Biased Beliefs and Reciprocal Behavior in Social Dilemmas

Von der Fakultät für Wirtschaftswissenschaften der
Rheinisch-Westfälischen Technischen Hochschule Aachen
zur Erlangung des akademischen Grades eines Doktors der
Wirtschafts- und Sozialwissenschaften genehmigte Dissertation

vorgelegt von

Dipl.-Ing. Anselm Hüwe

Berichter: Univ.-Prof. Dr. rer. pol. Wolfgang Breuer
Univ.-Prof. Dr. rer. pol. Christine Harbring

Tag der mündlichen Prüfung: 21.04.2015

Table of Contents

1 Introduction

We cannot understand the […] economy without having a
theory of how humans make decisions.
Herbert Gintis (2009), p. 2.

Adam Smith, who is usually considered to be the founding father of microeconomics, is most frequently cited for his image of the “invisible hand“, which expresses that everybody should intend only his own gain: By doing so, an “invisible hand” lets people simultaneously promote the public interest (Smith, 1937 [1776]). According to this concept, the existence of social dilemmas, where a conflict between self-interest and public interest exists, is denied. This “selfishness axiom” (Henrich et al., 2004) has for a long time prevailed in economic research. Typically, the selfishness axiom was (or still is) combined with the assumption of rational behavior, leading to a simple and often useful theory of human behavior (known as the homo economicus model, the neoclassical model, or the standard model). However, representations of social dilemmas, such as the public goods game (PGG), prove that selfishness is not always in the public interest. The selfishness axiom predicts that people cannot solve such social dilemmas. In contrast, experimental research in the last decades has proven that people often act in the public interest even when it causes harm to themselves. This finding, which – according to his less frequently cited book “Theory of Moral Sentiments” – even Adam Smith was aware of, has brought the understanding of social preferences for economic interactions back into the spotlight.

In addition, experimental research has discovered various deviations from the rationality assumption, which is an essential part of the homo economicus model. While giving up this assumption makes models more complicated, upholding it hampers attempts to gain important new insights. For example, an important finding of this dissertation is that social preferences and seemingly irrational biases in human thinking interact with each other. Therefore, they must be considered in an integrated way. More generally, the goal of this dissertation is to contribute to a better understanding of human decision making in social dilemmas.

To achieve this objective, we reject both the selfishness axiom and the rationality axiom. Instead, we modify the reciprocity model of Dufwenberg and Kirchsteiger (2004) (henceforth DK), and extend it by introducing biased and therefore irrational beliefs. We extract these two essential aspects of human decision making by using experimental methods: Indeed, people have reciprocal preferences, and they believe in more reciprocal behavior of others than is actually the case. We will call such people "reciprocal believers" in the following, and our results show that this belief bias only makes sense in combination with reciprocal preferences. Briefly, our findings can be summarized as follows. People

1. cooperate because they (to some degree: wrongly) believe that others cooperate as well,

2. trust because they (to some degree: wrongly) believe that others are trustworthy, and the fear of being betrayed does not diminish trusting behavior,

3. behave in a fair way because they (to some degree: wrongly) believe that they will be punished if they do not. Furthermore, fair – meaning equal – payoffs are achieved because people want to be kind to others. Reciprocating kind behavior in one situation does not necessarily mean that unkind behavior in another situation will be reciprocated as well.

The incorporation of these results into thinking about human decision making leads to a different way of designing corporations, institutions, and markets. Is it possible to auction goods anonymously and with little legal control via the internet? How much supervision is necessary to make people pay their taxes or for their bus tickets? Behavioral economists will come to very different conclusions compared to neoclassical economists: Reciprocal believers are much more successful in solving social dilemmas than homines economici, which is good news.

However, accepting that people are reciprocal believers is, at the same time, bad news because it implies that our economy is much more vulnerable than it would be if unbiased and selfish subjects were making economic transactions: Correct beliefs and selfishness are precisely defined. In contrast, biases are prone to framing. Akerlof and Shiller (2010) argue that people think with the help of "stories" (such as "house prices always rise"), and such stories can be significantly biased. Perceptions are influenceable or even manipulable positively as well as negatively (see Posten et al., 2014, to name only one example). With respect to social preferences, evolutionary analyses find oscillating or chaotic phases of cooperation and defection in social dilemmas (see Section 4.2). Thinking of people as reciprocal believers explains why economic developments can make sudden disruptions unjustified by "hard facts". Three of the five irrational "animal spirits" which drive human behavior according to Akerlof and Shiller (2010) (trust, fairness, biased beliefs) are addressed in this dissertation (we do not address corruption and money illusion), and their link between economic decision making of humans and macroeconomic phenomena, such as financial crises, illustrates how important it is to understand human decision making in social dilemmas much better than has been the case up to now.

2 Economic Experiments

Economists (unfortunately)… cannot perform the controlled experiments of chemists or biologists because they cannot easily control other important factors. Like astronomers or meteorologists, they generally must be content largely to observe.
Paul A. Samuelson and William D. Nordhaus,
Principles of Economics 12th ed. McGraw-Hill,
New York 1985, p. 8

Experimental economics is an 'exciting new development'.
Paul A. Samuelson and William D. Nordhaus,
Principles of Economics 14th ed. McGraw-Hill,
New York 1992, p. 5

Science can be theoretical, observational, or experimental. In economics, research has for a long time not been experimental, and it was believed that it is not possible to conduct economic experiments. While other (natural) sciences began to conduct experiments much earlier (starting with physics in the time of Galileo, followed by chemistry, biology, and no more than about one hundred years ago, psychology, compare Friedman and Sunder, 1994), the first economic experiments were not made until the second half of the last century. Since the mid-1980s, the number of experimental papers has greatly increased, but still account for no more than about 4 % of all the papers in economic top journals (Falk, 2009). Economic experiments are typically conducted in a computer laboratory, but they can also be made "in the field", in a universitarian classroom, or in a clinic using neuroscientific devices.

Experiments serve as a midpoint between theory and praxis: Elaboration of the theory, empirical phenomena, and laboratory experiments play complementary roles (Harstad and Selten, 2013): In contrast to empirical methods, experiments are able to allow tightly controlled variations of decision environments. "Controlled" means that "most factors which influence behavior are held constant and only one factor of interest (the 'treatment') is varied at a time" (Croson and Gächter, 2010). This tight control allows to make causal inferences between the explanatory factor and the dependent behavioral impact. Factors of relevance are typically available choices, decision makers' information sets, and the monetary incentive structure.

Theory is useful for designing experiments, because it tells the experimentator which variables to control or to test and it makes predictions about the experimental outcome. (Harstadt and Selten, 2013). Like theories, experiments cover (only) the most important aspects of economic behavior. Furthermore, like theories, experiments are abstract and accept descriptive inaccurateness (Croson and Gächter 2010), which is sometimes criticized as lacking realism.

Experiments are ideal for testing existing theories: Smith (1962) showed that the neoclassical model of competitive markets can be reproduced in experiments. Similarly, in our third paper "Using the Carrot Like the Stick? Theoretical and Experimental Insights Into Positive vs. Negative Reciprocity", henceforth UG-paper, we test the implication of the strong reciprocity model, which proposes that people who reward kind behavior also punish unkind behavior (we find this prediction to be not true). As in these examples, experiments allow theories an existence proof: If their predictions are not accurate in a carefully designed experimental environment, it is unlikely that they explain real phenomena in a better way. Testing theory with the help of observational data is not as easily done as using experiments, because the former approach jointly has to test whether the assumptions of the theory and its predictions hold. In contrast, in the lab, assumptions can be controlled very precisely, and only predictions have to be tested. Furthermore, experiments can measure and control *all* relevant variables, which observations cannot. Moreover, observational data can only prove comparative statics of a theory, while experimental data can serve to make point predictions (Croson and Gächter, 2010).

In turn, experimental results inspire new theories, such as the outcomes of UG experiments have led to the emergence of social preference models. In this way, on the basis of our experimental data, this dissertation supports the model of people as being reciprocal believers. Thus, theories are tested in experiments, and experiments provide new insights to formulate better theories. This dialectic process (Croson and Gächter, 2010, Friedman and Sunder, 1994) serves to capture relevant empirical phenomena.

Falk (2009) claims that "lab experiments are a major source of knowledge in the social sciences". First, experimental research can show how actual behavior deviates from neoclassical assumptions (which is especially of interest in German research, as most German experimentalists have been inspired by Reinhard Selten, who dedicates his research to exploring bounded rationality). Thereby, experiments can measure preferences (to which degree are people considering the well-being of others?) and parameters (what is the discount rate in an agent's time preference?). For the most part, this dissertation is dedicated to such questions. Second, experiments can test and improve institutions (so-called economic engineering): Aimone and Houser (2013) explain how well-functioning institutions can be built, bearing in mind that people may not trust others because they want to avoid the negative emotional feelings associated with the knowledge of being betrayed, known as betrayal-aversion. Andreoni and Samuelson (2006) show how formal and informal institutions can help to build up cooperation. Third, experiments can illustrate phenomena (people reward and punish, share equally, and their behavior depends on norms and culture). For that purpose, "standard" experiments have evolved which serve as behavioral models of basic social dilemmas: The public goods game (PGG) is the standard example of situations where collective self-interested behavior does not maximize the overall welfare; the trust game (TG) illustrates situations where contracts are incomplete; the ultimatum game (UG) displays how a social surplus is divided if one party has the whole bargaining power.

Some scientists argue that results of laboratory experiments must be viewed critically, because they are derived in a "non-realistic" environment (Levitt and List, 2007). This skepticism towards the external validity of experiments is not new: Galileo's critics did not believe that the motion of pendulums or balls had any relation to planetary motion (Friedman and Sunder, 1994). Experimental results can only offer inductive logic, meaning that one has to hope that behavioral regularities will persist outside of the lab as long as the relevant underlying conditions (formulated in theories) remain substantially unchanged.

To address concerns about external validity, experimentalists have started to compare behavior in the lab with behavior in the "real world" and found substantially consistent behavior. To name only a few: Baran, Sapienza and Zingales (2010) find that Chicago MBA students who return more to the sender in the proposer role of a TG played in their class at the beginning of their program also donate more to the university at the end of their program 18 months later. Franzen and Pointner (2013) measure the external validity of giving in the dictator game by sending "misdirected" letters with money to the experimental participants some weeks / two years after the experiment was conducted. Indeed, subjects who gave more in the dictator game were more likely to return the letter. Karlan (2005) finds that subjects who are more trustworthy in a TG are more likely to repay loans one year later.

As well, external validity can be addressed by testing the parameters of concern within the lab. In this context, Falk (2009) proposes that more experiments should be made instead of fewer: If you are afraid that unexperienced students behave differently from experienced experts, then invite experienced experts to the lab. If you believe that small payoffs do not capture decisions over large stakes, then increase the payoffs. If you doubt that small samples have enough statistical power, then raise the sample size (and so on). Indeed, for example, for the UG, all these points have been addressed in experiments, and the experiments have proven that the basic outcome of the ultimatum game (most proposers offer fair splits, which are accepted; unfair splits are often rejected) is remarkably robust (Samuelson, 2005).

When conducting economic experiments, certain rules of the economics discipline must be followed. These rules are different from those for psychological experiments because the questions that economists are interested in are different (Croson, 2005). Most importantly, economic experiments must establish a link between elicited decisions and monetary incentives. This means that subjects must be paid depending on their decisions, and they must understand the incentive structure. There is evidence that behavior differs without incentives, and one does not want to measure what people say but what they do. From this claim, it follows that subjects must not be deceived: Participants must deeply believe that their behavior is linked to payoffs, as explained. This belief is a public good which experimental economists carefully prevent from being exploited by individual researchers.

3 Game Theory and (Common) Belief in Rationality

Und er kommt zu dem Ergebnis:
"Nur ein Traum war das Erlebnis.
Weil", so schließt er messerscharf,
„nicht sein kann, was nicht sein darf!"
Christian Morgenstern, Die unmögliche Tatsache (1910)

To understand and predict economic phenomena and to give advice, scientists build models. Models which describe behavior in social dilemmas necessarily use *game theory*, because "game theory is about what happens when people – or genes, or nations – interact" (Camerer, 2003). In such models, one has to specify how outcomes are evaluated (preferences), how people process information, and how they view the world (beliefs), and how these assumptions translate into behavior ("solution concept", compare Croson and Gächter, 2010). For example, the neoclassical model assumes that people maximize their (expected) utility, that they have correct beliefs, and that they use – for example – the Nash equilibrium concept to determine behavior. In this dissertation, it is also worth mentioning that the utility is assumed to only depend on the outcome of a game, not on the way players achieve these outcomes.

Thus, the neoclassical model relies on extensive assumptions, which are made to simplify and to suggest normative appealing behavior. Nevertheless, even with such a "very rational" approach, the question of how to deal with irrationality cannot be avoided. For example, the subgame perfect equilibrium, which is a refinement of the Nash equilibrium, assumes that players will act rationally from a certain node in the game tree onwards, even if they have reached that node by making irrational moves. Such an assumption can be justified by arguing that people make mistakes, i.e. irrational moves, with small probabilities (leading to the trembling-hand equilibrium). However, observing seemingly irrational behavior in experiments can almost always lead to a very different conclusion: Subjects may have different preferences or beliefs, or may use a different solution concept than assumed by the researcher. For example, contributing to a public good is irrational for a homo economicus, and much of the early research on public goods tested whether subjects who contributed were "confused" (compare, e.g., Andreoni, 1995). In contrast, the same behavior is fully rational if one assumes that subjects have reciprocal preferences and believe that others contribute as well. While the former approach asks whether people behave according to the researcher's perception of rationality, the latter approach takes people's decisions as given and asks how such decisions can best be rationalized. Rationalization of decisions has become an important part of game theory, and in many parts, this dissertation can be attributed to this branch of research. It is also interesting to note that Reinhard Selten himself views his development of the concept of subgame perfect equilibria as a philosophical inquiry with no a-priori relevance for describing human economic behavior (Dufwenberg, 2001).

What is meant by being rational? Rational is often used in the sense of reasonable, leaving open how this is exactly defined (we also use the expression in the other sections of this Introduction in this unspecified way). In contrast, Perea (2012) differentiates between rational and reasonable choices, which will be useful to understand the modeling approach in our papers. Perea defines a choice as being *rational* if "there is *some* belief about the opponents' choices for which [a decision] is optimal" (without putting any restrictions on this belief). A rational choice is not necessarily a reasonable choice, and Perea does not define reasonable because this is subjective and depends on the solution concept one has in mind (Perea, 2012, p. 6 and 29). Perea's rationality definition is based on Aumann (1987), who calls such rationality *Bayesian rationality*. Aumann (1987) argues that the modern subjectivist, Bayesian view of the world is that players have a subjective probability distribution over every prospect (Savage, 1954), including that of players choosing certain strategies in certain games. It follows that rational playing only implies maximizing one's utility given these subjective distributions over the other players' strategy choices, without demanding that these beliefs are correct. If all players behave in this way, and if this behavior is common knowledge, equilibrium behavior unfolds, which Aumann (1987) calls "correlated equilibrium". This is obviously a more general equilibrium definition than Nash equilibrium.

If one assumes that players are rational, one may also assume that players believe that the other players are rational (1-fold belief in rationality), leading to the assumption that players believe that the others believe that players are rational (2-fold belief in rationality), and so on. If such higher-order-beliefs are rational ad infinitum, one speaks of *common belief in rationality*. Again, note that common belief in rationality (in contrast to the informal use of the expression rationality) does not imply that the belief hierarchies are correct (compare example 3.2 in Perea, 2012). Thus, common belief in rationality means that (everybody knows) that everybody maximize their utility, given their knowledge about the world. We assume such behavior in our papers.

If belief hierarchies are *correct*, they are called *simple*. This is, for example, assumed in the Nash equilibrium concept. As already mentioned, common belief in rationality is a far less restrictive solution concept than the Nash equilibrium and may, in many cases, not restrict possible strategies at all, but we show in the next section how this caveat can be overcome. Perea (2012) argues that the Nash equilibrium is "not a very plausible concept to use […], even though Nash equilibrium has played a central role in game theory for many years" (p. 134). He proposes that common belief in rationality is a better alternative. We cite Perea: "In fact, it would be an absolute coincidence if [your co-player] were to be correct about your belief. […] There is nothing wrong with believing that some of your opponents may have incorrect beliefs about your own beliefs. After all, your opponents cannot look inside your head, so why should they be correct about your beliefs?" (p. 146). We agree with Perea in the following way: While the concept of a Nash equilibrium may be a useful tool to provide a "benchmark case" and to make normative statements, our experimental data indeed show that severe descriptive mistakes can result from assuming simple belief hierarchies.

3.1 Boundedly Rational Behavior

The picture of rational decision making underlying most of contemporary economic theory is far away from observed behavior. It is therefore necessary to develop theories of bounded rationality.
Reinhard Selten (1998), p. 414

First, note that the expression *boundedly rational* (or *limitedly rational*) does not necessarily correspond to the definition of rationality presented in Perea (2012): According to Simon (1955), bounded rationality models include models which describe deviations from objectively optimal behavior by considering cognitive illusions (a behavior which Perea (2012) would still define as being rational), models which optimize under computational constraints, and models which consider that decisions must be made fast and simple (again, Perea's definition of rationality does not capture these two points). Bounded rationality is understood as rationality exhibited by actual human economic behavior (Selten, 1998), and is used in that sense in the following.

In this dissertation, we ask how observed behavior in experiments can best be understood. We present a model which assumes that players optimize, which is not self-evident: Instead, one can assume learning (reasoning-by-analogy), which includes trial and error (a survey on learning models can be found in Camerer, 2003, chap. 6; a learning model for the case of the public goods game is presented in Arifovic and Ledyard, 2012). At first glance, learning models may be more suitable to describe boundedly rational behavior, and some researchers indeed believe that such approaches may have the potential to become more successful than traditional optimizing approaches (Harstad und Selten, 2013). Nevertheless, optimization models have many advantages: Such models provide rigor, which allows the identification of key economic forces (in our case: biased beliefs and reciprocity); they are applicable in a context-free way; and they may make correct predictions even if they do not capture how people really think (as-if-approach). Rabin (2013) expresses this as follows: Such models reflect that people's reasoning "whittles away all but a few […] disastrous things all of us could do in virtually every new situation we face in life" (p. 536). Given an agent's knowledge, optimizing captures compelling behavior. Therefore, Rabin (2013) proposes to keep the existing neoclassical models and extend them such that both the neoclassical model and the limited-rationality model are embedded with the help of parameter values. Having done so, "the models can be compared and judged, in a fair fight, by establishing point estimates and confidence intervals on the parameter values" (p. 530). Incorporating such rationality limitations currently leads to rapid improvements of microeconomic theory. Furthermore, by defining explicitly irrationality parameters, assuming common belief in rationality no longer means that all possible strategies are part of a subject's strategy space. Instead, the models predict only one or only a few outcomes. The proposal of Rabin (2013) corresponds exactly to our proceeding, particularly in our first paper "Explaining Individual Contributions in Public Goods Games Using (only) Reciprocity and Overoptimism", henceforth PGG-paper: We add a belief bias ε to an existing reciprocity model. Furthermore, the reciprocity model adds a reciprocity parameter Y to the neoclassical model. By setting ε and Y equal to zero (in our second paper "Trust, Reciprocity, and Betrayal Aversion: Theoretical and Experimental Insights", henceforth TG-paper, ε must be set equal to one), our model collapses into the neoclassical one – and then makes clearly wrong predictions.

In line with the complementary role which theory plays in explaining experimental findings of irrational behavior, an increasing number of limited-rationality models have been built in recent years (compare Rabin, 2013, for examples). However, these models mostly refer to situations without a game theoretic context. In contrast, some game-theoretic models are presented in Crawford (2013), and in the following, we shortly want to comment on one of them, because it has striking similarities with our model. This especially holds for the PGG-paper: We comment on the cursed equilibrium model of Eyster and Rabin (2005).

Eyster and Rabin (2005) explain the winner's curse in auctions by assuming that players correctly predict the distribution of other players' actions, but underestimate the degree to which these actions are correlated with other players' information. Accordingly, to some degree, players neglect the informational content in other players' behavior. For example, a seller may know whether a used car is a worthless "lemon" or a valuable "peach", and, for a predetermined low price, sell only lemons. A rational buyer will realize that the seller only offers lemons, and will not buy, but a "cursed" buyer does not fully capture this interrelation and believes that both lemons and peaches are sold. As in our model, players optimize in the sense that they play a best response to their beliefs. As well, by setting their irrationality parameter to zero, the model collapses into the Nash equilibrium. Similarly to us, Eyster and Rabin (2005) propose that a natural generalization to their model is to allow different players to be "cursed" to different degrees. As well, their model could be interpreted as a theory where players believe other players to play suboptimally given their private information, which Rabin and Eyster do not find compelling: "Rather than say that Player A figures out Player B's optimal strategy but believes B does not figure this out, we say that A himself does not properly introspect about how B uses B's private information" (p. 1629). As in our model, their model leads to inconsistencies, which cannot be avoided if bounded rationality is modeled. In Rabin and Eyster's case, players underestimate the correlation between co-players' actions and co-players' information. In our case, players believe in too high contributions from their co-players, but do not realize that their beliefs are wrong. Rabin and Eyster point out that their model is conceptually troubling, but they justify their approach by arguing that players do “not (fully) think through the logic of the [model]” (p. 1632). Finally, like us, Rabin and Eyster have to estimate different values of their irrationality parameter for different experiments and for different players to fit the model to the data precisely.

However, with respect to modeling boundedly rational behavior in game-theoretic contexts, research is still in its infancy. As Rabin (2013) formulates: “Little has yet been done to integrate statistical errors, or models of how people are neglectful and irrational in extracting information from other economic actors in strategic and market contexts“. In that sense, this dissertation intends to contribute to a growing and fruitful research field.

3.2 Biased Beliefs

In the experimental data from the PGG and the TG, we find that subjects believe in too favorable outcomes, and we conclude that subjects show an overoptimistic, and therefore irrational, bias. However, believing in favorable outcomes induced by others may be an uncommon interpretation of overoptimism, as this expression typically refers to an overestimation of own capabilities and traits. In line with this skepticism, our UG-results raise doubt about the overoptimism interpretation: In the UG, subjects are pessimistic instead of optimistic. We propose in the UG-paper that all of these findings can be unified by assuming that people overestimate the reciprocity inclination of co-players instead of their own payoffs. We suggest in Section 4.2 that such an argumentation can be justified with evolutionary arguments.

With respect to the belief bias, an analogy to the research regarding social preferences comes to mind: Altruism and reciprocity both predict cooperating behavior in cooperation games, but come to different predictions in punishing games (not punish vs. punish). While it is still unknown how different social motives interact, experimentalists have started to research this question (compare the literature cited in the UG-paper). Analogously, overoptimism with respect to outcomes and with respect to reciprocal behavior are congruent in cooperation games, but make contrary predictions in punishing games. As far as we know, the relationship between these sometimes complementary belief biases has not been researched at all.

