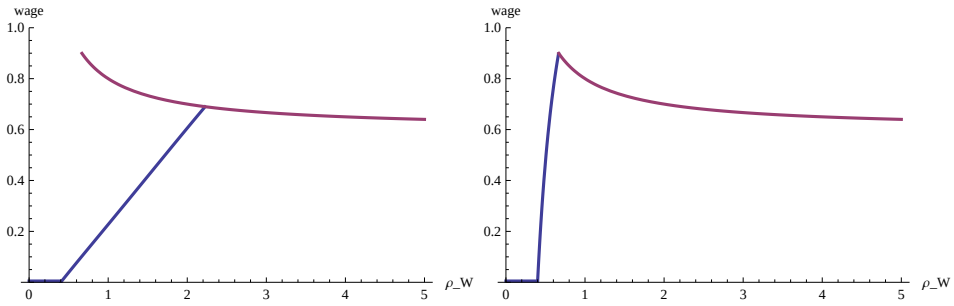
[2] For an exact and formal definition of all terms, see Falk and Fischbacher (1998, 2006).

[3] The exact expression $\tilde{w}(\alpha, \rho_E, \rho_W)$ is provided in the Appendix.

unkind. In such cases, the employer judges the unkindness as unintentional and with a sufficiently low ϵ_E, the employer nevertheless offers a comparably high wage. Formally, if $\bar{w}(\alpha, \rho_W) \leq 1$ and if

$$\epsilon_E \rho_E \leq \frac{\rho_W(-\rho_W + 2\alpha + 2\alpha\rho_W)}{2\alpha(-2\alpha - \rho_W + 2\alpha\rho_W)}, then\ w^* = \bar{w}(\alpha, \rho_W) \tag{3.4}$$

Figure 3.2 illustrate the behavior of the employer ($\alpha = .2$).



(a) $w^*(\rho_W)$ with $\rho_E = 3$ (b) $w^*(\rho_W)$ with $\rho_E = 0$

Figure 3.2: Employers behavior

Wage offers $\bar{w}(\alpha, \rho_W)$ are decreasing in ρ_W since workers with a higher reciprocal inclination provide the maximal effort for lower wages. For very low ρ_W, a zero wage is offered but beyond a threshold, wage offers $\tilde{w}(\alpha, \rho_E, \rho_W)$ strictly increase in ρ_W. If the reciprocal inclination of the employer is larger than zero, equilibrium wage offers may depend on ϵ_E (upper and lower branch in figure 3.2 a). If the reciprocal inclination of the employer is zero, only one equilibrium can exist (figure 3.2 b). Note that $\tilde{w}(\alpha, \rho_E, \rho_W)$ is strictly decreasing in ρ_E. This is due to the fact that the employer judges the worker as unkind.

3.3 Evolution of Wage Offers and Effort Decisions

In order to study the evolution of wage offers and effort decisions, I adopt an indirect evolutionary approach. Assume that there are two disjoint populations of workers and employers, each consisting of infinitely many agents. In each point in continuous time, $t \in [0, \infty)$, a one-shot version of gift-exchange game Γ is played. Assume:

(i) In each stage game, employers and workers are pairwise randomly matched to play the gift-exchange game Γ. Players have perfect information and behave fully rational, i.e., for given reciprocity parameters ρ_E, ρ_W, workers' behavior is guided by (3.2) and employers' behavior is guided by (3.3) or (3.4).

(ii) The stage games are embedded in an evolutionary process purely driven by pecuniary success. A fraction of the population (of workers, of employers) holding a specific reciprocity parameter will have a positive growth rate function whenever that parameter yields a payoff that is above population average (weakly payoff-positive selection dynamics, see Weibull, 1996, Definition 5.8).

Similar to Berninghaus et al. (2007), I construct a preference game $\hat{\Gamma}$ which allows to apply direct evolutionary tools. Preference game $\hat{\Gamma}$ is a $k \times k$ two-player game with the employers and workers as players. Individual strategies $\delta_i, i \in \{E, W\}$, in $\hat{\Gamma}$ are defined by the reciprocity parameters, i.e. $\delta_i \equiv \rho_i$. The outcome concern parameter ϵ_E is assumed to be a datum and there is no employer with $\epsilon_E = 0$. Both restrictions will be relaxed in the next section. The strategy space is assumed to be finite for each population, i.e. $\Delta_i \equiv \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \bar{P}\right\}$ with upper bound $\bar{P} > 0$ and $n \in \mathbb{N}$. Figure 3.3 illustrates preference game $\hat{\Gamma}$.

	ρ_W^1	ρ_W^2	$\cdots$	ρ_W^k
ρ_E^1	$\pi_E(\cdot), \pi_W(\cdot)$	$\pi_E(\cdot), \pi_W(\cdot)$	$\ldots$	
ρ_E^2	$\pi_E(\cdot), \pi_W(\cdot)$	$\pi_E(\cdot), \pi_W(\cdot)$		
$\vdots$	$\vdots$		$\ddots$	
ρ_E^k				

Figure 3.3: Preference Game $\hat{\Gamma}$

Given the above, strategy combinations $\delta \in \Delta^2$ in $\hat{\Gamma}$ uniquely determine equilibrium behavior in gift-exchange game Γ and payoffs in Γ determine the evolution on Δ^2 in preference game $\hat{\Gamma}$. Since there is a well-established connection between Nash equilibria and evolutionary stable states (see e.g. Weibull, 1996; Cressman, 2003), I check for Nash equilibria in $\hat{\Gamma}$. The payoff of the employers is given by:

$$\pi_E = e^*(\cdot) - w^*(\cdot) \tag{3.5}$$

First suppose that for given ρ_W, ρ_E is such that $w^* = \tilde{w}(\alpha, \rho_E, \rho_W)$, i.e. $e^*(\cdot) < 1$. Then

$$\frac{\partial \pi_E(\cdot)}{\partial \rho_E} < 0 \tag{3.6}$$

for all possible ρ_E, ρ_W, α. Thus, any ρ_E inducing a wage such that $e < 1$ cannot be best response. On the other hand, any ρ_E that supports $w^* = \bar{w}(\alpha, \rho_W)$ must be a best response since then $\pi_E(\cdot) = 1 - \bar{w}(\alpha, \rho_W)$ becomes independent of ρ_E.

Lemma 3 *The best response of the employers must satisfy*

$$\rho_E^* \in \{\rho_E | w = \bar{w}(\alpha, \rho_W)\} \tag{3.7}$$

In general, the best response of the employers lacks uniqueness. Further, for two employers with different ϵ_E, the set of ρ_E-values that ensures $w = \bar{w}(\alpha, \rho_W)$ may be different (those with a comparably high ϵ_E will have to choose lower ρ_E). By the exclusion of $\epsilon_E = 0$, however, $\rho_E = 0$ constitutes a unique best-response against $\rho_W = \frac{2\alpha}{1-2\alpha}$.

Now, I turn to the workers' behavior. Suppose that $w = \bar{w}(\alpha, \rho_W)$ such that $e = 1$. It is obvious that the lowest possible ρ_W that supports $\bar{w}(\alpha, \rho_W)$ is optimal since $\bar{w}$ decreases in ρ_W. With constant costs ($e = 1$), increasing ρ_W would lower the payoff for the workers. Now suppose that $w = \tilde{w}(\alpha, \rho_E, \rho_W)$. Then

$$\frac{\partial \pi_W(\cdot)}{\partial \rho_W} > 0, \tag{3.8}$$

i.e. the higher wage based on a slightly higher effort outweighs the additional cost of the extra effort.[4] Hence

Lemma 4 *The best response by the workers must satisfy*

$$\rho_W^* = \min \{\rho_W | w = \bar{w}(\alpha, \rho_W)\} \tag{3.9}$$

By Lemma 3 and 4, preference game $\hat{\Gamma}$ has many Nash equilibria. For given ρ_W, the maximal ρ_E which implies $w = \bar{w}(\alpha, \rho_W)$ is a best response by the employers and, given that choice, ρ_W is a best response by the workers. Since the employers earn the same payoff by choosing a lower ρ_E, equilibria are non-strict in general. An exception is the equilibrium $(\rho_E^*, \rho_W^*) = \left(0, \frac{2\alpha}{1-2\alpha}\right)$, since the employers cannot choose a lower ρ_E.

[4]Note also that since the workers earn a strictly positive payoff given that wages are positive, any ρ_W that induces $w = 0$ cannot be a best response either.

Proposition 4 *The equilibrium* $(\rho_E^*, \rho_W^*) = \left(0, \frac{2\alpha}{1-2\alpha}\right)$ *is the unique evolutionary stable strategy profile of preference game* $\hat{\Gamma}$*. It is asymptotically stable under any weakly payoff-positive selection dynamic and the unique asymptotically stable state.*

Proof 4 *First note that the strategy profile yields the unique strict Nash equilibrium of* $\hat{\Gamma}$*. According to (Weibull, 1996, Proposition 5.1), a strategy profile is evolutionary stable in multipopulation models* if and only if *it yields a strict Nash equilibrium. His Proposition 5.11 says that* every *strict Nash equilibrium is asymptotically stable in all weakly payoff-positive selection dynamics but in accordance with his Proposition 5.12, a* pure but non-strict *equilibrium is* not *asymptotically stable.*

With a mixed initial population and/or occasional mutations, one can expect that wage offers and effort decisions will converge toward $(w^*, e^*) = \left(1 - \frac{\alpha}{2}, 1\right)$. Long-run equilibrium payoffs are $(\pi_E, \pi_W) = \left(\frac{\alpha}{2}, 1 - \frac{3\alpha}{2}\right)$. The long-run equilibrium is efficient ($e^* = 1$) but characterized by strong inequity in favor of the workers.

3.4 Outcome Concern as a Strategic Variable

Alternative to the assumption that ϵ_E is a datum and non-zero, one can regard it as a strategic variable in $\hat{\Gamma}$. While Lemma 2 (workers behavior) is unaffected, the strategy space for the employers extends to $\Delta_E \equiv \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, \bar{P}\right\} \times \left\{0, \frac{1}{n}, \frac{2}{n}, \ldots, 1\right\}$. For any given ρ_W, employers can now ensure $w = \bar{w}(\alpha, \rho_W)$ either by lowering ρ_E or ϵ_E such that a set of ϵ_E, ρ_E combinations supports $e^* = 1$. Let $\bar{W} \equiv \{w | w = \bar{w}(\alpha, \rho_W)\}$ be the set of all wages associated with an effort of 1 and let $\hat{w} \in \bar{W}$ be a particular wage. Then define a set $\hat{V} \equiv \{(\epsilon_E, \rho_E) | w = \hat{w}\}$, i.e. $\hat{V}$ contains all ϵ_E, ρ_E combinations which support a particular wage $\hat{w}$. Since expression (3.6) still holds, $\tilde{w}(\alpha, \rho_E, \rho_W)$ is independent of ϵ, one can conclude

Lemma 5 *The best response by the employers must satisfy*

$$(\epsilon_E, \rho_E)^* \in \hat{V} \tag{3.10}$$

Now let $\hat{w}', \hat{w}''$ be two wages such that $\hat{w}' > \hat{w}''$. Then, for the associated sets $\hat{V}'$ and $\hat{V}''$, it must be true that $\hat{V}' \subset \hat{V}''$. If some ρ_E, ϵ_E combinations are small enough such that $w^* = \bar{w}(\alpha, \rho_W)$ for a given ρ_W, then they also support $w^* = \bar{w}(\alpha, \rho_W)$ for some higher ρ_W' (the higher ρ_W implies a lower $\hat{w}$). In this sense, $\hat{V}''$ is not unique. This non-uniqueness vanishes if $\rho_W = \frac{2\alpha}{1-2\alpha}$ as

discussed in the previous section. At this point, either $\rho_E = 0$ or $\epsilon_E = 0$ or both. Define this set as $\hat{V}^0 \equiv \left\{(\epsilon_E, \rho_E) \in \hat{V} | \epsilon_E \rho_E = 0\right\}$. For lower ρ_W, employers will never offer a wage that ensures an effort of one.

Evolution will select among the different sets just like it selects between different reciprocity parameters.

Corollary 1 *Given that ϵ_E and ρ_E are strategic variables in $\hat{\Gamma}$, $(\hat{V}^*, \rho_W^*) = (\hat{V}^0, \frac{2\alpha}{1-2\alpha})$ is asymptotically stable.*

3.5 Discussion and Summary

The results are somewhat ambivalent. On the one hand, the dilemma of gift-exchange can be resolved by reciprocity. A sufficient reciprocal inclination by the workers is stable such that an efficient outcome receives support. On the other hand, the disadvantage of the employers suggests that they may favor other mechanisms, like contract theory solutions including fines, in order to ensure high effort. Although such solutions are inefficient, they may become popular since they can ensure a higher share for those employers who make use of them. If one interprets the reciprocity solution as a form of norm-based efficiency wages in the sense of Akerlof (1982), then the result suggest that while efficiency wages might well work, they may be unstable over time.

Another question is whether the result will hold beyond the particular reciprocity model applied. Besides the specific cost function, the driving force behind the result is the fact that workers choose the maximal effort for less than maximal wages (which leads to a unique strict equilibrium in the preference game). If one adopts the view that workers might have some intrinsic upper bound such that even higher wages are judged as unnecessary or even unreasonable by them, and that this bound is likely to decline in the degree of reciprocal inclination, then the result seems to be qualitatively independent of the particular reciprocity model and hence, sufficiently general.

Part III

Experiments: Heterogeneity of Preferences

Chapter 4

Inconsistent People? An Experiment on the Impact of Social Preferences Across Games

4.1 Introduction

Theories of other-regarding preferences often implicitly suggest that it is possible to capture individual behavior in social settings by a few parameters and that these parameters are quasi preferences. A typical model defines utility as $u_i \equiv u_i(x_i, x_{-i}, \varphi_i, \gamma_i, \cdot)$, i.e. utility is a function of own and others' payoffs and some parameters scaling the impact of e.g. inequality or the desire to be reciprocal (Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Falk and Fischbacher, 2006; Fehr and Schmidt, 1999; Levine, 1998). While these theories do well in explaining patterns of behavior especially with respect to deviations from money maximization, it appears to be an open question whether or not it is possible to capture individual behavior by such models. The concern is driven by the fact that the theories are often context-free and suggest that the respective parameters are given and fixed for an individual just like preferences. For example, a person with a strong aversion against inequality should behave accordingly irrespective of the game being played. The alternative hypotheses are that social preferences are tailor-made, i.e. behavior depends on factors outside the models but in a systematic way, or that behavior across games is random in large parts. If the latter hypothesis holds and behavior is not consistent across games, then any form of inference on the individual level, relevant for example in

business relations or in the job-market, seems inadequate and even aggregate cross game inferences, for example relevant in politics, seem to stand on shaky ground. The aim and scope of this work is to test whether or not social preferences have a consistent impact on the individual level but across games and given a highly controlled experimental setting.

Despite studies regarding the general and aggregated impact of social preferences across games, see e.g. Camerer (2003) or Fehr and Schmidt (2006), and despite studies regarding a comparison of different pro-social motives, see e.g. Charness and Rabin (2002) or Engelmann and Strobel (2004), the question of individual consistency of social preferences has attained relatively little attention so far, at least if one goes beyond a comparison of just two games or decision environments. In an early work, Andreoni and Miller (2002) tested whether or not individuals behave consistently with respect to axioms of revealed preferences in several different versions of the dictator game. Their work is extended by Fisman et al. (2007). Both studies suggest that behavior across different versions of the same game is relatively consistent. Brosig et al. (2007) use a within-subject design to study the consistency of behavior with respect to theories of social preferences both with respect to variants of the same game as well as across games. They use a *take* and a *give*-version of the dictator game (each in four variants), and a prisoners' dilemma game (two variants). Individuals tend to behave consistent within the different variants of the same game but there is little consistency across games.[1] de Oliveira et al. (2008) estimate preferences for giving using a voluntary contribution mechanism and show a significant connection to actual donations in a field experiment. Due to the variation in their data, they conclude that while people *may have a stable preference to 'do the right thing'... observed behavior may vary by context because the perception of the 'right thing' would change (p.19).* The work that comes closest to the approach taken in this study is Blanco et al. (2011). The authors derive Fehr and Schmidt (1999) envy and guilt parameters from an ultimatum game (second-mover choice) and modified dictator games and use them to make predictions for an ultimatum game (first-mover behavior), sequential prisoner's dilemma game (both roles) and public goods game. The predictions are compared to actual behavior and while predictions at the aggregate level are mostly consistent with behavior, they are not at the individual level.

[1] Brosig et al. (2007) also tested the stability of preferences over time by repeating the experiment with the same subject pool twice with a one month delay in between. They observe an aggregate decline of pro-social behavior. Contrary to that, Volk et al. (2011) tested consistency over time in three repetitions of a public goods game and a 2.5 month delay in between. They find stable aggregate rates of cooperation but only partially, stability is due to individual consistency over time.

In this paper, the within-subject design literature is extended by three novelties. First, and most crucially, the analysis is not based on one-to-one comparisons of two games. The six games in the experiment, including a dictator, ultimatum, sequential prisoners' dilemma game and others, are quite diverse. Different motives concerning e.g. equity, efficiency, reciprocity (positive and negative), or the intentions behind an action, all well established in the literature, may thus be expected to trigger different individual reactions across different games and the results by Blanco et al. (2011) suggest that these *multiple behavioral forces* (p. 334) indeed cause great inconsistencies. However, while each force may cause inconsistencies under one-to-one comparisons, it is not clear that the inconsistencies carry over to an analysis which takes account off all decisions simultaneously. An individual could contribute to the fraction of inconsistent decisions in one one-to-one comparison but nevertheless show entirely consistent behavior across all other decisions (where the fraction of inconsistent decisions is then driven by other individuals). Therefore, the idea here is that individuals may have relatively robust character traits in general, but nevertheless deviate from those traits sometimes. The approach is more forgiving than one based on one-to-one comparisons but remains to be an individual level test for consistency. The second novelty is a strong focus on last mover decisions which reduces the potential for measuring inconsistencies in beliefs rather than in preferences. Finally, the third novelty is that the analysis is not based on any specific model of social preferences. Rather, via the size of the deviation from money maximization, it defines the impact of social preferences as being either low, medium, or high. The classification of each action yields individual preference profiles of the form $(\#low, \#med, \#high)$ which are then categorized from consistent to inconsistent. By that procedure, in some cases extended by survival analyses, results for different degrees of noise or several subclasses of games are easily accessible. Additionally, consistency is checked for type classifications such as conditional cooperators or unconditional defectors. Similar to Blanco et al. (2011), each player played both roles in each game without knowing his final payoff position (strategy method) and without any feedback upon outcomes during the experiment.

The paper is organized as follows. Section 4.2 describes the games, the experimental procedure, and provides an instrument check. In section 4.3, the classification of actions is summarized and compared to existing theories of social preferences. Section 4.4 provides a picture regarding the aggregated impact of other-regarding motives. Correlation analysis results are presented in section 4.5. Section 4.6 tests consistency with respect to all games as well as with respect to several subclasses of games. Section 4.7 checks for type consistency and section 4.8 discusses and concludes.

4.2 Experiment and Instrument Check

4.2.1 Games

This study rests on six games and seven decisions. In order to focus on belief-independent choices, sequential form games were selected. In order to check whether or not the design leads to a distortion of incentives, well-established games with a multitude of reference studies have been selected. In addition, several games are of similar strategic nature which allows testing whether or not consistency rates are different among different subclasses of games. Most of the games are very well known and therefore only very briefly described.[2]

In the *dictator game* (DG), player 1 splits one hundred tokens. Player 2 does not move. In the *ultimatum game* (UG), player 1 proposes a split of 100 tokens. If player 2 accepts, the proposed allocation is established. If player 2 rejects, both receive zero. Second movers had to state their minimal acceptable offer (MAO) and first movers were asked to state their belief regarding the MAO as well. In the *trust game* (TG), a variant of the game introduced by Berg et al. (1995), player 1 can send either $0, 30$, or 50 tokens to player 2. Player 2 receives three times the amount sent and can then return something to player 1. In the *third-party punishment game* (TPP), adopted from Fehr and Fischbacher (2004), two players play a DG as above and the dictator can allocate either $0, 20$, or 50 to player 2. A third player, endowed with 50 tokens, observes the outcome and can then assign deduction points to the dictator. A deduction point has a cost of 1 for player 3 but reduces the payoff for the dictator by 3. In the *gift-exchange game* (GE), player 1 can pay a wage $w \in \{0, 30, 50\}$ to player 2. After receiving the wage, player 2 decides on how much effort $e = 1, \ldots, 10$ with associated cost $c(e)$ to invest. Payoffs are $50 - w + 7.5e$ to the first mover and $50 + w - c(e)$ to the second mover. The associated costs $c(e)$ are:

e	1	2	3	4	5	6	7	8	9	10
$c(e)$	0	2	4	6	8	11	14	17	21	25

Table 4.1: Cost table GE

In the sequential *prisoners' dilemma game* (PD), player 1 can choose between three alternative actions $K1, K2, K3$ and player 2, after learning

[2] A full description of each game including the exact action sets is provided in the Appendix.

the choice of player 1, can also choose between three actions $K1, K2, K3$.[3] The associated payoffs are summarized in table 4.2.

		2^{nd} mover		
		K1	**K2**	**K3**
1^{st} mover	**K1**	(25,25)	(85,15)	(150,0)
	K2	(15,85)	(50,50)	(125,25)
	K3	(0,150)	(25,125)	(75,75)

Table 4.2: Payoff table PD

All participants made decisions in first and second mover roles such that each player found himself in 6 games, 11 positions, and made 18 choices.[4] Second mover choices where elicited using the strategy method. First mover choices in TG, GE, TPP (and PD) where restricted to three actions in order to avoid too many second mover decisions but also in order to avoid potential problems with intentionality if only in/out decisions were available. Choices were generally possible in decimal steps, i.e. $0, 10, 20, \ldots$ tokens, although the MAO and the number of deduction points in TPP had to be stated in steps of five tokens.

As pointed out, the analysis is based mainly on belief independent last mover decisions in order to avoid measuring inconsistencies in beliefs rather than preferences. Therefore, first mover choices in the TG, GE, PD, and TPP are not used in the analysis. However, the first mover choice in UG is included. On the one hand, this is due to the major reference aspect of that choice for theories of social preferences. On the other hand, it serves as a representative belief dependent choice in the baseline analysis.[5] The analysis of several subclasses of games does not include the ultimatum game. In the PD game, the possible responses to $K3$ are set to focal actions of either keeping everything, a return for the first mover that equals the amount at risk (25

[3]Each player had 3 actions to separate a low, medium, and high impact of social preferences.

[4]The 6 games and 11 positions are DG (1 position), UG (2 positions), TG (2 positions), GE (2 positions), PD (2 positions), and TPP (2 positions). DG contains 1 choice, UG 2 choices, TG 3 choices (one 1^{st} and two 2^{nd} mover, no choice for an investment of zero), and GE, PD, and TPP contain 4 choices (one 1^{st} and three 2^{nd} mover).

[5]The correlation between first and second mover behavior in UG (.44) is higher than the respective correlations in TG (.27), GE (.32), and PD (.32) (see the Appendix). The correlation in TPP (.56) is even higher but TPP is the only three player game and the only one dealing with indirect reciprocity. Therefore, TPP was not chosen as the representative first mover choice.

tokens), or an equal split. $K3$, in the sense of Cox et al. (2008), is *more generous than* all other available actions to the first mover (i.e. *most generous*). For matters of consistency, the analysis takes into account responses to the *most generous* first mover choices in TG and GE as well. For the TPP, where social preference based actions should be triggered by malevolent rather than benevolent choices, the reaction to the *most malevolent* action by the dictator is selected. Note that section 4.7 provides an analysis of the consistency of conditional cooperation which takes into account the reactions to alternative first mover choices as well. In total, the seven decisions included in the baseline analysis are DG, UG first mover, UG second mover, TG with full investment, GE with full wage, PD with $K3$ and TPP with a dictator keeping everything for himself. Table 4.3 summarizes all actions which are included in the baseline analysis.

Choice	**Abb.**	**Description**
dictator	DG	Dictator splits 100 tokens and keeps $100 - x$
ultimatum 1	UG1	1^{st} offers x out of 100 to 2^{nd} and states belief upon MAO
ultimatum 2	UG2	2^{nd} states minimal acceptable offer (MAO)
trust	TG	2^{nd} returns x out of 150 given 1^{st} sent 50 (highest investm.)
gift-exchange	GE	2^{nd} choses effort $\in \{1, ..., 10\}$ given 1^{st} paid 50 (highest wage)
prisoners' dilemma	PD	2^{nd} choses between $(\pi_{1^{st}}, \pi_{2^{nd}}) = (0, 150), (25, 125)$, or $(75, 75)$ given 1^{st} chose $K3$ (most cooperative)
third-party punishment	TPP	3^{rd} can deduce dictators pay in steps of 3 for a cost of 1 each given the dictator kept everything for himself (least generous)

Table 4.3: Choices and abbreviations (Abb.)

Five decisions (all except UG) are pure allocation tasks with the money maximizing prediction that the participant keeps the entire amount, or (in TPP) does not punish. A money maximizer in UG2 prefers any positive amount to zero and the first mover should send the lowest possible amount (if not gambling on acceptance in case of a zero offer). These predictions are unaffected by the fact that some games deal with positive (TG, GE, PD) and others with negative reciprocity (UG2, TPP). They are unaffected by potential efficiency gains (TG, GE, PD), or losses (TPP), and unaffected by a potential lack of intentionality (DG, TPP), or a difference between direct (UG2, TG, GE, PD), and indirect reciprocity (TPP).

The games have been scaled along several dimensions. For DG, UG, GE, and TPP, the average endowment per player is 50 tokens. For TG and PD, the average payoff is either 25 tokens (inefficient case) or 75 tokens (efficient

case). In all games, equity is established if a player gives up half of the potential gains from interaction. In six cases this means giving up half the endowment or allocation amount. In GE, equity is reached if the second mover incurs costs of 25 tokens which is a quarter of his total wealth when making his choice but it is half of what was paid to him by the first mover. All games with possible efficiency gains yield a payoff of 75 tokens for each player if equity is established.

4.2.2 Experimental Procedure

The experiment was conducted at the experimental lab of the University of Hamburg. Participants were students from various disciplines recruited using ORSEE, see Greiner (2004). 9 sessions with a total of 206 participants were run. The experiment took 66 minutes on average including instructions.

Upon arrival, participants randomly selected an envelope containing an id-code. Then the general instructions (see Appendix) were read and subjects could ask questions. Upon entering their id, the first decision was presented to them. The first screen always contained a description of the situation and two control questions. Subjects had to answer the control questions correctly in order to move on to the decision. If an answer was wrong, a separate screen appeared, containing a hint to answer the question correctly. The decision screens contained the descriptions of the situations and a list (one for each possible first mover choice) with all available actions from which players selected their choices. Once a decision task was finished, the next situation was presented. Participants did not receive any feedback on outcomes before the experiment was finished.

The order of decisions was partially set. The first bloc contained all choices relevant for the later analysis (see section 4.2.1). The order of decisions was random with the exception that the TPP choice was presented last. In an online based pre-test, it turned out that this decision is judged as the most complex one. The second bloc contained, in random order, the missing first mover decisions (TG, PD, GE, TPP) needed for payment. First mover choices were allocated to the second bloc just in case concentration suffered during the experiment.

The experiment was split up into two treatments. In treatment one (T1, 5 sessions, 118 participants), only one decision was relevant for payment whereas in treatment two (T2, 4 sessions, 88 participants), two decisions were relevant for payment. The second treatment was established in order to check for hints that a large number of games with only one decision paid distorts the incentives. In both cases, players were randomly matched in pairs once all participants had finished all decisions. Then, each pair was assigned

to one game and participants were assigned to either first or last mover role. Finally, payments were derived based on the previously stated choices. The exchange rate was 100 tokens = 10 Euro. Additionally, each player received a show-up fee of Euro 5. Total earnings were registered together with the id-code and participants could pick up their earnings at a separate office. Average earnings were 10.26 Euro (about 13.5 US dollar) per participant in T1 and 13.84 Euro (about 18.5 US dollar) in T2.

Participants were informed about the number of choices and about the fact that each particular situation occurs only once. They knew that only one game will be paid and that this game is selected at random (two games in T2). Further, they knew that matching partners stay anonymous and that the experimenter cannot match id-codes with names.

4.2.3 Treatment and Order Effects, Comparison to Previous Results

In order to check whether there are any hints to biased observations or diluted incentives, several tests have been carried out. At first, 1 data set was eliminated because it was incomplete. Second, Kruskal-Wallis tests were used to test for significant session, treatment, order and time (players need to make their decisions) effects. All 18 choices were tested and the null hypothesis of no significant difference is rejected between one and three times for each series of tests. Given the large number of choices and given that the differences do not seem to follow any systematic pattern, it is concluded that there are no relevant session, treatment, order or time effects.[6] Third, one or two key variables were selected for each game and compared to predictions based on the available literature.[7] Tests on significant differences with respect to distributions or fractions were done using chi-squared goodness-of-fit tests. Tests on mean differences were done using two-tailed t-tests. The threshold $p-$value was set to .1.

In the DG, the mean given is 28% and the distribution of choices in the intervals $[0, 10], [11, 30], [31, 50], [51, 100]$ is $(38\%, 19\%, 34\%, 9\%)$. Refereneces were calculated based on studies listed in Camerer (2003) taking into account all studies with entries in the respective intervals and studies which do not refer to e.g. communication possibilites. The mean is higher compared to the reference studies (24%, although there are observations with means as high as 28%) but the distribution is not significantly different. The fraction

[6] With respect to time effects, the data set was separated between those who need below and above median total time.

[7] In some cases, reference results had to be estimated from figures.

of people offering exactly zero is 31.2% in this experiment and the reference value is 33.5%.

With respect to UG1, the mean given is 41% and the distributions of offers is (same intervals as above) $(5\%, 21\%, 69\%, 5\%)$. Both the mean and the distribution are not significantly different to the predictions (mean 41%, distribution $(6\%, 15\%, 71\%, 8\%)$) which were again based on Camerer (2003). With respect to UG2, the mean MAO is 34% and not significantly different to the prediction (33%). The distribution of MAO's in intervals $[0, 10]$, $[11, 30]$, $[31, 50]$ is $(12\%, 24\%, 64\%)$ and upward shifted compared to the references $(35\%, 10\%, 55\%)$. Predictions for UG2 were based on Harrison and McCabe (1996); Larrick and Blount (1997); Weber et al. (2004) who report first round results obtained with the strategy method. However, the references do not take into account that players play both roles. Oxoby and McLeish (2004) have players play both roles and use the strategy method and they obtain a distribution of MAOs of $(10\%, 13\%, 77\%)$ which has an even higher fraction of players in the upper interval.

For the TG, the average return is -5% and the fraction of players with a positive return is 45%. Both results are not significantly different from predictions (-2%, 49%) which were derived from Berg et al. (1995); Bolle (1998); Burks et al. (2003); McCabe et al. (2003); Ortmann et al. (2000).[8] In the GE, the average effort is 37% and not significantly different from the prediction (36%). The fraction of players who choose the minimal effort is 49% but hardly comparable since it varies between 21% and 64% in the reference studies. References in this case were Fehr et al. (1997); Fehr and Gächter (2002); Fehr et al. (2007). With respect to the PD, 39% of players fully cooperate given that the first mover cooperates and this fraction is not significantly different from the references Blanco et al. (2011); Brosig et al. (2007); Clarc and Sefton (2001).

In the TPP, the average of assigned deductions is 5.8 and for each ten-token reduction in dictator giving (starting with a 50:50 split), punishment increases by 2 points. Both results are significantly different to the predictions (7 points, 2.8 points increase) but in this case, the only reference is Fehr and Fischbacher (2004) who have exactly 22 third-party observations compared to the 205 here.

In total, there are no hints to systematically biased observations.

[8] Of course, incomparable treatments as for example the unintentional treatment in McCabe et al. (2003) were not used.

4.3 Classification of Choices, Theoretical Predictions

Large parts of the upcoming analysis are based on classified actions rather than choices directly. The classification is based on the idea that social preferences explain deviations from money maximization and that the larger the impact of other regarding motives, the larger the deviation. More precisely, all actions are classified as made under an either low, medium, or high impact of other regarding motives depending on the cost an agent incurs compared to money maximizing behavior. By the classification in three categories, minor differences in behavior, e.g. switches between giving 40% and 50%, or between money maximization and giving 10%, are innocuous for consistency. At the same time, larger differences, e.g. giving between 20% and 50%, are nevertheless recognized as differences. Additionally, the classification makes behavior across games more comparable.

In order to classify actions, it is assumed that each game has a reasonable range of actions ranging from money maximization to the implementation of equity in payoffs. Given that the respective parameters are sufficiently high, equity is a focal prediction, not only in light of theories of inequity aversion, e.g. Bolton and Ockenfels (2000) and Fehr and Schmidt (1999), but also in light of theories of reciprocity which use equity as a reference to evaluate the kindness of the other players' action, e.g. Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006), or Rabin (1993). Let $c_i^{e,\Gamma}$ be the cost agent i has to incur in game Γ in order to establish equity, i.e. the amount of own payoff that must be given up in order to reach equity in payoffs. For $\Gamma = \{DG, UG2, TG, GE, PD, TPP\}$, one obtains $c_i^{e,\Gamma} = \{50, 50, 75, 25, 75, 25\}$.[9] For UG1, $c_i^{e,UG1} = 50 - belief$ since offering less than the belief upon the MAO is not a money-maximizing choice.

Now let $c_i^{a,\Gamma}$ be the cost an agent incurs by choice a in game Γ. In DG, TG, and PD, this is the amount transferred to the other person. In UG1, it is the amount offered minus the belief upon the MAO of the second mover, i.e. the true cost in excess of money maximization. In UG2, it is the amount the agent is willing to sacrifice in case the offer is too low. i.e. the MAO. In GE, it is the cost associated with effort e according to table 4.1 in section 4.2.1 and in TPP, it is the number of deduction points since each point has a cost of one.

[9] In the TPP, the reasonable range is defined up to the point where the third party establishes equity between himself and the dictator. This reference follows from e.g. Falk and Fischbacher (2006). Note that Bolton and Ockenfels (2000) would never predict punishment since the endowment by the third party matches the population average payoff.

Definition 1 *The relative impact costs incurred by agent i in game Γ are*

$$\gamma_i^{\Gamma} = \frac{c_i^{a,\Gamma}}{c_i^{e,\Gamma}} \tag{4.1}$$

For the classification, γ_i^{Γ} is compared to threshold values $\bar{\gamma}_k, k = 1, 2, 3$. In order to obtain a broader picture and to analyze how consistency depends on the thresholds, three different thresholds $\bar{\gamma}_k = \left\{\frac{1}{5}, \frac{1}{4}, \frac{1}{3}\right\}$ were chosen.

Definition 2 *The impact of other-regarding motives on individual behavior by player i in game Γ is*

- *low*, *if* $\gamma_i^{\Gamma} < \bar{\gamma}_k$
- *medium*, *if* $\bar{\gamma}_k \leq \gamma_i^{\Gamma} < 1 - \bar{\gamma}_k$
- *high*, *if* $\gamma_i^{\Gamma} \geq 1 - \bar{\gamma}_k$

The classification separates the cost interval associated with the reasonable range of actions into three subsets. With a threshold of $\bar{\gamma} = \frac{1}{4}$, for example, a low impact is assigned whenever the actual costs fall into the lower quartile of that cost interval. A high impact is assigned if costs are in the upper quartile and a medium impact is assigned for the remaining actions.

It needs to be pointed out how the classifications relate to certain theories of social preferences. Based on $\bar{\gamma} = \frac{1}{4}$, table 4.4 provides thresholds for an either low or high impact for the theories by Fehr and Schmidt (1999) (FS), Falk and Fischbacher (2006) (FF), and Charness and Rabin (2002) (CR, conceptual model).[10] Empty entries occur if either the threshold is unclear (belief dependency, UG1 low), not applicable (point estimation of one parameter needed, TPP for FS and UG2, TPP for CR), or unknown (equilibrium not calculated, GE for FF).[11]

The classification in categories low, medium, high is not strictly in line with the exemplary models. For example, a player with FS-utility and $\beta_i = .4$ would choose the money maximizing action in DG (low impact) but equity in GE (high impact). Similar, a player with FF-utility and $\rho_i = 1$ would choose a medium impact action in UG2 but a low impact action in TG. On the other hand, classical money maximizing players will have parameter values of, or close to, zero, inducing a low impact in all cases. At the other

[10]The FF-reciprocity equilibrium for the TPP game is available from the author. For the reciprocity equilibrium in the trust game, see Schliffke (2012b).

[11]Note that FF allows for point predictions while FS and CR, in general, do not. W.r.t. the similarity between FS and CR, see also Brosig et al. (2007). Finally, FS and CR predict either strict money maximization or strict equity, at least with respect to the dictator, trust, and prisoners' dilemma game.

	FS (1999)			**FF (2006)**			**CR (2002)**		
	par.	low	high	par.	low	high	par.	low	high
DG	β_i	$< .50$	$> .50$	$\rho_i\epsilon_i$	< 1.3	> 4.0	ρ	$< .50$	$> .50$
UG1	β_i		$> .50$	ρ_i	$< 1.3^b$	$> 4.0^b$	ρ		$> .50$
UG2	α_i	$< .17$	> 1.5	ρ_i	< 0.2	> 2.4	σ, θ		
TG	β_i	$< .50$	$> .50$	ρ_i	< 1.3	> 4.0	ρ	$< .50$	$> .50$
GE	β_i	$< .21^a$	$> .35^a$	ρ_i			ρ	$< .21^a$	$> .35^a$
PD	β_i	$< .50$	$> .50$	ρ_i	< 1.3	> 4.0	ρ	$< .50$	$> .50$
TPP	α_i, β_i			$\rho_i\epsilon_i$	< 0.9	> 2.7	σ, θ		

[a] exact threshold slightly above/below . [b] given a belief upon the MAO which equals zero.

Table 4.4: Classification thresholds in social-preference models

end of the scale, there is always one dominating threshold inducing equity in many games. For example, any FS-type (CR) with $\beta_i > .5$ ($\rho > .5$) should show a high impact of social preferences in DG, UG1, TG, GE, and PD. Similar, any FF-type with $\rho_i > 4$ should show a high impact in UG1, UG2, TG, PD, and potentially DG and TPP. The latter two cases would require that the other-regarding part of FF-utility is not downscaled too much by the potential impact of intentions captured via $\epsilon_i \in [0, 1]$.

The upcoming analysis will typically define consistency via a majority of choices which fall into the same category. Therefore, the questionable cases pointed out above must not necessarily affect consistency. In addition, the cases point to the fact that the classification here is purely choice based and treats each game equally. The latter is not true for the theoretical models where the thresholds follow from the interaction of the mathematical formulation of the model and the specific game.[12] Given the above, the link between formal models of social preferences and the costs approach taken here seems reasonably close.

4.3.1 Data Selection

All results presented in the forthcoming sections are based on a subset of 160 individuals. 34 individuals were removed since, on at least one occasion, they gave more than half the pie in DG, UG1, or TG.[13] 11 individuals were

[12] For example, the author is not aware of any fact why there should be different parameters in GE and TG. Given the FS-model and given both games with their respective specification, the different thresholds follow, but, a priori, it is not clear that they need to follow in a model of social preferences.

[13] Players who assign more than 25 deduction points in TPP were not removed. Assigning 25 points establishes equity between them and the dictator but assigning 35 points establishes equity between the dictator and the recipient. While this might not be in line

removed since they state a belief upon the MAO which is higher than their actual giving.

The decision to remove data sets is driven by the concern that any analysis of consistency should try to keep the amount of noise in the data as low as possible. Since the elimination of data sets is questionable, however, all results where also calculated taking the eliminated choices into account (classified as made under a high impact) and footnotes are used throughout the analysis to indicate whether or not the result are significantly different.

4.4 General Impact

The classification of actions provides a picture regarding the general impact of other-regarding motives. For each threshold $\bar{\gamma}_k$, 7 choices by 160 individuals (1120 decisions) were classified according to the definition in section 4.3. Figure 4.1 plots the aggregate impact distributions for all three threshold values $\bar{\gamma}_k$.

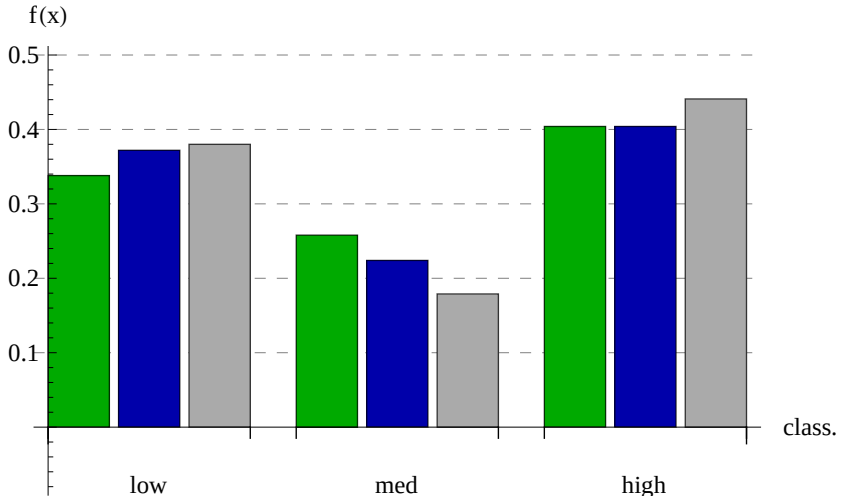


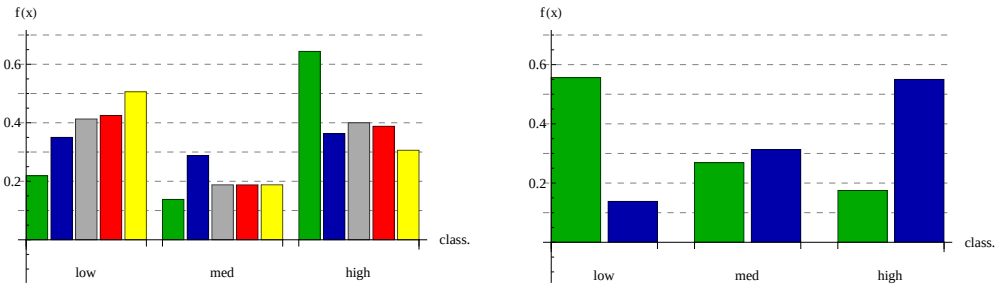
Figure 4.1: Aggregate impact distributions for $\bar{\gamma}_k = \left\{\frac{1}{5}, \frac{1}{4}, \frac{1}{3}\right\}$ (left to right)

On average, 36.4% of all choices are made under a low impact while 22.0% are associated with a medium and 41.6% with a high impact of other-regarding motives. The u-shaped distribution of classifications supports the assumption that besides money maximization, equity is focal for the games used in the experiment. Increasing the threshold $\bar{\gamma}_k$ has an expected effect as

with e.g. Falk and Fischbacher (2006), it can be in line with e.g. Fehr and Schmidt (1999).

the fraction showing a medium impact is strictly decreasing for an increasing threshold (increasing the threshold shrinks the cost interval associated with a medium impact). A χ^2-test on homogeneity rejects the null that the three distributions are realizations of one underlying true distribution ($\chi^2 = 21.75, d.f. = 4, p = .0002$). Similarity cannot be rejected for the pairwise comparison fifth vs. quarter-based ($\chi^2 = 4.49, d.f. = 2, p = .1060$), but for quarter vs. third-based ($\chi^2 = 7.73, d.f. = 2, p = .0210$), and fifth vs. third-based ($\chi^2 = 20.81, d.f. = 2, p = .0000$). Besides the differences in distributions, the fraction of the population showing a low impact is never larger than 38%, i.e. more than 60% of all choices are clear deviations from money maximization.

In order to illustrate differences across games, figure 4.2 plots the impact distributions for all seven games based on a quarter-based classification.[14]



(a) TG, GE, PD, DG, TPP (left to right) (b) UG1 and UG2 (left to right)

Figure 4.2: Game specific impact distributions, $\bar{\gamma}_k = \frac{1}{4}$

Figure 4.2 (a) contains all games with a u-shaped impact distribution. The average distribution for those five games is 38.3% low, 19.8% medium, and 42.0% high impact which is not significantly different from the overall quarter-based distribution ($\chi^2 = 1.99, d.f. = 2, p = .3704$). However, similarity within that subgroup is clearly rejected ($\chi^2 = 54.73, d.f. = 8, p = .0000$). Similarity between DG, GE, PD, and TPP is rejected at the 5%-level of significance ($\chi^2 = 12.60, d.f. = 6, p = .0498$) but the distributions DG, GE, and PD are not significantly different ($\chi^2 = 6.44, d.f. = 4, p = .1686$). The downward shift in the impact distribution of TPP is likely due to the fact that third-party punishment is an act of indirect reciprocity compared to TG, GE, and PD which refer to direct reciprocity. The upward shift in TG,

[14]For each game and each classification, impact distributions were also calculated based on the complete data set. χ^2-tests on homogeneity indicate no significant differences to the game specific distribution based on the reduced data set. On average, the change in the absolute fraction of one category is 2 percentage points.

on the other hand, is possibly due to a kind of endowment effect. If the first mover in TG did not transfer anything (hypothetical, the analyzed choice is the one where 50 tokens were invested), the second-mover in TG would have been left with zero tokens. In difference to that, the second mover in PD can always assure at least 25 tokens for himself and in GE, the second mover is endowed with 50 tokens. Thus, the hypothetical alternative outcome is worst in TG which might trigger the most positive response.[15]

Figure 4.2 (b) contains the impact distribution for UG1 and UG2 which both do not follow a u-shape. While the UG1 distribution is clearly downward shifted, the UG2 distribution is clearly upward shifted. The difference between the two is highly significant ($\chi^2 = 72.00, d.f. = 2, p = .0000$) and each distribution is also significantly different from all other five distributions ($p < .015$ in all cases). The downward shift in UG1 is clearly due to the fact that beliefs enter the classification. In fact, 48.8% of all individuals offer exactly their belief which corresponds 87.6% of all actions classified as low. The definition of impact in excess of the belief can hide an intrinsic concern to give which would otherwise suffice for a medium or high impact. The following analyses, however, will contain hints toward whether or not consistency rates depend on this specific definition.[16] Contrary, the upward shift in UG2 might be due to an experimental effect. It has been pointed out in section 4.2.3 that the distribution of MAO's (not the average) is upward shifted compared to other experiments but that such an upward shift has been observed experiments where players have to play both roles and know that in advance.[17]

[15]Note that neither e.g. Fehr and Schmidt (1999) nor e.g. Falk and Fischbacher (2006) would predict such an effect. Both theories essentially rely on current subgames and not on hypothetical alternative outcomes. The endowment effect could be in line with Cox et al. (2008) if one extends the definition of *more generous than* to cross-game situations. The difference between the maximal obtainable payoffs for the second mover and for all possible actions by the first mover is largest in TG (150 tokens). However, while this difference is lower in PD (125 tokens) and GE (50 tokens), the approach rather suggest a difference between PD and GE (difference of differences equals 75 tokens) than between TG and PD (diff. of diff. equals 25 tokens). Note further that framing is not likely to be responsible for the significant differences since it would rather suggest a difference between PD on the one (abstract, payoff-table framing) and TG and GE on the other hand (amount sent framing).

[16]Both the correlation analysis and the consistency analysis will also report results were the belief dependency is ignored and actions are taken as made, respectively, are classified as if the belief would equal zero. In the latter case, the impact distribution of UG1 would turn into a highly upward shifted one with 3.1% showing a low impact, 20.6% showing a medium impact, 76.3% showing a high impact. See section 4.5 and 4.6.

[17]Alternatively, the upward shift in UG2 might be due to the fact that players find themselves in the disadvantageous position. Fehr and Schmidt (1999), for example, assume

4.5 Correlations Across Games

If individual behavior across games is relatively consistent, then correlations should be significantly positive. Accordingly, table 4.5 reports the estimated Spearman rank correlation coefficients calculated on the basis of unclassified actions. For the ultimatum game first mover choice, both giving in excess of the belief (UG1), as it is used to classify actions, and absolute giving (UG0), i.e. excess giving under a hypothetical belief of zero, is taken into account.[18]

	DG	**UG0**	**UG1**	**UG2**	**TG**	**GE**	**PD**
DG							
UG0	.343***						
UG1	.168**	.102					
UG2	.244***	.442***	−.209***				
TG	.429***	.252***	.190**	.127			
GE	.338***	.119	.180**	.043	.474***		
PD	.394***	.165**	.267***	.070	.439***	.414***	
TPP	.377***	.086	.153*	.020	.393***	.360***	.405***

***,** ,* significantly different from zero at the 1, 5, 10%-level.

Table 4.5: Spearman rank correlations across games

First, with the exception of the correlation between UG2 and UG1, all correlations are positive and in most cases significantly different from zero. This supports the general idea that behavior is consistent in the way that giving in one situation is linked to giving in another situation. Second, DG behavior is positively correlated to all other games in a significant way. This is likely due to the fact that all deviations from money maximization must contain some element of other-regarding motives and that DG behavior may just reflect this empathy component of any pro-social behavior. Third, three of the five highest coefficients are found between the different variants of the

that the envy parameter α_i, scaling the impact of a payoff difference to the disadvantage of a player, is at least as high as the guilt parameter β_i which scales the impact of a positive payoff difference. While this does not necessarily imply that the impact as it is defined here is higher, it seems to be qualitatively in line. However, players are is the disadvantage position in TPP as well and TPP exhibits a strongly downward shifted distribution. A consistent interpretation would then require that the effect of being in the disadvantageous position is highly overcompensated by the fact that the TPP refers to indirect reciprocity.

[18] Based on the whole data set, the obtained picture is similar. All correlation have the same sign and, on average, coefficients by the complete data set are smaller by .0461. Some correlations loose one level of significance (DG/UG2, DG/TPP, UG0/PD). All other changes occur with respect to UG1, but this decision is highly affected by the elimination of those individuals who state a belief which is smaller than actual giving.

trust game (TG, GE, PD). This suggests that the more similar the games become, the more consistent becomes behavior. The latter explanation provides a link to Andreoni and Miller (2002) and Fisman et al. (2007) who report comparably high consistency rates based on several versions of the same game (DG in both cases). Fourth, the correlations between TPP on the one and TG, GE, and PD on the other hand are all positive, significant and comparably high. This suggest a relatively close link between direct and indirect reciprocity and in addition, a positive correlation between positive (TG, GE, PD) and negative reciprocity (TPP).

Very specific results are obtained for the ultimatum game. Similar to Andreoni et al. (2003); Bellemare et al. (2008); Blanco et al. (2011), there is a comparably strong and significant correlation between UG0 and UG2. One explanation would be an underlying positive correlation between positive and negative reciprocity (see above). In that case, however, UG2 behavior should also be correlated in a significant way to TG, GE, PD, and UG1, which it is not. In fact, all those correlations are insignificant which suggest a non-relationship between the willingness to give or punish and the amount people demand in UG2. A better explanation for the correlation between UG0 and UG2 is the consensus effect, i.e. those with a high MAO expect others to have a high MAO as well and therefore give, see Dawes (1989), and Mullen et al. (1985). Indeed, the correlation between UG2 and the belief first movers reported regarding their belief upon the MAO of the second movers can be calculated to be as high as .5231 and highly significant. The consensus effect can also explain the negative correlation between UG1 and UG2. The higher the own MAO and thus the belief, the higher is giving in first mover position. The higher the belief-induced giving, however, the less room remains for any excess giving leading to an overall negative relationship.

In addition to the insignificant correlations between UG2 on the one, and TG, GE, PD, and TPP on the other hand, also the correlations between UG0 and UG1 and the latter games remain of comparably low size. Both facts together suggest that the bargaining-type ultimatum game is judged in a very different manner compared to the reciprocity based games but also to the dictator game. With respect to the reciprocity games, the focus on excess giving (UG1) somewhat corrects this effect as the, on average, higher coefficients and levels of significance suggest.