4 Social Preferences

No matter how selfish you think man is, it's obvious that there are some principles in his nature that give him an interest in the welfare of others, and make their happiness necessary to him, even if he gets nothing from it but the pleasure of seeing it.
Adam Smith, Theory of Moral Sentiments, 2000 [1759], p. 1

While for a long time economists had forgotten that social preferences are a phenomenon worthy of consideration, their existence is undisputed nowadays. The relevance of social preferences has been shown in countless experiments. Moreover, researchers have started to decode how social preferences work in the human brain, a recent review can be found in Declerck, Boone, and Emonds (2013). Furthermore, a genetic basis for social preferences has been found, either with the help of evolutionary analysis (see below) or with twin studies (Sturgis et al., 2010).

Social preferences (alternatively: other-regarding preferences) can formally be defined as follows: "Individual i has social preferences if for any given [physical resource] x_i person i's utility is affected by variations of x_j, $j \neq i$ (Fehr and Schmidt, 2005). Therefore, social preferences are the opposed term to selfish preferences (being only interested in x_i) and build the generic term for preferences such as altruism (costly acts that confer economic benefits on other individuals, Fehr and Fischbacher, 2003), inequality aversion (willingness to give up some material payoff to move in the direction of more equitable outcomes, Fehr and Schmidt, 1999), quasi-maximin preferences (desire to maximize the minimal payoff in the group, Engelmann and Strobel, 2004), reciprocity (see below), or spiteful or envious preferences (always valuing the material payoff of relevant reference agents negatively, Fehr and Fischbacher, 2002).

4.1 When Social Preferences Should Be Assumed, and When Not

All models we are aware of capture social preferences by adding a social (in our case: reciprocal) utility component to a selfish (material) utility component. Both preferences are weighted against each other with the help of an additional parameter. Accordingly, social preferences need not be seen as a contradiction to selfish preferences, but as an extension. Given that – ceteris paribus – less parameters are better than more, the question arises when the social parameter can be set to zero, and when not.

Ockenfels and Raub (2010) list three arguments, in which cases the homo economicus model is still useful: It is useful as an "as-if"-interpretation, as a "worst case" scenario, and as a benchmark.

People may be socially orientated, but in markets behave like egoists, even if outcomes are highly unfair: In markets, several players compete for trade. Such situations are formally described in Fehr and Schmidt (1999), with the intuition being as follows: Accepting an unfair trade is better than making no trade, even with fairness considerations. Furthermore, making fair offers reduces inequality among *all* potential buyers only slightly, letting selfish considerations typically prevail. Smith (1962) showed that the neoclassical predictions are indeed precise if markets are competitive. Especially, a necessary condition for this result is that complete contingent contracts are traded (Schmidt, 2011): Incomplete contracts allow welfare-increasing actions after parties have agreed on a trade, making fair behavior beneficial even if players compete for trades. Contracts are obviously not complete in, for example, labor markets, making social preferences a relevant factor (Akerloff and Yellen, 1988, Fehr, Kirchsteiger, and Riedl, 1993). Broadly speaking, social preferences are the less relevant the more perfect a market situation is: They are extremely relevant in our experiments, where market forces are not at work at all. It is an open question as to whether they are relevant at all in “almost” perfect financial markets: While, for example, Breuer, Felde, and Steininger (2014) find that stock prices of firms are positively affected by a withdrawal from “sin states” (which is presumably due to the moral preferences of investors), others do not find lower yields of investing socially responsible (Riedl and Smeets, 2014 give a short overwiew). However, the following hypothesis opens room for a slight effect of social preferences on prices of financial securities: Unethical companies may have higher costs of capital (and are therefore traded at lower prices) because social investors are reluctant to hold such stocks in their portfolios, implying that they are extensively held by non-social investors. Consequently, these investors will demand a premium for their restricted possibilities to diversify (Heinkel, Kraus, and Zechner, 2001). Unfortunately, we are not aware of any experimental proof for this hypothesis.

Interestingly, Leibbrandt (2012) reports that sellers who are more pro-social in a laboratory experiment are also more successful in natural markets because they have superior trade relations and better abilities to signal trustworthiness to buyers. Henrich et al. (2005) discover a striking correlation between the degree of market integration in a society and its level of prosociality expressed in experimental games. These findings indicate that social preferences are a relevant factor in the real world even in market situations: Markets typically have a very limited degree of perfectness.

The homo economicus model may also serve as a worst case scenario when it comes to designing institutions and making economic policy decisions. Research on social preferences shows how these preferences help to overcome social dilemmas, and if an institutional design is working well among selfish players, it will certainly do even better among socially orientated ones.

Third, using the prediction of the homo economicus model as a benchmark allows to quantify the relevance of social preferences. This may be important for evaluating economic modeling. As well, such a benchmark may be directly relevant for the reasoning of people: Note that in reciprocal theories, a reference point must be determined to distinguish kind from unkind behavior. While it is still unclear how this reference point is actually determined, selfish behavior is an obvious candidate: Actions leading to higher payoffs compared to this benchmark may be perceived to be kind, whereas lower payoffs could be perceived to be a punishment.

4.2 Evolutionary Analysis

Evolutionary game theory merges population ecology (population ecology deals with the dynamics of species populations and asks how these populations interact with the environment) with game theory. It re-interprets game theory by using inheritable traits instead of optimal strategies, fitness (average reproductive success) instead of payoffs, and population members instead of players (Sigmund and Nowak, 1999). In simulations, selection (by inheritance or by social learning) leads to an increase in the frequency of strategies which grant higher fitness. Typically, this fitness depends on the frequency of a trait, leading to (ongoing) changes in the structure of the population. Evolutionary analyses are used in economic research to prove which advantages certain preferences (meant as stable determinants of a person’s strategy) have. Preferences link economic and evolutionary analysis because “we can […] expect our preferences and our decision-making to have been the products of evolution” (Samuelson, 2005). Evolutionary analysis can put findings in behavioral economics on more solid ground: The concept of maximizing (expected) utility can be criticized because it is basically a tautology: Utility is a theoretical construct, and it can only be operationalized by observing or measuring what people like and what not. Accordingly, utility maximization implies that such behavior maximizes utility which leads to the most preferable outcome. Evolutionary game theory helps to define utility by arguing that preferences must have developed to let people survive.

Accordingly, much of the early skepticism against the assumption of social preferences stems from the question of how such costly preferences should survive (or develop) in an environment where – according to Charles Darwin – only “the fittest survive”. Such skepticism was already expressed by Thomas Hobbes who argues that “homo homini lupus est”. Interestingly, experiments have revealed that Hobbes was wrong: People cooperate intuitively and are not predisposed towards selfishness (Rand, Greene, and Nowak, 2012). Moreover, socially oriented people seem to be more successful than selfish ones (Barr and Serneels, 2009, Dohmen et al., 2009). Evolutionary research supports such findings by identifying several plausible mechanisms which allow social preferences to increase “fitness” and to survive in human groups even without the existence of regulating institutions. We will briefly introduce such mechanisms in the following.

Martin Nowak summarizes research by himself and his colleagues in Nowak (2006) and outlines five mechanisms which lead to cooperative behavior (typically measured in repeated prisoner dilemma games where either cooperation or defection is possible). Such mechanisms are kin selection, direct reciprocity, indirect reciprocity, network reciprocity, and group selection.

Kin selection argues that behavior is determined by genes, and that a person’s genes also (partly) spread if a relative instead of the person survives. For example, two siblings share the same gene with a probability of 1/2.

Among unrelated individuals, cooperation spreads if they behave in a reciprocal way (introduced by Trivers, 1971). Most famously, Axelrod (1984) found that reciprocal behavior, so-called tit for tat is the winning strategy in repeatedly played prisoners’ dilemmas.

While direct reciprocity is successful in repeated interactions between the same two persons, reciprocity can also work even if the same persons never meet twice: So-called indirect reciprocity captures behavior where people are (un)kind to others who are in turn (un)kind to third parties. Such a strategy depends on the possibility to build reputation and on conditions where such reputation spreads by the contents of gossip.

The environments considered so far assume that people interact with each other equally likely. More realistically, one may assume that spatial structures exist, where some individuals interact more often than others. Such a possibility leads to clusters of subject types, where cooperation takes place in some networks and defection in others.

Finally, one can assume that selection acts not only on individuals but also on groups: While individual selection strengthens the fitness of individuals, it reduces the average fitness of the population in prisoners'-dilemma-environments. Accordingly, successful groups are those which contain many cooperating individuals, and such groups can crowd out defecting groups.

A typical finding in such evolutionary simulations is that the success of strategies is not constant (compare Nowak, 2004): Instead, it can oscillate or even be chaotic. The reason is that adaptions to environments also change the environment (in particular, the behavior of other people). For example, in repeatedly played prisoners' dilemmas, tit for tat can invade a population of defectors. Once tit for tat has been established, "generous tit for tat" can invade, which forgives accidential defections by responding to defection with cooperation from time to time. Such a population can be invaded by "always cooperate", which in turns makes "always defect" attractive, and so on. To analyze such situations, evolutionary game theory is a better framework compared to optimization techniques.

While the research mentioned above concentrates on the evolutionary advantage of reciprocal behavior in cooperative dilemmas, one may also ask with respect to our UG-paper how punishing unfair behavior can be evolutionary advantageous: Such an advantage may be less obvious, as cooperation makes the co-player better off, while punishments reduce payoffs for both parties. Nevertheless, preferences for fair outcomes can be explained in evolutionary ways. Gavrilets (2012) shows why third-parties, whose material payoff is first of all not affected by a game outcome, punish egalitarian norm violations, even if this is costly for themselves: In groups where individuals can take resources from others by force, interactions can be described by a hawk-dove-type game with either "do not fight" over the resource or "fight". As stronger individuals take away resources from weaker individuals and, as a result, have higher reproductive success, hierarchies develop where strong individuals usurp a disproportionally large share. In such environments, it is beneficial for oneself if *all others* are more equal. This makes a preference to help the weak against the strong an evolutionary advantageous one, and lets norms of inequality aversion evolve.

The papers we cited above assume that agents use simple and predefined strategies. It is more realistic to assume that strategies depend on beliefs, and as we find biased beliefs in our experiments, a natural question which arises is why these biases are evolutionarily advantageous. Typically, belief biases are justified as a heuristic which induces behavior that is almost optimal, but requires cognitively much less demanding calculations. The findings in this dissertation lead to a different idea: Among reciprocal players, belief biases can substitute reciprocity: The same forces that foster reciprocal preferences may lead to biases, which only make players *believe* that the co-player is a reciprocal type. Such beliefs induce welfare increasing, cooperative behavior even if the co-player is a selfish type. In the UG, biased beliefs with respect to the willingness of the responder to punish lead to fair behavior of the proposer even if the responder actually engages in welfare-destroying, punishing behavior only rarely. In that sense, belief biases and reciprocal preferences are interrelated with each other, which should be worth further exploration.

Although this dissertation does not use evolutionary game theory techniques, we have presented the excursion into this field for two reasons. First, we show that understanding the *causes* of social preferences is even more difficult than understanding their implications (which in turn are more difficult to understand than decisions in non-interactive environments). Interpreting human behavior in terms of "optimizing something" necessarily fails to capture feedback effects between environment and behavior, and will therefore not be successful in understanding the reasons for social behavior. While this dissertation primarily intends to discover *how* social behavior can be best described (and therefore relies on optimization models), subsequent research can ask *why* such behavior exists: Apart from the question of how biased beliefs about the social orientation of others can be advantageous, we also raise the question of among which circumstances a mixed strategy of sometimes acting reciprocally and sometimes acting selfishly performs better compared to a strict preference for *both* punishments and rewards.

4.3 Reciprocity

Tit for tat
A proverb

An individual behaves in a *reciprocal* way if "he responds to actions he perceives to be kind in a kind manner, and to actions he perceives to be hostile in a hostile manner. […] Thus, preferences do not only depend on material payoffs but also on intentions, i.e. on beliefs about why an agent has chosen a certain action" (Fehr and Schmidt, 2005). Strictly speaking, this definition refers to *direct* reciprocity, while *indirect* reciprocity means that a person is (un)kind to another person because he expects that a third person is (un)kind to himself. Research has shown that much of subjects' behavior in social dilemmas can best be described with the help of reciprocal preferences (see the Introductions of our papers), and for that reason we rely on a reciprocal theory in this dissertation to understand our experimental results.

(Direct) reciprocity can either be modeled as being intrinsic or as arising indirectly from other preferences. The model of Bolton and Ockenfels (2000) is an example of the indirect case, where reciprocal behavior stems from behaving according to inequity preferences. In a direct way, reciprocity is modeled by incorporating intentions into the utility functions: In that case, not only the outcome of a game becomes relevant, but also the beliefs on how these outcomes were achieved. Such games are called psychological games (Geanakoplos et al., 1989), and they add another layer of complexity to the analysis. Rabin (1993) developed a model of reciprocity where such intentions matter. DK and Falk and Fischbacher (2006) adapt Rabin's model to extensive-form games, and among other models which intend to capture reciprocal behavior (Charness and Rabin, 2002, Cox, Friedman, and Gjerstad, 2007, Levine, 1998, Segal and Sobel, 2007), these two are the most prominent ones. As Falk and Fischbacher (2006) capture both outcome concerns as well as intentional concerns, they need two parameters. In contrast, DK concentrate on purely reciprocal aspects and need only one parameter (if one abstracts from the fact that DK allow the modeling of the parameter co-player-dependent). This dissertation introduces two further parameters into the analysis (a belief bias parameter and a risk aversion parameter), and to not complicate the theory even further, we intend to capture social preferences with only one parameter. Accordingly, DK's approach is used in this dissertation.

5 Remarks on the experiments

All of our three papers have a joint experimental and theoretical basis: We always reproduce behavior in a well-known social dilemma in a control treatment and compare it to behavior in a second treatment, which modifies the standard game in order to gain insights about the motives which induce people's decisions. Also for that purpose, subject's beliefs are elicited.

We find that in each experiment, behavior can be explained with the help of reciprocal motives, complemented by the insight that subjects' beliefs are biased (in risky environments, risk aversion – of course – matters as well). Due to these homogeneous results, we can explain behavior in all three games in a very similar way. We have to make major adaptions to the model of DK because DK can neither explain UG results, nor PGG results, and they predict TG results only qualitatively. Interestingly, this is the case although reciprocal theories are the most powerful (and the most complex) models among the social preference models, and although experimental evidence is compelling that subjects' decisions are indeed influenced by reciprocal motives. We propose different modifications in each paper: Some of them result from the desire to explain the results not only qualitatively but also quantitatively (e.g., normalization of the strategy space in the UG-paper), others are made to simplify (e.g., normalization in the PGG-paper) and to assure analytical solvability (e.g., squaring reciprocal utility in the UG-paper). As well, some modifications are made to address questions which are relevant in one game, but play a minor role in other games (e.g., risk aversion in the TG vs. in the PGG). In Section 5.4, we present a summary of all modifications in this dissertation. This section also clarifies that the modifications of all three papers can be merged such that the results in all roles in all three games can be explained with the help of one theory.

5.1 Reciprocity in the Public Goods Game

Jedermann hat die sittliche Pflicht, für das Wohl des Ganzen zu wirken.
Preamble of the Constitution of the Free and
Hanseatic City of Hamburg

A *good* is defined as *public* if it "can be consumed by every group member regardless of the member's contribution to the good" (Fehr and Fischbacher, 2003). A player who contributes to a public good *cooperates* because he "increases the sum of all payoffs" (MacCrimmon and Messick, 1976). In contrast, competitive players would maximize comparative payoffs, meaning that the difference between payoffs is maximized. A homo economicus would always free ride on the contributions of others and would never cooperate. Therefore, public goods have difficulty to be provided or not to be depleted. Thereby, they build the contrary pole to goods traded on markets, which are provided in an efficient manner. Dietz et al. (2003) mention the public good example that "the global ocean has lost more than 90 % of large predatory fishes, with an 80 % decline typically occurring within 15 years of industrialized exploitation" (p. 1907). The PGG is the canonical representation of such situations, where selfish-interest is not in line with collective-interest, and can therefore be used to study collective action problems. Such situations are all around us: Its scale ranges from two persons (a couple with a joint bank account) over small groups (working for the success of a team) and large groups (making people pay their taxes) to the whole of mankind (reducing ozone-depleting substances).

Instead of endowing public goods with well-defined property rights such that they lose their public goods character, people are sometimes able to maintain informal institutions which are successfully able to govern the commons (Dietz et al., 2003). One aspect thereby is to appeal to the citizens to not deplete public resources, compare the Constitution of Hamburg. Finding mechanisms to solve the tragedy of the commons (Harding, 1968) substantially affects our way of life: Talhelm et al. (2014) argue that China has a more collectivistic culture than the West because farming rice makes people more collectivistic than farming wheat: While wheat grows through rainfall and is less labor-intensive, rice farmers must commonly build irrigation systems and help each other to harvest, building out higher cooperative cultures.

Typically, in the standard version of the PGG, subjects are initially willing to cooperate, but are not able to maintain high levels of cooperation. Our paper offers an explanation of why this is the case: We confirm that people want to cooperate, but only with a self-centered bias and only if others cooperate as well. As people's overoptimistic beliefs in the cooperating behavior of others can compensate the self-centered bias to some degree, high contributions can initially be established. Learning that their beliefs have been biased, subjects reduce their contributions, and cooperation breaks down.

5.2 Reciprocity in the Trust Game

A definition of *trust* was given by Coleman (1990): "An individual trusts if she voluntarily places resources at the disposal of another party without any legal commitment from the latter. In addition, the act of trust is associated with an expectation that the act will pay off in terms of the investor's goals. In particular, if the trustee is trustworthy the investor is better off than if trust were not placed, whereas if the trustee is not trustworthy the investor is worse off than if trust were not placed" (compare Fehr, 2009). The TG exactly captures this situation, where the receiver has no obligation to return money to the sender. As no contract is complete and fully enforceable in the "real world", peoples' ability to establish trusting relationships is essential for our welfare: La Porta et al. (1997) show that country measures of trust are favorably correlated with economic measures, such as GDP growth, inflation, or anticorruption (further studies with similar results are summarized in Nannestad, 2008, p. 429).

Thinking in the neoclassical framework, “to trust” means “to place a bet”. In that sense, Coleman (1990) specifies his definition of trust by arguing that rational (risk-neutral) players must “decide between not placing trust, in which case there is no change in his utility, and placing trust, in which case the expected utility relative to his current status is the potential [material] gain times the chance of gain minus the potential [material] loss times the chance of loss” (p. 99). However, assuming equivalently rational behavior on the receiver’s part, receivers would never turn out to be trustworthy, because there is no material gain from returning. Social preferences drive receiver behavior, and accordingly, the question follows as to how these social preferences shape the sender’s decision. This is also an empirically relevant question, as the success of economic interactions may depend on the degree to which an interaction is (framed as) a social one. For example, one might expect that people prefer to lend their money to friends and to relatives instead of lending it via financial institutions, because this saves transaction costs and reduces informational asymmetries. Instead, between 30 % (“30 % der Deutschen verleihen grundsätzlich kein Geld”, 2009) and 57 % (“Hört bei Geld die Freundschaft auf?”, 2012) of all people do not privately lend money at all, not even to friends. Researchers have introduced the expression “betrayal aversion” to indicate that subjects might rather prefer “gambling” to “trusting”. Understanding betrayal aversion is an important component of understanding economic exchange. Such an understanding can also be used to build well-functioning institutions: Institutions should offer the option to avoid knowing painful details of failed economic exchange (Aimone and Houser, 2013). Recent research also uses neuroscientific approaches to differentiate between gambling and trusting: There is evidence that risky decisions are processed differently in the human brain than trusting decisions (Aimone, Houser, and Weber, 2014).

In our TG-paper, we measure the effect of social preferences in the sender role if one controls for subjects' beliefs and risk aversion. While our "social treatment" is identical to the standard version of the TG, subjects place a bet in our "non-social treatment". A crucial point for the comparison of both treatments is that the probability distribution of receiver behavior is equal to winning chances in the lottery. In our design, we find no large behavioral differences between both treatments: Our results, if at all, contradict the idea of betrayal aversion. We model TG-behavior with the help of our reciprocal theory and indeed find that reciprocal and selfish preferences lead to similar behavior in the sender role: Trust if you believe that trust is reciprocated, and do not trust if you do not. However, as an additional unit of successfully exchanged money does not necessarily add the same quantity of utility to the material utility account and the reciprocal utility account, small differences between selfish and reciprocal behavior can exist. Especially, the distribution of receiver types can matter, opening up the possibility to explain situation-depending occurrence of betrayal aversion.

As in the two other papers, we find that beliefs about the behavior of co-players are significantly biased. Having a reciprocally-oriented pool of receivers is not sufficient to generate distinct proportions of trusting behavior. Additionally, senders must overestimate the receivers' trustworthiness. As stated in Section 1, this makes the foundations of trust much more fragile than assuming correct beliefs, because wrong beliefs are somehow framed. Having this result in mind, disruptive developments in the economy may be easier to understand than assuming players with perfect foresight.

5.3 Reciprocity in the Ultimatum Game

According to Samuelson (1996), the fundamental economic problem is how to divide a surplus. Assume that the surplus can only be consumed if the players are able to agree on how to divide the cake. In the UG, the simplest possible form of negotiation is implemented: There are only two players, and player one makes a proposal which player two accepts or rejects. In that sense, the UG displays situations where the division of welfare gains is not guided via market mechanisms but via social interaction, and, interestingly, standard economics "has virtually nothing to say about such situations" (Samuelson, 1996, p. 19). Nevertheless, non-market situations can be observed in the economy all the time: For example, parties in integrated supply chains may not easily have the possibility to exchange the business partner, but must achieve agreement over the division of payoffs from efficiency gains via negotiation. In such situations, the threat of punishment becomes relevant: If people feel they are being treated in an unfair way, they often punish their counterpart, even if this is associated with own costs: Taking revenge can be observed in all kind of human institutions, from workers who sabotage, to countries which impose sanctions on each other.

Our paper finds that punishing is a different character trait than rewarding (compare also Dohmen et al., 2009). This aspect has not been incorporated in models of reciprocal behavior so far, and it is also an open question why such "inconsistent" behavior can be observed. We show how punishing behavior can be modeled in a reciprocal way, and we outline some systematic differences between punishment and reward (men punish, while women reward; altruists reward, but do not punish). Again, both our experimental insights as well as our modeling approach show how important subjects' beliefs are in understanding punishing behavior. First, beliefs determine which behavior is seen as kind and which as unkind: If responders believe that proposers believe that responders do not accept unfair offers, offering low proportions is even more unkind than if unfair offers are believed to be accepted. Second, we again find that beliefs are systematically biased, which can explain why so many proposers offer the equal split.