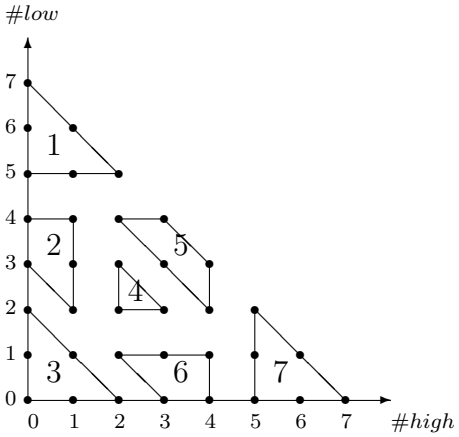
Finally, the results are quite similar to those found by Blanco et al. (2011) (BEN). BEN observe a correlation between UG first (UG0) and second mover behavior of .40 (.44 here, both highly significant), between PD second mover and DG behavior of .34 (.39 here, both highly significant), and insignificant correlations between PD first mover behavior and DG and UG0 (here insignificant as well, see the Appendix). UG2 behavior is insignificantly cor-

related to all other decisions except UG0 in both papers. The correlation between PD first and second mover behavior is .43 for BEN and .32 here, both highly significant. Differences are observed for the correlation between UG0 and DG (.13 and insignificant BEN, .34 and highly significant here) and between UG0 and PD second mover behavior (.49 and significant at 1% level BEN, .17 and significant at 5% level here).

4.6 Consistency of Individual Preference Profiles

4.6.1 All Games

The classification of all actions yields individual preference profiles of the form $(\#low, \#med, \#high)$. One of the central ideas of this work is that people may behave consistent in general, that is, they show a similar impact of social preferences in a clear majority of decisions while occasional deviations might occur. Accordingly, individual preference profiles can be classified as consistent or inconsistent depending on the individual impact distributions. Figure 4.3 summarizes the classifications.



Nr.	Category
1	consistent low
2	rel. consistent low/med
3	consistent medium
4	inconsistent
5	rel. inconsistent
6	rel. consistent med/high
7	consistent high

Figure 4.3: Preference profiles and classification

A preference profile is defined as consistent if five out of seven decisions are associated with the same impact of social preferences. In addition, a profile with at least six decisions in neighboring categories, which is nevertheless not consistent, is called relatively consistent. The profiles most dispersed, in which no type of impact is assigned to a majority of actions are called inconsistent. When choices are made under an either low or high impact

with no clear majority for one side, a profile is called relatively inconsistent. The definitions are chosen to strike a reasonable balance between what should be demanded from theories trying to capture individual behavior with single parameters and the acceptance of occasional deviations.

Table 4.6 summarizes the fractions of the population that fall into each category and for each threshold $\bar{\gamma}_k$.[19] Table 4.6 further contains the expected fractions for each category under the hypothesis that individuals randomly chose one of the possible three types of impact on each decision.[20]

Category	**fifth-based**		**quarter-based**		**third-based**		**random**	
con. low	15.6%		18.1%		20.6%		4.5%	
con. med	7.5%	46.9%	3.1%	45.0%	1.3%	48.8%	4.5%	13.5%
con. high	23.8%		23.8%		26.9%		4.5%	
rel. con. l/m	14.4%	27.5%	14.4%	26.3%	7.5%	17.5%	19.2%	38.4%
rel. con. m/h	13.1%		11.9%		10.0%		19.2%	
rel. inc.	17.4%	25.6%	20.0%	28.8%	23.1%	33.8%	19.2%	48.0%
inconsistent	8.1%		8.8%		10.6%		28.8%	

Table 4.6: Consistency in individual choices, all games

On the aggregated level - consistent, relatively consistent, and inconsistent plus relatively inconsistent - the distributions for the three different thresholds are not significantly different ($\chi^2 = 6.07, d.f. = 4, p = .1940$). Shrinking the range of the cost interval associated with a medium impact has an expected effect as the fractions of the category consistent medium and both relatively consistent categories shrink when moving from a fifth-based to a third-based classification.

The first observation is that all three distributions are different from the distribution which would realize if people chose the impact randomly ($d.f. = 2, p < .0000$ for all cases). The result is in line with the correlation analysis and supports the impression that behavior is not random.

[19] On the aggregated level - consistent, relatively consistent, and inconsistent plus relatively inconsistent - the distributions for the thresholds are not significantly different from the respective distributions obtained for the whole data set (χ^2-test on homogeneity, lowest $p = .7416$). For the reduced data set, the analysis was also carried out with UG0 choices under a hypothetical belief of zero. While this leads to an increase in fractions of categories based on a high impact, it leads to a comparable drop in fractions of categories based on a low impact. The aggregate distributions, and thus the overall rates of consistency, are insignificantly different to the ones presented in table 4.6 (lowest $p = .8997$).

[20] With a probability of 1/3 for each impact, the probability for any specific impact vector is $\left(\frac{1}{3}\right)^7$. The probability for any specific profile is then given by (7 over $\#low$) times ($7 - \#low$ over $\#med$) times $\left(\frac{1}{3}\right)^7$.

Besides the observation that behavior is non-random, it is not impressingly consistent either. For none of the thresholds, the fraction of individuals with a consistent preference profile exceeds 50%. Somewhat more reasonable fractions require that the categories consistent and relatively consistent are summarized but in that case, differences in giving of more than 100% are still accepted as consistent behavior (e.g. giving 20% or 50% in DG).[21]

Another possibility to increase consistency is an increase in the accepted rate of noise. The above results require that the same impact is shown not in a simple (4/7), but in a clear majority (5/7). Table 4.7 reports the fractions of the population which show the same impact on four, five, six, or all occasions based on a quarter-based classification.

	#4	#5	#6	#7
quarter-b.	34.4%	24.4%	16.3%	4.4%
cumulated	79.4%	45.0%	20.6%	4.4%

Table 4.7: Noise and consistency

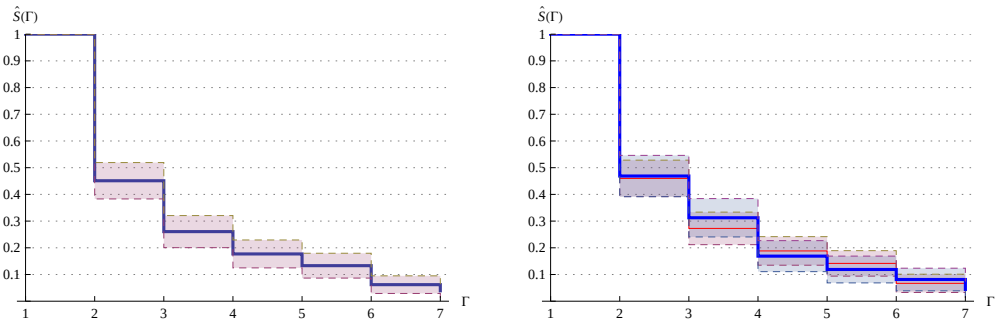
In fact, if consistency is defined via a majority of decisions under the same impact, 80% of the population behave in a consistent manner. Compared to the 45% consistent profiles with a noise rate of 29% (2/7), this is an increase of 76%. The other side of the coin is a consistency rate which is more than halved to only 20%, if the accepted rate of noise is decreased to 14% (1/7). If consistency would be defined in an absolute way, it almost vanishes as the fraction of people showing the same impact in every decision is as low as 4%. It is of course a matter of viewpoint, but if reasonable rates of consistency require an accepted noise rate of more than 40% (3/7), one may conclude that the results do not support the hypothesis of consistency in behavior.

In order to confirm and extend the descriptive results above, a survival analysis was run. Suppose that each player reveals his individual impact of social preferences in the first of all seven decisions. Then the individual impact *survives* whenever the same impact is shown in the second decision, third decision, and so forth. The Kaplan-Meier estimator $S(\Gamma)$ is the fraction of the population showing the same individual impact in Γ as in the initial decision. Since there is no natural order of games, the thick survival function

[21] Note here that consistency rates around 50% are also observed by Blanco et al. (2011) who estimate Fehr and Schmidt (1999) parameters, use them to predict behavior in other choices and then compare the predictions to actual behavior. The model fails in about half the cases for ultimatum game first mover choices and prisoners' dilemma first mover choices. It fails in about a third of prisoners' dilemma second-mover choices but in about two thirds of all cases for public good contributions.

in figure 4.4 (a) as well as 95% confidence bounds (Greenwoods formula) are calculated based on the mean survival frequencies for all 5040 possible permutations of seven games.[22] Truly consistent behavior would require a survival function which remains stable at 1.

A survival analysis further allows to test whether the results are likely due to experimental effects. One possibility is a *boredom-effect* where people start out with relatively stable behavior and then *do something else* (just for the fun of it?). The corresponding survival function should have a concave shape, i.e. be relatively flat at the beginning and steep for later positions. Another possibility is a sort of *learning-effect* in the sense that people realize the similarity of the tasks which then leads to similarity in behavior. In that case, the corresponding survival function should have a convex shape and the relative decline in $S(\Gamma)$ should flatten out for later decisions. For both cases, the survival function needs to be calculated based on the factual order of decisions for each individual in the experiment. The thick survival function in figure 4.4 (b) is based on exactly that order. The second, thin, survival function in figure 4.4 (b) is based on all possible 720 permutations such that the TPP game is always in last position (TPP was not rotated in the experiment). Both a boredom and a learning effect suggest significant differences between the order-based survival function and the permutation-based survival function.



(a) Impact survival, quarter-based

(b) Getting-bored-effect?

Figure 4.4: Survival analysis results

Figure 4.4 (a) confirms the impression regarding the relative inconsistency of behavior as the survival function is certainly not constant at 1. Secondly, and based on the fact that all possible permutations are taken into account,

[22]The function is calculated based on the quarter-based classification. The survival functions for the other classifications, and the one for the whole data set and a quarter-based classification, are not significantly different. Survival functions in figure 4.4 (b) are also based on the quarter-based classification.

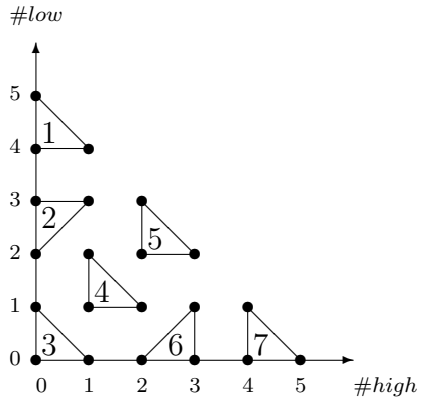
the initial drop is equivalent to the average probability that the same person shows a different impact of other-regarding motives in two randomly selected games. This probability is as high as 54.9%. Third, the survival function is monotonically decreasing, which is of course expected, but not flattening out towards the end. The relative drop between positions 5 and 6 is as high as -53.4% and between positions 6 and 7 as high as -37.1%. This yields a hint that drawing conclusions regarding consistency based on a comparison of possibly just two games might be misleading, since such a conclusion would require that behavior stabilizes after an initial drop.

Figure 4.4 (b) shows an order based survival function which is certainly not concave. In addition, it is convex, but, similar to the one in (a), not flattening out towards the end. The relative drop between positions 5 and 6 is as high as -31.6% and the relative drop between positions 6 and 7 is as high as -46.1%. Together with the insignificant differences to the reference line, there is no hint that observed inconsistency is due to some kind of *boredom-effect* or *learning-effect*.

4.6.2 Subclasses of Games

So far, the results do not support the idea of consistency. One possible reason is that some of the games are judged in a fundamentally different manner compared to others. The correlation analyses as well as the aggregate impact distributions certainly suggest such a fundamental difference between the allocation tasks on the one and the ultimatum game on the other hand. Ceteris paribus, one should thus expect an increase in consistency if the analysis is restricted to those choices where the second mover has full discretion over the pie and is directly responsible for the final allocation of payments. The correlation analysis further suggest an increase in consistency if the analysis is restricted to those games which explicitly deal with positive reciprocity as the correlations among those games are increased compared to others. In addition, the aggregate impact distributions suggest that consistency might be increased if the analysis is restricted to DG, GE, and PD as those are the only games with insignificantly different aggregated impact distributions. Accordingly, this section takes a look at all three subcategories of games motivated above. Figure 4.5 summarizes the applied categorizations for the five allocations tasks, figure 4.5 (a), as well as the reciprocity games and DG, GE, and PD, figure 4.5 (b). In figure 4.5 (b), profile refers to $(\#low, \#med, \#high)$. The categorization is discussed below.

The categorizations follow the same idea as the categorization for seven games. Consistency is assigned if more than a simple majority of choices is made under the same impact. For the allocations tasks, it is thus required



(a) Five games

Profile	Category
$(3,0,0)$	hom. low
$(2,1,0)$	rel. hom. l/m
$(1,2,0)$	rel. hom. l/m
$(0,3,0)$	hom. med.
$(1,1,1)$	het.
$(1,0,2)$	rel. het.
$(2,0,1)$	rel. het.
$(0,1,2)$	rel. hom. m/h
$(0,2,1)$	rel. hom. m/h
$(0,0,3)$	hom. high

(b) Three games

Figure 4.5: Categorizations for subclasses of games

that four out of five choices are made under the same impact. This corresponds to an accepted rate of noise of 20%, i.e. less than the accepted rate of 29% in section 4.6.1. The results should thus be compared not only to the ones reported in table 4.6, but also to the ones for seven games and a noise rate of 14% as reported in table 4.7. For the three-game categorization, the idea of a clear majority even implies an accepted noise rate of zero. The reference to judge the expected ceteris paribus effect of an increase in consistency is thus the consistency rate for seven games and a noise rate of zero, again reported in table 4.7. The decrease in the accepted rate of noise will, at least partially, compensate the expected increase in consistency due to an increase in the similarity of decisions. From the viewpoint of consistency, however, such compensation appears reasonable since, after all, it would be strange to accept large rates of noise in games which are very similar. Table 4.8 summarizes the fractions of the population that fall into each category based on a quarter-based classification.[23]

For all three subclasses of games, the obtained distributions are significantly different from the respective random distributions ($p < .0000$ in all cases).[24]

[23]For each of the three subclasses, χ^2-tests on homogeneity indicate no significant differences between the distributions for the three different classification thresholds $\bar{\gamma}_k$ (lowest $p = .2607$). For each subclass and each threshold, there are also no significant differences to results based on the complete data set (lowest $p = .1442$).

[24]For the allocations tasks, the expected distribution under random choice of impact is 16.6% consistent profiles, 32.9% relatively consistent profiles, and 53.5% inconsistent and relatively inconsistent profiles. For the remaining two subgroups with three games involved, the expected distribution under random choice of impact is 11.1% consistent pro-

Category	**Allocation**		**Reciprocity**		**DG, GE, PD**	
con. low	19.4%		15.6%		14.4%	
con. med	1.3%	45.6%	2.5%	39.4%	1.9%	33.8%
con. high	25.0%		21.3%		17.5%	
rel. con. l/m	14.4%	27.5%	9.4%	30.6%	20.6%	33.8%
rel. con. m/h	13.1%		21.3%		13.1%	
rel. inc.	16.9%	26.9%	19.4%	30.0%	16.9%	32.5%
inconsistent	10.0%		10.6%		15.6%	

Table 4.8: Consistency in individual choices, subclasses

The expected ceteris paribus effect of an increase in consistency for more and more similar games is indeed present in the data. For the allocation tasks, the obtained fraction of consistent profiles is almost unchanged compared to the fraction under all seven games while the accepted level of noise has been decreases by 30% from (2/7) to (1/5). Alternatively, starting out from 20.6% consistent profiles for seven games and a noise rate of (1/7), the fraction of consistent profiles increases by 121.4% to 45.6% while the accepted rate of noise is increased by only 40% from (1/7) to (1/5). For the reciprocity games and DG, GE, PD, the fractions of consistent profiles of more than one third are clearly much higher than the 4.4% observed for all seven games and a noise rate of zero.

If one accepts the countervailing procedure of lowering the accepted rate of noise, however, the overall picture remains very similar to the one obtained for all seven games. The differences to the distribution which was obtained for all seven games are insignificant for all three subclasses (lowest $p = .1087$ for DG, GE, PD) and for none of the subclasses, the fraction of people with a consistent preference profile exceeds 50%.[25] Hence, one may conclude again that while behavior is definitely non-random, it is not impressingly consistent either. It must be pointed out that the results are again heavily dependent on the accepted rate of noise. If the requirement for consistency is lowered to a simple majority of choices made under the same impact, the fraction of consistent profiles jumps up to 85.0% for the allocation tasks, 89.4% for the reciprocity tasks and 84.4% for DG, GE, PD.

Finally, the subclass DG, GE, PD reveals a specialty. Recall that for those

files, 44.4% relatively consistent profiles, and 44.4% inconsistent and relatively inconsistent profiles. Results are for χ^2-tests on homogeneity.

[25]Note that for the allocations tasks, the profile $(1, 3, 1)$ is classified as inconsistent which is not in line with the other categorization since one classification applies for a majority of actions. However, only 3.1% of the population reveal such a profile such that the overall picture does not drastically change if one applies a different categorization for that profile.

games, the aggregate impact distributions are insignificantly different from each other. However, only one third of this observation can be assigned to an underlying consistency in behavior since the fraction of consistent profiles is not higher than 33.8% for that subclass. The aggregate observation is thus based on a canceling-out effect between those who switch from a low to a high impact, and others who switch from a high to low impact, and so forth.

4.7 Type-Consistency

So far, a convincing picture regarding the consistency of behavior was neither found with respect to the overall sample of games nor for several subclasses of games. This section concludes the analysis by classifying individuals along conditional cooperation in order to test the hypothesis of type-consistency.

Conditional cooperation received most attention with respect to the public goods game, see e.g. Fischbacher et al. (2001); Frey and Meier (2004); Gächter (2007); Kocher et al. (2008), and refers to the fact that the own contribution is an increasing function of others' actual or believed giving. A relatively stable finding is that about half of all individuals are conditional cooperators and roughly 20% to 30% are free-riders. For the present experiment, conditional cooperation may matter in TG, GE, PD, and TPP.

In order to classify individuals, the definition of the relative impact costs γ_i^{Γ} (see section 4.3) is extended to $\gamma_i^{\Gamma,\phi}$ with $\phi = \{l, b, m\}$ such that "l" stands for the least benevolent first-mover choice, "m" for the most benevolent first-mover choice, and "b" for a benevolent choice. For the TPP, benevolent is replaced by malevolent.

Definition 3 *A conditional coordinator is characterized by*

$$\gamma_i^{\Gamma,l} < \gamma_i^{\Gamma,m} \tag{4.2}$$

and

$$\Delta\gamma_i^{\Gamma,l} \geq 0 \tag{4.3}$$

That is, the reaction to the most benevolent offer must be associated with higher relative costs than the reaction to the least benevolent offer and reactions must be weakly monotone in the generosity of offers.[26] Unconditional

[26]For $\Gamma = \{TG, GE, PD, TPP\}$ one obtains $c_i^{e,\Gamma,l} = \{-, 0, 0, 0\}$ and $c_i^{e,\Gamma,b} = \{45, 11, 35, 15\}$ in addition to the specified values in reaction to the most benevolent offer in section 4.3. Note that for TG and an investment of zero, the choice set for the second mover is empty. In GE with a wage of 30, the cost associated with the action that minimizes the payoff difference between both players was chosen as the reference.

behavior is present if $\gamma_i^{\Gamma,m} = \gamma_i^{\Gamma,b} = \gamma_i^{\Gamma,l}$ respectively $\Delta\gamma_i^{\Gamma,\phi} = 0$ for all ϕ.[27] Table 4.9 summarizes the classification.

Type	**TG**	**GE**	**PD**	**TPP**	**avg.**
conditional					
-cooperators	53.8%	43.8%	48.1%	35.0%	45.2%
unconditional					
-cooperators	13.8%[a]	02.5%[b]	01.3%[c]	00.0%	04.4%
-defectors	15.6%	25.0%	34.4%	36.9%	28.0%
unclassified	16.9%	28.8%	16.3%	28.1%	22.5%

[a] joint fraction of those who offer 50% and 33%.

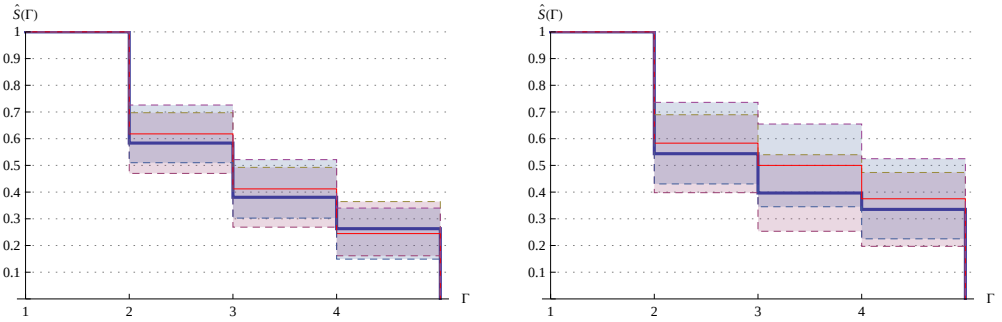
[b] 4 people strictly choosing $e = 10$. [c] 2 people choosing strictly $K3$.

Table 4.9: Type classifications

On average, the findings are in line with the stylized classification obtained from public goods experiments. Only very few people show unconditional cooperative behavior. Therefore, the consistency analysis is restricted to those who are either conditional cooperators or unconditional defectors. Consistency is checked by using Kaplan-Meier estimators similar to section 4.6. Since the TPP deals with indirect, negative reciprocity instead of direct, positive reciprocity, figure 4.6 plots both the weighted average survival functions for all possible permutations (thick line) and survival functions for all permutations given that TPP is in last position (thin line). Survival functions are calculated conditional on being a conditional cooperator or unconditional defector in the first game of a sequence. Confidence bounds are based on a 95%-interval.

If conditional cooperation or unconditional defection is a stable character trait, then the survival functions should be relative constant and close to one. Obviously, this is not the case. On average, 41.6% of all people who reveal to be a conditional cooperator in one game are not a conditional cooperator in some other game. Adding another game reduces that fraction by another 20.3% and only about a quarter (26.3%) of all players who revealed to be conditionally cooperative in one game are indeed conditional cooperators in general. For unconditional defectors, the survival function is almost identical except for the smaller drop between positions 3 and 4 such that about a third (33.5%) of all players who revealed to be unconditionally defective in one game are indeed unconditional defectors in general.

[27] With the exception of incurred costs of zero, this condition is hard to satisfy due to different scaling and limited action sets. Therefore, the table also reports fractions of those who unconditionally choose e.g. the highest effort.



(a) Conditional Cooperators (b) Unconditional Defectors

Figure 4.6: Type-consistency

The difference in tails is due to the relatively high fraction of unconditional defectors, and relatively low fraction of conditional cooperators, in TPP. Note, however, that both survival functions are not significantly different from each other. In addition, the different nature of the TPP game does have an impact on the survival functions but for none of the functions and no position, the impact is statistically significant.

4.8 Discussion and Conclusion

The aim of this paper was to shed new light on the consistency of other-regarding behavior. Subject in the experiment played six different games and their choices were categorized dependent on the strength of the deviation from money maximization and relative to equity in payoffs. The impact of other-regarding motives could either be low, of medium strength, or high, and for each individual, a preference profile of the form $(\#low, \#medium, \#high)$ was derived. A preference profile was defined as consistent whenever a clear majority of choices, in difference to a simple majority, fell into the same category. The preference profile based analysis was applied to all games in the experiment as well as to several subclasses of games. Additionally, it was extended by correlation and survival analysis. A survival analysis was additionally applied to analyze the consistency of conditional cooperation and unconditional defection.

The first result is that the data contains strong evidence in favor of the existence of social preferences as such. From more than 1100 analyzed choices, more than 60% are non-trivial deviations from the money maximizing predictions. The correlations between choices are positive and in many cases highly significant. Further, the consistency analysis reveals that behavior

is definitely non-random and the result holds irrespective of whether one looks at all decisions, only allocation tasks, only decisions related to positive reciprocity, or only at decisions in games with insignificantly different distributions regarding the impact of other-regarding motives.

The second result is, however, that the overall support for consistency is low. While correlations are generally positive and significant, they are of medium strength at most. In the analysis of preference profiles, the rate of consistent profiles drops to below 50% as soon as the requirement is set to a clear majority of choices in the same category. While there is a positive ceteris paribus effect on consistency rates if games become more similar, convincing rates cannot be obtained. This is true even among, for example, the trust, sequential prisoners' dilemma, and gift-exchange game which are essentially three different versions of the same decision problem. Only 40% of the subjects make sufficiently similar choices across all three decisions. The survival analysis reveals that the average likelihood that the same person shows a different impact of other-regarding motives in some other randomly selected choice is greater than 50%. Finally, the inconsistency of behavior extends into the domain of conditional cooperation or unconditional defection, i.e. into the domain of more underlying character traits.

As they stand, the results cast heavy doubt on the possibility to capture individual behavior by the assignment of one or two parameters scaling the impact of specific psychological motives such as inequality aversion or reciprocity. It is widely acknowledged that social preferences are heterogeneous in a horizontal way, i.e. while some players do not care at all about other-regarding motives, others do, and those who do, do so to different degrees. The findings here add a strong vertical heterogeneity within each individual. This is best illustrated by the comparison of the dictator, gift-exchange, and prisoners' dilemma game. Those three games share an identical aggregate impact distribution but individual consistency across all three games is found for only a third of the population.

The results are similar to those obtained in the closely related study by Blanco et al. (2011). It has been pointed out, however, that the approach here differs in several ways from their approach. The two crucial distinctions are a strong focus on last mover choices and a simultaneous consideration of all choices in difference to one-to-one comparisons across games. I have argued that both factors can have a positive impact on measured consistency. The former because it removes potential biases due to changing beliefs across games, and the latter because it is somewhat more forgiving than an one-to-one comparison because different underlying motives may cause deviating behavior in one other game but behavior may be consistent across many other games at the same time. Both arguments have been made on an a

priori basis but ex post, I do not observe increased rates of consistency.