5.4 An Integrated View of All Three Experiments

In this section, we summarize the modifications which are made to the original DK-model in our three papers, and we explain how these modifications correspond to each other. A good theory should be as simple as possible, as precise as possible, and as broad as possible. Obviously, trade-offs limit the achievability of these three goals. Our focal point is to describe observed behavior precisely, which comes with costs, at least in the domain of simplicity: We use an intention-based model instead of an outcome-based one, although this increases complexity. Furthermore, we assume uncertainty instead of certainty, and model biased beliefs. Third, we introduce curved utility functions to prevent corner solutions, which DK refrain from for the sake of simplicity. In turn, we get precise descriptions of observed behavior.

DK define their model very broadly (with respect, for example, to the determination of reference points). This allows us to apply their model to virtually every game-theoretic situation. Thereby, they are aware that they will not make precise predictions in each case. In contrast, paying the tribute to the broadness of our model implications, we make game-specific assumptions. We do not unify all modifications over the three papers, because we want to make as little modifications to the original DK-model in each of the papers as possible. As each game has its specific modeling difficulties, game-specific modifications arise. Nevertheless, as Croson and Gächter (2010) consider it as one out of 10 "commandments" not to "develop models in vain – no one needs a new model for every experimental or observational result" (p. 129), we show in the following that – accepting disadvantages at the domains of simplicity and preciseness – our experimental results can be explained in a unified way. Before, we summarize our modifications in Table 1.

<<<Insert Table 1 about here >>>

As just mentioned, we assume uncertainty instead of certainty, which induces several subsequent modifications (1.): First of all, utility maximization under uncertainty must be defined (1.1). We apply standard economics and assume that expected utility is maximized. Thereby, the only parameter which is uncertain is the reciprocity parameter of the co-player(s). With uncertainty, it is important to specify the curvature of the material payoff utility component, because this curvature defines how players value risk (1.2). While the curvature is typically be assumed to be linear under certainty (because stakes which can be earned in the lab are small), we again apply standard economics in the TG-paper and assume constant relative risk aversion. Not knowing with which type of co-player one is matched also implies that kindness cannot be made dependent on the co-player's personality (1.3), and, more importantly, that one's reciprocal intentions cannot depend on the co-player's reaction (1.4). In contrast, under certainty, it is consistent that DK assume that one takes the co-player's reaction into account to determine one's own kindness (because one can foresee this reaction). While this point need not be addressed in the PGG-paper (because kindness does not depend on the behavior of the co-player) and in the TG-paper (because it is not qualitatively relevant), it does become relevant in the UG-paper, where we assume that kindness is determined by using one's belief about the average expected behavior of co-players.

As well, modifications are made with respect to the determination of kindness (2.). Importantly, we ask where to set the reference point to separate kindness from unkindness (2.1), and we find that the suggestion of DK – simply use the average between the kindest and the unkindest efficient strategy – is not always applicable. We adopt the DK approach in the UG-paper, but choose different reference points in the two other papers. Using game-dependent reference points reveals that research has not been able to determine generally valid reference points so far. Actually, we are not aware of any research on that question at all. The reference points chosen in our papers can be justified by the experimental results: Using different ones, theory and observation could no longer be reconciled. Unfortunately, our experimental data are not helpful in detecting the underlying causes of how to choose reference points. Different explanations are possible, which however partly contradict each other: One could argue that people consider it to be kind (unkind) if behavior of others in the game leads to positive (negative) payoffs. In that sense, the "status quo" would determine the reference point, meaning that contributing in the PGG and sending money / returning more money than what one has received is kind in the TG. However, agreeing on an unfair division should then be seen as kind as well, which is contradicted by our finding in the UG-paper that unfair divisions are shrunk. In turn, one may assume that it is kind (unkind) if a player *i*'s strategy grants higher (lower) payoffs to a co-player *j* than the strategy which maximizes *i*'s material utility component given *j*'s expected reaction. In that way, selfish behavior can justify the reference points in the PGG and to some degree in the UG, but this definition comes with an interesting implication in the TG: Assume that senders believe that sending money is believed to result in a payoff loss (our data reveal that such a belief is not common). In that case, keeping the money as a sender would be selfish, and sending money would be kind. Such kindness would be reciprocated by receivers, meaning that sending money becomes a profitable strategy, which would also be pursued by selfish senders. In that case, sending money would no longer be kind. As selfish strategies would not be reciprocated by receivers, sending money would result in payoff losses again. Thus, no equilibrium behavior could be derived among reciprocal players. This consideration explains that Camerer (2003) robustly finds that "the return to trust is around zero": If this were not the case, trust would no longer be trust. We summarize as follows: As it was not possible so far to find a simple, precise, and coherent definition of reference points across different games, more research is needed to solve this problem.

In the PGG-paper, we normalize kindness with respect to the marginal per capita return (*MPCR*) and the group size. In the UG-paper, we normalize it with respect to the strategy space. In the former case, this is simply done to standardize the reciprocity parameter. In the latter case, such "fine-tuning" allows the capturing of full shrinking of close to zero offers (compare Section 5 in the UG-paper). This standardization has also been applied in Rabin (1993), which is the basis for the DK-paper. However, the question of whether kindness somehow has to be normalized does not affect the qualitative predictions of our model.

Third, we make important modifications with respect to belief formation (3.). In contrast to DK, we assume that beliefs can be wrong. We assume a systematic bias (3.1), meaning that players overestimate the reciprocal inclination of co-players. This systematic bias is modeled disproportionally in the PGG-paper, meaning that the degree of overoptimism depends on the actual kindness of the co-players. Simpler, overoptimism is assumed to be proportional to actual kindness in the TG. In the UG, the bias is only identified, and not modeled.

As well, we allow for unsystematic errors (3.2), which are found to be shaped by the false consensus effect (Ross, Greene, and House, 1977). By eliciting actual beliefs instead of assuming modeled equilibrium-beliefs, behavior can be predicted much more precisely. In the TG-paper, these actual beliefs are used to determine subject-dependent overoptimism biases.

Last, we make changes to the form of the reciprocal utility function (4.). DK multiply the players' kindnesses (denoted as *k*, respectively their unkindness *u*, respectively their neutrality *n*). For example, player one being kind and player two being unkind results in the *k*/*u*-outcome with a reciprocal utility of $u \cdot k$. In the DK model, reciprocal utility is maximal if agents respond to the others' (un)kindness with the most extreme reciprocal reaction possible. We prevent corner solutions in the PGG-paper and in the TG-paper by curving utility with the help of the root function, which is proposed by DK themselves (4.1). In the UG-paper, we deviate from this solution due to mathematical convenience. More importantly, the UG reveals that the preference order implied by DK's assumption of multiplying kindnesses is implausible (4.2):

u/u-outcomes should not be preferred to n/(·) ones. Again, this problem primarily concerns the UG, and we will show that it is solved by assuming gradual reciprocation.

Table 1 shows that all game properties create their own modeling difficulties, but it also makes clear that all games can be explained in a unified way if one is willing to give up some accuracy which results from our game-specific modifications: Using the modifications which are printed in bold, all three games can be captured by a single set of modifications (abstracting from the issues we discussed above for modification 2.). These bold-printed modifications are explained in detail in the UG-paper. There, we also show that behavior in the TG can be explained with that set of assumptions (compare equations (11) and (12) in the UG-paper). To complete, we now show that behavior in the PGG can also qualitatively be explained by using these modifications.

For illustrative purposes, we abstain from normalizing the reciprocal utility component in the following (modification 2.2). Accordingly, utility from playing the PGG is described by (compare Sections 2 in the UG- and in the PGG-papers)

$$E(U_i) = E\big(U_i(\pi_i)\big) - Y_i \cdot E\Big(\big(\kappa_{ij} - \lambda_{iji}\big)^2\Big)$$

$$= 1 - (1 - MPCR) \cdot g_i + \textstyle\sum_{j=1}^{J} MPCR \cdot g_{ij,o} - Y_i \cdot E\Big(\textstyle\sum_{j=1}^{J}\Big(\big(MPCR \cdot g_i - g_{rp}\big) -$$

$$\big(MPCR \cdot g_{ij,o} - g_{rp}\big)\Big)^2\Big). \qquad (1)$$

Thereby, g_{rp} denotes the contribution which defines the reference point. g_i ($g_{ij,o}$) defines *i* contribution (*i*'s overoptimistic belief over *j*'s contribution). For simplicity reasons, assume in the following (as in the PGG-paper) that all other group members contribute the same amount. In that case, (1) simplifies to

$$U_i = 1 - (1 - MPCR)\cdot g_i + J\cdot MPCR\cdot g_{ij,o} - Y_i \cdot J \cdot \left(MPCR\cdot g_i - MPCR\cdot g_{ij,o}\right)^2 . \tag{2}$$

Maximizing (2) over g_i leads to

$$1 - (1 - MPCR)\cdot g_i - 2 \cdot Y_i \cdot J \cdot \left(MPCR\cdot g_i - MPCR\cdot g_{ij,o}\right) \cdot MPCR = 0 \tag{3}$$

$$\Rightarrow g_i = \begin{cases} 0, & \text{if } Y_i = 0, \\ \max\left\{0;\, g_{ij,o} - \frac{1-MPCR}{J \cdot Y_i \cdot 2 \cdot MPCR^2}\right\}, & \text{if } Y_i > 0. \end{cases} \tag{4}$$

This captures behavior observed in PGGs: Selfish subjects do not contribute, while reciprocal subjects want to contribute like the others, but with a “self-serving” bias. Importantly, as U_i does not depend on g_{rp} in equation (2), the question of where to locate the reference point is irrelevant in the PGG.

In sum, this dissertation shows that it is worth expanding the homo economicus model to a reciprocal believers model. This allows an understanding of the behavior observed in prominent social dilemmas much better than was previously the case. At the same time, new questions arise: How have such biased beliefs and reciprocal preferences emerged and how can they be influenced? What are the consequences for institutional designs and why is observed behavior not stable across games? These are interesting questions, which are worthy of future research.

References

Aimone, J. A., and Houser, D. (2013). Harnessing the Benefits of Betrayal Aversion. *Journal of Economic Behaviour and Organization*, 89, 1-8.

Aimone, J. A., Houser, D., and Weber, B. (2014). Neural Signatures of Betrayal Aversion, an fMRI Study of Trust. *Proceedings of the Royal Society B: Biological Sciences*, 281, 1471-2954.

Akerlof, G. A., and Yellen, J. L. (1988). Fairness and Unemployment. *American Economic Review*, 78, 44-49.

Akerlof, G. A., and Shiller, R. J. (2010). *Animal Spirits: How Human Psychology Drives the Economy, and Why It Matters for Global Capitalism*. Princeton, NJ: Princeton University Press.

Andreoni, J. (1995). Cooperation in Public-Goods Experiments: Kindness or Confusion? *American Economic Review,* 85, 891-904.

Andreoni, J., and Samuelson, L. (2006). Building Rational Cooperation. *Journal of Economic Theory*, 127, 117-154.

Arifovic, J., and Ledyard, J. (2012). Individual Evolutionary Learning, Other-Regarding Preferences, and the Voluntary Contributions Mechanism. *Journal of Public Economics*, 95, 808-823.

Aumann, R. J. (1987). Correlated Equilibrium as an Expression of Bayesian Rationality. *Econometrica*, 55, 1-18.

Axelrod, R. (1984). *The Evolution of Cooperation*. New York: Basic Books.

Baran, N. M, Sapienza, P, and Zingales, L. (2010). Can We Infer Social Preferences from the Lab? Evidence from the Trust Game. NBER Working paper series 15654.

Barr, A., and Serneels, P. (2009) Reciprocity in the Workplace. *Experimental Economics*, 12, 99-112.

Bolton, G. E., and Ockenfels, A. (2000). A Theory of Equity, Reciprocity, and Competition. *American Economic Review*, 90, 166-193.

Breuer, W., Felde, M., and Steininger, B. (2014). Financial Impact of Firm Withdrawals from State Sponsor of Terrorism Countries. Working Paper.

Camerer, C. (2003). *Behavioral Game Theory. Experiments in Strategic Interaction.* Princeton, NJ: Princeton University Press.

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

Coleman, J. S. (1990). *Foundations of Social Theory*. Cambridge, MA: The Belknap Press of Harvard University Press.

Cox, J. C., Friedman, D., and Gjerstad, S. (2007). A Tractable Model of Reciprocity and Fairness. *Games and Economic Behavior*, 59, 17-45.

Crawford, V. P. (2013). Boundedly Rational versus Optimization-Based Models of Strategic Thinking and Learning in Games. *Journal of Economic Literature*, 51, 512-527.

Croson, R. (2005). The Method of Experimental Economics. *International Negotiation*, 10, 131-148.

Croson, R. and Gächter, S. (2010). The Science of Experimental Economics. *Journal of Behavior and Organization*, 73, 122-131.

Declerck, C. H., Boone, C., and Emonds, G. (2013). When Do People Cooperate. *Brain and Cognition*, 81, 95-117.

Dohmen, T., Falk, A., and Sunde, U. (2009). Homo Reciprocans, Survey Evidence on Behavioural Outcomes. *Economic Journal*, 119, 592-612.

Dufwenberg, M. (2001). Modeling Bounded Rationality, Ariel Rubinstein, Book Review. *Economics and Philosophy*, 17, 121-145.

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

Dietz, T., Ostrom, E., and Stern, P. C. (2003). The Struggle to Govern the Commons. *Science*, 302, 1907-1912.

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

Eyster, E., and Rabin, M. (2006). Cursed Equilibrium. *Econometrica*, 73, 1623-1672.

Falk, A. (2009). Lab Experiments Are a Major Source of Knowledge in the Social Sciences. *Science*, 326, 535-538.

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

Fehr, E. (2009). On the Economics and Biology of Trust. *Journal of European Economic Association*, 7, 235-266.

Fehr, E., Kirchsteiger, G., and Riedl, A. (1993). Does Fairness Prevent Market Clearing? An experimental investigation. *The Quarterly Journal of Economics*, 108, 437-459.

Fehr, E., and Fischbacher, U. (2002). Why Social Preferences Matter – The Impact of Non-Selfish motives on Competition. *Economic Journal*, 112, C1-C33.

Fehr, E., and Fischbacher, U. (2003). The Nature of Human Altruism. *Nature*, 425, 785-791.

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

Fehr, E., and Schmidt, K. M. (2005). 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* (pp. 615-691). Amsterdam: Elsevier.

Franzen, A., and Pointner, S. (2013). The External Validity of Giving in the Dictator Game. *Experimental Economics*, 16, 155-169.

Friedman, D., and Sunder, S. (1994). *Experimental Methods. A Primer for Economists.* Cambridge, MA: Cambridge University Press.

Gavrilets, S. (2012). On the Evolutionary Origins of the Egalitarian Syndrome. *Proceedings of the National Academy of Sciences*, 109, 14069-14074.

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

Gintis, H. (2009). Animal Spirits or Complex Adaptive Dynamics? A Review of George Akerlof and Robert J. Schiller – Animal Spirits. *Journal of Economic Psychology*, 30, 511-515.

Harding, G. (1968). The Tragedy of the Commons. *Science*, 162, 1243-1248.

Harstad, R. M., and Selten, R. (2013). Bounded-Rationality Models: Tasks to Become Intellectually Competitive. *Journal of Economic Literature*, 51, 496-511.

Heinkel, R., Kraus, A., and Zechner, J. (2001). The Effect of Green Investment on Corporate Behavior. *Journal of Financial and Quantitative Analysis*, 36, 431-449.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., and Gintis, H. (2004). *Foundations of Human Sociality: Economic Experiments and Ethnographic Evidence from Fifteen Small-Scale Societies*. New York: Oxford University Press.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., McElreath, R., Alvard, M., Barr, A., Ensminger, J., Henrich, N. S., Hill, K., Gil-White, F., Gurven, M. Marlowe, F. W., Patton, J. Q., and Tracer, D. (2005). "Economic Man" in Cross-Cultural Perspective: Behavioral Experiments in 15 Small-Scale Societies. *Behavioral and Brain Sciences*, 28, 795-815.

"Hört bei Geld die Freundschaft auf?" (2012), Postbank, https://www.postbank.de/postbank/postbank_pd_0212_geld_und_freundschaft.html, accessed 28 August 2014.

Karlan, D. S. (2005). Using Experimental Economics to Measure Social Capital and Predict Financial Decisions. *American Economic Review*, 95, 1688-1699.

La Porta, R., Lopez-de-Silanes, F., Shleifer, A., and Vishny, R. W. (1997). Trust in Large Organizations. *American Economic Review,* 87, 333-338.

Leibbrandt, A. (2012). Are Social Preferences Related to Market Performance? *Experimental Economics*, 15, 589-603.

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

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

Maccrimmon, K. R., and Messick, D. M. (1976). A Framework for Social Motives. *Behavioral Science*, 21, 86-100.

Morgenstern, C. (1910). Palmström. Berlin: Bruno Cassirer.

Nannestad, P. (2008). What Have We Learned About Generalized Trust, If Anything? *Annual Review of Political Science*, 11, 413-436.

Nowak, M. A. (2006) Five Rules for the Evolution of Cooperation. *Science*, 314, 1560-1563

Nowak, M. A., and Sigmund, K. (2004). Evolutionary Dynamics of Biological Games. *Science,* 303, 793-799.

Ockenfels, A., and Raub, W. (2010). Kontroverse "Altruismus, Egoismus, Rationalität". *Kölner Zeitschrift für Soziologie und Sozialpsychologie*. Sonderheft, 50, 119-153.

Perea, A. (2012). *Epistemic Game Theory – Reasoning and Choice*. Cambridge, MA: Cambridge University Press.

Posten, A. C, Ockenfels, A., and Mussweiler, T. (2014). How Activating Cognitive Content Shapes Trust: A Subliminal Priming Study. *Journal of Economic Psychology*, 41, 12-19.

Rand, D. G., Greene, J. D., and Nowak, M. A. (2012). Spontaneous Giving and Calculated Greed. *Nature*, 489, 427-430.

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

Ross, L., Green, D., and House, P. (1977). The "False Consensus Effect": An Egocentric Bias in Social Perception and Attribution Processes. *Journal of Experimental Social Psychology*, 13, 279-301.

Riedl, A., and Smeets, P. (2014). Social Preferences and Portfolio Choice. Working Paper.

Samuelson, L. (1996). Bounded rationality and Game Theory. *The Quarterly Review of Economics and Finance*, 36, 17-35.

Samuelson, L. (2005). Economic Theory and Experimental Economics. *Journal of Economic Literature*, 43, 65-107.

Samuelson, P. A., and Nordhaus, W. D. (1985), *Principles of economics.* 12th ed. New York: McGraw-Hill.

Samuelson, P. A., and Nordhaus, W. D. (1992), *Principles of economics.* 14th ed. New York: McGraw-Hill.

Savage, L. J. (1954). *The Foundations of statistics*. New York: Wiley.

Segal, U., and Sobel, J. (2007). Tit for Tat: Foundations of Preferences for Reciprocity in Strategic Settings. *Journal of Economic Theory*, 136, 197-216.

Selten, R. (1998). Features of Experimentally Observed Bounded Rationality. *European Economic Review*, 42, 413-436.

Sigmund, K., and Nowak, M. A. (1999) Evolutionary Game Theory. *Current Biology*, 9, R503-R505.

Simon, H. A. (1955). A Behavioral Model of Rational Choice. *The Quarterly Journal of Economics*, 69, 99-118.

Smith, A. (2000) [1759]. *The Theory of Moral Sentiments*. New York: Prometheus Books.

Smith, A. (1937) [1776]. *The Wealth of Nations.* New York: Modern Library.

Smith, V. L. (1962). An Experimental Study of Competitive Market Nehavior. *The Journal of Political Economy*, 70, 111-137.

Schmidt, K. M. (2011). Social Preferences and Competition. *Journal of Money, Credit and Banking*, 43, 207-231.

Sturgis, P., Read, S., Hatemi, P. K., Zhu, G., Trull, T., Wright, M. J., and Martin, N. G. (2010). A Genetic Basis for Social Trust? *Political Behavior*, 32, 205-230.

Talhelm, T., Zhang, X., Oishi, S., Shimin, C., Duan, D., Lan, X., and Kitayama, S. (2014). Large-Scale Psychological Differences within China Explained by Rice Versus Wheat Agriculture. *Science*, 344, 603-608.

Trivers, R. L. (1971). The Evolution of Reciprocal Altruism. *The Quarterly Review of Biology*, 46, 35-57.

“30 Prozent der Deutschen verleihen grundsätzlich kein Geld“ (2009). Press release at 08/19/2009, Comdirect Bank, https://www.comdirect.de/cms/ueberuns/de/presse/cori1088_0462.html, accessed 28 August 2014.

		Dufwenberg, Kirchsteiger (2004)	Public Goods Game paper	Trust Game paper	Ultimatum Game paper
1. certainty / uncertainty	1.1 uncertainty	modelling under certainty	**maximizing expected utility, belief about a probability distribution of reciprocity parameters in the subject pool**		
	1.2 curvature of utility from material payoffs	linear		**constant relative risk aversion**	not specified in the proposer role, linear in the responder role (simplification)
	1.3 co-player dependent reciprocity parameter	yes	**no** (simplification; not necessary under uncertainty)		
	1.4 kindness intention and co-player's reaction	kindness depends on strategies of actual co-players (no uncertainty about their reciprocal inclination)	kindness not dependent on co-players' strategies	kindness depends on strategies of actual co-players (modeling like in the UG paper is possible)	**believed average strategy of possible co-players from the subject pool**
2. kindness deter-mination	2.1 reference point for kindness to the co-player	Average between maximal and minimal efficient kindness	status quo or selfish behavior (meaning: keeping the money)	status quo (meaning: sending no money / returning as much money as received)	Average between maximal and minimal efficient kindness / selfish behavior (meaning: offering 0.5 (which is a simplification of selfish behavior) /
	2.2 normalisation	no normalization ("reciprocity propotional to the gift"), but done in Rabin 1993, which is the basis for DK's model	by MPCR (simplification, no normalization may be more realistic)	no normalization	by strategy space (compare Rabin 1993; full shrinking of zero-offers can be captured)
3. beliefs	3.1 systematic error	modeling under certainty, correct beliefs	**overoptimistic beliefs, disproportional to co-players' kindness** (necessary to predict non-zero contributions) **false consensus effect modeled**	subject-dependent overoptimistic beliefs, proportional to co-players' kindness (simplification) false consensus effect in data identified	pessimistic belief in data identified
	3.2 unsystematic errors		**incorporated**; others' beliefs (about others' beliefs about…) are estimated by using one's own belief		
4. reciprocal utility function	4.1 curvature of reciprocal utility	linear (proposition for application purposes: root function)	root function (to prevent corner solutions). Reference point related; unkind domain not specified, concave in the kind domain	root function (to prevent corner solutions). Reference point related; convex in the unkind domain, concave in the kind domain	**quadratic** (to prevent corner solutions and for simplification reasons; root function is also possible)
	4.2 degree of reciprocation and preference order	k/k = u/u preferred to k/n = n/n = u/n preferred to u/k	unkindness not possible, k/k preferred to k/n=n/n	u/u, n/n, u/n not possible, k/k preferred to k/n preferred to u/k	**k/k preferred to n/k, preferred to u/k; n/n preferred to k/n = u/n; u/u preferred to n/u preferred to k/u**
		harshest possible reaction to (un)kindness (if abstracted from material costs)			**gradual reciprocation**

Table 1: Summary of modifications performed to the DK-model

Explaining Individual Contributions in Public Goods Games Using (only) Reciprocity and Overoptimism

by

Wolfgang Breuer[*] and Anselm Hüwe[**]

September 2014

Abstract: We explain contributions in public goods games with the help of the reciprocity model of Dufwenberg and Kirchsteiger (2004) by applying some plausible modifications: Most importantly, we assume that subjects overestimate the kindness of their group members. In combination with the finding that subjects are typically imperfect conditional cooperators, equilibrium contributions to public goods can be derived. We test our model experimentally and find robust evidence for our modifications: In the experiment, we directly link reciprocal preferences elicited in contribution schedules to beliefs and show that behavior is indeed primarily driven by reciprocity and overoptimism. Although we find distinctly heterogeneous behavior on the subject level, our model can predict such behavior if subjects' reciprocal inclinations are known. Thereby, the false consensus effect additionally fosters cooperation because it lets conditional cooperators overestimate the level of reciprocity in the subject pool.

Keywords: public goods games, reciprocity, overoptimism, conditional cooperation, false consensus effect

JEL classification: C72, C91, C92, D84, H41

[*] Prof. Dr. Wolfgang Breuer
RWTH Aachen University
Department of Finance
Templergraben 64
52056 Aachen
Germany
Phone: +49 241 8093539
Fax: +49 241 8092163
E-mail: wolfgang.breuer@bfw.rwth-aachen.de

[**] Dipl.-Ing. Anselm Hüwe
RWTH Aachen University
Department of Finance
Templergraben 64
52056 Aachen
Germany
Phone: +49 241 8093505
Fax: +49 241 8092163
E-mail: anselm.huewe@bfw.rwth-aachen.de

1 Introduction

Why do people cooperate in public goods dilemmas? In recent years, researchers have made progress in explaining such behavior by assuming that subjects have social preferences. But how exactly do preference functions look like which describe behavior in public goods games (henceforth PGGs)? We shed light on these questions by using the reciprocal utility model of Dufwenberg and Kirchsteiger (2004, henceforth DK) which we combine with the assumption that subjects' beliefs have an overoptimistic bias.

What do we know about subject behavior in PGGs? The most important stylized facts stemming from extensive research can be summarized as follows (compare Chaudhuri 2011 or Holt and Laury 2008 for reviews):

1. Average contributions start at around 50 % of the social optimum and decline steadily with repetition.

2. There are distinct types of players who differ in their social preferences and/or beliefs about their peers. Accordingly, individual contributions range from 0 % to 100 %. Many participants are conditional cooperators, who only want to contribute if the others contribute as well.

3. Higher marginal per capita returns (henceforth MPCRs) from the public good lead to higher contribution levels.

4. Increased group size leads to higher contribution levels, at least for low MPCRs and low-to-moderate group sizes.

5. A surprise restart of the game at the end of a session induces an increase in contributions, known as the "restart effect". Contributions thereby do not reach the initial levels.

While fact 1. and 3. to 5. refer to aggregated behavior, fact 2. can only be understood with the help of a theory which addresses the individual level. Applying such a theory, we are able to explain individual contributions. Other researchers have also tried to do this (Ambrus and Pathak 2011; Arifovic and Ledyard 2012; Chaudhuri 2011; Dijkstra 2012; Klumpp 2012; Ledyard 1995), but even those models which have been solely designed to explain behavior in PGGs cannot explain all the empirical findings mentioned above. Why are explanations so difficult? The standard assumption of selfish behavior obviously falls short because it predicts free riding for all subjects. Theories of learning, which may explain decreasing contributions, have trouble with the fact that there is also cooperation among experienced subjects and "unconfused" subjects. The assumption that people have social preferences cannot adequately account for declining contributions. Attempts have been made to explain contributions with the help of signaling strategies, but such models ignore the fact that there is also significant contribution in the strangers setting, where – in contrast to the partners setting – no signaling is possible, because subjects interact with the same partner only once. Based on recent findings that most subjects are conditional cooperators, reciprocal theories seem to be promising for solving the puzzle: Subjects contribute because others do so as well. However, more precisely, the experimental finding is that most conditional cooperation is imperfect, meaning that subjects try to cooperate less than others do (see, for example, Fischbacher, Gächter and Fehr 2001; Herrmann and Thöni 2009). As it is of course impossible for everybody to contribute less than the others, the only equilibrium strategy is to contribute nothing. It is this mechanism which results in an equilibrium of zero contributions in the reciprocity model of Falk and Fischbacher (2006).

In this paper, we suspect that overoptimism, which is a commonly known bias in human reasoning, is essential for understanding contributions to public goods. Chaudhuri (2011) formulates this idea as follows: "Conditional cooperators with [we add: over]optimistic beliefs regarding the contributions to be made by their peers will contribute to the public account." Such overoptimism has already been found in PGGs: Andreoni's conclusion that the decline of cooperation in his experiment is "due to frustrated attempts at kindness" (Andreoni 1995) implies that subjects are overoptimistic at the beginning of a game. Furthermore, overoptimism is reported in Croson (2007) for some subjects in the first game (albeit for zero subjects in the second one), in Neugebauer et al. (2009), in Fischbacher and Gächter (2010, henceforth FG), and in Ambrus and Pathak (2011) (compare their Table 4 on page 508). While all of these papers mention overoptimism, none of them incorporate it into a model. We do so and thereby pursue an argumentation similar to that of Orbell and Dawes (1991): Cooperators may establish high levels of cooperation simply by believing that others are cooperators as well, which can be evolutionary advantageous in certain circumstances. We test our theory by reducing the FG-design of the PGG by one parameter: FG elicit conditional contribution preferences in so-called contribution schedules and compare these preferences with actual contributions and beliefs in the PGG. As the contribution schedule already links beliefs to contributions, requesting both of these parameters in the PGG results in "too much" data: Either many subjects answer inconsistently or they consider unknown – and therefore uncontrolled – aspects in their contribution decision. For example, subjects might intend to signal cooperativeness in the first rounds to induce higher conditional contributions of co-players in later rounds. We reduce such noise in our one-parameter design: In our PGG, subjects cannot contribute directly. Instead, they are only asked for their beliefs about the contributions of their co-players. Knowing these beliefs, and knowing the contribution schedule, we can compute preferred contributions directly. Due to the mechanism of imperfect conditional cooperation, contributions that are different from zero are inevitably linked to overoptimistic beliefs in this design.

The contribution of this paper is threefold: Firstly, we propose a model which explains individual behavior in PGGs. Our model is based on that of DK, but extends their approach by allowing for overoptimistic, boundedly rational behavior. Secondly, we experimentally demonstrate that our model does capture the relevant determinants of subject behavior. Thirdly, using our model to predict actual behavior, we find that predictions are precise for most subjects, and we explain why they are imprecise for some subjects.

The rest of the paper is structured as follows: In Section 2, we present our model and formulate hypotheses. In Section 3, we explain our experimental design in more detail. Section 4 presents the experimental results and tests our hypotheses. In Section 5, we report how well our model can predict individual contributions. Section 6 concludes.

2 The Model

2.1 The Dufwenberg/Kirchsteiger Model

The DK theory of sequential reciprocity assumes that people want to reciprocate kindness with kindness. Applying their model to a linear PGG, a player's utility function is as follows:

$$U_i = 1 - (1 - MPCR) \cdot g_i + MPCR \cdot \sum_{j=1}^{J} g_{ij} + \sum_{j=1}^{J} Y_{ij} \cdot \left[MPCR \cdot \left(g_i - 0.5\right)\right] \cdot \left[MPCR \cdot \left(g_{ij} - 0.5\right)\right]. \tag{1}$$

DK assume that i's utility function consists of two terms, weighted against each other with exogenously given non-negative reciprocity parameters Y_{ij}. If i is a free rider ($Y_{ij} = 0 \ \forall \ j$), utility is equal to the monetary payoff (which we normalize to 100 % of the initial endowment). The monetary payoff is calculated given that own contributions to the public good have a negative yield of 1 – *MPCR*, whereas each of the J co-players' contributions yield the *MPCR* to player i. Thereby, g_i (g_{ij}) describes the percentage of the endowment that i is contributing (that i believes j will contribute). Free riders want others to contribute, but will not contribute themselves. In contrast, a reciprocal subject will contribute if Y_{ij} is sufficiently high and if he believes that $g_{ij} > 0.5$: Such co-players' contributions are considered to be kind, because they are higher than the reference point of 0.5, and will therefore be reciprocated: Being kind yourself then pays off in the form of reciprocal utility. In this case, reciprocal players will contribute their whole endowment. In turn, with $g_{ij} \leq 0.5$, i will never invest into the public good. However, these predictions only qualitatively mirror behavior that is typically observed in the lab: Subjects contribute significant amounts if the contributions of the others are below 50 %, and they typically do not contribute their whole endowment even when their beliefs are above that threshold.

More importantly, DK assume that players' beliefs are correct. However, this assumption does not fit the experimental finding that contributions are positive and that subjects only want to contribute a fraction of the others' contributions (Fischbacher et al. 2001): In equilibrium, players cannot contribute less than the others. To be able to predict positive contributions, we make the following modifications to the DK model, whereby the model does not lose its predictive power for games other than PGGs. In contrast, for example, behavior in trust games can also be predicted more precisely if our modifications are used (Breuer and Hüwe 2014).

2.2 Modifications

1. To simplify, we model the reciprocity parameter Y_{ij} to be independent of j: Subjects in the lab play anonymously and have no possibility to condition their strategies on individual group members.

2. To determine kindness, we use a different reference point than DK do, who themselves admit that their reference point was chosen without deep justification. DK measure the kindness of i to j by comparing j's material payoff with the average of the highest and the lowest material payoff that i can grant to j. Instead, our reference point relies on the status quo: We consider it to be kind if a co-player is made better off compared to his situation before the game starts. Accordingly, the reference strategy is to contribute

nothing, meaning that a small contribution is already regarded as (slightly) friendly. This is a plausible assumption, as each contribution is costly and comes with the risk of being exploited. Therefore, the kindness of i to a co-player j is $\kappa_{ij} = MPCR \cdot g_i$, and unkindness need not be considered in the following.

3. Most importantly, we assume that players are overoptimistically biased. We suggest formally considering overoptimism by adding a factor ε, resulting in biased beliefs about j's kindness. Accordingly, i's belief about j's contribution, g_{ij}, is no longer equal to j's actual contribution, g_j, We differentiate between the correct belief about j's contribution, $g_{ij,c}$, and the overoptimistic one, $g_{ij,o}$ (furthermore, i's belief about j's overoptimistic belief about k's contribution is denoted as $\mathrm{g}_{ijk,o}$), and we model $\mathrm{g}_{ij,o}$ as

$$g_{ij,o} = g_{ij,c} + \left(1 - g_{ij,c}\right) \cdot \varepsilon = g_j + \left(1 - g_j\right) \cdot \varepsilon, \tag{2}$$

with $0 \leq \varepsilon \leq 1$. We formulate (2) such that ε increases $g_{ij,c}$ dependent on the latter's distance to the maximal possible contribution. This implies that overoptimism diminishes when contributions are already high, and it prevents beliefs from being larger than the maximal possible contribution. We model ε as identical for all subjects, which simplifies the formal analysis significantly. Assuming different degrees of overoptimism for different players would be a natural generalization of our approach.

4. Whereas kindness has been linear in g_i so far, we will assume in the following that it is concave. As suggested in DK, p. 291, the square root will therefore be used. Thus, i's utility function now has the following form:

$$U_i = 1 - (1 - MPCR) \cdot g_i + \sum_{j=1}^{J} MPCR \cdot g_{ij,o} + Y_i \cdot \sum_{j=1}^{J} \sqrt{MPCR \cdot g_i} \cdot \sqrt{MPCR \cdot g_{ij,o}}. \tag{3}$$

5. While DK assume that the reciprocity parameters of co-players and their strategies are known, subjects typically remain anonymous in experiments. Therefore, *i* has to estimate $g_{ij,o}$. To keep the model simple, we only consider the average of co-players' contributions to be relevant for *i*'s utility,

$$U_i = 1 - (1 - MPCR) \cdot g_i + J \cdot MPCR \cdot g_{ij,o} + Y_i \cdot J \cdot \sqrt{MPCR \cdot g_i} \cdot \sqrt{MPCR \cdot g_{ij,o}}, \quad (4)$$

whereas $g_{ij,o}$ represents the overoptimistically expected average contribution level in the subject pool. This can be justified by assuming that all co-players contribute equally. More complexly, one might assume that subjects estimate a probability distribution of contributions within the subject pool, and that they maximize their expected utility. We show such an approach in Appendix A (appendices in this paper are available from the authors upon request), but will not apply it to our experimental data for simplicity reasons.

6. Believing in only one (average) co-player type, maximizing (4) over g_i results in (see Appendix B)

$$g_i = \left(\frac{J \cdot Y_i \cdot MPCR}{2 \cdot (1 - MPCR)}\right)^2 \cdot g_{ij,o}. \quad (5)$$

Equation (5) implies that *i* c.p. contributes more in settings with higher MPCRs and more group members, which is typically the case (compare stylized facts 3. and 4.). As we vary none of these parameters, we simplify equation (5) by defining $\hat{Y}_i := \left(\frac{J \cdot Y_i \cdot MPCR}{2 \cdot (1 - MPCR)}\right)^2$:

$$g_i = \hat{Y}_i \cdot g_{ij,o}. \quad (6)$$

$\hat{Y}_i$ will also indicate *i*'s reciprocity parameter in the following, because it is simply an *MPCR*- and *J*-dependent transformation of Y_i. Equation (6) states that subjects want to contribute a constant proportion of the believed overall contribution level. This coincides with the experimental results reported in the literature (FG; Fischbacher et al. 2001; Fischbacher et al. 2012; Neugebauer et al. 2009), and with our own findings, see Section 4.2. To be more precise: It is found that people want to contribute with a selfish bias, meaning that $\hat{Y}_i \leq 1$. Furthermore, $\hat{Y}_i \geq 0$, because contributions cannot become negative. Subjects with $\hat{Y}_i = 0$ are free riders, whereas subjects with $\hat{Y}_i = 1$ are perfect conditional cooperators. Subjects in-between are imperfect conditional cooperators.

2.3 Deriving Equilibrium Contributions

Our equilibrium definition relies on Aumann (1987), who defines behavior as being in equilibrium if each player maximizes his expected utility, given his information. This does not necessarily imply that this information is correct, which allows the derivation of *overoptimistically biased* equilibria. In turn, in the following, we will refer to equilibria where players' beliefs are correct as *unbiased* equilibria. In the biased case, the following is assumed to be known: First, players are optimists, compare equation (2). Second, as players have reciprocal preferences, utility-maximizing behavior is given by equation (6). Combining these two equations, a biased equilibrium can be found where overoptimism and imperfect conditional cooperation balance each other out. Note that we use the variable $\hat{Y}_{ij}$ in the following to denote *i*'s belief about *j*'s reciprocity parameter. The index specifies that the variable refers to *i*'s belief, and does not, unlike DK do, define *i*'s co-player-dependent reciprocity parameter towards *j*.

$$g_{ij,o} = g_{ij,c} + \left(1 - g_{ij,c}\right) \cdot \varepsilon = \hat{Y}_{ij} \cdot g_{ijk,o} + \left(1 - \hat{Y}_{ij} \cdot g_{ijk,o}\right) \cdot \varepsilon = \hat{Y}_{ij} \cdot g_{ij,o} + \varepsilon - \hat{Y}_{ij} \cdot g_{ij,o} \cdot \varepsilon$$

$$\Leftrightarrow g_{ij,o} = \frac{\varepsilon}{1-(1-\varepsilon) \cdot \hat{Y}_{ij}}, \tag{7}$$

$$g_i = \hat{Y}_i \cdot \frac{\varepsilon}{1-(1-\varepsilon) \cdot \hat{Y}_{ij}}. \tag{8}$$

Equation (7) denotes *i*'s belief with respect to the contribution level of the group. All players have identical beliefs, which follows from the fact that all players have identical information. The higher the overoptimism parameter ε and the higher *i*'s belief with respect to the reciprocity parameter of *j*, $\hat{Y}_{ij}$, the higher the perceived contribution level of co-players is. Subjects reciprocate this level dependent on their own reciprocity parameter $\hat{Y}_i$, compare equation (8). With $\varepsilon = 0$, our model collapses into the unbiased equilibrium of contributing nothing. We now want to comment on the properties of our model.

First, it is essential that, according to equation (2), players overestimate contributions, not reciprocity, and that this is done in a disproportional way. In contrast, as $\hat{Y}_i$ and g_{ij} are multiplied in equation (6), assuming the belief bias to be proportional to g_{ij} would have the same effect as overestimating $\hat{Y}_j$. However, only overestimating $\hat{Y}_j$ does not lead to positive contributions as long as $\hat{Y}_j$ is believed to be smaller than one. We therefore assume that contributions are overestimated disproportionally. Overestimating contributions instead of overestimating reciprocity also implies that too high contributions can be observed even when players know the reciprocal inclination of their co-players. That is what is tested in Wolff (2013): His experimental design makes public the unbiased equilibrium, which results from the contribution schedules. Still, players believe that co-players will contribute more than is preferable according to the latter's conditional contribution preferences. This overoptimistic bias appears although subjects themselves play a best response to their beliefs. Wolff (2013) emphasizes that subjects even believe those co-players will contribute who are, according to their elicited preferences, free riders. This also follows from our model, which can be seen by setting $\hat{Y}_{ij} = 0$ in equation (7).

Second, as subjects' beliefs are wrong, our model necessarily comes with inconsistencies. This shortcoming must inevitably be accepted if one models equilibria in boundedly rational models, compare, e.g., Eyster and Rabin (2005), who discuss inconsistencies in their "cursed equilibrium" model. In our case, players assume that everybody shares the same belief with respect to the contribution level ($g_{ijk,o} = g_{ij,o}$), but simultaneously know that players want to contribute less than the others. In other words, players "hope" that co-players will contribute more than is optimal, but "know" that this will not be the case. This inconsistency appears in the first-order belief as well as in all higher-order beliefs. For example, *i* knows that *j* hopes that *i* will contribute too much, but knows that this will not be true. As mentioned above, the results of Wolff (2013) show that this logic does reflect how people think in PGGs.

Third, as in Eyster and Rabin (2005), we justify our approach by arguing that players do "not (fully) think through the logic" of the model. Alternatively, our model may be seen as an "as-if" model, which makes correct predictions, but captures human thinking in a conceived and therefore unrealistic way: For example, instead of assuming that players optimize but have biased beliefs, the same contributions as in our model can be derived by assuming that players have correct beliefs but somehow fail to optimize. As well, one may argue that players do not build higher-order beliefs in infinite depth. This opens room for proposing somehow unjustified but positive first- or second-order beliefs, inducing positive contributions: The findings of Nagel (1995) indicate that such considerations do play a role in human thinking. Arifovic and Ledyard (2012) and Dijkstra (2012) follow this path, which entails logical inconsistencies for sophisticated subjects as well.

We claim that our model provides significant help for understanding the stylized facts. In contrast to Ambrus and Pathak (2011), who rely on signaling and therefore predict no contributions in the strangers setting, we can explain such contributions (our model is compatible with such strategic considerations, but must be extended in order to capture them). Our model also enables an explanation of both beliefs and contributions, even in the first round (in contrast to Dijkstra 2012, who assumes first beliefs to be given due to social norms or introspection). Most importantly, we can make point predictions of individual behavior (in contrast to Arifovic and Ledyard 2012, who assume reactive learning by subjects based on a random "trial and error" process). Assuming a learning algorithm (compare Section 5.1), we can also explain typically decreasing contributions in repeated-round games. Summarizing, to our best knowledge, this is the first paper which reports the fit of individual point predictions to experimental data of PGGs.

2.4 Hypotheses

To prove our assumptions and to test the implications of our model, we formulate the following hypotheses:

Hypothesis 1: Subjects are either (im)perfect conditional cooperators or free riders.

Hypothesis 1 has already been supported by Fischbacher et al. (2001). Thus, we simply expect to replicate their findings. As Hypothesis 1 is essential to our model, we will nevertheless discuss our data with respect to this question. In a strict sense, Hypothesis 1 implies that nobody ever wants to contribute more than the others. We expect that this will not hold perfectly true. We will refer to contributions which exceed the believed average contribution level as *hyper-conditional contributions*, and we are well aware of the fact that they justify unbiased contribution equilibria above zero. We will address this point with Hypothesis 4.

Hypothesis 2: Subjects correctly predict the average reciprocal inclination of co-players.

According to equation (8), i's belief about average reciprocity in the subject pool determines his decision. In this context, we test whether $\hat{Y}_{ij}$ is built correctly or whether it is biased as well.

Hypothesis 3: Subjects are overoptimistic with respect to co-players' contributions. This is true even when subjects are not participating in the game.

Our second essential assumption is that subjects estimate in a systematically biased way. We will be able to measure overoptimism directly by comparing the beliefs of the participating subjects with the true contribution levels. Nevertheless, we introduce a robustness check for overoptimism in our experiment: We use a control treatment, where subjects take an outsider position and only estimate the others' contributions without themselves having a stake in the game. In contrast to the so-called "participants", one can be confident that these "estimators" are unaffected by any strategic considerations.

Hypothesis 4: Subjects will contribute more than predicted by the unbiased equilibrium.

Our modeling implies that any contribution is due to overoptimism. As already mentioned, some subjects may unconditionally (meaning: irrespective of beliefs) or hyper-conditionally contribute, justifying unbiased but positive equilibria. In that case, our overoptimism hypothesis will imply higher contribution levels than predicted with unbiased beliefs. Such a premium will be more meaningful in our one-parameter treatment (compare the introduction) than in the original FG-design because it can *only* be due to overoptimism.

Hypothesis 5: Subjects will contribute in the one-parameter treatment to the same extent as in the original FG-design.

Hypothesis 5 tests whether aspects other than overoptimism and reciprocity must be considered to explain contributions. Our model captures these two parameters only, and if we find no differences between the two treatments, we can conclude that these two parameters are the key determinants of PGG behavior.