In addition, one may argue that the findings are not truly surprising given other results, for example on inconsistencies found comparing lab versus field settings. Levitt and List (2007) survey several articles on such comparisons and argue that *there is only weak evidence of cross-situational consistency of behavior* (p. 160). Camerer (2011), in a reply, surveys an even richer set of studies and points out that *between-situation correlations of trait-like behavior are often not much larger than* .30. However, he also questions whether this is actually a low correlation and argues that *there may be a practical upper limit on how much lab-field consistency we expect within people* (both p. 32) and draws a much more positive overall conclusion regarding the experimental method and the generalizability of experimental results. With respect to a potential practical upper limit on correlations, note that this need not transfer to laboratory environments. The set of potential exogenous factors that influence behavior, e.g. the weather, sounds, other people around, etc., is much richer in a field setting compared to a lab setting. Nevertheless, the observed correlation remain in the domain between .2 to .4 here as well.

What are the reasons? One obvious one would be that preferences are simply not stable. This would be a dead-end argument but it is also rejected by the data because while consistency rates are low, behavior is nevertheless far from random. A related argument would be that preferences are, or at least behavior is, stochastic, as in a recent approach by Oppenheimer et al. (2011). An explanation based of stochastic behavior keeps up the possibility to systemize and thereby understand behavior. However, such a systemization is much more demanding compared to a situation where preferences can be assumed to be more or less stable. It would require, for example, sorting out mixed strategies and occasional but otherwise random errors, and it would require that each individual is observed in each decision environment several times which causes other problems (see below). Obviously, the explanation put forward by Blanco et al. (2011) that behavior is guided by a multiplicity of norms which may be uncorrelated within an individual can explain the results here as well. It has been argued that such different motives could be restricted to cause only occasional deviation from otherwise consistent behavior but as pointed out above, the deviations are not restricted to be occasional.

In addition, the specific experimental design could be unsuitable to some degree. In his book *Lack of Character*, Doris (2002) summarizes several experiments conducted in psychology and points to the massive impact seemingly unimportant facts like finding a dime, or simply telling people that they are in a hurry, can have on cooperative behavior. Such obvious manipulations of the decision environments were not part of this experiment.

There are, however, fundamental design features which could make a difference: Choices are elicited using the strategy method, players play both roles, and they do not know their final payoff position. All three aspects are discussed in a recent article by Brandts and Charness (2011). The main topic of the article is a comparison of results obtained under the strategy versus the direct response method and the authors found more studies in favor of no significant difference between both ways of eliciting choices. There are counter examples though, especially in the domain of punishment decisions. Punishment is more common under the direct response method and this fact is mainly attributed to stronger emotions triggered by a *hot* treatment. If anything, however, short run emotions are probably likely to increase the variance in behavior such that in reverse, the strategy method is the more conservative approach to study consistency. With respect to playing both roles, a common argument is that *..., they* (subjects) *are likely to undertake greater self-reflection right from the beginning*, see Brosig et al. (2003) (p. 85). The experimental evidence on whether there is an effect of playing both roles or not seems very mixed but in any case, greater *self-reflection* should imply choices more in line with underlying preferences as it, again, reduces the impact of short run emotions. Finally, Iriberri and Rey-Biel (2011) provide an example that role uncertainty makes a difference in modified dictator games where selfish behavior is much more common if a player knows that he will be in the dictators position. More evidence seems to be lacking. Note, however, that role uncertainty is a common feature of all games in the experiment and without any hint that role uncertainty matters in some but not in other games, it at least yields no definite reason to assume that it has an impact on consistency.

A more fundamental possibility is that the results are driven by issues regarding the experimental method in general. Subjects in the experiment face each position is each game once and although they have to answer control questions correctly, the decision situations remain abstract and players may lack a sound knowledge which preferences should guide their behavior in such unfamiliar situations. This leads to the approach that preferences are ad hoc *constructed* in the laboratory, see e.g. Lichtenstein and Slovic (2006); Borgloh et al. (2010). Constructed preferences will be inconsistent if the factors that influence the construction a diverse across games even though the outside conditions are highly controlled. One potential factor that might affect the construction of preferences are scrutiny or demand effects, see e.g. Levitt and List (2007); Zizzo (2010). Demand effects occur, for example, if the subjects in the lab try to meet whatever they believe that the experimenter wants to observe. This will cause inconsistencies whenever the believed expected behavior is different across games which itself could simply arise do to the

complexity of finding out what another person believes. Another factor are framing effects. While all decisions were framed relatively neutral and in addition, each possible choice was listed together with the resulting payoffs, which arguably makes things very comparable, framing effects can, of course, not be excluded. However, Tversky and Kahneman (1987) argued more than 30 years ago that large framing effects question the rational choice approach as an explanation of actual behavior in a fundamental way and consequently called for a descriptive analysis of choice.

Overall, the most likely explanation for the findings is a combination of both a multiplicity of norms together with potentially stochastic behavior and the presence of weaknesses of the experimental method. While it has been argued that e.g. stochastic behavior is hard to systemize, and while e.g. framing effects can never be ruled out entirely, other driving forces behind the results allow for further investigation. For example, the suggestion by Blanco et al. (2011) of a multiplicity of norms which are uncorrelated within individuals can be tested against an approach of game specific social preferences where similar motives work across different games but potentially imply different levels of choices. Evidence in favor of this hypothesis is presented in Schliffke (2012a). A way to reduce problems of unfamiliar situations but also demand effects is the repeated measurement of decisions. For the domain of consumer goods, e.g. Hoeffler and Ariely (1999) argue that repeated measurement is a way of obtaining more *stable* preferences. For social preferences, the problem seems more severe though. For example, contributions to public goods typically decline under repeated measurement but this is not necessarily due to changed, or stabilized, preferences but can be the result of subsequent belief updating, see Fischbacher and Gächter (2010). Mechanisms that prohibit the decline in contributions, like e.g. punishing options, cause other problems because they alter the strategic incentives. Repeated measurement in the domain of social preferences may thus cause problems of sorting out different effects, but if a method is suitable to sort out different effects in a rigorous way, than it is the experimental method. One should add that even if the experimental method causes some problems with respect to measuring the consistency of behavior, this does not imply that it is questionable, for example, with respect to the discovery of treatment effects.

Finally, and returning to the fact that the data yields weak support for individual consistency, Schotter (2006) points out that a theory which is very strong is also very likely to be wrong. For example, the assumption that utility is purely dependent on monetary outcomes is very strong. The fact that it turned out wrong in many experimental tests was one of the triggers for the development of theories of other-regarding motives. In this case, the strong but wrong prediction led to a great improvement in the understanding

of behavioral patterns. However, Schotter (2006) further argues that a theory should not only be right, it should be right for the right reason. Given the large degree of observed vertical heterogeneity, one may doubt that the assignment of specific psychological motives is the right explanation. Of course, e.g. Fehr and Schmidt (2010) argue with respect to their theory that its objective was to structure data and to obtain testable hypotheses, i.e. not to predict individual behavior. This is in line with Schotter (2006) who defends the rational choice approach as the only approach which yields clear cut theoretical predictions. While the predictions may be falsified, their failure and especially the reasons for the failure are the source of new knowledge and thus fundamental for the progress in our understanding of behavior.

Chapter 5

Game Specific Social Preferences: Different Types and a Canceling-Out Effect

5.1 Introduction

The goal of this work is to answer the question of whether or not social preferences are likely to be game specific both at the aggregate and the individual level. The question originates in the evolution of preferences for reciprocity as formalized by Falk and Fischbacher (2006) in the dictator, ultimatum and trust game. Berninghaus et al. (2007) have shown that an infinitely large reciprocal inclination implying equal splits is stable in the ultimatum game while money maximization evolves in the dictator game. In contrast, Schliffke (2010) established that a medium reciprocal inclination and neither equal splits nor money-maximization is stable in the trust game. What remains unclear is what happens if all three games are jointly analyzed. Since the Falk and Fischbacher (2006) model contains an extra parameter capturing the lack of intentionality of *acceptance* in the dictator game, money maximization remains to be the prediction for the dictator game. On the other hand, both trust game returns as well as ultimatum game minimal acceptable offers are intentional acts and with the default assumption of one parameter scaling the impact of reciprocity, similar behavior should arise in both games. However, if one allows reciprocity parameters to be game specific instead, the predictions for each game studied in isolation would carry over to the joint *game-of-life*. The evolutionary predictions are thus assumption dependent and one goal of this work is to establish which assumption, a universal parameter or game specific ones, is correct.

The deeper motivation to tackle the question is the fact that game specificity has a great impact on the a priori possibility of cross game inference both at the aggregate and the individual level. At the aggregate, the incorporation of social preference theory might help to design more efficient contracts, it may improve charitable giving campaigns, or help to improve tax schedules or social security systems. On the individual level, cross game inference and the consistency of behavior is a direct issue of understanding human behavior. Without, or with little game specificity, it seems accessible to capture human behavior reasonably well not necessarily by one, but potentially by a few models representing some key motivational factors like inequality aversion or reciprocity. Given that, it also seems possible to derive relatively stable parameter distributions which can be used for accurate cross game inference. With, and especially with high degrees of specificity, both tasks become much more complex if not even impossible.

The evidence both with respect to aggregate inference and individual consistency is mixed so far.[1] Blanco et al. (2011), for example, study the consistency of behavior both at the aggregate and the individual level across several well known games as the dictator, ultimatum, public good and sequential prisoners' dilemma game. While they generally find stable aggregate outcomes, individual consistency is found for only about half of their subjects. Andreoni and Miller (2002) and Fisman et al. (2007) obtain better results with respect to consistency but they do not study behavior across different games but across different versions of the same game. On the other hand, Schliffke (2012b) studies individual consistency across six different games and finds consistency rates similar to those by Blanco et al. (2011). Another result pointing to low individual consistency is Camerer (2011) who surveys several within-subject studies comparing field and laboratory behavior, i.e. different domains, and points out that correlations become hardly larger than .3. With respect to aggregate inference, e.g. Fehr and Schmidt (2004) and Fehr et al. (2005, 2007) obtain predictions for three different contract games based on a simplified parameter distribution for the Fehr and Schmidt (1999) model of inequality aversion, which is derived from ultimatum game behavior, and argue that these predictions fit the data quite well. However, Binmore and Shaked (2010) heavily criticize the authors and argue that the predictions are not at all in line with the data. As it turns out, the different views are

[1] In the following discussion, I focus on results either from explicit within-subject designed experiments which study the consistency of behavior, or on results which make explicit cross game predictions. There exists a quite large additional literature comparing e.g. different motives behind behavior. For example Charness and Rabin (2002) and Engelmann and Strobel (2004) compare inequity and efficiency, or Falk et al. (2008), McCabe et al. (2003) and Stanca (2010) analyze the impact of intentions.

based on a different understanding of how close the link between theory and behavior needs to be. Fehr and Schmidt (2010) argue that their interpretation is based on the fact that average behavior is close to the prediction of average behavior while Binmore and Shaked (2010) argue, among other things, that the predictions are based on parameter distributions which are not found in the data. Game specificity is potentially in line with both individual and aggregate results and both sides in the argumentation. Specificity can certainly cause low consistency rates.[2] Therefore, it can also explain the failure of cross game inference if the focus is on distributions of behavior. At the same time, specificity can be in line with stable aggregate outcomes or accurate predictions. Ideally, in the sense of reliability, aggregate stability is the result of individual consistency but this is of course not a necessity. Given the above, the link between individual and aggregate behavior is one focus of this paper.

The method to search for game specificity is an across-game, within-subject, multiple rounds experiment with information transmission. Participants play both roles in a dictator, ultimatum and trust game. All three games are repeated several times and appear in random order. Crucially, players receive information regarding the past average behavior of their current matching partner. The analysis then checks whether or not behavior across games is significantly different from each other, whether or not preference parameters and distributions of preference parameters are significantly different from each other, and whether or not aggregate observations are the result of consistent individual behavior.

The approach differs from previous works in several dimensions. It differs from the literature on consistency (see above) by the fact that players play multiple rounds. While the multiple rounds approach is inspired by the background in evolutionary game theory, its main advantage is that players receive the chance to reconsider, and potentially adapt their behavior both with respect to own success and with respect to others' behavior. Of course, this kind of learning behavior is likely to be present in the *real* world all the time but replicating it in the lab allows to disentangle to some degree inconsistency from specificity. Additionally, the replication may help to sort

[2]Game Specificity is not the same as inconsistency though. Specificity refers to the idea that behavior is systematically different across different games, or more generally, across different domains. True inconsistency, on the other hand, is different from that because it additionally removes any systematic behind differences across domains. In principle it is thus possible to capture specificity in a (complex) model, something that is not possible with true inconsistency. From a practical perspective, however, the differentiation is very blurry since models trying to capture large degrees of heterogeneity become intractable quite fast.

out experimental effects like an experimenter demand effect, see Zizzo (2010), or even the *construction of social preferences in the lab*, see Borgloh et al. (2010) and, on both effects, Levitt and List (2007). An across-game, within-subject design including the ultimatum and trust (and battle-of-the-sexes) game has been analyzed by Schotter and Sopher (2004, 2006, 2007). In their approach, games are also played for multiple rounds and there is some information transmission. However, the focus is on intergenerational advice and while each single player encounters all games, she does not play each game for multiple rounds but is replaced by another player after finishing each game and can advise the following subject on how to play the game. Thus, the setup does not provide subjects with information regarding their current matching partner and one of the consequences is, for example, that trust game returns are so low on average that first movers should not trust (which is in line with theoretical predictions from evolutionary games). Given that players do receive individual specific information, the approach here is also linked to experiments on the evolution in repeated games, see e.g. Dal Bó and Fréchette (2011) for a recent work including an extensive literature review. Crucially, however, subjects in this experiments cannot identify the other player, for example by the subject number, but simply observe the average past behavior of the *other* player. This rules out the application of repeated game strategies like tit-for-tat, see Axelrod (1984), such that the link to the repeated game literature is rather weak as well.

The paper proceeds as follows. Section 5.2 discusses the theoretical background with respect to Falk and Fischbacher (2006) preferences for reciprocity and the evolutionary predictions. Section 5.3 deals with the experimental design. The results are first presented with respect to choices in section 5.4 and then in terms of preference parameters in section 5.5. Additional results with respect to signaling issues and correlations across games and within types are presented in section 5.6. Section 5.7 discusses and concludes.

5.2 Theoretical Background and Predictions

5.2.1 Reciprocal utility and one-shot predictions

The work is founded in the reciprocity model by Falk and Fischbacher (2006) who define utility by[3]:

$$u_i \equiv \pi_i(\cdot) + \rho_i \varphi_j(\cdot)\sigma_i(\cdot) \tag{5.1}$$

i.e. as an additive connection of material rewards $\pi_i(\cdot)$ and, in this case, reciprocal utility $\varphi_j(\cdot)\sigma_i(\cdot)$ scaled by the reciprocity parameter $\rho_i \in \mathbb{R}_+$. The model is founded in psychological game theory, see Geanakoplos et al. (1989), and all calculations of expected payoffs are based on a second-order belief structure, i.e. both the belief by a player regarding the other players' action and the belief about the other players' belief regarding the own action are taken into account.

Whenever $\rho_i = 0$, the model collapses to the baseline assumption $u_i = \pi_i$. Otherwise, reciprocal utility matters and is obtained by the interaction of the kindness term, $\varphi_j(\cdot)$, and the reciprocation term, $\sigma_i(\cdot)$. The kindness term is composed of two factors. The first factor is the difference in expected payoffs and by definition, the other player is perceived as kind (unkind) whenever a player expects a higher (lower) material gain compared to the other player, i.e. equity is the reference standard to evaluate kindness. The second factor is an evaluation of the intentionality of the other players' expected action. By definition, an action is perceived as intentional whenever the other player has a *true* alternative. A true alternative is given when the co-player can be more or less kind without moving from the domain of kind (unkind) actions into the domain of unkind (kind) actions. Consequently, an action is unintentional whenever such an alternative does not exist. Crucially, whenever an action is unintentional, a second parameter kicks in which is the so called outcome concern parameter $\epsilon_i \in [0, 1]$. The outcome concern parameter is multiplicatively connected to reciprocal utility such that any $\epsilon_i < 1$ implies that overall reciprocal utility is downscaled whenever the other player acts unintentionally.[4] The other term $\sigma_i(\cdot)$ is the reciprocation term and defined via the comparison of the ex ante expected payoff to the other player and the ex post payoff once player i chooses a particular action. It is thus the

[3]The treatment here is kept relatively short for reasons of space. The interested reader is kindly asked to search the original sources for deeper insights.

[4]In more detail: the kindness term is itself a multiplicative connection defined by $\varphi_j \equiv \vartheta_j \Delta_i$ where Δ_i is the outcome term defined over the payoff difference $\pi_i - \pi_j$ and a term ϑ_j capturing intentionality. If the co-player acts intentionally, one obtains $\vartheta_j = 1$, but if the action is unintentional $\vartheta_j = \epsilon_i$.

impact of i choice on j's payoff relative to expectations. Overall, player i will obtain positive reciprocal utility either if he reacts in kind to a kind co-player (kindness and reciprocation term positive) or if he punishes unkind co-players (both terms negative). Whether and to which degree this extra utility affects actual choice is then dependent on the parameters ρ_i and ϵ_i and their interaction.

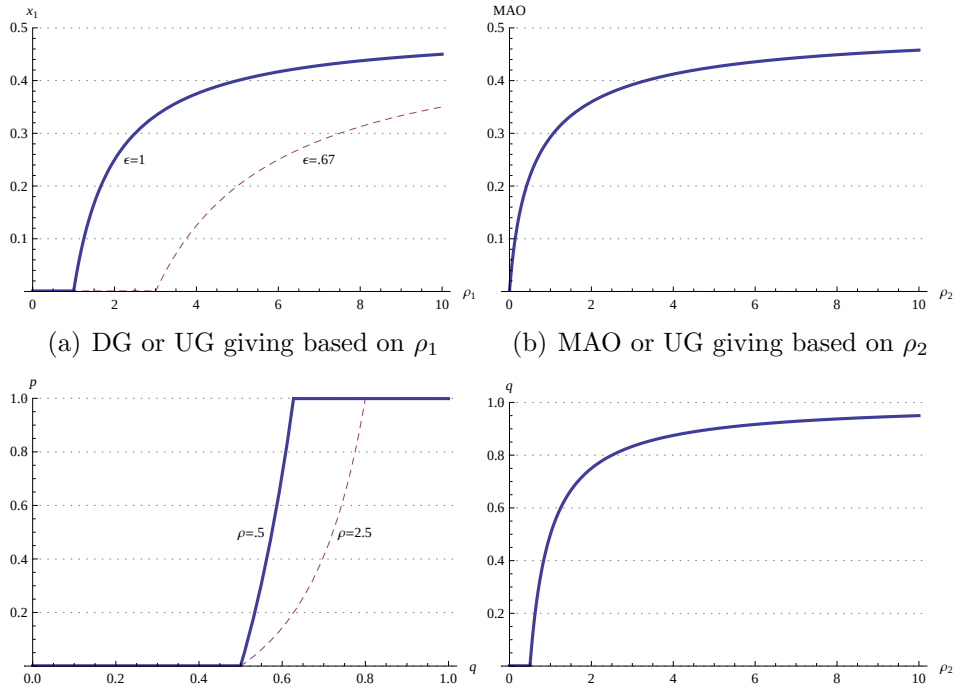
For the current study, the above model is applied to a dictator, an ultimatum, and a trust game. In the dictator game (DG), the first mover can split a pie of size 1 by allocation some x_1 to the receiver who has no choice. Payoffs are $(\pi_1, \pi_2) = (1 - x_1, x_1)$. In the ultimatum game (UG), the first mover can split a pie of size 1 but the second mover can reject the offer. If the offer is not rejected, payoffs are those of the dictator game. Otherwise, both receive 0. In the trust game (TG), the first mover can either trust or not trust. Not trusting ends the game and yields some default payoff for both players (e.g. $\frac{1}{3}$ each). If trust is shown, the second mover can reward, in which case both players receive an equal amount which is greater than the default payoff (e.g. $\frac{2}{3}$ each), or he can exploit, in which case the first mover receives zero and the second mover some payoff which is larger than the payoffs under reward (e.g. 1). In this version of the trust game, choice is defined as the probability to trust, p, and the probability to reward, q. Table 5.1 summarizes the equilibrium predictions for each game of the Falk and Fischbacher (2006) model. For the ultimatum game, the expression x_2^* is the minimal acceptable offer (MAO). For the trust game, the expressions are those for the example payoffs provided above.

Γ	1^{st} mover	2^{nd} mover
DG	$x_1^* = \begin{cases} 0 & if \quad \epsilon_1\rho_1 < 1 \\ \frac{1}{2} - \frac{1}{2\epsilon_1\rho_1} & if \quad \epsilon_1\rho_1 \geq 1 \end{cases}$	–
UG	$x_1^* = \max\left[x_2^*, \frac{1}{2} - \frac{1}{2\rho_1}\right]$	$x_2^* = \frac{1+3\rho_2-\sqrt{1+6\rho_2+\rho_2^2}}{4\rho_2}$
TG	$p^* = \begin{cases} 0 & if \quad q < \frac{1}{2} \\ \min\left[1, \frac{2q-1}{\rho_1(2-3q+q^2)}\right] & if \quad q \geq \frac{1}{2} \end{cases}$	$q^* = \begin{cases} 0 & if \quad \rho_2 < \frac{1}{2} \\ 1 - \frac{1}{2\rho_2} & if \quad \rho_2 \geq \frac{1}{2} \end{cases}$

Table 5.1: Equilibrium behavior in DG, UG, TG (one-shot)

A close examination of each expression reveals that the outcome approaches equal splits whenever $\rho_i \to \infty$. This follows as equity is the reference standard to evaluate kindness. One consequence is, for example, that ultimatum game first mover behavior can be guided by an own inclination to behave reciprocally (the term containing ρ_1), but whenever the second mover

has a MAO in excess of the own inclination to offer, the first mover will match that MAO. The outcome concern parameter ϵ_i matters in the dictator game only. *Acceptance* is clearly non-intentional since the receiver simply has no choice. Contrary, MAOs and trust game returns are intentional acts since a player can vary both choices without necessarily moving the co-player from the advantageous to the disadvantageous position or the other way around. Figure 5.1 additionally illustrates the functional forms given in table 5.1 for different parameter values.



(a) DG or UG giving based on ρ_1 (b) MAO or UG giving based on ρ_2

(c) TG giving based on q and ρ_1 (d) TG return based on ρ_2

Figure 5.1: Equilibrium behavior illustration for DG, UG, TG

Dictator giving, ultimatum game behavior in both roles, and trust game returns are qualitatively similar in the sense that beyond a threshold, choice is strictly increasing in the reciprocal inclination ρ_i but at a diminishing rate.[5]

[5]The diminishing rate of increase follow from the formalization of the model in terms of second-order beliefs. In equilibrium, beliefs must be correct. Then, an increased second-order belief implies a lower expected payoff difference and thus less kindness, for example. In addition, the possibility to surprise the recipient by actual giving in excess of the belief is ceteris paribus lower for a higher second-order belief as well. Taken together, and with both terms multiplicatively combined such that squared terms appear, reciprocal utility is decreasing in the second-order belief and higher and higher ρ_i values are necessary for an additional increase in choice.

A different shape is found for trust probabilities. Here, trust is never shown if the first mover expects a payoff which is lower than the default payoff (which happens for $q \leq .5$ in the example). Beyond the threshold, the probability is strictly increasing in q and reaches $p = 1$ for $q < 1$. However, p is decreasing in ρ_1, respectively the q-values needed to ensures $p = 1$ are increasing in ρ_1.

5.2.2 Assumption dependent evolutionary predictions

Given the one-shot predictions, Berninghaus et al. (2007) study the evolution of Falk and Fischbacher (2006) preferences in the dictator and ultimatum game. The analysis follows the indirect evolutionary approach, see e.g. Güth and Kliemt (1994), or Huck and Oechssler (1999), which assumes that each end-node of a game is evaluated in a dual fashion. First, the potential outcomes are subjectively evaluated by each individual who might or might not take other than pecuniary motives into account. The subjective evaluations, in this case guided by preferences for reciprocity, imply choices and outcomes. Second, the outcomes determine the evolutionary fitness of specific preferences and may lead to an increase or decrease of certain preferences over time depending on their relative fitness. A changing composition of the population will then feedback into the subjective evaluations by the subjects such that there is an indirect link between preferences and evolution and evolution and preferences. Schliffke (2010) uses the same approach with the same set of assumptions and obtains the corresponding prediction for the trust game. Table 5.2 summarizes the predictions.

Γ	parameter	choice	$\Delta = \pi_1 - \pi_2$
DG	$\rho_i\epsilon_i \to [0,1]$	$x_1 \to 0$	$\Delta \to 1$
UG	$\rho_i \to \infty$	$x_1, x_2 \to \frac{1}{2}$	$\Delta \to 0$
TG	$\rho_i \to 2.5$	$p \to 1, q \to .8$	$\Delta \to -.2$

Table 5.2: Evolutionary predictions, single games

Positive giving is never stable in the dictator game. A higher $\rho_i\epsilon_i$-product implies higher giving and accordingly lower payoffs. Additionally, since dictators do not (have to) condition their behavior on the second movers' preference parameter, there is no room for any support in favor of high $\rho_i\epsilon_i$-combinations and, thus, $\rho_i\epsilon_i$ goes to $[0,1]$ (note that $x_i = 0$ for $\rho_i\epsilon_i \leq 1$) and the evolutionary stable outcome is money maximization accompanied with a maximal payoff difference in favor of the dictator. Very differently, ρ_i will approach infinity in the ultimatum game implying equal split offers and a zero payoff difference. Obviously, first mover payoffs are decreasing

in ρ_i similar to the dictator game but crucially, first movers will match any MAO (of less than half the pie) such that second mover payoffs increase in ρ_i. The equilibrium prediction is then driven by the fact that the latter effect outweighs the former. Since all actions are intentional in the ultimatum game, there is no evolutionary pressure on ϵ_i and each initial distribution of the parameter will be stable. This is also true in the trust game, but with respect to ρ_i, a yet different outcome emerges. In the trust game, a higher ρ_i implies higher returns and thus lower second mover payoffs causing downward pressure on ρ_i. On the other hand, it is always profitable for second movers if first mover fully trust ($p = 1$) but since p is decreasing in ρ_i, higher ρ_i values by first movers favor higher responder parameters, i.e. there is upward pressure on ρ_i. In equilibrium, up and downward pressure cancel out and a medium reciprocal inclination is stable. First movers fully trust but the return probability is less than 1 such that the outcome is neither characterized by money maximization, nor by equity, but in between with some advantage for the second-movers.