3 Experimental Design

To test the hypotheses and to prove the predictive power of the individual model predictions, we conducted a PGG, the design of which we will present in the following (more details, such as the experimental instructions and several screenshots, are available in Appendix D). For the experiment, we used instructions and parts of the design of FG as far as these were published in their paper and in the corresponding web appendix, which ensures comparability. Accordingly, we conducted a standard PGG played over 10 rounds with MPCR = 0.4 and group size of four in the strangers setting (thus, contributions to the public good are multiplied by 1.6 and distributed among the four group members equally). Subjects received an endowment of 20 tokens per round, which was worth 80 euro-cents. As in FG, in each round, subjects had to estimate the others' average contributions. They also had to fill in a contribution schedule before the game started, which served to measure subjects' reciprocal preferences (compare Fischbacher et al. 2001, or Fischbacher, Gächter and Quercia 2012). The contribution schedule is based on the strategy method of Selten (1967), which enables – rather than measuring single contribution decisions – the revealing of subjects' complete strategies: For all possible average contribution levels of co-players (rounded to integers), subjects have to state how much they are willing to contribute. In contrast to FG, after contribution schedules had been completed, subjects were assigned to three different treatments: In Treatment 1 (called "standard treatment" in the following) subjects were confronted with a standard PGG, like that in FG. In Treatment 2 (the "belief treatment"), subjects did not have the possibility to decide on contributions during the game. Instead, contributions were directly computed from their belief regarding the average contribution level in the respective round and contribution preferences according to their contribution schedule. This novel design inextricably connects beliefs and contributions: If subjects had stated that they were imperfect conditional cooperators in their contribution schedules on average and believed in positive contributions on average (which we expected), then positive contributions had to be due to overoptimism. In Treatment 3, subjects did not participate in the game. Instead, their only task was to estimate the others' contributions, either in the standard treatment or in the belief treatment. We will now justify Treatment 3 in more detail.

In all treatments, we decided to incentivize beliefs regarding the co-players contributions in a noticeable form because there is evidence that incentivizing beliefs in PGGs significantly increases their accuracy, compare Gächter and Renner (2010). In each round, subjects were compensated with 50 euro-cents for each correct estimation. The compensation was reduced by 1 euro-cent for each percentage point (computed in relation to the solution space) of deviation from the correct value. For example, if the correct answer was 1 and a subject had estimated 3, payments were reduced by 10 euro-cents (the solution space was 0 to 20). Payments could not become negative. We chose such comparably high incentives to be sure that any overoptimism which might be found is in fact a robust bias. However, there is concern that incentivizing beliefs might somehow affect subjects' decisions: Being incentivized, beliefs themselves become part of the payoff-relevant action space (Armantier and Treich 2013; Blanco et al. 2010; Gächter and Renner 2010). For example, risk-averse participants might hedge lower than expected contribution levels with lower than true beliefs. Thus, even if subjects state their correct beliefs, their true beliefs might be overoptimistic. We therefore introduced Treatment 3, where estimators' beliefs served as a neutral benchmark for participants' beliefs. Since estimators were not engaged in the game, we consider their beliefs to be unbiased by any strategic considerations. As they did not receive endowments, we compensated them with an additional payment of 5 euros each.

FG conducted a "P-experiment" to give subjects an incentive to indicate their true preferences in the contribution schedule. In our design, all subjects were informed prior to the contribution schedule stage that they would be randomly assigned to one of the 3 treatments and that in Treatment 2 their contribution schedule would predetermine their contribution behavior. Thus, all subjects had an incentive to fill out the schedule correctly, because the schedule turned out to be irrelevant in Treatments 1 and 3, but was payoff-relevant in the case of a subject being assigned to Treatment 2. Proceeding as described, we did not ask for contribution preferences conditional on *real* contribution levels (unlike FG did) but on *believed* ones. This procedure is incentive-compatible, and subjects should elicit their true preferences in the contribution schedule: The belief treatment restricts subjects' strategy space, but, if the schedule is filled out truthfully, it only eliminates suboptimal strategies. For example, unconditional cooperators should enter the same number in each field of their contribution schedules, because this will result in constant contributions irrespective of their beliefs, which is exactly what unconditional cooperators want. Furthermore, as the belief elicitation was incentivized as well, it was optimal for subjects to state their true beliefs.

There is much debate in the literature about the degree of confusion in PGGs. To test whether subjects had understood the incentive structure of the game, after explaining the PGG, we asked (due to time constraints) 6 out of 10 control questions presented in the appendix of FG. However, it might have been possible that subjects who had understood the incentives gave wrong answers because they miscalculated. Others with little understanding might have answered correctly. Thus, after contribution schedules had been completed, we asked subjects to explain their inputs in the contribution schedules in two or three sentences in written form.

Afterwards, subjects had to estimate average values of the others' contribution schedules. We incentivized these estimations, which we call the "belief schedule", with up to 50 euro-cents. This schedule thus directly asks for subjects' beliefs about the average degree of reciprocity within the subject pool.

Having completed the schedules, subjects were randomly assigned to their roles and the PGG started. In each round, estimators could see the history of their guesses and of the others' average contributions. Participants could additionally see their own contributions, their money privately kept, and the repayments from the public good for all past rounds. Roles did not change between rounds, but group members did change within the two participant pools, and subjects were repeatedly made aware of that fact. We conducted four sessions with 30 or 29 subjects each. 44 subjects were pooled to play the standard treatment and 48 played the belief treatment; 13 subjects served as estimators in the standard treatment, and 14 estimators were assigned to the belief treatment, meaning that one to two estimators were assigned to each group of participants. All experiments were computerized, using the software z-Tree (Fischbacher 2007). The experiments were conducted in the Laboratory for Experimental Economics at RWTH Aachen University in August 2012 and March 2013. Participants were – apart from a few exceptions – students from various disciplines, with the majority studying business administration or business administration with engineering.

4 Experimental Results

4.1 Descriptive Statistics

We start with descriptive statistics. Table 1 shows average parameter values and gives an overview of how the model variables translate into the experimental parameters. Fig. 1 displays average entries in the contribution schedules and in the belief schedules.

<< Insert Fig. 1 and Table 1 about here >>

The "contribution lines" in Fig. 1 show how much subjects are, on average, willing to invest into the public good, depending on the estimated average contribution of the co-players. We present average contributions both for all subjects and without "other" subjects (see below): Omitting the "other" subjects makes the contribution line steeper and displays much fewer hyper-conditional contributions at low contribution levels. The contribution lines are similar to those of Fischbacher et al. (2001) and Fischbacher et al. (2012), although their contribution schedules are dependent on *actual* average contributions, while ours are dependent on *believed* ones. We conclude that our method does not distort the results. Especially, on the individual level, we do not find serrated schedules. With such schedules, subjects could to some degree undermine our one-parameter design and keep some control over their contributions, because in serrated schedules, different contribution levels are assigned to very similar beliefs.

The “belief line” in Fig. 1 shows how much all subjects believe that others want to contribute (always on average), again depending on the estimated contribution level of the co-players. Qualitatively, it looks very similar to the contribution line. We will discuss this in more detail when testing Hypothesis 2.

Figures 2A and 2B display average results of the PGG from rounds 1 to 10. They show contribution levels and beliefs of estimators and participants in the standard treatment (Fig. 2A) and in the belief treatment (Fig. 2B).

<< Insert Fig. 2A and 2B about here >>

Both figures replicate the typical behavior of subjects in PGGs: Average contributions start at around 50 % of the endowment and decline with repetition (although the decline is less pronounced than typically reported in the literature). At a first glance, behavior in both treatments seems to be very similar. Also note that the presentation of average behavior hides the large amount of disparity which can be observed between different subjects, groups, and even sessions: We find (average) contributions of 0 to 20, 1.5 to 17.8, and 5.38 to 11.8 tokens on the individual level, group level, and session level in round 1 and respective values of 0 to 20, 0.75 to 16, and 1.0 to 10.2 tokens in round 10. This large variability underlines the necessity for explanations of individual behavior and its interaction on the group and session levels.

4.2 Hypotheses Testing

Hypothesis 1: Subjects are either (im)perfect conditional cooperators or free riders.

We classify subjects according to the scheme proposed in Fischbacher et al. (2001). Hence, 69 % of our subjects are conditional cooperators (their schedules are increasing und (weakly) monotonic, or not strictly monotonic but show a highly significant and positive Spearman rank correlation coefficient between own and others' contribution). 8 % are free riders (schedules contained '0' in all 21 entries). Moreover, we find only 6 % "hump-shaped" patterns (increasing own contributions at low average contribution levels and decreasing contributions for high levels). The kink in the schedules of these hump-shaped subjects is on average at a believed contribution level of 12.3 tokens. No schedule kinks before a contribution level of 9 tokens. Thus, for typical contribution levels, hump-shaped subjects behave like conditional cooperators. 16 % of all subjects do not fall into these three categories, but show "other" patterns. Therefore, our results are very similar to those of FG and Fischbacher et al. (2001). We assert that equation (6), which defines conditional contribution preferences in our model, captures the most of the elicited contribution schedules.

Most of the "other" patterns correspond to flat contribution schedules, which could be interpreted as a preference for unconditional cooperation. However, both our control questions and the written check of subjects' understanding indicate that "other" subjects were not fully aware of the incentive structure of the game, meaning that their contribution and belief schedules contain little information: "Others" answered only 2.6 out of 6 control questions correctly (compared to 3.7 correct answers for the remaining subjects). Also, their written explanations revealed a limited understanding (typical answers were: "values are chosen arbitrarily", "did not understand the task", "always invested everything because this maximizes my payoffs"). Therefore, with respect to the proportion of *meaningful* preference elicitations, the proportion captured by our model might be even larger.

Irrespective of classifications, Hypothesis 1 implies that subjects want to contribute less than the others. According to Fig. 1, this is on average only true for contribution levels above five tokens. At this point, "estimated contributions of the others" equal the average "own contribution": Unbiased subjects should believe in average contributions of five tokens, because by doing so, subjects will actually contribute five tokens on average. Thus, at a first glance, significant contributions are possible even without assuming overoptimistic beliefs. However, Fig. 1 also shows that most of the hyper-conditional contributions stem from “other” subjects: Excluding these, the unbiased contribution level is one token, which almost matches our zero-token prediction. We are therefore interested in whether the contribution schedules are precisely predicting actual contributions. To answer this question, we analyze behavior in the standard treatment in the following, as only this treatment allows deviations from the schedules.

<< Insert Table 2 about here >>

Contributions in the standard treatment have already been analyzed in FG. Regression (1) in our Table 2 replicates their results (with our R² being 26 percentage points higher), compare their model (3) in their Table 2, p. 549. FG conclude that subjects contribute a weighted average of “predicted contribution” (which is the contribution calculated from a subject’s “belief” and his contribution schedule) and “belief”, meaning that their willingness to conditionally cooperate is higher than predicted from the contribution schedules. In the following, we want to disentangle this result on the subject-type level. In Regression (2), we display how “contribution” actually depends on “belief”: “Contribution” is slightly smaller than “belief” plus the “constant”. Compared to the average contribution schedule in Fig. 1, we observe two differences: First, subjects do not contribute in a hyper-conditional way, even when beliefs are low. Second, “contribution” almost equals “belief”, meaning that subjects are almost perfect conditional cooperators. We find that this is not due to an increase in subjects’ willingness to *conditionally* cooperate. While it is true that subjects contribute more than stated in their contribution schedules (for conditional cooperators, the effect is 0.76 tokens per subject per round on average), this behavior is not conditional on “belief”: For conditional cooperators, the difference between “predicted contribution” and the actual contribution is not correlated with “belief” (Pearson, ρ = 0.024, p = 0.637). Instead, subjects with a high willingness to cooperate also believe that others will contribute on high levels: “Belief” and $\widehat{Y}_i$, measured as the “slope of the contribution schedule” (according to a linear regression without constant) are significantly correlated (Pearson, ρ = 0.239, p = 0.000). This “false consensus effect” (Ross, Greene and House 1977) explains why Regression (2) is steeper than the average contribution line in Fig. 1. This fosters cooperation: The higher $\widehat{Y}_i$ is, the larger the effect of overoptimism on contributions is. Regression (3) confirms that “others” do not contribute hyper-conditionlly. Instead, they behave like imperfect conditional cooperators – note that “contribution” is close to but below “belief” in Regression (3). Furthermore, the contribution schedules of “others” are meaningless – the coefficient of “predicted contribution” is small and insignificant. We interpret the results as follows: “Others” had not indicated that they were conditional cooperators in their schedules, but behaved as such in the PGG, either because they had not understood the contribution schedule, but did understand the intuition of the PGG, or because they were simply imitating the behavior of their group members.

Regression (4) differs from Regressions (1) and (3) in the way in which we predict contributions: Instead of using “predicted contribution”, we simplify this variable to “pred. contribution slope”, which is calculated by multiplying “belief” by the slope in subjects’ contribution schedules. We propose such a simplification in our model, $g_i = \widehat{Y}_i \cdot g_{ij}$. As the contribution schedule of “others” has been found to be meaningless, we replace their reciprocity parameters in the following: In such cases, we use $\widehat{Y}_i = 0.73$, which is the average reciprocity parameter of all non-“other” subjects. As already mentioned, this can be justified by the observation that “others” contribute similarly to the rest of the subject pool. Doing so, Regression (4) underlines our main point of this section: Contributions of subjects can be well described by using only $\widehat{Y}_i \cdot g_{ij}$. Thus, our results support Hypothesis 1.

Hypothesis 2: Subjects correctly predict the average reciprocal inclination of co-players.

As already mentioned, on average, subjects predict the average contribution schedule of the others quite precisely. On the individual level, the *average* absolute mistake per entry in the belief schedule is 3.28 tokens. While we do not discuss whether this is precise, we find that positive and negative mistakes do not balance out; on average, subjects overestimate each entry by 0.44 tokens. This is not significantly different from zero (*t*-test, p = 0.124). We are also interested in whether the slope of the contribution schedule is misestimated: While the average contribution schedule has a slope of 0.74, subjects believe in 0.81 (difference significant, *t*-test, p = 0.012; p = 0.002 if "others" are excluded). This is also visible in Fig. 1: The belief schedule is steeper than the contribution schedule. According to our model (compare equation (8)), this bias also fosters cooperation.

Again, on the individual level, we detect a distinct false consensus effect: $\hat{Y}_i$ and $\hat{Y}_{ij}$ are significantly correlated (Pearson, ρ = 0.595, p = 0.000). People believe others to be similar to themselves. While it does not matter for contributions whether free riders underestimate the reciprocal inclination of others, contributions rise if conditional cooperators overestimate others' willingness to conditionally cooperate. Thus, Hypothesis 2 is not supported by our results, as systematic biases can be found. However, these biases foster cooperation, compare also Section 5.2.

Hypothesis 3: Subjects are overoptimistic with respect to co-players' contributions. This is true even when subjects are not participating in the game.

Indeed, according to Figures 2A and 2B, subjects continuously overestimate contributions. Participants overestimate by 0.67 tokens per round in the standard treatment (0.62 tokens in the belief treatment); estimators overestimate contributions in their groups by 0.97 (0.60) tokens.

In all cases, estimation mistakes are significantly different from zero in all treatments (*t*-test, all *p*-values < 0.061)

The estimator groups allow us to conduct a robustness test for our assumption that subjects have overoptimistic beliefs: First, one might be concerned that risk-averse participants in the standard treatment hedge lower than expected contributions of the co-players with lower than true beliefs. However, with respect to the numbers reported above, mistakes do not differ significantly between participants and estimators, meaning that hedging considerations play no important role in our design. In any case, the hedging argument implies that the incentivizing of participants' beliefs, if at all, leads to an *under*estimation of their true overoptimism. Second, estimators may differ from participants, because participants are emotionally involved in the game, while estimators are not: As participants directly benefit from contributions of others, they hope that others will contribute and may, due to wishful thinking, believe in higher contributions than estimators do. This argument implies that estimators should have lower beliefs than participants, which is also not the case. Third, one may notice that estimators have beliefs about the average contribution of four participants in each round, whereas participants only estimate contributions of their three co-players. Therefore, it is easier for estimators to be precise on average contribution levels, as their estimations are less vulnerable to outliers in their groups. As estimators are overoptimistic as well, our overoptimism hypothesis is also robust with respect to this argument. We summarize that our data support Hypothesis 3.

Hypothesis 4: Subjects will contribute more than predicted by the unbiased equilibrium.

We defined an equilibrium to be unbiased if all players' beliefs, connected to the contributions by the contribution schedules, are correct. Contributions depend on beliefs in a subject-specific way, resulting in group-specific equilibria. Knowing the group assignment mechanism, we can ex post derive all equilibria. Due to hyper-conditional entries in some contribution schedules, equilibria different from zero exist. In most groups, we find only one equilibrium. In other groups, no exact equilibrium exists because none of the possible belief combinations is associated with its predicted contributions. In these rare cases, we define those beliefs to be in equilibrium which result in the smallest possible belief mistake. Furthermore, several equilibria per group are possible: For example, among solely perfect conditional cooperators, each contribution level can be in equilibrium. In such cases, we determine the smallest and the largest possible equilibrium in that group, and we will analyze both of them below.

Proceeding as described, we can compare actual contributions to equilibrium ones. If we use the lowest possible equilibria, we find that subjects contribute 2.13 tokens more on average than predicted by the equilibrium (actual: 7.22 tokens; predicted: 5.09 tokens). The difference is significant (*t*-test, $p = 0.000$). However, with respect to the highest possible equilibria, subjects contribute 0.05 tokens too short (predicted: 7.27 tokens; difference not significant, $p = 0.808$). Thus, one could conclude that subjects do not estimate too high: Instead, they estimate too low, because with slightly higher beliefs, equilibrium contributions could be realized as well. However, requesting equilibrium play implies that subjects can always coordinate on the highest possible equilibrium. This is impossible, as neither the contribution schedules nor the random group composition was announced: In our data, we find group equilibria between 0 and 20 tokens. Thus, simply increasing all beliefs by 0.05 tokens would of course not result in equilibrium play. Instead, it would increase belief mistakes: Computing contributions from the contribution schedules, subjects' actual beliefs result in an overestimation of contributions by 3.39 tokens on average (not coordinating on an equilibrium typically implies too high beliefs, as most entries in the contribution schedules are below the bisecting line). Increasing *each* belief (if the belief is not already at 20 tokens) worsens their average preciseness. Instead, lowering actual beliefs (if possible) improves the preciseness: Beliefs are most precise if lowered by 3 tokens. This procedure indicates that subjects indeed have too high instead of too low beliefs. Thus, our results confirm Hypothesis 4. Also note that the discussed equilibria can be seen as an upper bound for unbiased behavior: As it is impossible for subjects to predict contributions on the group level, one may assume that they try to reciprocate the contribution level of the whole subject pool, rather than the levels of their co-players. In that case, we can make a clear equilibrium prediction, visible in Fig. 1: Beliefs above five tokens are biased. With that benchmark, beliefs – and therefore contributions – are significantly too high. Furthermore, one may consider that high equilibria are almost always only realized because “other” subjects claim to contribute hyper-conditionally. However, the analysis of Hypothesis 1 has shown that such behavior does not appear, at least not in the standard treatment. Excluding “other” subjects from the analysis (almost) results in equilibrium predictions of zero, compare Fig. 1 again. Referring to the actual relationship between beliefs and contributions justifies an equilibrium of zero as well, because according to regression model (2), actual contributions match actual beliefs most precisely at a contribution level of zero.

Hypothesis 5: Subjects will contribute in the one-parameter treatment to the same extent as in the original FG-design.

Our data confirm Hypothesis 5: Subjects in both treatments start with average contributions of between 8 and 9 tokens and end up contributing between 5 and 6 tokens. In all ten rounds, contributions in both treatments do not differ significantly from each other (*t*-test, all *p*-values > 0.235).

In the preceding analysis, we have identified effects which lead to differences between both treatments: In the standard treatment, subjects overcontribute (defined as contributing above the prediction of the contribution schedule), and “others” do not behave at all in the way that their contribution schedules had predicted. Such behavior is not possible in the belief treatment. However, these effects are small, or they balance each other out. Thus, both treatments are comparable with each other, and as overoptimism is the decisive cause for contributions in the belief treatment, we conclude that this is true for the standard treatment as well. Furthermore, in the belief treatment, the fact that subjects cannot contribute less than the others is made more explicit. Unbiased subjects should therefore start contributions at a lower level and decrease them faster than in the standard treatment (note that most belief schedules demonstrate beliefs in imperfect conditional cooperation). In contrast, biased subjects can believe that they will contribute less than the others, making it irrelevant if they play the standard treatment or the belief treatment. Finding no difference between both treatments supports the idea of subjects being biased. Thus, our results confirm that contributions to PGGs can be explained using only reciprocity and overoptimism.

5 Prognostic Power of the Model

5.1 Model Application

While we have presented statistical analyses to support our hypotheses, we will now report the prognostic power of our model that was presented in Section 2. Our model abstracts from phenomena such as overcontributions or hyper-conditional contributions. Therefore, the question is to what degree are we nevertheless able to capture individual behavior in PGGs. To answer this question not only for the first round, we have to clarify how we can capture the repeated-round structure of our PGG setting. As we use the strangers setting, each round can be seen as an independent game: Although this has not been perfectly true, we assume that subjects meet their co-players only once, meaning that no signaling and no repeated reciprocal exchange with the same person were possible. Rounds are independent, except that subjects learn about the reciprocity inclination of the subject pool as the game proceeds.

Our model explains both beliefs and own contributions. Nevertheless, let us first assume that beliefs are given and that we only predict contributions: This is easily be done by applying equation (6). Therefore, in our Model (1), we multiply subjects actual beliefs by their reciprocity parameter in order to predict contributions. This corresponds to Regression (4) in Table 2, but sets the constant of the regression to zero and replaces the coefficient of “pred. contribution slope” by one. We regard Model (1) as an upper bound for the predictive power of the following two model variants, because we have to expect that predictions will be less accurate if we do not use our information regarding subject’s beliefs.

We also want to endogenize beliefs. To do so, we use our biased equilibrium concept as stated in equation (7). For an application, we must clarify how players determine ε and $\hat{Y}_{ij}$. As we assume biased thinking, it is conceivable that these variables are subject-dependent parameters. Furthermore, with repetition, players may learn and reduce their biases over time. We will proceed as follows: We will hold ε constant for all subjects and all rounds. In turn, we will use subject-dependent $\hat{Y}_{ij}$-values, which are derived from i’s belief schedule prior to the first round. In later rounds, $\hat{Y}_{ij}$will be updated, which we will explain below. We do not update ε because this would complicate results and we could not disentangle it from the updating process of $\hat{Y}_{ij}$. Allowing individual parameters for $\hat{Y}_{ij}$, we automatically incorporate our finding that the false consensus leads to a correlation between $\hat{Y}_i$ and $\hat{Y}_{ij}$, especially at the beginning of the PGG. While holding ε constant simplifies the model application, we are aware of the fact that it may be more realistic – but more complex – to let $\hat{Y}_{ij}$ converge towards the true reciprocal inclination of the subject pool, and to let ε decline simultaneously. In any case, since $\hat{Y}_{ij} \leq 1$ must be assumed in order to avoid implausible equilibrium beliefs, we lower $\hat{Y}_{ij}$ for some subjects to 1. This affects 15 (out of 92) participants.

From FG, we know that after round one, i builds his belief by taking the average of his previous belief and the observed contribution level of the previous round. By reproducing their regression analysis, we confirm their finding (results not displayed). Accordingly, we use this belief building process in the following. We also report our observation that contribution levels prior to the last round actually do not additionally affect beliefs significantly (regression not displayed). In our model context, with ε being known, the belief building process can be interpreted as follows: Observed contributions from previous rounds are used to calculate co-players' believed reciprocity inclination of the subject pool. This is done by solving equation (7) for $\hat{Y}_{ij}$. Thus, if *i* observes lower than expected contributions, he will conclude that the others believe in less reciprocation than he himself does and will partly adopt their belief. Using this procedure, $\hat{Y}_{ij}$ does not converge towards the true reciprocity inclination of the pool: Instead, as subjects always observe lower than expected contributions on average, $\hat{Y}_{ij}$ will converge to zero, corresponding to the minimum expected contribution level of $g_{ij} = \varepsilon$. Contributions cannot increase in the whole subject pool, but single subjects will increase their contributions if these subjects were matched with highly reciprocal subjects before.

We will apply this belief building process in Models (2) and (3). In Model (2), we will use actual contributions from the experimental data to update *i*'s belief. Thus, Model (2) will make round-to-round predictions. In contrast, in Model (3), we will use no actual subject data from the PGG. Instead, we will update beliefs by using the *modeled* previous contributions. Thus, we will predict contributions for all rounds with the information given before round 1 starts: We will only use the contributions schedules, the belief schedules, and the group assignment mechanism.

We assess the preciseness of our model variants by reporting "R^2" (1 minus quotient of average squared prediction mistakes to squared variations of all contributions). R^2 can be computed for the whole subject pool, but also for subgroups of four individual subjects: In the latter case, for example, "average squared prediction mistakes" refers to the prediction mistakes with respect to the 10 contribution decisions of a subject. "All contributions" is defined as contributions of all participants in both treatments: As we find both treatments to be very similar, we merge the data. R^2 (of all participants) will also be used to calibrate ε: We determine

ε such that R^2 in Models (2) and (3) is maximized. We will show below that our results are not driven by this optimization procedure.

5.2 Results

We present our results as follows: Table 3 displays a summary: For all three models, we present R^2 on an aggregated level, and on the treatment level. Additionally, we explain the results of Model (2) in more detail. Here, we show which subjects can be well explained, and which not. A complete survey of model results and its determinants for each subject can be found in Appendix C.

<< Insert Table 3 about here >>

Calibrating ε as described above, we get $\varepsilon = 0.225$. We will comment on the calibration process below, and start the analysis of the results by discussing Table 3: It shows that Model (1) explains contributions of all participants with an R^2 of 0.68. R^2 of the standard treatment is, with 0.54, only slightly smaller than R^2 in Regression (4) in Table 2. Accordingly, omitting the constant from Regression (4) and setting the regression coefficient to one almost does not affect the results. In the belief treatment, R^2 is higher because subjects cannot deviate from their schedules. This result defines an upper bound for our model: 20 % (100 % − 80 %) of the predictive power is lost because subjects do not exhibit strictly proportional contribution schedules. Another 26 % (80 % − 54 %) is lost if subjects can deviate from their schedules.

Comparing R^2 in Model (2) with that of Model (1), we find that the predictive power drops by 25 percentage points (0.68 − 0.43) if beliefs are endogenized. Another 9 percentage points (0.43 − 0.34) are lost if contributions are not predicted on a round-by-round basis, but, as done in Model (3), for all 10 rounds in one go. We will now turn to Fig. 3, where we present Model (2) predictions and actual contributions of selected subjects graphically. The subjects are selected such that the factors which drive our outcomes can be exemplarily explained. The results hold similarly true for Model (3).

<< Insert Fig. 3 about here >>

Fig. 3 immediately makes it clear how heterogeneous the contribution patterns are. Contributions can be high, low, increasing, decreasing, or constant. Our model explains these patterns by using subjects' reciprocal inclination, their beliefs with respect to the reciprocal inclination of others, and their information with respect to previous contribution levels. The diagrams in Fig. 3 are sorted downwards by R^2. We start with subject 26, who is a free rider assigned to the standard treatment. As she sticks to her contribution schedule, our predictions perfectly match her actual contributions. Only one of our seven free riders contributes, leading to bad predictions in that case (not displayed). Subject 73 is that non-selfish subject, whose behavior we can predict most precisely. This example demonstrates that contributions only decline on average. Single subjects can considerably increase their contributions if they are unexpectedly matched with others who contribute on high levels. Such increasing contributions are also visible for subject 111. Subjects 25 and 101 are examples that contributions on both low levels as well as high levels can correctly be predicted. Subject 1 is classified as "others". We already pointed out in Section 4 that "others" do not stick to their schedule if they are allowed to: In the standard treatment, "others" behave like conditional cooperators. Subject 1 is an example of the fact that behavior of such subjects in Treatment 1 can be well-predicted by using $\hat{Y}_i = 0.73$, which is the average reciprocity parameter of non-"other" subjects.

Subjects 111 and 58 demonstrate nicely that subjects react on previously observed contribution levels. We do not display beliefs, but contributions are always predicted to be a subject-dependent fraction of beliefs; thus, contributions only change if beliefs change. According to the updating process of beliefs, contributions rise (decrease) if beliefs in the previous round were below (above) the actual contribution level. Subject 111 was pessimistic in rounds 7 and 8, leading to increasing contributions in rounds 8 and 9. More typically, subjects are optimistic, which induces declining contributions: Subject 58 is a perfect conditional cooperator, who believes in a contribution level of 20 tokens in round 1. Four times in a row, he has do adapt his overoptimistic beliefs, leading to a rapid decline of his contributions (in the comment field at the end of the experiment, this subject actually expresses frustration with his co-players). Subject 77 has hump-shaped preferences. As long as these subjects estimate contribution levels within the increasing part of their schedule, contributions can be predicted quite precisely. As the linear regression underestimates the slope of this schedule part, contributions are underestimated. Subject 101 is an example of some subjects overcontributing: His actual beliefs fluctuate between 10 and 16 tokens, meaning that contributions of 20 tokens cannot be predicted.

Subject 107 is categorized as “others”: Like many “other” subjects, subject 107 entered a flat contribution schedule; in this case, at 20 tokens. The graph shows that such subjects do have to contribute 20 tokens in the belief treatment. But we predict belief-dependent contributions, which lead to high prediction mistakes. Similarly, there are cases where conditional cooperators stick to their schedule, but their predictions are bad because these schedules are not linear: Subject 46 wants to contribute everything if the others do so as well, but wants to contribute zero for beliefs below 10. However, none of her beliefs exceed 10 tokens, and consistently, subject 46 contributes nothing. Due to the proportional relationship of beliefs and contributions in our model, we instead predict positive contributions. Two other features can be observed for subject 46: First, we predict too high beliefs in the first two rounds. This happens if subjects state high $\widehat{Y}_{ij}$, but do not actually believe in high initial contributions. Second, we always predict $g_{ij} \geq \varepsilon$, compare equation (7). Therefore, with ε being positive, we will never predict zero contributions for conditional cooperators. The last but one graph shows that contributions can be explained very precisely on an aggregated level. This is true for Model

(3) as well, compare the last graph. This is not surprising: If individual contributions can be predicted, aggregated results are precise as well.

Also note that our results are quite robust with respect to the calibration of ε: Calibrating single sessions, we get $0.125 \leq \varepsilon \leq 0.375$. First of all, this indicates that ε fluctuates from session to session, occurring from the fact that the subject pool is very inhomogeneous and that we only have between 20 to 24 participants per session. However, our results do not react sensitively on ε: With R^2 being optimized over all sessions, R^2 equals 0.384 in Models (2) and (3). Using ε = 0.125 (0.375), R^2 only drops to 0.365 (0.297). Thus, our results do not depend on an exact determination of ε. Finally, note that the calibrated ε is, with 0.225, much higher than the ε, which can be computed from actual contributions and beliefs: This value equals only 0.05, compare Table 1. Thus, the calibration process captures the fact that the contribution schedules underestimate subjects' willingness to cooperate.

Table 3 summarizes our results: Generally, we can group our subjects as follows: In the standard treatment, those subjects can be predicted who do not deviate from their contribution schedules very much: In Table 3, we filter such subjects who deviate from their schedule by less than three tokens on average. In turn, predictions for the remaining subjects are bad. In the belief treatment, as subjects cannot deviate from their schedule, predictions are precise if subjects have linear conditional preferences (meaning that the regression which determines $\hat{Y}_i$ is precise with $R^2 \geq 0.8$), and if they express more or less precise beliefs (belief mistake < 4 tokens on average).

Model (3) can also serve to disentangle the sources of contributions. As already mentioned, $\varepsilon > 0$ is a necessary condition for positive contributions in our framework. However, contributions would be lower if $\hat{Y}_{ij}$ was not positively biased. This effect is reinforced by the fact that due to the false consensus effect, this bias particularly occurs for subjects with high reciprocal inclinations. We now want to report how contribution levels change if these parameters are ceteris paribus varied. In round 1 (10), participants are predicted to contribute 11.1 (4.5) tokens on average. With $\varepsilon = 0$, zero contributions will be predicted. Therefore, in what follows we take subjects' overoptimism $\varepsilon > 0$ regarding the other players' contributions as given. Now reducing each $\hat{Y}_{ij}$ by 6 %, which retains the dispersion in individuals' beliefs with respect to the reciprocity of others, but conforms to the true reciprocity level of $\hat{Y}_i = 0.75$ on average, reduces contributions from 11.1 (4.5) tokens to 9.8 (4.2). With all subjects having identical but biased beliefs about reciprocity, $\hat{Y}_{ij} = 0.80 \ \forall \ i$, we abstract from the dispersion in reciprocity beliefs and focus on the corresponding pure overoptimism component. This decreases contributions from 11.1 (4.5) tokens to 8.9 (4.0). Assuming that each subject correctly estimates $\hat{Y}_j$ – which eliminates both biases – 8.1 (3.8) tokens are predicted. Thus, while the false consensus effect and overoptimism with respect to $\hat{Y}_{ij}$ do foster contribution, they are not the main drivers. Instead, contributions are primarily induced by an overestimation of the co-players' contributions.

6 Conclusion

This paper has shown that empirically observed contributions to a PGG can formally be explained with the help of the reciprocity model of Dufwenberg and Kirchsteiger (2004) if the reference points are changed and if overoptimism with respect to the contributions of others is incorporated. Such overoptimism leads to the prediction of positive contribution levels, which are then reciprocated. Learning that their beliefs were overoptimistic, subjects adjust them, and their contributions decline. We are thereby able to explain prominent stylized facts. Especially, we can explain individual contributions within an equilibrium framework.

Our experimental design confirms the assumption of overoptimistic subjects. We observe two effects: First, subjects overestimate contributions of others, which allows us to model equilibrium beliefs at positive contribution levels. Additionally, reciprocal subjects overestimate, on account of the false consensus effect, the degree of reciprocity in the subject pool. We build two control treatments to test the robustness of our model. Subjects in the estimator treatment are only estimating beliefs and are not playing. In spite of this, they display overoptimism as well. Contrary to the standard treatment of the public goods game, subjects in our belief treatment have to stick to their contribution plan because contributions are directly calculated with the help of subjects' beliefs and contribution schedules. Nevertheless, contributions in the belief treatment are about the same as in the standard treatment, and the contribution level in the belief treatment can *only* be explained with the help of overoptimism.

In our setting, we assume that all subjects are equally overoptimistic. However, it might be interesting to research whether overoptimism develops only in certain strategic situations, how it is influenced by framing, and on which personal character traits and cognitive capabilities it depends. For example, the concept of the social exchange heuristic used in Dijkstra (2012) proposes that reciprocal subjects are overoptimistic as opposed to free riders, who should have unbiased beliefs, and this prediction could be tested experimentally.

Our theory covers the most relevant factors which drive cooperation. These are: reciprocity, overoptimism, and the false consensus effect. Additionally, more aspects could be regarded: Our formal game-theoretic approach allows the incorporation of signaling effects into the analysis. It may also be expanded to cover a preference for indirect reciprocity (i.e., being (un)kind to a player if this player is (un)kind to a third party) or for unconditional altruism. Nevertheless, our findings based on reciprocity and overoptimism are challenging enough: Merely switching from the concept of *homo oeconomicus* to *homo reciprocans* will not solve the puzzle of the cooperation dilemma unless a cognitive overoptimism bias is considered as well.

Acknowledgements

We gratefully acknowledge the useful comments provided by participants of the 6th Maastricht Behavioral and Experimental Economics Symposium 2013 and the ESA European Meeting 2013 in Zurich. Especially, we thank Irenaeus Wolff for his helpful and extensive comments.

References

Ambrus, A., and Pathak, P. A. (2011). Cooperation over Finite Horizons: A Theory and Experiments. *Journal of Public Econonomics*, 95, 500–512.

Aumann, R. J. (1987). Correlated Equilibrium as an Expression of Bayesian Rationality. *Econometrica*, 55, 1-18.

Andreoni, J. (1995). Cooperation in Public-Goods Experiments: Kindness or Confusion? *American Economic Review*, 85, 891-904.

Arifovic, J., and Ledyard, J. (2012). Individual Evolutionary Learning, Other-regarding Preferences, and the Voluntary Contributions Mechanism. *Journal of Public Economics*, 96, 808-823.

Armantier, O., and Treich, N. (2013). Eliciting Beliefs: Proper Scoring Rules, Incentives, Stakes and Hedging. *European Economic Review*, 62, 17-40.

Blanco, M., Engelmann, D., Koch, A., and Normann, H. T. (2010). Belief Elicitation in Experiments: Is There a Hedging Problem? *Experimental Econonomics*, 13, 412-438.

Breuer, W., and Hüwe, A. (2014). Trust, Reciprocity, and Betrayal Aversion: Theoretical and Experimental Insights. Working Paper.

Chaudhuri, A. (2011). Sustaining Cooperation in Laboratory Public Goods Experiments: A Selective Survey of the Literature. *Experimental Economics*, 14, 47-83.

Croson, R. T. A. (2007). Theories of Commitment, Altruism and Reciprocity: Evidence From Linear Public Goods Games. *Economic Inquiry*, 45, 199-216.

Dijkstra, J. (2012). Explaining Contributions to Public Goods: Formalizing the Social Exchange Heuristic. *Rationality Society*, 24, 324-342.

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

Eyster, E., and Rabin, M. (2005). Cursed Equilibrium. *Econometrica*, 73, 1623-1672.

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

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

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

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

Fischbacher, U., Gächter, S., and Quercia, S. (2012). The Behavioral Validity of the Strategy Method in Public Good Experiments. *Journal of Economic Psychology*, 33, 897-913.

Herrmann, B., and Thöni, C. (2009). Measuring Conditional Cooperation: A Replication Study in Russia. *Experimental Economics*, 12, 87-92.

Holt, C. A., and Laury, S. K. (2008). Theoretical Explanations of Treatment Effects in Voluntary Contributions Experiments. In C. R. Plott, V. L. Smith (Eds.), *Handbook of Experimental Economics Results* (pp. 846-856). New York, Elsevier Press.

Klumpp, T. (2012). Finitely Repeated Voluntary Provision of a Public Good. *Journal of Public Economic Theory*, 14, 547-572.

Ledyard, J. O. (1995). Public Goods: A Survey of Experimental Research. In J. H. Kagel, A. E. Roth (Eds.), *The Handbook of Experimental Economics* (pp. 111-194). Princeton, Princeton University Press.

Nagel, R. (1995) Unraveling in Guessing Games: An Experimental Study. *American Economic Review,* 85, 1313-1326.

Neugebauer, T., Perote, J., Schmidt, U., and Loos, M. (2009). Selfish-biased conditional cooperation: On the decline of contributions in repeated public goods Experiments. *Journal of Econonomic Psychology*, 30, 52-60.

Orbell, J., and Dawes, R. M. (1991). A 'Cognitive Miser' Theory of Cooperators' Advantage. *American Political Science Review*, 85, 515-528.

Ross, L., Greene, D., and House, P. (1977). The 'False Consensus Effect': An Egocentric Bias in Social Perception and Attribution Process. *Journal of Experimental Social Psychology*, 13, 279-301.

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, J.C.B. Mohr (Paul Siebeck).

Wolff, I. (2013). When Best-Replies Are Not in Equilibrium: Understanding Cooperative Behaviour. Working Paper.

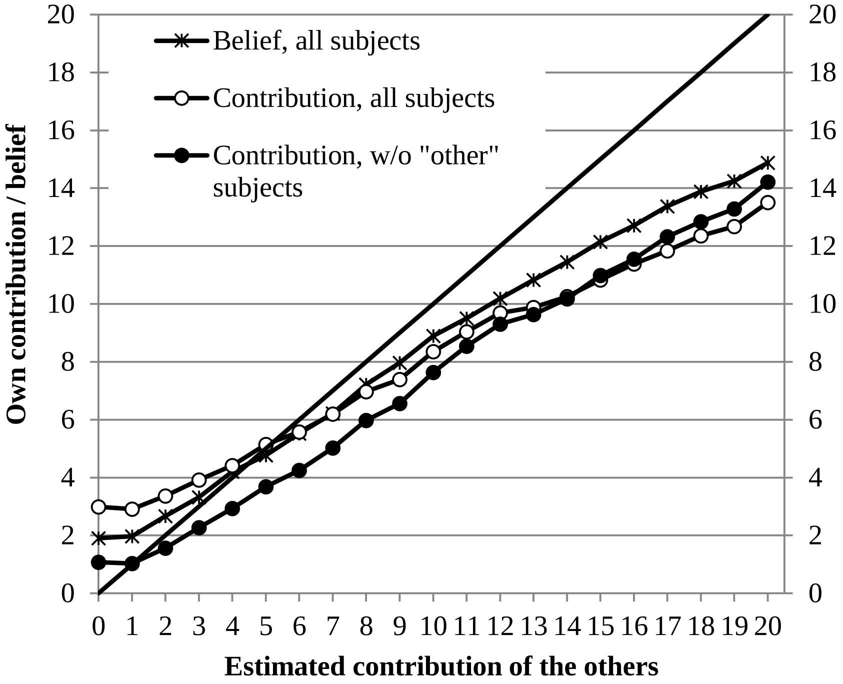


Fig. 1: Average contribution schedule and belief schedule. The chart shows the average own contribution / belief depending on the estimated average contribution level of other subjects.

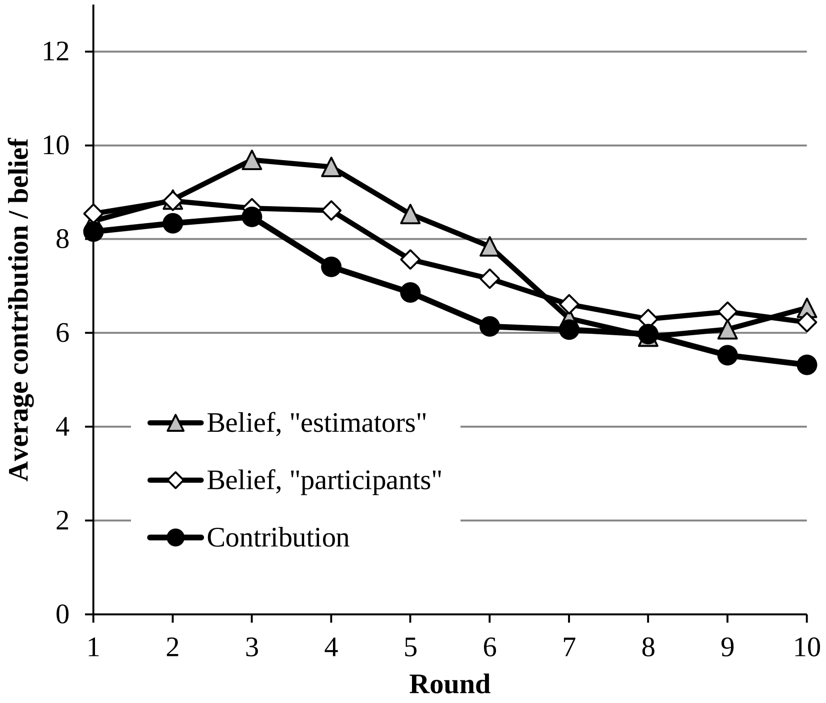


Fig. 2A: Average contribution and belief levels in the “standard treatment”

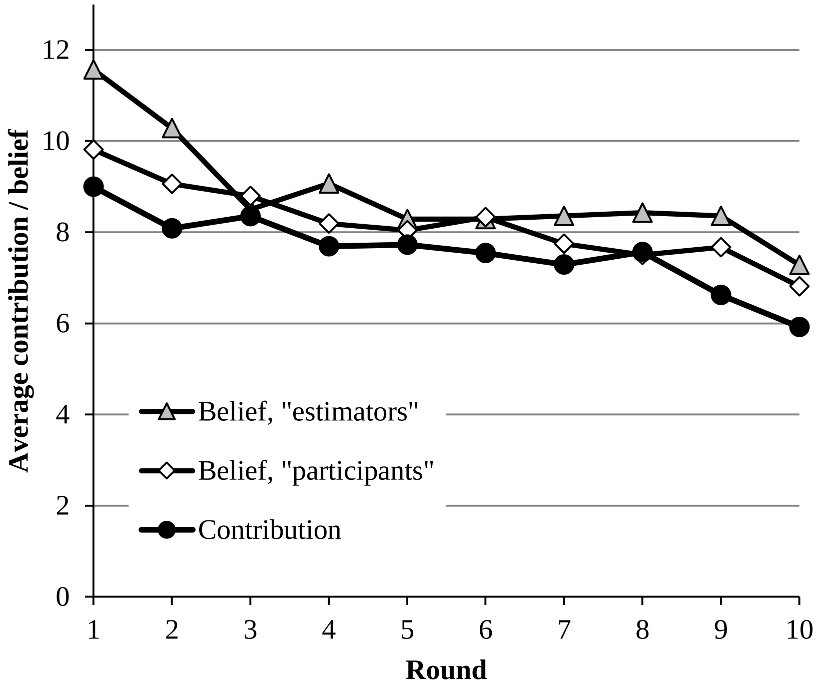


Fig. 2B: Average contribution and belief levels in the “belief treatment”

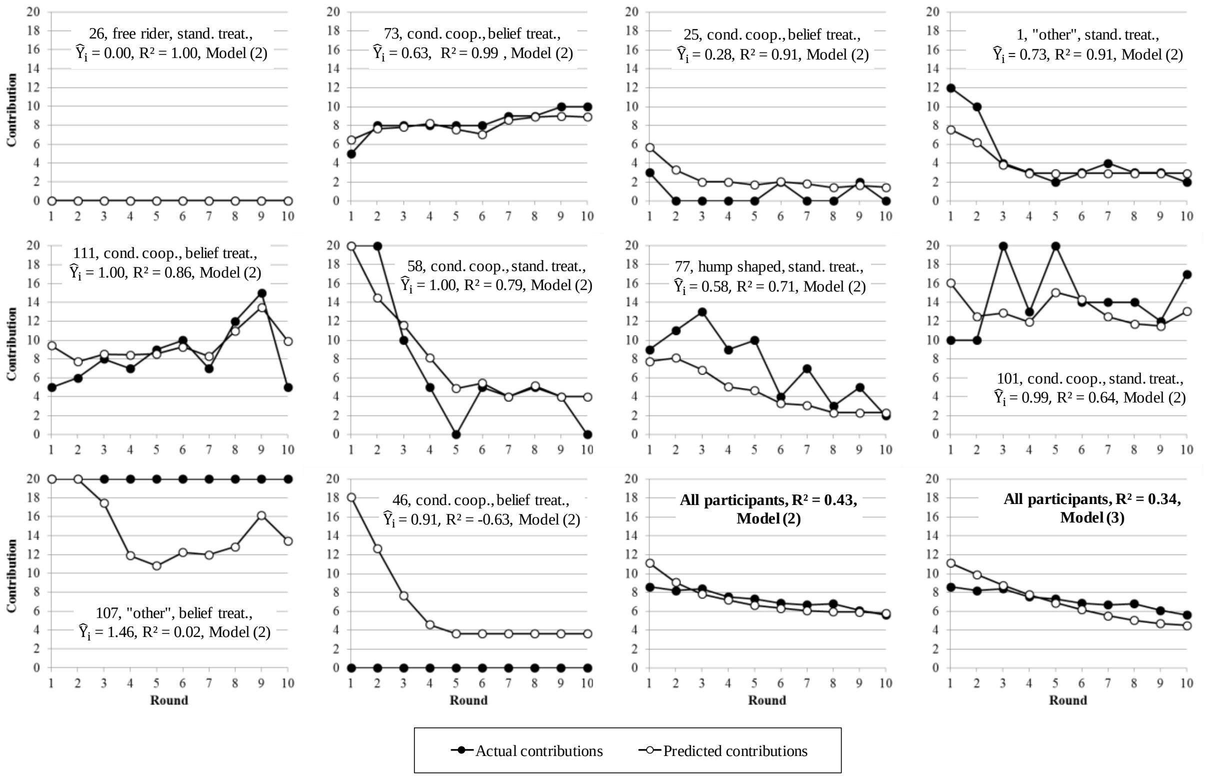


Fig. 3: Actual contributions and model predictions

Table 1 - Definitions and descriptive statistics

Variable	Meaning	Experimental representation	Average value
g_i	*i*'s contribution	contribution	7.2 tokens
g_{ij}	*i*'s belief over average contribution of other subjects	belief	8.0 tokens
ε	overoptimism parameter	difference between g_{ij} and g_i, as defined in equation (2)	0.05
$\hat{\gamma}_i$	*i*'s reciprocity parameter	slope of the contribution schedule [1)]	0.74
$\hat{\gamma}_{ij}$	*i*'s belief over the average reciprocity parameter of other subjects	slope of the belief schedule [1)]	0.81

# Subjects: 119
  # Participants: 92
  # Estimators: 27

[1)] Slope, according to a linear regression without constant.