All above predictions are based on an evaluation of each game in isolation. Likely, the *game-of-life*, as Berninghaus et al. (2007) put it, is composed of a more or less random sequence including all of the above games (and others). In their paper, they additionally study a convex combination of the ultimatum and dictator game. The result can be easily transferred to a combination of the trust and the dictator game but as table 5.3 summarizes, it is unclear what will happen if all three game are played at once.

Γ	ρ_i	ϵ_i
DG $\times$ UG	$\rho_i \to \infty$	$\epsilon_i \to 0$
DG $\times$ TG	$\rho_i \to 2.5$	$\epsilon_i \to [0, .4]$
DG $\times$ UG $\times$ TG	$\rho_i \to$???	$\epsilon_i \to$???

Table 5.3: Evolutionary predictions, *game-of-life*

The predictions for each game studied in isolation carry over to a combination of the dictator and one other game. The reason can be simply found in the interaction of ρ_i and ϵ_i. Any ρ_i parameter is in line with dictator offers of zero given that ϵ_i is sufficiently low, i.e. $\rho_i \to \infty$ and $\epsilon_i \to 0$ imply money maximization in the dictator game and equity in the ultimatum game at the same time. As dictator giving is zero once $\epsilon_i \rho_i \leq 1$, ϵ_i receives no pressure to fall below .4 given that $\rho_i = 2.5$ in the trust game.

The above logic still applies with respect to the dictator game if all three games are summarized, i.e. dictator giving must approach zero. However, the predictions with respect to the ultimatum and trust game become assump-

tion dependent. Clearly, if one assumes that individuals hold specific ρ_i, ϵ_i parameters which are allowed to evolve independent of each other, the predictions are those summarized in table 5.2 for each game studied in isolation. If one follows the default assumption of no specificity instead, a universal ρ_i parameter must evolve.[6] Which assumption is correct is of course a question that cannot be answered on theoretical grounds.

5.3 Experimental Design

The possibility of game specific social preference parameters is analyzed with an experiment which took place at the experimental lab of the University of Hamburg. The experiment was programmed with z-tree, see Fischbacher (2007), and participants were recruited via ORSEE, see Greiner (2004), from a subject pool containing students from various disciplines. In total, there were 3 sessions with 6 groups and 72 participants facing 90 to 96 choices each. Sessions lasted between 100 and 110 minutes on average including instructions. Average earnings were 18.46 Euro.

The experiment consisted of a pre-stage and a main stage. In both stages, subjects interacted in a dictator, ultimatum, and trust game and faced both positions of each game. In the dictator and ultimatum game, the pie size was set to 100 ECU, decisions were possible in integer amounts, and second movers in the ultimatum game were asked to state their MAO. In the trust game, the first mover was endowed with 30 ECU and could sent either $0, 10, 20$, or 30 ECU to the second mover. The sent amount was tripled and the second mover could return any integer amount between zero and the tripled amount. The returned amount was additionally doubled for the first

[6] If only one parameter is accepted, of course one parameter must evolve. It is less clear though which parameter will evolve. The evolutionary predictions rely on the construction of normal form *preference games* where the different reciprocity parameters are the strategies. Asymptotically stable states then correspond to symmetric and strict Nash equilibria in the preference games. Since the equilibria are strict, the best response given the equilibrium parameter of the other *player* in the preference game is unique, i.e. the payoff function has a singleton peak. These peaks need not vanish simply because the payoff space is reconstructed by adding another game (although they can depending on the frequency of each game and the payoff scaling). Therefore, in a combination of the trust and ultimatum game, each equilibrium found for each single game can also be an equilibrium in the joint game accompanied by a third equilibrium between both reference outcomes. Each equilibrium will then be associated with some basin of attraction around it and which one evolves will depend on the initial condition. Clearly, however, for a given initial condition (and frequencies, and payoffs), only one equilibrium will evolve and this still contrast the specificity prediction.

mover.[7] Importantly, the strategy method, see Selten (1967), was applied, i.e. trust game responders had to decide on how much to return for any possible positive investment and without knowing the actual amount sent.[8]

After all participants arrived at the lab, they were informed that the experiment consists of two parts but they did not receive any information regarding the content of part 2, which is the main stage. Then the instructions of part 1 were read and players entered the pre-stage.[9] During the pre-stage, each player faced each game and each position in each game exactly once, i.e. went through all six possible positions. The position assignment was random but secured that each position appears in each slot equally often and each game preceded each other game equally often. Players did not receive any feedback on their choices during the pre-stage. They were informed that one position of the pre-stage will be payoff relevant at the end of the experiment but they did not receive any feedback on outcomes before finishing the main stage as well. The pre-stage ended once all players had finished all positions. The purpose of the pre-stage was to elicit players' behavior unconditional on them knowing that information transmission will matter. This information is used later to test for possible signaling effects.

The main part of the experiment consisted of the same games as the pre-stage but now played for multiple rounds. In each round, a player faced each game once, but only one position in each game, i.e. three position assignments per round. All players were separated into two groups of 12

[7]This setup closely matches the setup of the game used for the evolutionary prediction. It differs from the classic setup by Berg et al. (1995) especially because the returned amount is doubled for the first mover such that the second movers choice is efficiency enhancing. The efficiency enhancing second mover choice is a close relation to gift-exchange games such that results are likely to expand into this domain.

[8]There are two reasons to apply the strategy method. First, it allows all players to decide at the same time which yields more decisions per subject in a given time frame. Second, the experimental implementation of the trust game is different from the game used for the theoretical predictions. Since subjects in the experiment can invest less than their entire endowment, which is not possible in the game in section 5.2, they may possess some residual endowment. This leads to different theoretical predictions for the one-shot game which, however, have been found to be inaccurate in several pre-sessions (and which are inaccurate here as well, see section 5.4). The differences between the theoretical predictions vanish if the first mover invests the entire pie and with the strategy method, a response to full investment is generated whenever a subject faces the second mover role in the trust game. Note that in a recent meta study, Brandts and Charness (2011) report that a majority of studies find no significant difference between the strategy and the direct response method. An exception are games with explicit punishment options as e.g. Brosig et al. (2003) who study a version of the trust game where the first mover can be punished if he does not trust and where punishment is more likely if it is relatively cheap and elicited under the direct-response method.

[9]The instructions for both parts are provided in the Appendix.

and matching and position assignment was random within each group.[10] Crucially, players received information about others' behavior which was updated at the end of each round. The information is provided in terms of a matrix with a structure as displayed in table 5.4. 'DG' refers to the dictator game, UG1 (TG1) to ultimatum (trust) game first mover position, and UG2 (TG2) to the respective second mover position. The information for TG2 is subdivided depending on whether the first mover sent an amount of $10, 20$, or 30 ECU. In the experiment, the different games were labeled as situation I (DG), situation II (UG), and situation III (TG).

	DG	UG1	UG2	TG1	TG2-10	TG2-20	TG2-30
OWN	*xx.x*	*xx.x*	*xx.x*	*xx.x*	*xx.x*	*xx.x*	*xx.x*
OTHER	*xx.x*	*xx.x*	*xx.x*	*xx.x*	*xx.x*	*xx.x*	*xx.x*

Table 5.4: Information Matrix

Besides the structure given in table 5.4, players could never observe the entire matrix. Rather, they were only shown the entries with respect to their current game. A subject assigned to the dictator game could only observe column DG, a player in the trust game observed the four last columns (TG1 to TG2-30), and a player in the ultimatum game the remaining columns UG1 and UG2. The entries in the matrix were moving averages over the last three choices by a player in the respective position. Displayed were absolute amounts. Overall, an individual could thus observe his average past behavior (OWN) plus the average past behavior of his current matching partner (the OTHER player) with respect to the current situation. If it was not possible to calculate the averages (less than three decisions), 777 was displayed. Note that subjects could observe average behavior but they did not receive any information on who the other player actually is. This ruled out repeated game strategies. In difference to the pre-stage, players now received feedback on their choices and outcomes (of both players) after each choice.

At the end of the experiment, one choice in each game (not in each position) played in the main part was determined to be payoff relevant. Overall earnings were the sum of earnings from the pre-stage and the main stage. The exchange rate was 100 ECU = 10.00 Euro and there was an additional

[10] Friedman (1996) reports that for group sizes of ≤ 4, groups end up in the cooperative solution of prisoners' dilemma games much more frequently than for larger groups. This indicates that small groups play the game more like a repeated rather than an evolutionary game. It is unclear how this result transfers to a multiple game setting with individual rather than mean information, but a group size of 12 appears reasonable far from the threshold. Players were already split into groups of 12 in the pre-stage as well, but this was done for reasons of consistency across both parts.

minimal payment of 10 Euro. The end of the experiment in the first session was reached close to the time limit of two hours. The other sessions ended once the same number of rounds as in the first session was reached.[11] Participants just knew that there is *a stopping rule* such that from their perspective, the experiment had an uncertain end. Players knew, however, that the maximum lab time does not exceed two hours.

5.4 Results: Choices

The results are presented first with respect to choices and then, in section 5.5, with respect to preference parameters. Some preliminary comments are necessary. First, while aggregated results for all choices are presented at the beginning of the analysis, I then focus on dictator giving and ultimatum and trust game second mover behavior, i.e. first mover choices are left aside. First mover choices are troublesome since they are belief dependent and even though subjects receive a good anchor to form their beliefs by the information structure, it is not clear on how they actually form them. One could tackle the problem by belief elicitation mechanisms but even then, choices may be additionally affected by e.g. risk preferences. Given a question regarding the specificity of preferences, I avoid this kind of ambiguity.

Second, while results are presented with respect to all rounds and with respect to the 2^{nd} half of the experiment whenever necessary, the interpretation is based on 2^{nd} half results. In the 2^{nd} half of the experiment, round 16+, players are expected to have a well understanding of each choice and are thus in a position to make quite conscious choices given that they have to benchmark their behavior, and thus preferences, against past experience and observed behavior by others. In some cases, especially where the presentation would suffer too much, I focus on 2^{nd} half results entirely.

Third, there exists a *trust game problem* which is easiest illustrated by an example. Suppose the first mover sent 10 ECU. Then the second mover received 30 while the first mover still possesses 20. Given that the Falk and Fischbacher (2006) model relies on equity as the reference, any second mover would return more than necessary to establish equity. In the limit of $\rho_i \to \infty$, the return would approach 3.33 and final payoffs become 26.67 for both. Relative to the amount sent, the return would thus approach $\frac{1}{3}$. However, the aggregate relative return for an investment of 10 is 42.6%, i.e. above the theoretical prediction, and, on the individual level, 48.6% of choices violate

[11] The number of rounds was 30 in sessions 1 and 2 but only 29 in session 3. The missing round in session 3 is due to two computers, which, for some reason, shut down in round 30 such that it could not be finished.

the prediction. Without manipulating the model, the available alternative is to assume that players ignore the residual endowment, i.e. to use the predictions presented table 5.1 such that the reasonable maximal amount to return is always the received amount. An average of 43% is in line with this but it is also likely to underestimate returns. This in turn is likely to yield game specificity where, in fact, there is none. In order to solve the puzzle, the discussed trust game returns are always those for an investment of 30. In this case, there is no residual endowment and both theoretical approaches yield the same prediction, i.e. the data is in line with theory by default. As a sort of robustness check, I also calculated the average realized return, i.e. for actual investment levels. For the whole data set, it is 63.66% which is just a percentage point different to the 62.55% found for investments of 30. The difference is far from significant (Mann-Whitney, $p = .5829$).

Fourth, in order to make choices comparable across games, choices are normalized by the maximal choice predicted by theory. For the dictator and ultimatum game with a pie size of 100, this means that each absolute amount is divided by 50. For trust game investments, the actual amount sent is set relative to the full endowment of 30 and for trust game returns, the absolute return is set relative to the amount sent (30 as well, see above). In the analysis, a choice of 1, or 100%, then corresponds to establishing or requiring equity except for trust game investments.

Finally, tests are always two-sided and I indicate the test procedure on the first application of a series of tests.

5.4.1 Aggregated Choice Results

To begin with, table 5.5 provides choice means, standard deviations (std.), means of within-subject standard deviations (w-std.), and observations over all periods as well as for the 2^{nd} half of the experiment (round 16+).

		DG	UG1	UG2	TG1	TG2
all periods	mean	19.84%	70.37%	65.30%	50.62%	62.55%
	std.	32.65%	25.24%	30.81%	38.60%	39.03%
	w-std.	17.37%	15.23%	15.64%	30.63%	23.73%
	count	1068	1068	1068	1068	1068
2^{nd} half	mean	15.76%	72.72%	69.13%	52.27%	64.82%
	std.	28.06%	24.94%	29.10%	42.67%	39.91%
	w-std.	10.29%	11.75%	11.37%	27.99%	20.01%
	count	528	528	528	528	528

Table 5.5: Average choices: all positions

Dictator game behavior is clearly distinct from all other choices with averages at least 30 percentage points below all other averages. This highlights that the inclusion of a parameter capturing the non-intentionality of acceptance in the dictator game is indeed necessary. Ultimatum game giving is higher compared to the average MAO but the gap of $3-5$ percentage points is comparably small. This is in line with the theoretical predictions as well since ultimatum game first movers may give based on an own concern for reciprocity but should always match the MAO of the other player, which is likely to yield a positive but comparably small gap between first and second mover behavior. On average, trust game returns are sufficiently high to yield a positive return on investment for first movers (which requires $> 50\%$) but first movers, on average, invest only half of their endowment. The latter result is probably due to the high variance and within-subject variance in the data. In the trust game, both measures are high, and higher compared to the other games, pointing to a) a large type heterogeneity, and b) a comparably large likelihood for players to change their behavior. Both factors make the displayed return averages less reliable and may thus cause relatively cautious investment behavior.

With respect to game specificity, dictator game averages are clearly below ultimatum and trust game second mover averages (sign tests, $p < .0000$ in all cases). On the other hand, while average MAOs are higher compared to trust game returns, the difference is neither significant for all periods ($p = .3421$), nor for the 2^{nd} half of the experiment ($p = .1257$).

In order to obtain an impression with respect to the evolution of behavior, figure 5.2 plots the average choices per round. The dark solid line plots UG2 behavior, the dark dashed line UG1 behavior. The light solid line plots TG2 behavior, the light dotted line TG1 behavior. The medium colored line (at the bottom) plots DG behavior.

The figure confirms the general impression obtained from aggregated choice averages. In addition, it highlights that while dictator game giving starts out at a level typical for experimental research (close to .4, i.e. 20% of the pie), the clear discrimination against the ultimatum and trust game evolves during the course of the experiment. On the other hand, ultimatum game offers are lower compared to typical experimental results (averages of .8 (40% of the pie) are not unusual), although the figure reveals a small positive time trend and later-period-outcomes are relatively close to the expected values.[12]

In order to verify the results with respect to specificity, I ran several regressions with *choice* in DG, UG2, and TG2 as the dependent variable,

[12]Time trends are more closely evaluated in section 5.6.

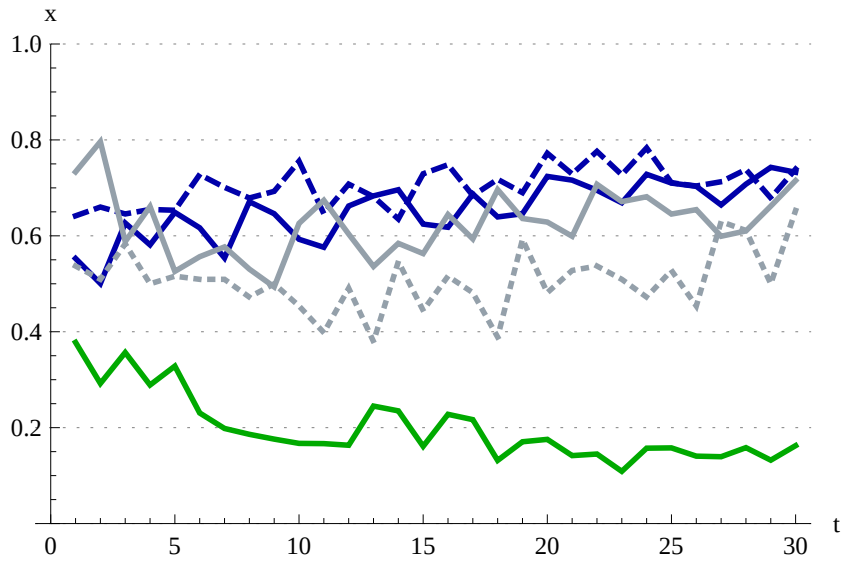


Figure 5.2: Evolution of average behavior

and a constant plus dummies for DG ($dum - dg$) and UG2 ($dum - ug$) as independent variables. Alternative specifications include the displayed own average past behavior ($own - av$) and the displayed other players' average ($oth - av$) as additional independent variables.[13] Table 5.6 presents the results for a random effects specification.[14] Standard errors were clustered at the group level.

The results confirm that while there is intention-based specificity, there is none between trust and ultimatum game behavior. Additionally, besides the non-negligible within-subject variance in choices reported in table 5.5, individual behavior is strongly path dependent but not independent of others' displayed behavior as well. The relation between the coefficients $own - av$ and $oth - av$ is well in line with the fundamental assumption that behavior is preference driven in principle, but that players may reconsider their behavior in light of others' behavior as well.

[13] Players could only observe the averages for one game at a time. This allows to summarize them into one variable capturing the general path dependence of individual behavior ($own - av$) and capturing the general impact of others' behavior on own non-conditional, respectively preference based, choices ($oth - av$).

[14] Alternative model specifications included pooled OLS, pooled tobit, and fixed effects, again with robust errors clustered at the group level. Sign and significance levels are identical in models I and III for all specifications. In II and IV, $dum - dg$ and *const.* are significant at the 5% level under pooled OLS. In IV, additionally $oth - av$ looses 1 level of significance. In II and IV, *const.* is insignificant under tobit but signs and significance levels for the coefficients remain unchanged.

choice	all periods		2^{nd} half	
	I	*II*	*III*	*IV*
const.	.6239***	.0928***	.6455***	.1073***
	(.0405)	(.0297)	(.0432)	(.0274)
$dum - ug$	.0342	.0110	.0516	−.0011
	(.0339)	(.0134)	(.0400)	(.0217)
$dum - dg$	−.4287***	−.0905***	−.4880***	−.0934***
	(.0108)	(.0272)	(.0219)	(.0238)
$own - av$		.7729***		.7887***
		(.0333)		(.0230)
$oth - av$		.0934***		.0743***
		(.0159)		(.0120)
R^2	.2688	.6385	.3530	.6467
χ^2	1845.11***	> 9999.9***	526.32***	> 9999.9***
n	3204	2416	1584	1584

Robust standard errors clustered at the group level in parenthesis.

***,** ,* indicates significance at the 1%, 5%, and 10% level respectively.

Table 5.6: Random effects regressions: Δ UG2, TG2, DG

5.4.2 Individual Types

The findings from the last section rely on aggregated choices. Clearly, the fact that trust and ultimatum game second mover behavior is not different at the aggregate does not imply that there is no specificity on the individual level. It could be the results of a canceling out effect, such that types with specificity in the theoretically expected direction ($UG2 > TG2$, *expected*-types) live in a population that also contains types with unexpected specificity ($TG2 > UG2$, *unexpected*-types). In order to separate types, I calculated the mean choice, $\bar{x}_i$, together with the mean absolute deviation of the mean, mad_i, for each game and each individual. Then, the pairwise relation between UG2 and TG2 is calculated such that $\Gamma_1 > \Gamma_2$ if either $\bar{x}_{i,1} - mad_{i,1} > \bar{x}_{i,2}$ or $\bar{x}_{i,1} > \bar{x}_{i,2} + mad_{i,2}$, or both, and $\Gamma_1 = \Gamma_2$ otherwise (the *equal*-types).[15] Table 5.7 summarizes the classification.

Very few subjects belong to the class with no game specific behavior, the equal-types.[16] The largest fraction, a slight majority of the population

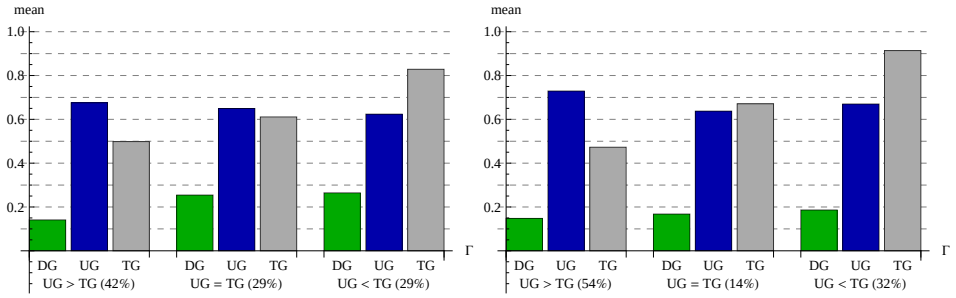
[15] As alternatives, I used individual medians and median absolute deviations of the median as well as individual means and standard deviations for the classification. The resulting fractions for the 2^{nd}-half differ by 5.5 percentage points at most and, more importantly, subsequent results are very similar such that the overall obtained interpretation is the same under both alternatives.

[16] 14% is a dramatically low rate of *consistency* but to some degree, it is the consequence of the decision to assign a type difference whenever one of the conditions presented before

	$UG2 > TG2$ expected	$UG2 = TG2$ equal	$UG2 < TG2$ unexpected
all periods	41.67%	29.17%	29.17%
2^{nd} half	54.17%	13.89%	31.94%

Table 5.7: Types dependent on $UG2 \lessgtr TG2$

in the 2^{nd} half, are expected-types but there is also about a third of subjects with higher trust game returns compared to ultimatum game MAOs, i.e. unexpected-types. Figure 5.3 summarizes the average (over individual averages) behavior for each type in each game.



(a) all periods (b) 2^{nd} half

Figure 5.3: Type dependent choice averages

If the type classifications are reasonable, the within differences across games should be significant except for the comparison of $UG2$ and $TG2$ for the equal-types. This is indeed the case. Dictator game giving is significantly lower for all groups and all periods as well as for the 2^{nd} half (sign tests, highest $p = .0039$). For the expected-types, ultimatum game MAOs are significantly larger compared to trust game returns (both $p < .0000$) while for the unexpected-types, trust game returns are significantly higher (both $p < .0000$). For the remaining equal-types, the difference between $UG2$ and $TG2$ is indeed insignificant (lowest $p = .3593$).

Especially figure 5.3 (b) yields the impression that type differences are mainly due to differences in trust game returns. While ultimatum game

table 5.7 is satisfied. If both means are required to lie outside the $\pm mad$ confidence bound of the respective other mean, there are 33.3% expected-types, 23.6% unexpected-types, and 43.1% equal-types. While a consistency rate of 43% is still low, it is in line with the findings by Blanco et al. (2011) and Schliffke (2012b). The reason to apply the weaker criterion for type differences is that it is sufficient to obtain clearly separated types (see below) and that the consistency rate as such is not the focus of this work.

MAOs are quite similar across types, trust game returns are not. The impression can be validated via cross game tests. For all periods, average MAOs are weakly different between the expected and equal-types (Mann-Whitney, $p = .0797$) (besides very similar means) but neither between the expected and unexpected-types nor between the equal and unexpected-types (lowest $p = .2063$). For the 2^{nd} half, all UG2 MAOs are insignificantly different across types (lowest $p = .1530$). This is not true for trust game returns. The difference in returns is highly significant for all periods and the 2^{nd} half between the expected and unexpected-types (both $p < .0000$). Each type is also weakly significantly different from the equal-type in the 2^{nd} half (highest $p = .0992$). For all periods, the difference between the equal and unexpected types is significant ($p = .0003$) although the one between the equal and expected-types is not ($p = .2206$). In addition, dictator game giving is not significantly different across types in the 2^{nd} half (lowest $p = .3548$) but it is for all periods and between the expected and unexpected-types ($p = .0150$) and between the equal and unexpected types ($p = .0248$).

The type classification supports the possibility that the non-specificity found in aggregate behavior is due to a canceling out effect. About half of all individuals are of the expected type under specificity but almost a third is of the reverse, unexpected type. While ultimatum and dictator game behavior does not differ across types, the decisive element of discrimination are trust game returns. The expected-types return 47.28% in the 2^{nd} half of the experiment which is just below the threshold for a zero return on investment for first movers and does not reveal any particular disposition to behave reciprocally. In contrast to that, the unexpected-types return 91.38% in the 2^{nd} half of the experiment which is close to the equity implying choice of 100% and thus a strongly reciprocal response.

5.5 Results: Preference Parameters

In this section, I estimate ρ_i, ϵ_i-parameters that best fit the data. The question is whether or not the estimates for each single game are significantly different from the joint estimates over all choices. If this is not the case, it would be justified to use one parameter distribution irrespective of the specific game. Otherwise, model-based cross game inference would need to discriminate given that the goal is to obtain relatively accurate predictions and even if choices are very similar. Note that insignificant differences do not necessarily follow from the non-specificity in choices since the choice functions dependent on ρ_i, ϵ_i reported in table 5.1 are different across games, i.e. parameters may be different while choices are not.

The best fit is understood here as those parameter values which minimize the residual sum of squares, i.e. those ρ, ϵ-values that solve

$$\min_{\rho,\epsilon} \left[\frac{1}{n-1} \sum_i \sum_\Gamma \left[y_i^\Gamma - \hat{y}^\Gamma(\rho, \epsilon) \right]^2 \right] \tag{5.2}$$

with y_i^Γ the actual and non-normalized choice averages by each individual in the respective situations Γ, and $\hat{y}^\Gamma(\rho, \epsilon)$ the choice in the respective situation implied by a specific ρ, ϵ combination according to the equilibrium solutions presented in table 5.1. Note that ρ, ϵ, and $\hat{y}^\Gamma$ have no subscript i as parameters are estimated for the entire sample or subgroups of it, but not for individuals. The method was a grid search with two digit precision for each parameter and reported standard errors are the standard deviations of parameter estimates obtained from 1000 bootstrapped samples each. Since $\rho_i \to \infty$ for equity implying actions, I set $y_i = .495$ if $x_i > .99$. Further, since there are two parameters to estimate, I always jointly estimate ultimatum MAOs and dictator giving (UG-model), trust game returns and dictator giving (TG-model), and all three choices at once (FULL-model). Finally, I focus on 2^{nd} half results for ease of presentation.

5.5.1 Aggregate Outcomes

As a first approach, table 5.8 reports the estimated parameters that best capture the behavior of the entire population.

$\Gamma =$	$\{UG, DG\}$	$\{TG, DG\}$	$\{UG, TG, DG\}$
$\hat{\rho}$	1.73	1.36	1.44
	(.2585)	(.1330)	(.1299)
$\hat{\epsilon}$	0.69	0.88	0.83
	(.1026)	(.0831)	(.0780)
$\sqrt{rss/n}$	.2271	.2295	.2578
n	144	144	216

Bootstrapped standard errors (1000 iterations) in parenthesis.

n is the number of fitted observations.

Table 5.8: ρ, ϵ-estimates, entire population, 2^{nd} half

Comparing the UG and TG-model estimates of ρ and ϵ, the parameters are different from each other but the difference is not significant (z-tests, $p_\rho = .2040, p_\epsilon = .1498$).[17] Given that, also the parameters estimated for the

[17] The applied formula to calculate the z-test statistic is $z = (\rho_1 - \rho_2)/\sqrt{se_1^2 + se_2^2}$ which

FULL-model are not significantly different to each single estimate (lowest $p_\rho = .3174, p_\epsilon = .2758$) and hence, the non-specificity found in choice data carries over to preference parameters. Additionally, while there is some loss is precision estimating the FULL-model with all three games, the increase in root RSS per observation of about 3 percentage points is comparably low in relation to the overall level of variance in the data and relative to the increase in fitted observations.

As a second approach, I seek parameter values that capture some of the heterogeneity within the sample. The goal is a parameter distribution that characterizes the population in terms of fractions with comparably low, medium and high parameter values which could be applied for aggregate cross-game predictions. Importantly, cross game predictions at the aggregate do not require any within-subject consistency. Therefore, all individual choice averages are independently sorted in ascending order before the estimation. Table 5.9 summarizes the results for the population subdivided into four quartiles (18 subjects each) with the 1^{st} quartile the lowest averages, the 2^{nd} quartile the second lowest averages, and so on.[18]

$\Gamma =$	$\{DG, UG\}$		$\{DG, TG\}$		$\{DG, UG, TG\}$	
	ρ	ϵ	ρ	ϵ	ρ	ϵ
1^{st} quartile	0.40	$[0, 1]$	0.63	$[0, 1]$	0.61	$[0, 1]$
	(.0651)		(.0260)		(.0357)	
2^{nd} quartile	1.26	0.80	1.08	0.93	1.12	0.90
	(.0574)	(.0367)	(.0476)	(.0393)	(.0432)	(.0353)
3^{rd} quartile	2.93	0.40	2.49	0.47	2.66	0.44
	(.2225)	(.0305)	(.2625)	(.0485)	(.1877)	(.0309)
4^{th} quartile	39.76	0.05	24.85	0.08	33.14	0.06
	(10.6802)	(.0136)	(6.7017)	(.0208)	(6.9882)	(.0120)
$\sqrt{rss/n}$	.0805		.0831		.0912	
n	144		144		216	

Bootstrapped standard errors (1000 iterations) in parenthesis.

Table 5.9: ρ, ϵ-estimates, distribution, 2^{nd} half

For all three specifications, the within differences from one quartile to

does not correct for the within-subject nature of the data. This is motivated by the fact that the bootstrap samples were independently drawn for each game such that the bootstrapped parameter estimates used to calculate standard errors are indeed uncorrelated. With joint bootstrap samples, and thus some positive correlation, p-values will be lower, i.e. some results barely insignificant would turn significant. On the other hand, whenever significant differences are found, they are more robust given the approach chosen here.

[18] I also estimated a third-based division of the population. The results are qualitatively very similar and yield the same overall conclusion.

the other with respect to ρ are highly significant (highest $p = 0.0006$). The within differences with respect to ϵ are highly significant (highest $p = .0046$) except for the TG-model between the 1^{st} and 2^{nd} quartile where the difference is weakly significant ($p = .0750$).

With respect to across model comparisons, the TG and FULL-model are not significantly different (lowest $p = .3898$ for all parameter comparisons). For the UG-model, the parameters in the 3^{rd} and 4^{th} quartile are not significantly different from those of the FULL and the TG-model (lowest $p = .2006$). However, significant differences occur between the ρ estimates in the 1^{st} and 2^{nd} quartile for both comparisons UG vs. FULL-model and UG vs. TG-model (highest $p = .0512$) and between the ϵ estimates in the 2^{nd} quartile and both model comparisons (highest $p = .0500$).

Although there are some significant differences between the UG-model and the FULL-model, the FULL-model is nevertheless the suitable model. Similar to the results obtained if only one parameter is estimated for the entire population, there is some loss in precision as the root RSS per observation increases by 11.5% compared to the average of the UG and TG-model. However, the increase of 50% in fitted observations outweighs this effect. Additionally, the FULL-model yields an adequate representation of behavior. It predicts average dictator giving of 16.27% compared to actual 16.24%. It predicts an average MAO of 70.70% compared to actual 69.5%, and it predicts TG returns of 63.27% compared to 63.01% actual returns. The average misrepresentation of choices is thus 0.5 percentage points. This is a lot higher compared to an average misrepresentation in the UG-model which can be calculated to be .04 percentage points, and compared to the .01 percentage points which can be calculated for the TG-model, but nevertheless a very close, and seemingly sufficient match.

5.5.2 Individual Types

Best fit parameters are also estimated for the two main types identified in section 5.4.2. Figure 5.4 displays the estimated ρ, ϵ-parameters for ultimatum game MAOs and trust game returns together with bars indicating 95%-confidence bounds for the expected ($UG2 > TG2$) and unexpected-types ($UG2 < TG2$).[19]

[19]For the expected-types, the estimated parameters (standard errors) for ultimatum game MAOs are $\rho^{UG} = 2.09$ (.4537) and $\epsilon^{UG} = 0.56$ (.1207) and for trust game returns are $\rho^{TG} = 1.12$ (.0962) and $\epsilon^{TG} = 1.00$ (.0036). For the unexpected-types, the respective values are $\rho^{UG} = 1.52$ (.2751), $\epsilon^{UG} = 0.81$ (.1324), $\rho^{TG} = 4.24$ (1.5472) and $\epsilon^{TG} = 0.29$ (.0850).

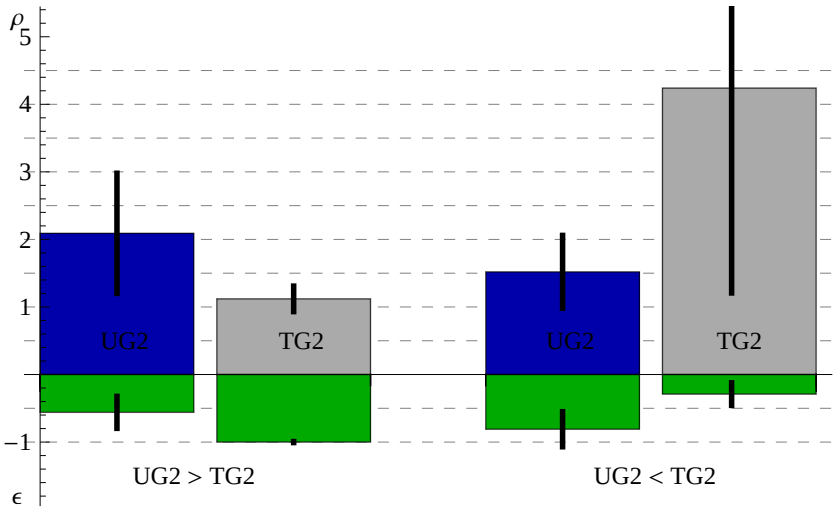


Figure 5.4: ρ, ϵ-estimates, type differences, 2^{nd} half

For both types, the relation of parameter estimates is as expected. For the expected-types, the ρ^{UG} estimate for ultimatum game MAOs is larger compared to the trust game return estimate, ρ^{TG}, while for the unexpected-types, it is the other way around. Both differences are significant (z-tests, expected-types, $p = .0366$, unexpected-types, $p = .0836$). For the expected-types, the ϵ^{UG} estimate is smaller compared to the ϵ^{TG} estimate while for the unexpected-types, it is, again, the other way around. Both differences are significant (expected-types, $p = .0003$, unexpected-types, $p = .0009$). Thus, the significant differences found in choice averages within each type carry over to parameter estimates.

The choice analysis revealed that types are mainly differentiated with respect to trust game return behavior while ultimatum game behavior is similar. This can be validated for the parameter based analysis as well. Across types, the ρ^{UG} parameters are not significantly different ($p = .2846$) and neither are the associated ϵ^{UG} parameters ($p = .1615$). On the other hand, the differences in ρ^{TG}, ϵ^{TG} estimates are significant across types ($p_\rho = .0444, p_\epsilon < .0000$).

Finally, the difference in types is further validated if the FULL-model is estimated for each type. For the expected-types, the estimated parameter values (standard errors) are $\rho = 1.15$ (.0606) and $\epsilon = 1$ (.0277). For the unexpected-types, the estimated parameter values are $\rho = 2.56$ (.3173) and $\epsilon = 0.48$ (.0643). The difference across types is highly significant for both parameters (both $p < .0000$).

5.6 Additional Results

5.6.1 Signaling

Subjects in the experiment cannot identify their matching partner which rules out the application of repeated game strategies. Nevertheless, the information transmission protocol can be used for signaling purposes. For example, second movers in the ultimatum game may state higher than preference induced MAOs given the belief that first movers will match those MAOs, i.e. offer more.[20] The problem is that if first movers indeed match the MAO, then second movers do not have an incentive to revise them and play is in a self-confirming equilibrium at a non-preference based level.

In order to check for signaling effects, the pre-stage was implemented where subjects make exactly one choice in each decision position and do not know that there is a second part with multiple encounters of each game and information transmission. Figure 5.5 displays the average choices in DG, UG2, and TG2 both for the pre-stage and over the very first decision of each subject in the respective position in the second part of the experiment.

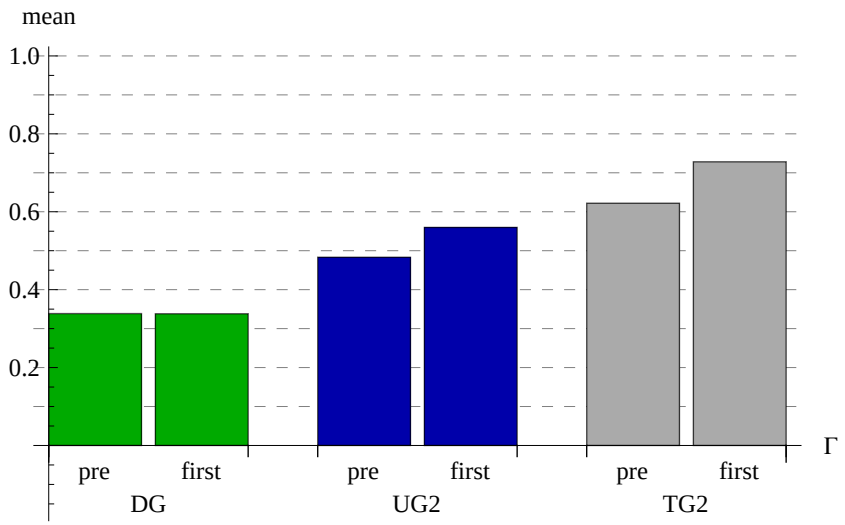


Figure 5.5: Signaling: pre-stage vs. first choice

There is not really a possibility to exploit signaling in the dictator game.

[20]Such a belief is correct in general. The rank-correlation between displayed MAOs and ultimatum offers is .5253 and highly significant ($p < .0000$). Similar, the correlation between displayed returns in the trust game (for an investment of 30) and investments is .4438 ($p < .0000$). Both relations are easily verified by regression and are very robust to model specification.

Hence, it is not surprising that no difference between the pre-stage and first choices is found (sign-test, $p = .4408$). On the other hand, were signaling can be useful, first choice averages in the main part with information transmission are indeed higher compared to averages in the pre-stage. For the ultimatum game, the difference of +7.67 percentage points is significant ($p = .0660$) while the higher difference in trust game returns of +10.65 percentage points is not significant ($p = .1996$).

For the trust game, another argument against an impact of signaling is the overall choice average of 62.55% which is of almost exact size as the pre-stage average of 62.18%. Even if the first-choice average is increased, and besides that the increase is not significant, the potentially signaling induced level of returns is not sustainable. The situation in the ultimatum game is different. The average MAO over all rounds and choices is 65.30% and thus higher compared to both pre-stage results (48.31%) and first choices (55.97%). A positive time trend, however, is an argument against signaling as well. Before the discussion of the argument, the time trend should be verified. Table 5.10 reports ols and tobit time-trend estimates. The dependent variable is choice in the respective game and the independent variable is period. For a more complete picture, results for all three games are displayed. Since there are quite drastic changes especially in TG2 behavior within the first rounds (compare figure 5.2), I eliminated rounds 1 to 5 from the analysis.[21]

choice	DG		UG2		TG2	
	ols	*tobit*	*ols*	*tobit*	*ols*	*tobit*
const.	.2247***	−.0745	.5795***	.5724***	.5346***	.5088***
	(.0546)	(.1784)	(.0630)	(.0644)	(.0686)	(.0807)
period	−.0010*	−.0022**	.0017**	.0017***	.0016*	.0017**
	(.0005)	(.0010)	(.0006)	(.0006)	(.0007)	(.0009)
R^2	.0052	.0026	.0142	.0036	.0077	.0055
F	4.33*	4.82**	8.09**	8.51***	4.65*	3.90**
n	888	888	888	888	888	888

Robust standard errors clustered at the group level in parenthesis. Pseudo-R^2 for tobit.

***,** ,* indicates significance at the 1%, 5%, and 10% level respectively.

Table 5.10: Time trend regressions

As expected from previous results, the time trend in the dictator game is negative. The comparably small size is due to the elimination of rounds 1 to 5 where the largest drop occurs. For the trust and ultimatum game, time trends are positive, and significant, which matches the impressions based on

[21] With three choices in each round, period is thus running from 16 to 90.

figure 5.2. The fact that the trust game trend is less significant compared to the ultimatum trend is due to the higher variance in trust game returns (compare table 5.5).

The crucial argument with respect to signaling is the following. In principle, stating higher MAOs or higher returns can be profitable given that first movers indeed react to the displayed averages. Ex post, the Pearson correlation between ultimatum payoffs and displayed MAOs is .1582 and between trust payoffs and displayed returns is .3758 (both $p < .0000$).[22] However, during the experiment, the exact relation between displayed averages and first mover reaction is unknown to the players. Somebody with a low average has no, or very little knowledge, about investments or offers under high averages. Given that, the benefits from an attempt to signal remain very uncertain whereas the costs are much less uncertain. With each encounter of a position being equally likely to be paid, the expected costs of an increased return in the trust game are fully determined. In the ultimatum game, the costs are less certain but, on the other hand, the consequences of rejection are much more severe. This sort of asymmetry renders signaling relatively unattractive on a priori ground.[23] The fact that players nevertheless adapt to higher return and MAO levels can then have two reasons. Either, players are unhappy with the current situation, for example due to their current payoff situation, or due to the observation of others' play. Unhappiness, however, is another word for suffering disutility and is thus a preference based explanation directly. Or, players indeed learn that higher returns or MAOs yield a sufficient chance for higher payoffs and then adapt to that. In the lab, where players and strategies do not die out (at least, not in this case), true preference evolution cannot be implemented. However, the adaptation to payoff opportunities represents the same driving forces responsible for preference evolution in theory.

Overall, significant time trends rule out self-confirming equilibria at non preference based levels. On top of that, time trends are either preference based itself, or they are in line with the mechanisms of preference evolution transferred to the lab. While this does not rule out that strategic behavior is a better explanation for the behavior by some players, the overall influence on results is either small, or it is non-separable from a preference based interpretation.

[22]The respective rank correlations are .2837 in the ultimatum game and .3698 in the trust game (both $p < .0000$).

[23]Note further that displayed averages are moving averages. Players thus cannot signal their type by stating e.g. high returns on early encounters and then simply rely on that like in the case of a monopolist signaling to be the tough type and then making monopoly profits for many periods because no one enters.

5.6.2 Cross-Game and Within-Type Correlations

Finally, I look at the correlations across games both for the entire population and the two main types. The analysis seeks to sort out whether the findings of similar aggregated outcomes but different types is more likely a result of chance or the results of game specific preferences indeed. In case of game specific preferences, the observed correlations should be positive, significant, and comparably strong such that the type differences are the consequence of level differences with choices nevertheless not being relatively independent of each other. Table 5.11 summarizes Spearman correlations across all choices based on individual averages in the 2^{nd} half of the experiment.

	DG	UG1	UG2	TG1	TG2
DG	–	–	–	–	–
UG1	.3190***	–	–	–	–
UG2	.0904	.6320***	–	–	–
TG1	.2668**	.3269***	.4136***	–	–
TG2	.2163*	.3607***	.5170***	.5286***	–

***,** ,*significance at 1%, 5%, and 10%.

Table 5.11: Spearman correlations

The lowest correlations and weakest in terms of significance are found between the dictator game and ultimatum game MAOs (in fact insignificant, $p = .4502$) and the dictator game and trust game returns (10%-level, $p = .0680$). To some extent, this is expected since dictator giving in the 2^{nd} half is very low in general with no type differences either. On the other hand, both dictator correlations to first mover choices are significant at least on the 5%-level, although of comparably low size either. A potential explanation for this may be found in a recent result by Brook et al. (2012) who establish a significant predictive power of giving in a standard dictator game for giving in risky environments and, first mover choices clearly entail some risk. This effect may also explain a significant correlation between trust and ultimatum game first mover choices.[24]

[24]The focus here is on second mover and dictator game correlations such that I discuss the remaining first-mover correlations only briefly. High correlations between first and second mover choices are often explained by a consensus effect, see Mullen et al. (1985) and Dawes (1989), i.e. players extrapolate from their own type to others. However, with information transmission this may be much less a good explanation. Rather, another effect known as the *positive self-image* effect may be the driving force, see e.g. Farwell and Weiner (1996), or Singh et al. (1998). Under this effect, people want to maintain a positive self-image which seems arguably hard, for example, if someone frequently gives less than he himself expects. Given that trust game returns and ultimatum game MAOs

Besides the weak or insignificant correlation of dictator giving and second mover choices in TG and UG, the correlation between ultimatum game MAOs and trust game returns is of medium size and highly significant ($p < .0000$).[25] The results becomes even more clear if one separates both main types identified in section 5.4.2. Figure 5.6 plots all individual data pairs for the expected types, i.e. those with higher average MAOs than trust game returns (dark, round), and the unexpected types where it is the other way around (light, squared).

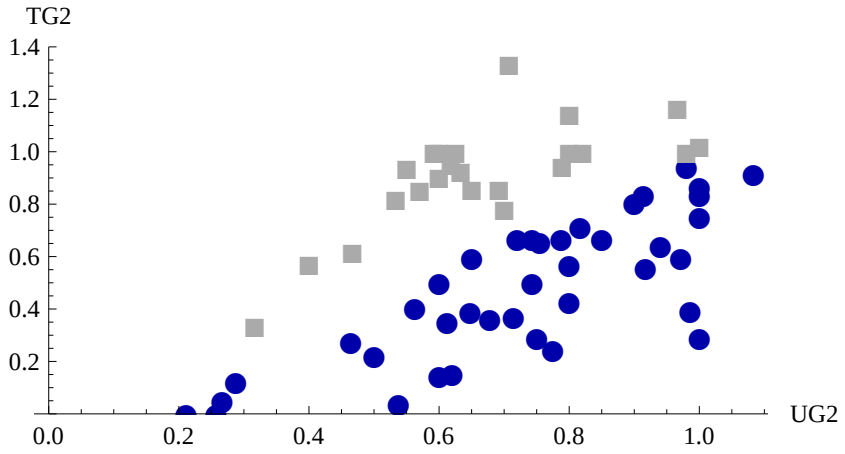


Figure 5.6: UG2 and TG2 correlations within types

For the expected types alone, the correlation between UG2 and TG2 is as high as .7691 ($p < .0000$) and for the unexpected types it is .7182 ($p = .0001$).[26] For both types there is thus strong evidence that the main difference between both choices is a level difference but beyond that, the two choices are well connected. The fact that the overall correlation is smaller by more than .2 is again a canceling out effect as it is present between average choices.

---

are strongly correlated, this may also explain cross game correlations between first and second mover choices which are positive and significant (UG1 and TG2, and TG1 and UG2). Likely, the effect will be weaker but so are the correlations.

[25]The high correlation here stands in sharp contrast to the correlation between ultimatum game responses and sequential prisoners' dilemma game responses found by Blanco et al. (2011) of .19 (insignificant) and between UG responses and trust game returns found by Schliffke (2012b) of .12 (insignificant). Potentially, the difference is due to the different experimental design and especially the repetition of play. In the pre-stage of this experiment, the correlation between UG2 and TG2 is .3006 ($p = .0103$) which is larger and significant, but much closer to the reference values than .52.

[26]In addition, the correlation between DG and UG2 is .0246 ($p = .8821$) for the expected and .0367 ($p = .8680$) for the unexpected types. The correlation between DG and TG2 is .1613 ($p = .3265$) for the expected and .0164 ($p = .9406$) for the unexpected types.

5.7 Discussion and Summary

The aim of this work was to answer the question of whether or not social preferences are game specific both at the aggregate and the individual level. The answer turns out to depend heavily on the viewpoint. Dictator game behavior has been found to be clearly different compared to trust and ultimatum game behavior both at the aggregate and at the individual level. While giving starts out at levels typical for experimental research (close to 20% giving), it declines to less than 10% giving during the course of the experiment and is, if at all, very weakly correlated to ultimatum game MAOs and trust game returns. Additionally, there is no difference in dictator game behavior across the different types of players identified in the paper. This result is very much in line with the theoretical background which assumes a discount on giving due to the lack of intentionality behind *acceptance* in the dictator game and which predicts a decline of giving over time.

The crucial distinction with respect to specificity is the one between ultimatum game MAOs and trust game returns as both are fully intentional choices. Here, a clear distinction between the aggregate and the individual level is found. On aggregate, there are no differences in choices across both games. Trust game returns start out at higher levels compared to MAOs but drop initially and after round 5 (of 30) of the experiment, both choices evolve more or less parallel with slightly, but insignificantly higher MAOs. The non-specificity carries over to the estimation of Falk and Fischbacher (2006) parameters for reciprocity which best fit the data. Especially, the between-subject heterogeneity in the population is captured by a distribution of preference parameters and those parameters obtained for the full model including all three games are, in general, not significantly different from those parameters obtained if the trust and ultimatum game are independently fitted together with the dictator game. In addition, the jointly estimated parameters of the full model predict choices in each single game very well with an average difference between predicted and observed averages of .5 percentage points.

On the individual level, however, specificity is indeed found and the aggregate outcome is the result of a canceling-out effect between two distinct types. The type distinction is based on the relation of ultimatum game MAOs and trust game returns. The evolutionary background suggest that MAOs should approach half the pie under specificity while trust game returns need to ensure a positive return on investment for first movers but need not approach equity implying actions. The expected relation is thus that MAOs are larger than returns and for a slight majority in the population (54%), such a relation is indeed found. However, a third of individuals are of the

reverse type with returns non-trivially higher than MAOs. The distinctive element between types are trust game returns. MAOs are not significantly different across types both with respect to choices and preference parameters. However, while the expected types with higher MAOs reveal no specific inclination to behave reciprocal in the trust game - on average, they just miss the threshold for a positive first mover return on investment - the unexpected reverse types behave strongly reciprocal with an average return close to the equity implying return. Quite crucially, the within-type correlation between MAOs and returns is greater than .7 for both types. I take this as evidence in favor of a *preference* based explanation in the sense that game specificity implies level differences across games, but the different domains are nevertheless evaluated on a common ground - like a specific utility model - and thus not independent of each other.

The results have two direct consequences. The non-specificity found at the aggregate suggests that cross game inference may well be possible. However, the individual level specificity strongly questions the robustness of any inference. While aggregate stability would follow from individual consistency, the reverse does not hold and the canceling-out effect that drives aggregated non-specificity here need not prevail under other circumstances, or with different populations. For example, even for a given population with known behavior under a specific incentive scheme, e.g. an exchange orientated with effort, consideration, and control, one may not be able to make any inference with respect to a change in the incentive scheme, e.g. towards a trust based scheme. Picking up the debate between Binmore and Shaked (2010) (BS) and Fehr and Schmidt (2010) (FS) mentioned in the introduction, the aggregated outcomes here support the FS view that aggregate inference can yield predictions which are true on aggregate. But the results support also BS who point out that the predictions are based on a distribution of preferences which is simply not found in the predicted games. This difference is likely to become particularly relevant once specific mechanisms target at specific subgroups of a population.

The second consequence is that game specificity offers an explanation, for recent results on the inconsistency of social preferences. Blanco et al. (2011), who find individual consistency for, on average, half of their subjects comparing dictator, ultimatum, public goods and sequential prisoners' dilemma decisions, argue that the low levels of consistency are probably due to a multiplicity of social norms, that different situations trigger different norms, and that different motives may be weakly correlated within each subject. They continue stating: *We would hence expect a model calibrated on decisions in one type of game to yield reliable predictions only within the class of games where the same motives dominate. Since this is difficult to know ex-ante,*

deriving predictions for new games appears to be problematic. (p. 334). The results here, with types clearly distinct along their (reciprocal) reactions in the trust game, yield direct evidence for different games triggering different motives on the level of the individual. For a majority of players in this experiment (either 86% or 57% dependent on the strictness of the definition of a difference), behavior in one game is a bad predictor for behavior in the other game. However, the results here are not based on the possibility that different motives are only weakly correlated within individuals. Contrary to that, I found a comparably strong correlation between MAOs and trust game returns and, within each type, truly strong ones. This suggests that inconsistency is not the result of motives which are relatively independent of each other, but the consequence of related motives which nevertheless trigger different levels of behavior across games.