Table 2 - Contributions in the "standard treatment"

Dependent variable	Contribution			
Regression	(1)	(2)	(3)	(4)
Included subjects	all	all	"others"	all
Predicted contribution	0.291**		0.009	
	(0.075)		(0.040)	
Pred. contribution slope				0.941***
				(0.084)
Belief	0.717***	0.974***	0.960***	
	(0.093)	(0.042)	(0.132)	
Constant	-0.415**	-0.473	-1.241	1.282
	(0.120)	(0.225)	(0.975)	(0.854)
Observations	440	440	50	440
R²	0.597	0.554	0.542	0.564

Significant at the 1 percent (***), 5 percent (**), 10 percent (*) level.
Notes: OLS regressions with robust standard errors (clustered on sessions) in parentheses.

Table 3 - Simulation results

Model	Treatment	Grouping	N	R²
(1)	Standard	-	44	0.54
	Belief	-	48	0.80
			92	0.68
(2)	Standard	avg. deviation from schedule < 3	27	0.70
		remaining subjects	17	-0.04
			44	0.42
	Belief	R² of schedule regression ≥ 0.8 and avg. belief mistake < 4	29	0.70
		remaining subjects	19	0.05
			48	0.44
			92	0.43
(3)	Standard	-	44	0.32
	Belief	-	48	0.36
			92	0.34

Appendix – not for publication, only for referees' information

A: Optimal Behavior Under Uncertainty

If co-players contribute unequally, risk aversion with respect to reciprocal utility becomes relevant because utility from the co-players' kindness is curved concavely. Thus, *i* should not reciprocate the *expected* kindness of the group members, but reduce it by a risk discount which depends on the variance of the co-players' uncertain kindness. Cheung (2013) provides the experimental proof for this implication by showing that subjects indeed want to contribute less if the others contribute more unequally. Therefore, we show the following extended modeling approach (which is not relevant for explaining our experimental results in the paper because we do not vary the diversity of the subject pool).

Based on the probability distribution of $\hat{Y}_j$, $\tilde{g}_j$ denotes the uncertain contribution of group member *j* (depending on her type) and $\tilde{g}_{ij,o}$ denotes the corresponding probability distribution of *i*'s belief regarding *j*'s behavior. If subjects maximize expected utility, with *MPCR* = 0.4, the following equation holds (compare equation (3)):

$$E(U_i) = E\left(1 - (1 - MPCR) \cdot g_i + \textstyle\sum_{j=1}^{J} MPCR \cdot \tilde{g}_{ij} + Y_i \cdot \textstyle\sum_{j=1}^{J} \sqrt{MPCR \cdot g_i \cdot MPCR \cdot \tilde{g}_{ij,o}}\right)$$

(A.1)

with Y_i set to $\frac{2\cdot(1-MPCR)}{J\cdot MPCR}\cdot\sqrt{\hat{Y}_i}$, compare Appendix B, (A.1) simplifies to

$$E(U_i) = 1 - (1 - MPCR) \cdot \mathrm{g_i} + MPCR \cdot \textstyle\sum_{j=1}^{J} E\left(\tilde{g}_{ij,o}\right) + \frac{2\cdot(1-MPCR)}{J} \cdot \sqrt{\hat{Y}_i \cdot g_i} \cdot E\left(\textstyle\sum_{j=1}^{J} \sqrt{\tilde{g}_{ij,o}}\right)$$

(A.2)

$$\frac{\mathrm{d}E(U_i)}{\mathrm{d}g_i} = 0$$

$$\Rightarrow (1-MPCR) = \frac{2\cdot(1-MPCR)}{J} \cdot \sqrt{\hat{Y}_i} \cdot E\left(\sum_{j=1}^{J}\sqrt{\tilde{g}_{ij,o}}\right) \cdot \frac{1}{2\cdot\sqrt{g_i}}$$

$$\Rightarrow g_i = \hat{Y}_i \cdot \frac{1}{J^2} \cdot E^2\left(\sum_{j=1}^{J}\sqrt{\tilde{g}_{ij,o}}\right) = \hat{Y}_i \cdot \left(\left(E\sqrt{\tilde{g}_{ij,o}}\right)^2 - \mathrm{Var}\left(\sqrt{\tilde{g}_{ij,o}}\right)\right) \quad \text{(A.3)}$$

Thus, *i* does not reciprocate the expected kindness of the group members, but reduces its squared value by its variance. In the following, we will denote $\frac{1}{J^2} \cdot E^2\left(\sum_{j=1}^{J}\sqrt{\tilde{g}_{ij,o}}\right)$ as *EQ*, such that (A.3) can be simplified to

$$g_i = \hat{Y}_i \cdot EQ. \quad \text{(A.4)}$$

EQ can be determined (numerically) in such a way that equation (A.4) is fulfilled and no subject has an incentive to deviate. For exemplary reasons, assume the following: The population solely consists of free riders and perfect conditional cooperators in equal parts, meaning that $E(\hat{Y}_{ij}) = 0.5$; furthermore, $\varepsilon = 0.2$ and $J = 3$. While free riders will not contribute, perfect conditional cooperators will choose to contribute $g_i = 0.3125$, which can be proven to be optimal by using equations (A.3) and (2):

$$g_i = 1 \cdot \frac{1}{9} \cdot \left(\frac{1}{8} \cdot 3 \cdot \sqrt{0.2} + \frac{3}{8} \cdot \left(1 \cdot \sqrt{0.3125 \cdot 0.8 + 0.2} + 2 \cdot \sqrt{0.2}\right) + \frac{3}{8} \cdot \left(2 \cdot \sqrt{0.3125 \cdot 0.8 + 0.2} + 1 \cdot \sqrt{0.2}\right) + \frac{1}{8} \cdot 3 \cdot \sqrt{0.3125 \cdot 0.8 + 0.2}\right)^2 \approx 0.3125. \quad \text{(A.5)}$$

As an approximate solution, equilibrium contributions can be derived analytically if subjects are assumed to reciprocate expected average contributions, $E\left(\tilde{g}_{ij,o}\right)$, instead of *EQ*, compare equations (7) and (8): $E\left(\tilde{g}_{ij,o}\right)$ equals $\frac{1}{3}$ in our example and thus is almost identical to the true equilibrium outcome of 0.3125. Equations (8) and (A.4) predict the same contribution if all subjects contribute equally. The less exactly *i* can reciprocate the contributions of the group members (contributions may be known, but they are unequal), and the less *i* knows about the contributions (the variance increases), the less he will contribute. However, both aspects are not varied in our experiment, and the difference between EQ and $E\left(\tilde{g}_{ij,o}\right)$ can to some degree be compensated by adapting $\hat{Y}_i$. Furthermore, according to our example, the mistake of equation (8) does not exceed 2 % of the solution space in PGGs with group size of four.

B: Maximizing (4) Over g_i

$$U_i = 1 - (1 - MPCR) \cdot g_i + J \cdot MPCR \cdot g_{ij} + Y_i \cdot J \cdot \sqrt{MPCR \cdot g_i} \cdot \sqrt{MPCR \cdot g_{ij,o}}, \qquad \text{(B.1)}$$

$$\frac{dU_i}{dg_i} = 0$$

$$\Rightarrow 1 - MPCR = Y_i \cdot J \cdot \sqrt{MPCR} \cdot \sqrt{MPCR \cdot g_{ij,o}} \cdot \frac{1}{2 \cdot \sqrt{g_i}}$$

$$\Leftrightarrow \sqrt{g_i} = \frac{J \cdot Y_i \cdot MPCR}{2 \cdot (1 - MPCR)} \cdot \sqrt{g_{ij,o}}$$

$$\Leftrightarrow g_i = \left(\frac{J \cdot Y_i \cdot MPCR}{2 \cdot (1 - MPCR)}\right)^2 \cdot g_{ij,o} \qquad \text{(B.2)}$$

C: Modeling Results and Modeling Determinants

Subject[1]	R² [2]	Treatment[3]	Session	Classification[4]	$\hat{Y}_i$	Inconsistency[5]	R² schedule regression[6]	Belief mistake[7]
26	1.00	1	1	2	0.00	0.0	1.00	0.2
57	1.00	1	2	2	0.00	0.0	1.00	1.1
18	1.00	2	1	2	0.00	0.0	1.00	0.7
47	1.00	2	2	2	0.00	0.0	1.00	0.7
49	1.00	2	2	2	0.00	0.0	1.00	0.4
53	1.00	2	2	2	0.00	0.0	1.00	1.2
73	0.99	2	3	1	0.63	0.0	0.91	2.2
54	0.98	1	2	1	0.24	0.9	0.94	2.7
72	0.97	2	3	1	0.79	0.0	0.94	1.3
7	0.96	2	1	1	0.29	0.0	0.90	0.4
74	0.95	1	3	1	0.47	1.8	0.10	0.6
113	0.94	2	4	1	0.74	0.0	0.99	0.1
90	0.93	2	4	1	0.86	0.0	0.87	3.5
59	0.93	1	2	1	0.20	0.0	0.36	2.7
25	0.91	2	1	1	0.28	0.0	0.90	0.7
1	0.91	1	2	4	0.73	7.4	-	1.7
109	0.91	1	4	0	1.00	0.8	1.00	1.7
43	0.89	1	2	4	0.73	2.4	-	0.3
103	0.89	2	4	1	0.82	0.0	0.94	3.0
23	0.88	1	1	1	0.06	2.0	0.75	1.2
71	0.88	2	3	0	1.00	0.0	1.00	0.5
60	0.88	1	2	1	0.86	1.2	0.98	2.7
114	0.88	1	4	1	0.65	1.4	0.78	1.1
30	0.86	1	1	0	1.00	0.2	1.00	1.0
111	0.86	2	4	0	1.00	0.0	1.00	0.5
51	0.86	2	2	1	1.01	0.0	0.96	3.0
9	0.86	1	1	1	0.37	0.9	0.83	0.8
52	0.86	1	2	1	0.90	0.8	0.96	0.5
84	0.84	2	3	4	0.32	0.0	-	2.4
14	0.84	2	1	4	0.37	0.0	-	1.0
66	0.82	1	3	1	0.93	1.6	0.99	0.1
27	0.81	1	1	1	0.78	0.9	0.99	0.1
58	0.79	1	2	0	1.00	1.3	1.00	2.2
28	0.79	2	1	0	1.00	0.0	1.00	2.8
22	0.78	1	1	1	0.48	2.5	0.99	1.0
97	0.78	1	4	4	0.73	7.1	-	0.0
65	0.78	2	3	4	1.10	0.0	-	0.3
42	0.76	2	2	4	0.37	0.0	-	0.9
75	0.76	2	3	1	0.91	0.0	0.98	0.5
19	0.75	1	1	3	0.66	2.0	0.38	1.3
76	0.75	2	3	4	1.46	0.0	-	1.1
83	0.75	2	3	1	1.19	0.0	0.95	2.9
85	0.75	2	3	1	1.07	0.0	0.96	2.0
41	0.73	2	2	1	0.72	0.0	0.83	0.2
37	0.72	2	2	1	1.01	0.0	0.99	3.3
77	0.71	1	3	3	0.58	2.2	0.16	0.6
110	0.70	2	4	4	0.87	0.0	-	1.3
16	0.68	2	1	1	0.64	0.0	0.75	0.5
45	0.65	2	2	0	1.00	0.0	1.00	0.8
101	0.64	1	4	1	0.99	2.2	0.96	0.7
70	0.64	1	3	1	0.86	2.8	0.98	0.3
13	0.63	2	1	4	0.44	0.0	-	0.1
5	0.61	2	1	1	0.87	0.0	0.95	2.0
118	0.57	1	4	1	0.93	1.7	1.00	0.4
21	0.57	2	1	0	1.00	0.0	1.00	0.6
81	0.51	2	3	4	0.46	0.0	-	0.5
94	0.50	2	4	1	1.00	0.0	1.00	2.4
17	0.48	1	1	1	0.72	0.6	0.83	0.7
32	0.39	1	2	4	0.73	5.9	-	0.4
50	0.36	2	2	1	1.07	0.0	1.00	0.0
93	0.34	1	4	1	0.91	4.3	0.98	0.9
55	0.33	1	2	1	1.31	5.4	0.75	1.7
119	0.30	1	4	1	1.17	2.2	0.96	4.6
108	0.30	2	4	3	0.21	0.0	0.07	1.9
24	0.29	1	1	1	0.87	0.6	0.99	0.2
31	0.27	2	2	0	1.00	0.0	1.00	1.4
104	0.25	1	4	1	1.14	3.6	0.96	0.7
6	0.24	1	1	1	0.98	3.0	0.79	2.4
12	0.17	2	1	4	0.73	0.0	-	0.1
34	0.16	1	2	1	1.13	3.6	0.97	1.6
79	0.16	1	3	1	0.88	2.3	0.83	1.2
80	0.12	1	3	1	0.48	3.8	0.55	3.5
115	0.12	1	4	1	1.02	4.8	1.00	1.3
29	0.10	2	1	0	1.00	0.0	1.00	1.5
15	0.09	2	1	4	0.40	0.0	-	4.1
38	0.09	2	2	1	0.93	0.0	1.00	1.5
107	0.02	2	4	4	1.46	0.0	-	2.3
78	0.01	2	3	4	1.04	0.0	-	1.7
117	-0.16	1	4	1	0.67	5.3	0.79	1.3
116	-0.17	1	4	1	0.85	4.7	1.00	2.3
56	-0.26	1	2	1	1.25	6.2	0.91	0.3
39	-0.27	1	2	1	0.54	8.0	0.75	2.0
61	-0.39	1	3	4	0.73	11.7	-	1.9
67	-0.39	2	3	1	0.93	0.0	0.79	5.6
106	-0.41	2	4	4	1.46	0.0	-	3.0
82	-0.43	1	3	1	0.91	4.5	0.33	1.9
46	-0.63	2	2	1	0.91	0.0	0.82	1.6
20	-0.64	1	1	1	0.97	1.3	0.95	0.7
105	-1.03	2	4	3	0.70	0.0	0.00	0.7
112	-1.04	2	4	1	0.97	0.0	0.95	9.1
91	-2.67	1	4	2	0.00	11.6	1.00	1.8
102	-3.35	2	4	0	1.00	0.0	1.00	9.8
Avg.	0.43					1.49	0.86	1.6

[1] 92 Participants. Missing subjects to N=119: Estimators. Subjects printed in bold are displayed in the paper.

[2] R²: 1 - average(prediction mistake)²/average(contribution variation of participants)².

[3] 1: Standard treatment. 2: Belief treatment.

[4] 1: Perfect conditional cooperator. 2: Imperfect conditional cooperator. 3: Free rider.

[5] Average deviation from the contribution schedule per round, in tokens.

[6] R² of a regression of entries in the contribution schedule on the believed contribution level; regression without constant. R² of "other" subjects cannot be computed.

[7] Average absolute difference between beliefs and actual contributions of the three co-players.

D: The Experiment

In the following, we present a translation of the most important extracts of the experiment, which was conducted in German originally.

Remark: Note that the experiment presented in the paper was preceded by a question part, which intended to measure subject's individual overoptimism, respectively overconfidence, in a framework independent of the public goods context. As this part turned out to deliver little insight with respect to the model and the analysis of the hypotheses presented in the paper, we abstained from presenting its design and the corresponding analysis in the paper. However, both is available from the authors upon request.

General Information

Welcome to the experiment. You participate in a study about individual decision making in the context of experimental economics research. The experiment will last about 60 minutes. You can always ask questions to the instructor. However, you are not allowed to communicate with other participants until the experiment ends.

During the experiment, you decide completely anonymously and the results of this session will only be used anonymized for research purposes.

The experiment is divided into two part: In the first part, you will be asked to forecast several future events and to estimate their probabilities. You will also be asked to assess some personal skills and to solve three small brain teasers. In the second part, you will interact with other participants (anonymously). Further information on this will follow later.

In this session, you can earn between 3.20 and 25.60 euros. Your actual earnings will depend on your answers, on your decisions and on the decisions of other participants. Detailed information about the payment structure will follow. The money will be paid to you in cash at the end of this session.

[…]

Questions On Future Life Events (Questions Taken From Weinstein, 1980)

Thank you. We now ask you to estimate how your own chances of experiencing the following events deviate from the chances of the other participants in this room. For example, if you believe that the probability to experience the event named in the following is 40 % higher for you than for the other participants, you should click on the 40 %-button.

[Possible answers (in comparison to the average probability of the subjects pool): -100 % (impossible), -80 %, -60 % -40 %, -20 %, 0 (same probability), +20 %, +40 %, +60 %, + 80 %, +100 %, +200 %]

1. Being in a hospital in the next 5 years
2. Like postgraduation job
3. Victim of burglary within the next five years.
4. Dying in an accident.
5. Owning your own home within the next 10 years
6. Having mentally gifted children

Assessing Own Character Traits and Skills

Thank you. You are now asked to assess some character traits and personal skills. With the sliders, you can define how much you agree with the following statements. Please position the slider in the middle if you believe that you are as good as the other participants in this room. Accordingly, move the slider to the left or to the right if you believe to be worse or to be better.

1. I can predict a person’s trustworthiness.

2. I am a cooperative person.

3. I can tell if someone is lying to me.

[…]

Explanation of the Prisoner’s Dilemma

With the following question, subjects are asked to solve a dilemma: [The two] subjects can either choose “cooperate” or “not cooperate”. If both, subject A and subject B choose “cooperate”, they receive payments of 2 euros each. If, on the other hand, both subjects choose “not cooperate”, they receive 1 euro each. If one subjects chooses “cooperate” and the other subject chooses “not cooperate”, the first subject receives 3 euros and the co-player receives nothing. Thus, because **no matter what the other subject does, “not cooperate” is better than “cooperate”, many subjects choose “not cooperating” and only receive 1 euro**. The reason is:

Assume, the co-player “cooperates”. If you choose “not cooperate”, you will receive 3 euros, which is more than 2 euros in the case of “cooperation”. Instead, assume that the co-player chooses “not cooperate”. In this case, you receive 1 euro, which is still better than receiving 0 euro in the case of “cooperating”. Thus, “not cooperate” is always the best choice. Despite this incentive scheme, there are subjects who choose “cooperate” because they hope that the other subject is doing the same.

How much is the proportion of subjects who choose “cooperate”? (**in %**) □

[…]

Third Brain Teaser

3. Question

You and 49 other readers participate in a contest, carried out by a newspaper. The task which the readers have to solve is the following: “Please send a number between 0 and 100 to the editorial office (inclusive 0 and 100). The winner of the contest is that reader, who sends the number which is closest to two third of the average of all numbers which have been sent in. If several readers send the correct answer, the winner will be chosen among them by lot. Which number should you choose if all other participants (and you) solve the puzzle? (You have two minutes time.)

The number is: □

[…]

Public Goods Game Instructions

Thank you. We will now start the interactive part of the experiment. We explain it first and ask some control questions to be sure that you understand the experiment. The amount of money you can earn depends on your own decisions and on the decisions of your co-players. Regard, that we will not calculate with euros in the following but with “play money” because divisibility will be better. One “play money unit” (PM) will be worth 4 euro-cents.

The experiment, which we will describe on the next page, will be conducted 10 times, meaning, that you will play 10 rounds. Before the experiment starts, different variants of the experiment will be assigned to you by lot, detailed information will follow within a short time.

The experiment will be conducted in **groups of four**, theses groups will be **reassigned in each round arbitrarily**.

[next page]

The experiment will be conducted as follows: In each round, you (and the other group members as well) will receive 20 units of "play money" (corresponding to 80 [euro-]cent). You can keep this money or invest it (fully or partially) in a project. Every unit which you do not invest, remains with you in your so-called "private account".