The results need to be put into perspective both with respect to the question of whether or not the specificity found is expected given the selected games and with respect to the specific experimental method. The ultimatum game as an abstract representation of bargaining situations is clearly distinct from the trust game capturing reciprocity and constituting a social dilemma situation. A crucial difference is, for example, that the responder in the ultimatum game is in the disadvantageous position but the returnee in the trust game is in the advantageous position. This implies that e.g. different parameters are used to capture behavior if the Fehr and Schmidt (1999) model is applied. It also implies that the ultimatum game responder choice is an act of negative reciprocity while the trust game returnee choice is an act of positive reciprocity. Differences between positive and negative reciprocity have been incorporated theoretically by e.g. Charness and Rabin (2002) (with an extra parameter for *misbehavior*). Experimental comparisons of both are provided for example by Pereira et al. (2006) and Al-Ubaydli et al. (2010) who test positive and negative reciprocity in different adaptations of the gift-exchange game. Both studies find that both positive and negative reciprocity are present in their samples but while the former report higher fractions of negative reciprocal acts given an environment that favors negative reciprocity, the latter report higher fractions of positive reciprocity if the environment favors positive reciprocity (both appears very natural). However, in the baseline treatment of Al-Ubaydli et al. (2010) which does not favor one of both possible traits, positive and negative reciprocal acts appear in similar frequencies which is at least qualitatively in line with the aggregate findings here. On the other hand, Dohmen et al. (2009) report questionnaire evidence for a representative sample of 20.000 German citizens and find a) negative reciprocity to be more dispersed than positive reciprocity and b) a literally tiny .01 albeit 5%-level significant correlation between the

two which stands in sharp contrast to the results here. Overall, and besides mixed experimental and empirical results, there are thus potentially good reasons to expect specificity which relativizes the actual finding of individual level specificity. On the one hand, however, the non-separation of both games was causal for the recent debate between FS and BS and plays a major role in recent findings on inconsistencies as well, i.e. a clear separation does not appear as the current consensus for applications. On the other hand, to the best of my knowledge, this is the first investigation into game specificity using this specific experimental design. The goal was therefore to establish some benchmark results and for this purpose, the use of well-known games with clearly distinct predictions under specificity seems justified.

An absolutely crucial question is whether the design is useful for the analysis of social preferences at all. A typical experiment investigating questions related to social preferences has a clear cut money maximizing prediction and properly set pecuniary incentives such that deviations from the money maximizing predictions call for the existence of other than purely selfish and money orientated preferences. This is not true in this experiment since strategic considerations and the use of signaling cannot be ruled out on a priori ground. The crucial argument in favor of a preference based interpretation is that while signaling can motivate *some* deviations from the money maximizing predictions of the one-shot games, it does not imply any *specific* levels of returns or MAOs. Both returns and MAOs exhibit significant positive trends over time which rules out self-confirming equilibria at non preference based levels. In addition, the time trends either have a preference based cause directly, or they are based in the adaptation to payoff opportunities which amounts to an explanation at least non-separable from the assumptions of preference evolution.

Assuming that the design is suitable for the measurement of preferences, it has a clear advantage. Subjects encounter each position multiple times and are thus very familiar with each game and the consequences of their behavior in later rounds. The expectation is thus that choices in later rounds are made much more conscious and are much less affected by potential misunderstanding, simple errors, or experimental effects. These are common arguments in favor of looking at later period results. The information transmission protocol and the possibility to condition on information prevents, however, that e.g. cooperation breaks down which is otherwise expected for theoretical reasons, see Güth and Kliemt (1994, 1998), but also very common and best established experimentally for public goods games where contributions strongly decline if the game is played for multiple rounds. As discussed above, the information transmission protocol introduces strategic considerations which question a preference based explanation to some degree. It is clear, however,

that no information transmission does not generate unbiased preference based results either. For example, Fischbacher and Gächter (2010) show how the decline in contributions to a public good can be explained by the interaction of beliefs, contributions, updated beliefs, contributions, and so on, because the belief updating mechanism and contributions imply outcomes frequently worse than the current belief such that a downward belief updating process sets in which drives contribution close to the money maximizing prediction. Contrary, classically Fehr and Gächter (2000), or recently Khadjavi et al. (2012) obtain stable contributions given individual type information and the option to punish players. Punishment options set strategic incentives as well but they also help to stabilize, this is an assumption, preference induced levels of cooperation which would otherwise not be stable. Overall, there is thus a trade-off. Either one uses data from one-shot encounters which may be biased due to inexperience or experimental effects, or one used data from repeated measurements where necessarily other considerations like strategic considerations or belief updating mechanisms matter and may cause other biases.

There is further work to do. As pointed out, the individual specificity found here may not be surprising given the truly different games such that an exploration of specificity across other games appears necessary. As argued, the canceling-out effect on the aggregate need not be robust given individual specificity but more evidence seems desirable with respect to the a priori possibility of stable cross game inference. Finally, while I do believe to have good reasons for a preference based interpretation, more effort to sort out strategic effects and preference based explanations, and potentially experiments designed for that particular task, are needed.

Part IV

Appendix

Appendix A

Would You Trust Yourself? - On the Long Run Stability of Reciprocal Trust

A.1 Proof of Proposition 1

Let q denote the probability that trust is rewarded: q' is the first-movers belief about q and q'' is the second-movers belief about q'. In decision node n of player 2, player one has chosen to trust with probability 1. The expected payoff for player 2 is $\pi_2(\cdot) = q''r + 1 - q''$. The expected payoff for player 1 is $\pi_1(\cdot) = q''r$ such that $\varphi_1(\cdot) = 1 - q''$. If player 2 chooses to reward, the payoff for player 1 is r such that $\sigma_2(\cdot)_{q=1} = (1 - q'')r$. If player 2 chooses to exploit, then $\sigma_2(\cdot)_{q=0} = -q''r$. Overall utilities are

$$u_2(\cdot)_{q=1} = r + \rho_2(1 - q'')^2 r$$
$$u_2(\cdot)_{q=0} = 1 - \rho_2(1 - q'')q''r$$

Setting both expressions equal, one obtains a critical value $q''_{crit} = 1 - \frac{1-r}{\rho_2 r}$. Exploitation yields a larger utility than rewarding whenever $q'' > q''_{crit}$ in which case consistency of beliefs requires $q'' = 0$. Thus, exploitation is rational whenever $\rho_2 < \frac{1-r}{r}$. If $q'' < q''_{crit}$, player 2 would reward for sure and consistency of beliefs would require $q'' = 1$. This would require that $r - 1 > 0$ which is impossible since $1 > r$. Finally, if $q'' = q''_{crit}$, player 2 is indifferent between reward and exploitation and in equilibrium $q = q'' = q''_{crit}$. This case applies for $\rho_2 \geq \frac{1-r}{r}$. Note that $q = 0$ if $\rho_2 = \frac{1-r}{r}$. Summarizing yields expression (2.1) in Proposition 1.[1]

[1] Player 2 perceives 1's action as fully intentional. First, player 2 is in the advantageous position since $\pi_2(\cdot) > \pi_1(\cdot)$. The alternative payoff for 2 is s which is strictly less than

To analyse first-mover behavior, note that the acceptance probability is q^*. The expected material reward for player 1 is either s or q^*r. Let p denote the probability that player 1 chooses to trust; p' is the belief by 2 about p and p'' is the belief held by 1 on p'. The expected payoff for player 1 is $\pi_1(\cdot) = p''q^*r + (1-p'')s$. The expected payoff for player 2 is $\pi_2(\cdot) = p''(q^*r + 1 - q^*) + (1-p'')s$ such that $\varphi_2(\cdot) = -p''(1-q^*)$. If 1 chooses not to trust, 2 simply gets s such that $\sigma_1(\cdot)_{p=0} = -p''(q^*r - s + 1 - q^*)$. Further, if 1 chooses to trust, then $\sigma_1(\cdot)_{p=1} = (1-p'')(q^*r - s + 1 - q^*)$. Overall utility sums up to

$$u_1(\cdot)_{p=1} = q^*r - \rho_1 p''(1-p'')(1-q^*)(q^*r - s + 1 - q^*)$$
$$u_1(\cdot)_{p=0} = s + \rho_1 (p'')^2(1-q^*)(q^*r - s + 1 - q^*)$$

Setting both equations equal yields $p''_{crit} = \frac{q^*r - s}{\rho_1(1-q^*)(q^*r - s + 1 - q^*)}$. If $p'' > p''_{crit}$, then the rational choice is not to trust which implies $p = 0$ in equilibrium. This is the case if $q^* < \frac{s}{r}$. If $p'' < p''_{crit}$, trust is rational such that $p = 1$ in equilibrium. Solving $1 < p''_{crit}$ for q^* yields a rather nasty square root expression with no intuition. Note, however, that if $p'' = p''_{crit}$, then, in equilibrium, $p = p'' = p''_{crit}$ such that $p^* = \min\{1, p''_{crit}\}$ whenever $q^* > \frac{s}{r}$. If $q^* = \frac{s}{r}$, then $p''_{crit} = 0$ which is associated with no trust. Summarizing yields expression (2.2) in Proposition 1.[2]

any mixing over $r > s$ and $1 > s$ such that 1 is kind and could have been less kind. This corresponds to case (a) of the Ω-function in the Appendix to Falk and Fischbacher (2006).

[2]Player 1 will always consider player 2's behavior as fully intentional. First note that if $p'' = 0$, then player 1 does not expect 2 to move and it does not matter whether he is kind or unkind. If $p'' > 0$, note that 2 is in the advantageous position for any $q' > 0$ such that 1 judges him as unkind. However, player 2 could play $q = 1$ which would be less unkind but still does not move 2 in the disadvantageous position because both players would then receive an equal split of r. Thus, 2 is unkind and has a true alternative to be less unkind. This corresponds to case (c) of the Ω-function in the Appendix of Falk and Fischbacher (2006).

Appendix B

The Co-Evolution of Reciprocity-Based Wage Offers and Effort Choices

B.1 Functional Expressions

If $\rho_E = 0$, then

$$\tilde{w}(\alpha, 0, \rho_2) = \frac{-4\alpha^2 - 4\alpha\rho_W + 3\rho_W^2}{8\alpha\rho_W^2} \tag{B.1}$$

If $\rho_E > 0$, then

$$\tilde{w}(\alpha, \rho_E, \rho_W) = \frac{3}{16\alpha} + \frac{3}{4\rho_E} + \frac{3\alpha}{4\rho_W^2} + \frac{3}{4\rho_W} + \frac{\rho_W}{8\alpha\rho_E} + \frac{\rho_W^2}{16\alpha\rho_E^2} - \ldots$$

$$\frac{(6\alpha\rho_E + 3\rho_E\rho_W + \rho_W^2)\sqrt{4\alpha^2\rho_E^2 + 4\alpha\rho_E^2\rho_W + 12\alpha\rho_E\rho_W^2 + \rho_E^2\rho_W^2 - 2\rho_E\rho_W^3 + \rho_W^4}}{16\alpha\rho_E^2\rho_W^2} \tag{B.2}$$

Appendix C

Inconsistent People? An Experiment on the Impact of Social Preferences Across Games

C.1 Instructions

Experimental Instructions

Welcome and thank you for showing up. You will now participate in an experiment on decision making. Since the experiment has now begun, we ask you to stop communicating and to turn off your mobile phones. If you have questions, either now or during the experiment, please raise your hand.

Please follow the instructions of the experimental team at all times. If you do not follow the instructions or disrupt the experiment in any other kind, we can ask you to leave and you will receive no payment.

General Course of the Experiment

First, we would like to explain the envelope that you chose. The envelope contains three sheets. One is a copy of these general instructions. On the

second you find a password and a personal id-code. You need the password to log on to the system. Following a welcome screen, you are asked to type in your id-code. Your id-code is saved together with your decisions and is very important to determine your earnings. The third sheet will serve as a receipt for your earnings. Please do not write your id-code on the receipt and do not write your name on the sheet with the id-code.

You will encounter 6 different decision situations during the experiment. 5 of the 6 decision situations require choices in two different positions, that is, you will have to make decisions in 11 situations. Some of the situations require up to three decisions given three possible actions of the other person. The total number of decisions is 18.

The decision situations will appear one after the other on your screen. In principle, each situation is composed of a description, 2 control questions, a position assignment and your choice. The first screen contains the description and the control questions. You have to answer the control questions correctly in order to proceed. If an answer to a control question is wrong, another screen will appear containing a hint toward answering the questions. If you need further advice, please raise your hand and we will help you. Some later decisions will not contain control questions given that the situation has been finished in one position already. Upon answering the control questions correctly, you will encounter a second screen. It repeats the description of the situation and you will be informed in which position you are asked to make your choice. All possible choices are listed and you have to mark the box next to your choice. Please note that answers may be unsorted. Once all decision situations are finished, some questions regarding your age, gender, etc. follow as well as some statements which you can approve more or less.

Some decision situations are similar, but they are all different. Please read the instructions carefully. You will encounter each situations exactly once in each position, that is, 5 of the 6 situations appear twice but once you have to decide in one position and then in the other. Your answers have no influence on the number of decisions. Further, your answers do not affect the possible answers in other situations, i.e. all choices are independent of each other. In addition to that, the descriptions of each decision situation are complete. That is, the descriptions contain all relevant information. Also, the information provided here is complete. Neither during the experiment, nor later, you will encounter new information nor will be asked to do something different. We do not cheat on you.

Payment

Once all participants have finished the experiment, your payments will be determined in a four step process. In step one, all participants are matched in pairs, that is you and one other person will be a pair. You will not know who the other person is. In step two, each pair gets assigned to exactly one game relevant for payment. In step three, positions in that game are assigned. In step four, your and the other persons' payoff is determined given your choice and the other persons' choice in the respective game. Payments are recorded together with the id-codes and transferred to the payment office. Please note

- At the end of the experiment, exactly one of your decisions is relevant for payment. It is not clear yet, which one it will be.
- Upon payment you can receive the information which game was assigned to you, which position you were in and how your payoff realized. Nevertheless, you will not get to know the other person in your pair and that person does not get to know you.
- The payment will be done by a different person at a separate office. There you will have to turn in the sheet with your id-code and the receipt. Since your id-code and name are separated, it is not possible to match your name with your choices.

Since you have shown-up on time, you will receive a guaranteed payment of 5 Euros. You can earn up to 15 additional Euros during the experiment. That is, your minimal payment is 5, your maximal payment is 20 Euros. Your payment depends on your choices and/or on the choices of the other person in your pair. In the decision situations, the word "tokens" is used rather than Euros. The exchange rate is 100 token = 10 Euro. If, for example, you and the other person each earn 50 tokens during the experiment, then each of you will receive 5 Euros for showing up plus 5 Euros experimental earnings, that is, each person will receive a total of 10 Euros.

Finally, if you finish the decision situations before all others, please remain at you seat and be calm. We will inform you once all participants have finished the decision situations. The calculation of payoffs and documentation will take about 5 minutes. Afterwards we will ask you to head to the payment office.

Please open the envelopes and start with the decision tasks. Thank You.

C.2 Decision Situations

Dictator Game

There are two persons A and B. Person A is asked to allocate 100 tokens between himself and person B. Person B is not making a decision, i.e. person B simply receives the amount person A allocates to her.

Decision: You are in position A, i.e. person A. Which allocation do you choose?

___ Person B: 0 token, You: 100 token
___ Person B: 10 token, You: 90 token
...

Ultimatum Game

There are two persons A and B. Person A is asked to allocate 100 tokens between himself and person B. Person B has to decide whether to accept or reject the allocation. If person B accepts the proposal, person A and B each receive the proposed amount. If person B rejects the proposed allocation, each person gets 0 token.

Person B has to state the minimal amount that must be allocated to her for acceptance. If A allocates exactly the amount which B states as minimal, the proposal is reckoned as accepted.

Decision: You are person A. Which allocation do you propose?

___ Person B: 0 token, You: 100 token
___ Person B: 10 token, You: 90 token
...

or

Decision: You are person B. What is the minimal amount that person A needs to allocate to you in order for you to accept the allocation.

___ 0 token
___ 5 token
...
___ 50 token

Trust Game

There are two persons A and B. Person A has 50 tokens. Person A can send either 0, 30, or 50 tokens to person B. The amount sent by person A is triplet. Person B has to decide on how to allocate the triplet sum between person A and herself, person B.

Decision 1: You are person B. If person A sends 30 tokens, you can allocate 90 tokens between yourself and person A. Which allocations do you choose?

___ Person A: 0 token, You: 90 token
___ Person A: 10 token, You: 80 token
...
___ Person A: 45 token, You: 45 token
...

Decision 2: You are person B. If person A sends 50 tokens, you can allocate 150 tokens between yourself and person A. Which allocations do you choose?

___ Person A: 0 token, You: 150 token
___ Person A: 10 token, You: 140 token
...
___ Person A: 75 token, You: 75 token
...

Gift-Exchange Game

There are two persons A and B. Person A and B each have 50 tokens. Person A can pay an amount w of either 0, 30, or 50 tokens to person B. Person B receives the money and then decides among a total of 10 alternatives e. The greater the alternative person B picks, the greater the final payment for person A. On the other hand, the payment for person B is reduced depending on the alternative chosen. The following table provides an overview on the alternatives available to person B and the associated costs $c(e)$.

e	1	2	3	4	5	6	7	8	9	10
$c(e)$	0	2	4	6	8	11	14	17	21	25

Final payments are calculated as follows. For person A, the final payment is $50 - w + 7,5e$ and for person B the final payment is $50 + w - c(e)$.

Decision 1: You are person B. Person A paid an amount of 0 token. Which e do you choose?

___ $e = 1$, Person A: 57,5 token, You: 50 token
___ $e = 2$, Person A: 65,0 token, You: 48 token
...

Decision 2: You are person B. Person A paid an amount of 30 token. Which e do you choose?

___ $e = 1$, Person A: 27,5 token, You: 80 token
___ $e = 2$, Person A: 35,0 token, You: 78 token
...

Decision 3: You are person B. Person A paid an amount of 50 token. Which e do you choose?

___ $e = 1$, Person A: 7,5 token, You: 100 token
___ $e = 2$, Person A: 15,0 token, You: 98 token
...

Prisoners' Dilemma Game

There are two persons A and B. Each person has to decide between three alternatives $K1, K2, K3$. Person A decides first. Person B observes the decision by person A and then decides. The following table summarizes the payoffs for person A and B for each possible combination of alternatives. The first entry in brackets is the payment for person A, the second entry is the payment for person B.

		Player B		
		K1	**K2**	**K3**
Player A	**K1**	(25,25)	(85,15)	(150,0)
	K2	(15,85)	(50,50)	(125,25)
	K3	(0,150)	(25,125)	(75,75)

Decision 1: You are person B. Which alternative do you choose if person A has chosen $K1$?

___ $K1$, Person A: 25 token, You: 25 token
___ $K2$, Person A: 85 token, You: 15 token
___ $K3$, Person A: 150 token, You: 0 token

Decision 2: You are person B. Which alternative do you choose if person A has chosen $K2$?

___ $K1$, Person A: 15 token, You: 85 token

...

Decision 3: You are person B. Which alternative do you choose if person A has chosen $K3$?

___ $K1$, Person A: 25 token, You: 25 token

...

Third Party Punishment Game

There are three persons A, B, and C. Person A has 100 token and is asked to split that amount between himself and person B. Person A can either allocate 0, 20, or 50 tokens to person B. Person B makes no decision and receives exactly the amount allocated to her by person A. Person C has 50 tokens and observes the allocation by person A. Then person C can decide to assign deduction points to person A. Each deduction point reduces the payment for person C by 1 token but reduces the payment for person A by 3 token. The minimal final payment for person A is 0 token.

Decision 1: You are person C. Person A allocated 0 token to person B. How many deduction points do you assign to person A? ___ 0 deduction points, payments: A=100, B=0, You=50 token

___ 5 deduction points, payments: A=85, B=0, You=45 token

...

___ 35 deduction points, payments: A=0, B=0, You=15 token

Decision 2: You are person C. Person A allocated 20 token to person B. How many deduction points do you assign to person A? ___ 0 deduction points, payments: A=80, B=20, You=50 token

___ 5 deduction points, payments: A=65, B=20, You=45 token

...

___ 30 deduction points, payments: A=0, B=20, You=20 token

Decision 3: You are person C. Person A allocated 50 token to person B. How many deduction points do you assign to person A? ___ 0 deduction points, payments: A=50, B=50, You=50 token

___ 5 deduction points, payments: A=35, B=50, You=45 token

...

___ 35 deduction points, payments: A=0, B=50, You=30 token

Appendix D

Game Specific Social Preferences: Different Types and a Canceling-Out Effect

D.1 Instructions

Experimental Instructions

Welcome to the experimental lab and thank you for showing up. You will now participate in an economic experiment and, dependent on your decisions and/or dependent on the decisions of the other participants, you can earn a non-negligible amount of money.

Please switch off your mobile phones and do not communicate during the experiment. If you have any questions, please raise your hand through the curtain. We will help you individually. Please obey the instructions by the experimental team at all times. Any violation of the instructions or any other disruption of the experiment will cause your exclusion from the experiment and the exclusion from all payments.

General Course of the Experiment

This experiment consists of two parts. Both parts are independent of each other. Especially, your earnings from part 1 and part 2 are independent of

each other. The instructions for part 2 will be read once part 1 is finished. At the end of part 2, we will further ask you a couple of questions regarding your age, gender, subject of study, etc.

In part 1 of the experiment, you will face 3 different decision situations. In each decision situation, two people interact. Each person will be allocated to one of two positions. There is a total of $3 \times 2 = 6$ positions. In part 1, you will face each position exactly once.

The respective situation, your position in this situation, and your decision options will be displayed at your computer screen. Once all participants made their decisions, the next position assignment follows until all participants have finished all positions.

All attendees are split into two groups with 12 participants each. In each position, you will encounter another person of your group. That means, you will never interact with the same person twice. You will encounter each decision situation twice, since each situation consists of two positions, but the respective other person will not be the same. Given the above restrictions, your actual matching partner is randomly selected by the computer.

All your decisions stay anonymous. You will not get to know your matching partners and they will not know who you are. Further, all matchings' remain secret at all later times, especially in part 2 of the experiment.

Payment

All situations deal with the assignment of money amounts. All amounts are measures in so called *ECU* (Experimental Currency Unit). The rate of exchange is $100\ ECU = 10 EURO$.

At the end of part 1, one person will be asked to randomly choose one out of six cards containing the numbers from 1 to 6. The selected number determines which of the six position assignments is relevant for payment. If, for example, the 1 is selected, your choice in the first position assignment is relevant for payment. If 3 is selected, then the third position assignment is relevant, and so on. For each person, the payment is then calculated dependent on the specific position and dependent on the own decision, the other persons' decision, or on both decisions.

The actual payment will take place at the end of part 2. Beforehand, you will not receive any information regarding your earnings in part 1. Your final

payment will be the sum of earnings in part 1 and part 2. Again, note that the payment for each part is independent of the respective other part of the experiment. Especially, you cannot lose your earnings from part 1 in part 2.

The payment at the end of the experiment is not done by a member of the experimental team. Rather, a laboratory staff member will receive a list with your seat number and your earnings and will pay you at the separate office next to the laboratory. By that procedure, your identity is entirely kept private, i.e. neither can we match your name with your decisions, nor do we see how much you earn.

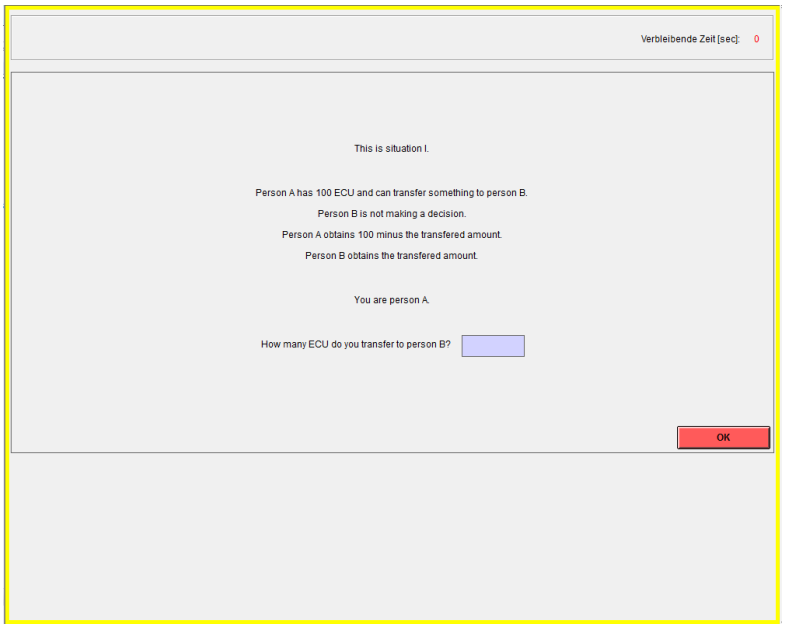
Decision Situations

We now explain the three different decision situations. A description of each situation will also appear on your computer screen during the experiment.

The computer screen is separated into three parts. At the upper boundary, you can see how much time you have left for your decision. The field in the middle of the screen contains information regarding the situations you currently face (including a description of that situation) and regarding your position in that situation. Further, you make your decisions in that field by either entering integer amounts or by selecting your choice from a list. The field at the lower boundary is empty.

Please notice that you always need to hit the "OK"-button to approve your decision.

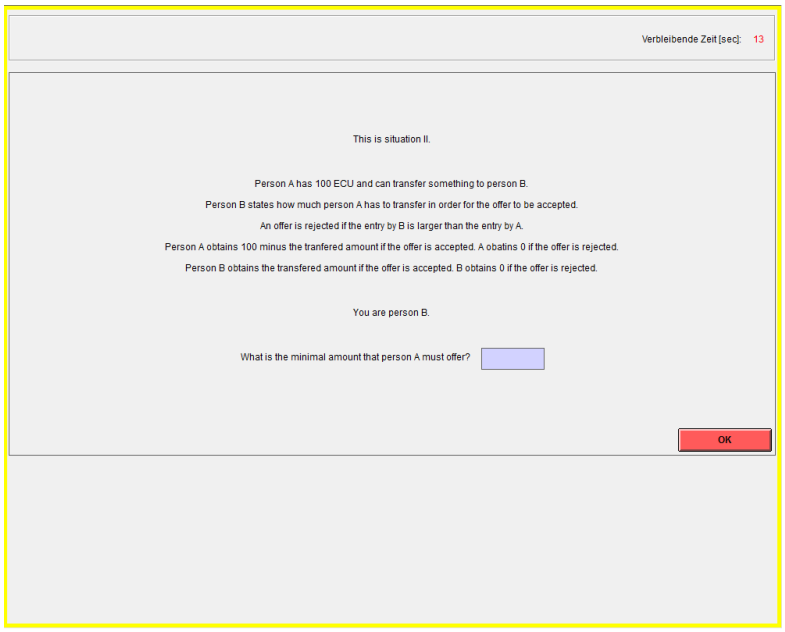
Situation 1



At first, you can observe that you are in situation 1. Second, the description of situation 1 follows. Third, you get to know your position. In the figure, you are person A which, in this case, is the only person actually making a choice. You enter your decision in the blue field. For person B, there appears no field to enter a decision in this case.

Payment Example: If person A enters 100, then person B obtains 100 ECU and person A keeps 0 ECU. If person A enters 35 ECU, than person B obtains 35 ECU and person A keeps 65 ECU, and so on.

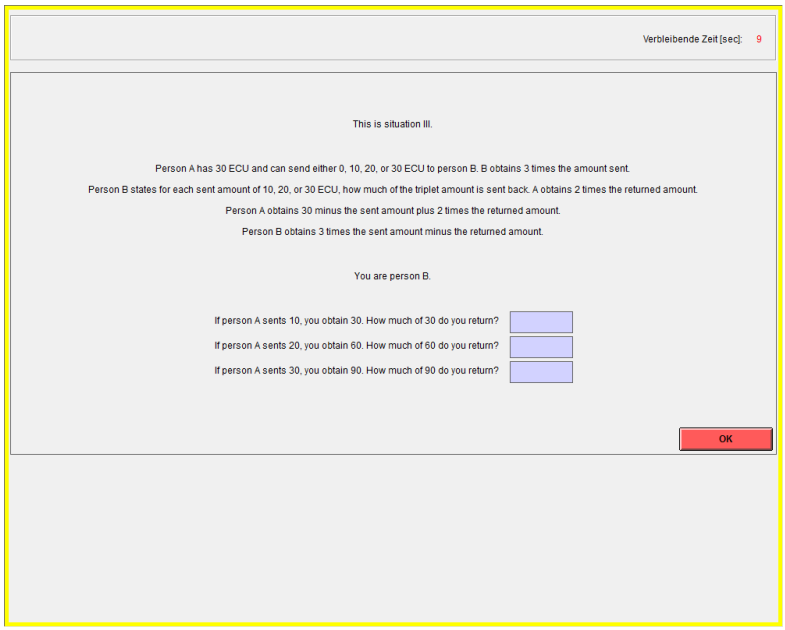
Situation 2



The decision by person A in situation 2 is almost identical to the decision by person A in situation 1. In this case, however, person B states how much person A has to offer such that the offer is accepted. Whenever the amount stated by B is larger than the amount offered by A, both players receive 0 ECU.

Payment Example: Suppose person A transfers 30 ECU and person B enters 0 ECU as the minimal acceptable offer. In this case, the choice by B is lower than the choice by A. The offer is accepted and person B obtains 30 ECU while person A keeps 70 ECU. If person B enters 30 ECU as the minimal acceptable offer, the offer is also accepted since the entry by B is not larger than the entry by A. If, however, person B enters e.g. 50 ECU, the offer is rejected and both person A and person B receive 0 ECU.

Situation 3



Person A is provided with a list with the amounts 0, 10, 20, and 30 ECU and selects which amount is transferred to person B. Person B has to make a decision for each possible transfer by A which allows a decision by B. While making the decisions, person B does not know which amount is actually transferred. The triplet amount sent by A, for each case, is shown next to the fields were the decision is entered. Please recall, however, that the returned amount is additionally doubled for A.

Payment Example: Suppose A sent 10 ECU and B stated an amount to return of 10 ECU for this case. Person A will obtain a payment of 30 (endowment) - 10 (sent) + 2 $\times$ 10 (returned) = 40 ECU. Person B will obtain a payment of 3 $\times$ 10 (received) - 10 (returned) = 20 ECU.

Suppose A sent 20 ECU and B stated an amount to return of 10 ECU for this case as well. Person A will obtain a payment of 30 (endowment) - 20 (sent) + 2 $\times$ 10 (returned) = 30 ECU. Person B will obtain a payment of 3 $\times$ 10 (received) - 10 (returned) = 50 ECU.

Part 2

After part 1 is finished, we now proceed with part 2 of the experiment. The main differences are as follows:

- You will again encounter the three decision situations faced in part 1. In this part, however, you will play many rounds, i.e. you will face each position multiple times.
- You will receive information regarding the previous behavior of you current matching partner and your matching partner receives information regarding your previous behavior.
- You will receive feedback regarding your and your co-players' payoff after each decision.

The rounds

Part 2 consists of many rounds. In every round, you will face each of the three decision situations, but only one of both possible positions in each situation. In every round, you face three decisions.

Which three positions you encounter is randomly determined. Also, your matching partner for each decision is randomly determined. It is secured, however, that you never face the same person twice within one round.

Once a round is over, the next one starts. The experiment ends once the criterion of a stopping rule is satisfied. This criterion is related to the number of decisions by each person in each position. Since the position assignments are random, however, it is not yet determined in which round the experiment ends.

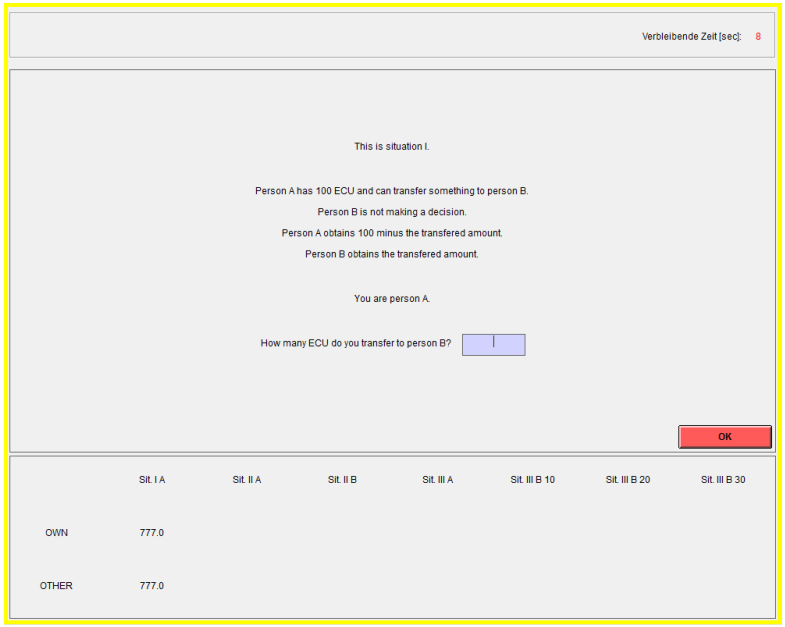
Information

In this part of the experiment, an information matrix appears in the field at the lower boundary of your screen. The information matrix generally contains entries for YOU and the respective OTHER person for each of the

three possible decision situations. Actually displayed is the information with respect to the current decision situation.

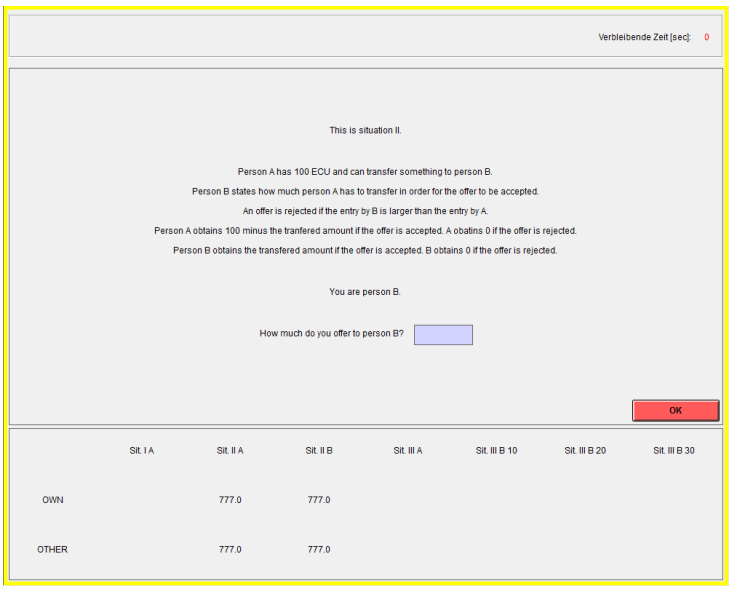
Displayed is always the average behavior with respect to the last three decisions in a position. Whenever this average cannot be calculated, since a person made less than 3 decisions in a position, 777 is displayed (i.e. at least in the first three rounds, there is always 777 displayed). Once you made more than three decisions in a specific position, older decisions will no longer enter the calculation of the average.

Example 1:



In this case you face situation 1. The displayed part of the information matrix contains entries only for Sit. I A, since person B does not make a decision in situation 1. Nevertheless, you can observe the average over the last three choices of yourself as person A in situation 1 as well as the respective average of your current matching partner (all examples always contain 777 entries).

Example 2:
In this case you are in situation 2 and person A. You can observe the respective averages of behavior for both positions. As person A you can observe how much your current matching partner (row OTHER) offered as person A on average (column Sit. 2 A) as well as her or his average minimal acceptable offer over the last three decisions made as person B (column Sit. II B).



Example 3:

In this case you are in situation 3 and person A. Since person B decides for each amount sent, either 10, 20, or 30 ECU, there is a total of four columns displayed (person A + 3 $\times$ person B).

Payment

At the end of the experiment, exactly one round will be relevant for payment. Note that you face each situation in each round such that each situation is paid exactly once. Within the randomly chosen round relevant for payment, the earnings from each situation are added up. The rate of exchange is again $100\ ECU = 10 EURO$.

Then, we sum your earnings from part 1 and part 2 of the experiment. Since you showed up on time and since the experiment will last up to two hours, there is an additional minimal payment of 10 Euro. That is, whenever the sum of earnings from part 1 and part 2 is less than 10 Euro, you will nevertheless receive 10 Euros.

You will have to sign a receipt for you earnings. You will be informed about your earnings by the staff member of the laboratory once you enter the payment room. At the end of the experiment, we will ask you to proceed to the payment room one after another.

Part V

Bibliography

Bibliography

Akerlof, G. A. (1982). Labor Contracts as Partial Gift Exchange. *Quarterly Journal of Economics 97*, 543–569.

Al-Ubaydli, O., U. Gneezy, M. S. Lee, and J. A. List (2010). Towards an Understanding of the Relative Strengths of Positive and Negative Reciprocity. *Judgment and Decision Making 5*, 524–539.

Altmann, S., T. Dohmen, and M. Wibral (2008). Do the Reciprocal Trust Less? *Economics Letters 99*, 454–457.

Andreoni, J., M. Castillo, and R. Petrie (2003). What do Bargainers' Preferences Look Like? Experiments with a Convex Ultimatum Game. *American Economic Review 93*, 672–685.

Andreoni, J. and J. Miller (2002). Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism. *Econometrica 70*, 737–753.

Arrow, K. J. (1974). *The Limits of Organization*. Norton.

Axelrod, R. M. (1984). *The Evolution of Cooperation*. Basic Books.

Bellemare, C., S. Kröger, and A. van Soest (2008). Measuring Inequity Aversion in a Heterogeneous Population using Experimental Decisions and Subjective Probabilities. *Econometrica 76*, 815–839.

Berg, J., J. Dickhaut, and K. McCabe (1995). Trust, Reciprocity, and Social History. *Games and Economic Behavior 10*, 122–142.

Berninghaus, S. K., C. Korth, and S. Napel (2007). Reciprocity - An Indirect Evolutionary Analysis. *Journal of Evolutionary Economics 17*, 579–603.

Binmore, K. and A. Shaked (2010). Experimental Economics: Where Next? *Journal of Economic Behavior and Organization 73*, 87–100.

Blanco, M., D. Engelmann, and H. T. Normann (2011). A Within-Subject Analysis of Other-Regarding Preferences. *Games and Economic Behavior 72*, 321–338.

Bolle, F. (1998). Rewarding Trust: An Experimental Study. *Theory and Decision 45*, 83–98.

Bolton, G. E. and A. Ockenfels (2000). ERC: A Theory of Equity, Reciprocity, and Competition. *American Economic Review 90*, 167–193.

Borgloh, S., A. Dannenberg, and B. Aretz (2010). On the Construction of Social Preferences in Lab Experiments. *ZEW Discussion Paper 10-085*.

Brandts, J. and G. Charness (2011). The Strategy Versus the Direct-Response Method: A First Survey of Experimental Comparisons. *Experimental Economics 14*, 375–398.

Brook, M. J., A. Lange, and E. Y. Ozbay (2012). Dictating the Risk - Experimental Evidence on Giving in Risky Environments. *American Economic Review forthcoming*.

Brosig, J., T. Riechmann, and J. Weimann (2007). Selfish in the End? An Investigation of Consistency and Stability of Individual Behavior. *FEMM Working Paper 5*.

Brosig, J., J. Weimann, and C.-L. Yang (2003). The Hot Versus Cold Effect in a Simple Bargaining Experiment. *Experimental Economics 6*, 75–90.

Burks, S., J. P. Carpenter, and E. Verhoogen (2003). Playing Both Roles in Trust Games. *Journal of Economic Behavior and Organization 51*, 195–216.

Camerer, C. F. (2003). *Behavioral Game Theory*. Princeton University Press.

Camerer, C. F. (2011). *The Promise and Success of Lab-Field Generalizability in Experimental Economics: A Critical Reply to Levitt and List*. working paper.

Charness, G. and M. Rabin (2002). Understanding Social Preferences with Simple Tests. *Quarterly Journal of Economics 117*, 817–869.

Clarc, K. and M. Sefton (2001). The Sequential Prisoners' Dilemma: Evidence on Reciprocation. *Economic Journal 111*, 51–68.

Cox, J. C., D. Friedman, and V. Sadiraj (2008). Revealed Altruism. *Econometrica 76*, 31–69.

Cressman, R. (2003). *Evolutionary Dynamics and Extensive Form Games*. MIT Press.

Dal Bó, P. and G. R. Fréchette (2011). The Evolution of Cooperation in Infinitely Repeated Games: Experimental Evidence. *American Economic Review 101*, 411–429.

Dawes, R. (1989). Statistical Criteria for Establishing a Truly False Consensus Effect. *Journal of Experimental Social Psychology 25*, 1–17.

de Oliveira, A., R. Croson, and C. Eckel (2008). *Are Preferences Stable Across Domains? An Experimental Investigation of Social Preferences in the Field*. working paper.

Dohmen, T., A. Falk, D. Huffman, and U. Sunde (2009). Homo Reciprocans: Survey Evidence on Behavioral Outcomes. *Economic Journal 119*, 592–612.

Doris, J. M. (2002). *Lack of Character*. Cambridge University Press.

Dufwenberg, M. and G. Kirchsteiger (2004). A Theory of Sequential Reciprocity. *Games and Economic Behavior 47*, 268–298.

Engelmann, D. and M. Strobel (2004). Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments. *American Economic Review 94*, 857–869.

Falk, A., E. Fehr, and U. Fischbacher (2008). Testing Theories of Fairness - Intentions matter. *Games and Economic Behavior 62*, 287–303.

Falk, A. and U. Fischbacher (1998). A Theory of Reciprocity. *IEER working paper 6*.

Falk, A. and U. Fischbacher (2006). A Theory of Reciprocity. *Games and Economic Behavior 54*, 293–315.

Farwell, L. and B. Weiner (1996). Self-Perceptions of Fairness in Individual and Group-Contexts. *Personality and Social Psychology Bulletin 22*, 867–881.

Fehr, E. and U. Fischbacher (2004). Third-Party Punishment and Social Norms. *Evolution and Human Behavior 25*, 63–87.

Fehr, E. and S. Gächter (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review 90*, 980–994.

Fehr, E. and S. Gächter (2002). Do Incentive Contracts Undermine Voluntary Cooperation? *IEER working paper 34*.

Fehr, E., S. Gächter, and G. Kirchsteiger (1997). Reciprocity as a Contract Enforcement Device: Experimental Evidence. *Econometrica 65*, 833–860.

Fehr, E., A. Klein, and K. M. Schmidt (2007). Fairness and Contract Design. *Econometrica 75*, 121–154.

Fehr, E., S. Kremhelmer, and K. M. Schmidt (2005). Fairness and the Optimal Allocation of Ownership Rights. *Economic Journal 118*, 1262–1284.

Fehr, E. and K. M. Schmidt (1999). A Theory of Fairness, Competition, and Cooperation. *Quarterly Journal of Economics 114*, 817–868.

Fehr, E. and K. M. Schmidt (2004). Fairness and Incentives in a Multi-Task Principle-Agent Model. *Scandinavian Journal of Economics 106*, 453–474.

Fehr, E. and K. M. Schmidt (2006). The Economics of Fairness, Reciprocity, and Altruism - Experimental Evidence and New Theories. In S.-C. Kolm and J. M. Ythier (Eds.), *Handbook of the Economics of Giving, Altruism and Reciprocity - Volume 1*. Elsevier, North-Holland.

Fehr, E. and K. M. Schmidt (2010). On Inequity Aversion: A Reply to Binmore and Shaked. *Journal of Economic Behavior and Organization 73*, 101–108.

Fischbacher, U. (2007). z-Tree: Zurich Toolbox for Ready-Made Economic Experiments. *Experimental Economics 10*, 171–178.

Fischbacher, U. and S. Gächter (2010). Social Preferences, Beliefs, and the Dynamics of Free Riding in Public Goods Experiments. *American Economic Review 100*, 541–556.

Fischbacher, U., S. Gächter, and E. Fehr (2001). Are People Conditionally Cooperative? Evidence from a Public Goods Experiment. *Economics Letters 71*, 397–404.

Fisman, R., S. Kariv, and D. Markovits (2007). Individual Preferences for Giving. *American Economic Review 97*, 1858–1876.

Frey, B. S. and S. Meier (2004). Social Comparisons and Pro-Social Behavior: Testing "Conditional Cooperation" in a Field Experiment. *American Economic Review 94*, 1717–1722.

Friedman, D. (1996). Equilibrium in Evolutionary Games: Some Experimental Results. *Economic Journal 106*, 1–25.

Gächter, S. (2007). Conditional Cooperation: Behavioral Regularities from the Lab and the Field and their Policy Implication. In B. S. Frey and A. Stutzer (Eds.), *Economics and Psychology. A Promising New Cross-Disciplinary Field*. MIT Press.

Geanakoplos, J., D. Pearce, and E. Stacchetti (1989). Psychological Games and Sequential Rationality. *Games and Economic Behavior 1*, 60–79.

Greiner, B. (2004). An Online Recruitment System for Economic Experiments. In K. Kremer and V. Macho (Eds.), *Forschung und Wissenschaftliches Rechnen. GWDG Bericht 62*, pp. 59–73. Gesellschaft für Wissenschaftliches Rechnen, Göttingen.

Güth, W. and H. Kliemt (1994). Competition or Co-Operation: On the Evolutionary Economics of Trust, Exploitation and Moral Attitudes. *Metroeconomica 45*, 155–187.

Güth, W. and H. Kliemt (1998). The indirect evolutionary approach: Bridging the gap between rationality and adaptation. *Rationality and Society 10*, 377–399.

Güth, W. and S. Napel (2006). Inequality in a Variety of Games - An Indirect Evolutionary Analysis. *Economic Journal 116*, 1037–1056.

Harrison, G. W. and K. McCabe (1996). Expectations and Fairness in a Simple Bargaining Experiment. *International Journal of Game Theory 25*, 303–327.

Herold, F. and C. Kuzmics (2009). Evolutionary Stability of Discrimination under Observability. *Games and Economic Behavior 67*, 542–551.

Hoeffler, S. and D. Ariely (1999). Constructing Stable Preferences: A Look into Dimensions of Experience and their Impact on Preference Stability. *Journal of Consumer Psychology 8*, 113–139.

Huck, S. and J. Oechssler (1999). The Indirect Evolutionary Approach to Explaining Fair Allocations. *Games and Economic Behavior 28*, 13–24.

Iriberri, N. and P. Rey-Biel (2011). The Role of Role Uncertainty in Modified Dictator Games. *Experimental Economics 14*, 160–180.

Khadjavi, M., A. Lange, and A. Nicklisch (2012). *Transparency & Accountability: Substitutes or Complements?* Mimeo.

Kocher, M. G., T. Cherry, S. Kroll, R. J. Netzer, and M. Sutter (2008). Conditionally Cooperation on Three Continents. *Economics Letters 101*, 175–178.

Larrick, R. P. and S. Blount (1997). The Claiming Effect: Why Players are More Generous in Social Dilemmas than in Ultimatum Games. *Journal of Personality and Social Psychology 72*, 810–825.

Levine, D. (1998). Modeling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics 1*, 593–622.

Levitt, S. D. and J. A. List (2007). What Do Laboratory Experiments Measuring Social Preferences Tell Us About the Real World? *Journal of Economic Perspectives 21*, 153–174.

Lichtenstein, S. and P. Slovic (2006). The Construction of Preference: An Overview. In S. Lichtenstein and P. Slovic (Eds.), *The Construction of Preference*. Cambridge University Press.

McCabe, K. A., M. L. Rigdon, and V. L. Smith (2003). Positive Reciprocity and Intentions in Trust Games. *Journal of Economic Behavior and Organization 52*, 267–275.

Mullen, B., J. Atkins, D. Champion, C. Edwards, D. Hardy, J. Story, and M. Venderklok (1985). The False Consensus Effect: A Meta-Analysis of 115 Hypothesis Tests. *Journal of Experimental Social Psychology 21*, 263–283.

Oechssler, J. and F. Riedel (2001). Evolutionary Dynamics on Infinite Strategy Spaces. *Economic Theory 17*, 141–162.

Oppenheimer, J., S. Wendel, and N. Frohlich (2011). Paradox Lost: Explaining and Modeling Seemingly Random Individual Behavior in Social Dilemmas. *Journal of Theoretical Politics 23*, 165–187.

Ortmann, A., J. Fitzgerald, and C. Boing (2000). Trust, Reciprocity, and Social History: A Re-examination. *Experimental Economics 3*, 81–100.

Ostrom, E. (2010). Beyond Markets and States: Polycentric Governance of Complex Economic Systems. *American Economic Review 100*, 641–672.

Oxoby, R. J. and K. N. McLeish (2004). Sequential Decision and Strategy Vector Methods in Ultimatum Bargaining: Evidence on the Strength of Other-Regarding Behavior. *Economics Letters 84*, 399–405.

Pereira, P. T., N. Silva, and J. A. e Silva (2006). Positive and Negative Reciprocity in the Labor Market. *Journal of Economic Behavior and Organization 59*, 406–422.

Rabin, M. (1993). Incorporating Fairness into Game Theory and Economics. *American Economic Review 83*, 1281–1302.

Schliffke, P. (2010). Would You Trust Yourself? - On the Long-Run Stability of Reciprocal Trust. *Universität Hamburg, Beiträge zur Wirtschaftsforschung 170*.

Schliffke, P. (2012a). *Game Specific Social Preferences: Different Types and a Canceling Out Effect*. working paper.

Schliffke, P. (2012b). *Inconsistent People? An Experiment on the Impact of Social Preferences Across Games*. working paper.

Schotter, A. (2006). Strong and Wrong: The Use of Rational Choice Theory in Experimental Economics. *Journal of Theoretical Politics 18*, 498–511.

Schotter, A. and B. Sopher (2004). Social Learning and Coordination Conventions in Intergenerational Games: An Experimental Study. *Journal of Political Economy 111*, 498–529.

Schotter, A. and B. Sopher (2006). Trust and Trustworthiness in Games: An Experimental Study of Intergenerational Advice. *Experimental Economics 9*, 123–145.

Schotter, A. and B. Sopher (2007). Advice and Behavior in Intergenerational Ultimatum Games: An Experimental Approach. *Games and Economic Behavior 58*, 365–393.

Selten, R. (1967). Die Strategiemethode zur Erforschung des Eingeschränkt Rationalen Verhaltens im Rahmen eines Oligopolexperiments. In H. Sauermann (Ed.), *Beiträge zur Experimentellen Wirtschaftsforschung* , pp. 136–168. Tübingen: Mohr.

Singh, R., W. M. Choo, and L. L. Poh (1998). In-Group Bias and Fair-Mindedness as Strategies of Self-Presentation in Intergroup Perception. *Personality and Social Psychology Bulletin 24*, 147–162.

Stanca, L. (2010). How to Be Kind? Outcomes Versus Intentions as Determinants of Fairness. *Economics Letters 106*, 19–21.

Tversky, A. and D. Kahneman (1987). Rational Choice and the Framing of Decisions. In R. Hogarth and M. Reder (Eds.), *Rational Choice*. University of Chicago Press.

Volk, S., C. Thöni, and W. Ruigrok (2011). Temporal Stability and Psychological Foundations of Cooperation Preferences. *University of St.Gallen Discussion Paper 2011-01*.

Weber, R. A., C. F. Camerer, and M. Knez (2004). Timing and Virtual Observability in Ultimatum Bargaining and "Weak Link" Coordination Games. *Experimental Economics 7*, 25–48.

Weibull, J. W. (1996). *Evolutionary Game Theory*. MIT Press.

Zizzo, D. J. (2010). Experimenter Demand Effects in Economic Experiments. *Experimental Economics 13*, 75–98.

Eidesstattliche Versicherung:

Hiermit erkläre ich, Philipp Schliffke, an Eides statt, dass ich die Dissertation mit dem Titel *Essays on the Evolution, Stability, and Heterogeneity of Social Preferences* selbständig und ohne fremde Hilfe verfasst habe.

Andere als die von mir angegebenen Quellen und Hilfsmittel habe ich nicht benutzt. Die den herangezogenen Werken wörtlich oder sinngemäß entnommenen Stellen sind als solche gekennzeichnet.

Hamburg, 31. Mai 2012

Philipp Schliffke