The **"play money" from your "private account" will be paid out to you** at the end of the experiment (converted into real euros). No one except you earns something from your private account.

On the other hand, if you (or another group member) invest into the project, **the instructor will multiply this money by 1.6**. Afterwards, the money will be **paid out to each of the four group members equally**. For example, if three group members invest 2 units and one member 20 units, altogether, 26 units are invested which are increased to 41.6 units. Accordingly, each group member receives 10.4 units back. Thus, the other group members profit from the amount you invest into the project, on the other hand, you profit from contributions of the others **(irrespective of your own payment)**.

Furthermore, before each round, you are requested to estimate the average contributions of the other group members. For precise estimations, you will, as before, be rewarded with 50 [euro-] cent.

Thus, your total income is the sum of your income you kept on your “private account”, of payoffs from the project, and of rewards for your estimations.

Click on “next” to see the control questions. If you want, you can click on according buttons to use a calculator or to see these instructions again. I will not receive any money for this task. Nevertheless, please take care in answering the questions correctly.

Control Questions

Each group member has 20 units of “play money”. Assume that none of the four group members (including you) contributes anything to the project.

What will your income (from the “private account” and the project) be?

What will the income (from the “private account” and the project”) of the other group members be?

[next page]

The correct answer is: Each group member will earn 20 units (20 units from the “private account” and 0 units from the project).

[next page]

Each group member has 20 units of “play money”. You invest 20 units in the project. Each of the other three members of the group also contributes 20 units to the project.

What will your income (from the “private account” and the project) be?

What will the income (from the “private account” and the project) of the other group members be?

[next page]

The correct answer is: 4 group members times 20 units = 80 units. 80 * 1.6 = 128. Thus, 128/4 = 32 will be paid out to each group member from the project and 0 units are in the “private accounts”.

[next page]

Each group member has 20 units of “play money”. The other three members contribute a total of 30 units to the project.

What will your income (from the “private account” and the project) be, if you invest 10 units?

What will your income (from the “private account” and the project) be, if you invest nothing?

[next page]

1. question: (30+10)*1.6 = 64. 64/4 = 16. Thus, 16 units are paid out to you. As you kept 10 units, your total income is 16+10 = 26.

2. question: (30+0)*1.6 = 48. 48/4 = 12. Thus, 12 units are paid out to you. As you kept 20 units, your total income is 12+20 = 32.

Explanation of the Different Treatments

Thank you. Now, we will explain the different variants of the experiment, which you will be assigned to by lot within a short time in detail:

1. variant: “estimator”

In the “estimator” variant, you will not participate in the experiment actively. Instead, you are asked to estimate the contributions of the other participants.

2. variant: “participant”

As a “participant”, you will on the one hand estimate the contributions of your group members. On the other hand, you will participate in the experiment and in every round you can invest any amount of money from your “private account” into the project you want.

3. variant: “contribution schedule”

In the third variant, you will estimate the contributions of the others as well. Your investments to the project will be calculated automatically afterwards with the help of both, your estimation and a so-called contribution schedule, which you will determine before starting the experiment. We will explain the functionality of the contribution schedule on the next page.

In each of the 10 rounds, new group members will be randomly assigned to you. For sure, your group members will participate **in the same experiment variant** like you. Your variant will be determined shortly. Afterwards, you can be sure that your group members will be assigned to the same variant.

Click on “next” to see the explanation of the contribution schedule.

The Contribution Schedule

If you are assigned to the 3. variant, your contribution schedule which you are requested to fill in on the next page, will be applied in the experiment. In the schedule, you determine how much you want to invest in the project **depending on the estimated average contributions of**

the other three group members. For example, if you always want to contribute 20 units irrespective of the estimated contributions of the others, you should enter “20” into all input boxes of the schedule. More general, if you want to contribute irrespective of the actions of the others, you should **enter the same number into all input boxes**. In contrast, if you want to contribute about as much as the other group members, your entries should **increase** in the boxes 0 to 20. For example, if you always want to contribute 2 units more than the average of the others, then enter a 3 next to the 1, a 4 next to the 2, and so on. You can also contribute less if the others contribute more. If this is the case, your entries should **decrease** from 0 to 20. Of course, all other entries are possible as well, the examples above were chosen randomly. You are not informed about the schedules of the others, accordingly, your schedule remains secret as well.

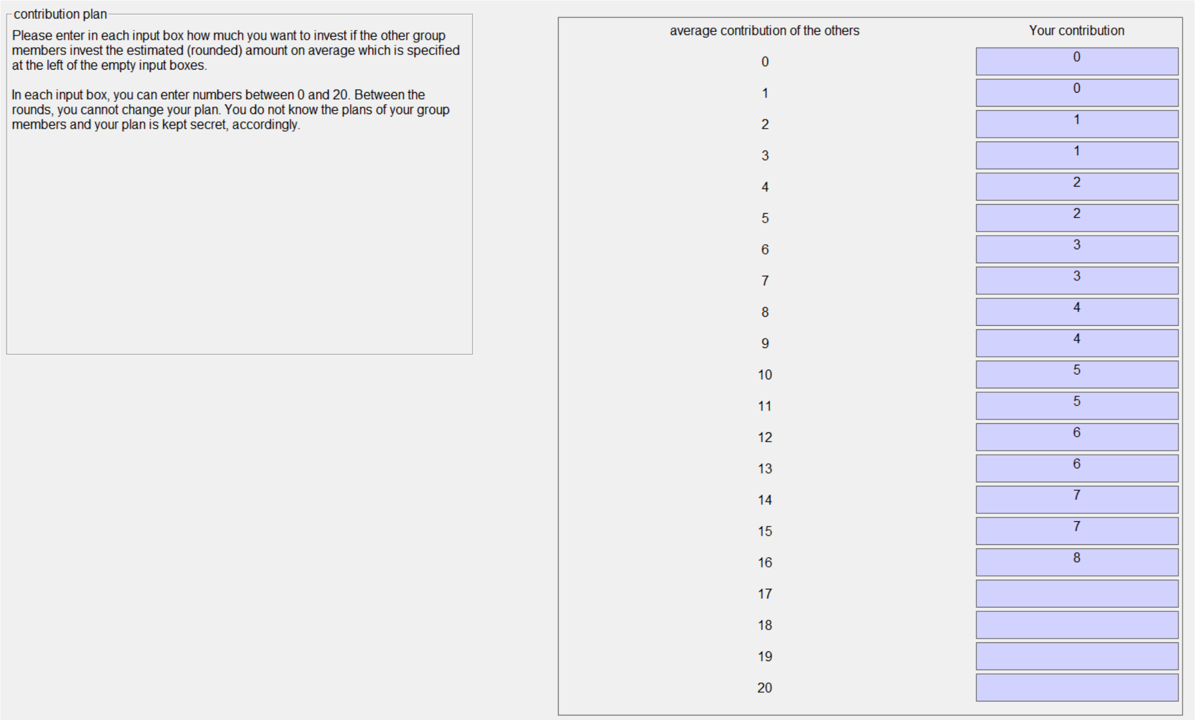
An example: You state in your contribution schedule that you want to contribute 18 units if your group members contribute 15 units. If you estimate that the other three group members contribute 15 units on average in a certain round, 18 units will be taken from your “private account” automatically and will be invested in the project.

You will fill in your contribution schedule only once; it will be applied for all 10 rounds. In contrast, you can make **new estimations after each round**.

You will be informed whether you are assigned to the 3. variant after you filled in the contribution schedule. Please think about your decisions carefully because the contribution schedule will predefine your behavior in the experiment decisively.

Please click on “next” to fill in your contribution schedule.

[next page]



contribution plan

Please enter in each input box how much you want to invest if the other group members invest the estimated (rounded) amount on average which is specified at the left of the empty input boxes.

In each input box, you can enter numbers between 0 and 20. Between the rounds, you cannot change your plan. You do not know the plans of your group members and your plan is kept secret, accordingly.

average contribution of the others	Your contribution
0	0
1	0
2	1
3	1
4	2
5	2
6	3
7	3
8	4
9	4
10	5
11	5
12	6
13	6
14	7
15	7
16	8
17	
18	
19	
20	

[next page]

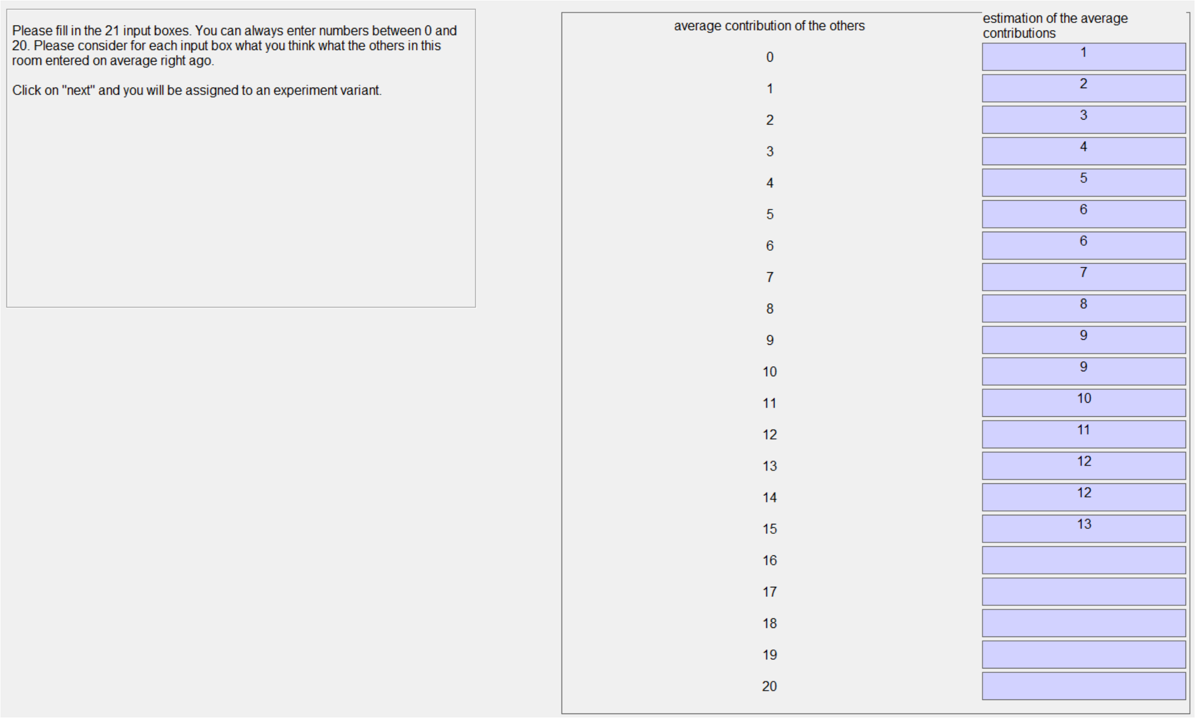
Thank you. Please explain in two or three sentences, why you filled in the contribution schedule specifically in the chosen way. Please use the sheet of paper lying on your desk. Of course, these data will only be analyzed anonymously as well.

The Belief Schedule

Thank you. Please now **estimate the average contribution schedule of the other participants in this room**. Thus, for each of the 21 input boxes, you should estimate what the other participants in this room filled in on average just now. For example, start with the first input box: Consider, what the others entered into this box. Meaning: How much does the others want to contribute on average, if the other group members want to contribute 0?

For a correct estimation of the average contribution schedule, you will receive 50 [euro-] cents at the end of this experiment. For each percent which you deviate from the correct answer, the 50 cent are reduced by 2 percent. Your estimation will be kept secret and **has no effect on the course of play!**

[next page]



Please fill in the 21 input boxes. You can always enter numbers between 0 and 20. Please consider for each input box what you think what the others in this room entered on average right ago.

Click on "next" and you will be assigned to an experiment variant.

average contribution of the others	estimation of the average contributions
0	1
1	2
2	3
3	4
4	5
5	6
6	6
7	7
8	8
9	9
10	9
11	10
12	11
13	12
14	12
15	13
16	
17	
18	
19	
20	

[subjects are assigned to the different treatments now]

Instructions Before Round 1 of the Public Goods Game

“Participants” in the “Standard Treatment”

The “participant” role was assigned to you and your contribution schedule will not be applied!

Accordingly, you will determine in each round how much you want to invest in the project, your group members will be doing the same. Units which remain on your “private account” and which you earn from the project will be credited to you. Additionally, you will estimate the contributions of your group members, for each estimation you will receive up to 50 [euro-] cents. The estimation will be kept secret. In contrast, you will be informed about the average contribution of group members (anonymously). Afterwards, new group members will be assigned to you and the next round starts.

Your potential group members have received exactly the same information.

Please click on “next” to start with round 1!

“Participants” in the “Belief Treatment”

The “participant” role was assigned to you and your contribution schedule will be applied!

Accordingly, you will estimate the contributions of your group members. Based on your contribution schedule, your own contribution will be determined. Units which remain on your “private account” and which you earn from the project will be credited to you. Additionally, for each estimation you will receive up to 50 [euro-] cents. The estimation will be kept secret. In contrast, you will be informed about the average contribution of group members (anonymously). Afterwards, new group members will be assigned to you and the next round starts.

Your potential group members have received exactly the same information.

Please click on “next” to start with round 1!

“Estimators” in the “Standard Treatment” (“Belief Treatment”)

The “estimator” role was assigned to you, accordingly, your contribution schedule will not be applied. Likewise, (in contrast,) for “participants” in your group, the contribution schedule will (not) be applied.

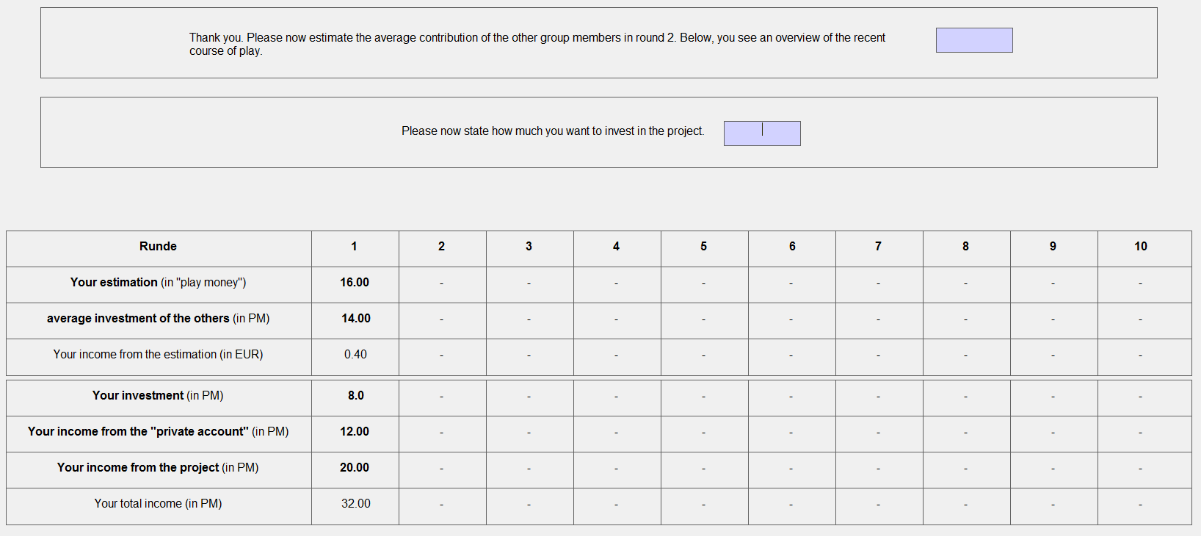
Accordingly, in each round, “participants” in your group will state how much they want to invest in the project (estimate how much the other group members will contribute. Based on this estimation and based on their contribution schedules, contributions of each “participant” will be determined”). Now, it is your task to estimate the average of these contributions. Each of your estimation will be rewarded with up to 50 [euro-] cents. The estimation will be kept secret. Therefore, you will not be able to influence the course of play. You and the “participants” will be informed about the average contribution of group members. Afterwards, new “participants” will be assigned to you and the next round starts.

As you are in the “estimator” role, you will probably earn less than the “participants”. Therefore, you will now receive a fixed payment of 5 euros.

Please click on “next” to start with the first round!

[…]

Screenshot of Round 2, “Standard Treatment”



Thank you. Please now estimate the average contribution of the other group members in round 2. Below, you see an overview of the recent course of play.

Please now state how much you want to invest in the project.

Runde	1	2	3	4	5	6	7	8	9	10
Your estimation (in "play money")	**16.00**	-	-	-	-	-	-	-	-	-
average investment of the others (in PM)	**14.00**	-	-	-	-	-	-	-	-	-
Your income from the estimation (in EUR)	0.40	-	-	-	-	-	-	-	-	-
Your investment (in PM)	**8.0**	-	-	-	-	-	-	-	-	-
Your income from the "private account" (in PM)	**12.00**	-	-	-	-	-	-	-	-	-
Your income from the project (in PM)	**20.00**	-	-	-	-	-	-	-	-	-
Your total income (in PM)	32.00	-	-	-	-	-	-	-	-	-

“Participants” in the “belief treatment” cannot state own investments (second box). Additionally, “estimators” do not see the lowermost box.

References

Cheung, S. L. (2013). New Insights Into Conditional Cooperation and Punishment From a Strategy Method Experiment. *Experimental Economics*, 17, 129-153.

Weinstein, N. D. (1980). Unrealistic Optimism About Future Life Events. *Journal of Personality and Social Psychology*, 39, 806-820.

Trust, Reciprocity, and Betrayal Aversion: Theoretical and Experimental Insights

by

Wolfgang Breuer[*], Anselm Hüwe[**]

October 2014

Abstract:

We propose that there are three determinants of sender behavior in trust games: Beliefs regarding the amounts returned, risk aversion, and reciprocity. Particularly, we are interested in the role of reciprocity because the possibility of negative expected reciprocal utility may lead to betrayal-averse sender behavior, i.e. to a situation where reciprocal subjects send less money than solely selfish ones. In our experiment, most subjects show distinct social preferences in the receiver role, but in the sender role they do not distinguish between a standard trust game with a human partner and a non-social, lottery-like setting, where a computer plays the role of the receiver. This means that the relevance of reciprocity considerations under uncertainty might be fundamentally different from those when no uncertainty is present. Furthermore, we find that sendings are mainly driven by overoptimistic beliefs. We use a modified version of the reciprocity model of Dufwenberg and Kirchsteiger [Games Econ. Beh. 47 (2004) 268] to explain our results and thereby show why reciprocal and selfish subjects almost do not differ in their sending behavior if one controls for beliefs and risk aversion. For other experimental settings, our model does predict differing behavior, which is in line with findings reported in the literature.

Keywords: investment game, trust game, reciprocity, overoptimism, betrayal aversion, false consensus effect

JEL classification: C72, C91, D03, D81, D84

[*] Prof. Dr. Wolfgang Breuer
Chair of Business Administration and Finance
RWTH University
Department of Finance
Templergraben 64
52056 Aachen
Germany
Phone: +49 241 8093539
Fax: +49 241 8092163
E-mail: wolfgang.breuer@bfw.rwth-aachen.de

[**] Dipl.-Ing. Anselm Hüwe
Chair of Business Administration and Finance
RWTH University
Department of Finance
Templergraben 64
52056 Aachen
Germany
Phone: +49 241 8093505
Fax: +49 241 8092163
E-mail: anselm.huewe@bfw.rwth-aachen.de

1 Introduction

In this paper, we analyze whether reciprocal motives play a role for senders in the trust game. Despite a large body of literature on this question, it is still not fully clear what drives sender behavior. Probably, subjects send money to receivers because they expect at least some of them to return money. These sendings may on the one hand be selfishly motivated, as they can yield a positive return. On the other hand, sendings may be influenced by reciprocal preferences, as they offer the possibility of making both players better off. Which of these motives prevails is an important question, because human interactions following the incentive scheme of trust games can be observed ubiquitously in economic life. Knowing the motives of senders allows us to better understand when such welfare increasing interactions can be established. For example, reciprocally oriented business people might prefer a “hand-shake”-environment to a formal and contractually secured one, even when the monetary consequences are identical. People might prefer to lend money to acquaintances instead of putting it into a bank account, even when the yields are the same. However, preferences might be exactly the other way round, because a failure of an investment hurts particularly if it comes along with reciprocal interaction. If the latter view is true, people might also be willing to pay more for insurance against risks if these risks are believed to have reciprocal elements. Up to now, the findings of the experimental literature answering such questions are not consistent: Several papers conclude that reciprocally oriented senders act in a betrayal-averse way, suggesting that they send less money than if they were purely selfishly oriented (Aimone and Houser 2012; Bohnet and Zeckhauser 2004; Bohnet et al. 2008; Hong and Bohnet 2007). In contrast, other researchers state that sendings in social environments are higher than in non-social ones (Fetchenhauer and Dunning 2009 and 2012), which corresponds to the finding that subjects with distinct social preferences send higher amounts than selfish subjects (Altmann et al. 2008; Ashraf et al. 2006;

Cox 2004). Such comparisons of the literature are hindered by the fact that researchers have not agreed on a formal definition of trust and betrayal aversion yet. Trust may be associated with certain levels of sendings (which is a behavioral definition based on Coleman 1990) or with the belief that the sender has about the receiver's reaction (see, e.g., Cox 2004). Kazuhiro (2009) has suggested a formal definition of betrayal aversion, which introduces a betrayal discount on the sender's utility if the monetary payoff stems from a social interaction. This definition however has not been adopted by other researchers so far. In this paper, we derive betrayal aversion from the reciprocity model of Dufwenberg and Kirchsteiger (2004), DK henceforth, and propose that subjects act in a betrayal-averse way if expected reciprocal utility from an interaction is negative. Betrayal-averse behavior is thus a consequence of negative reciprocal marginal utility, as risk-averse behavior is a consequence of decreasing monetary marginal utility.

To measure betrayal-averse behavior, we use an experimental design which allows us to compare sending decisions in the standard version of the trust game (trust game, henceforth) with investments in a computer setting, where the distribution of returns to the sender is known to be identical, but is determined by a non-social computer draw (investment task, henceforth). As well, we precisely control for senders' beliefs, risk aversion, and social preferences. Typically, previous studies have not done so (see Fehr 2009, for an overview), but have only focused on one or on two of these three factors: Researchers have controlled for trust and social considerations, but not for risk aversion (Cox 2004), for risk and social preferences, but not for beliefs (Houser et al. 2010), or for risk and beliefs, but not for social preferences (Eckel and Wilson 2004). In contrast, Ashraf et al. (2006) address all three parameters mentioned above and find a small positive influence of social preferences. However, they do not have a non-social setting which would allow them to precisely disentangle social from non-social motives. Other studies have compared behavior in social and in non-social settings and have thereby taken care to ensure that results cannot be biased by beliefs about different return distributions:

A design using minimal acceptable winning probabilities (MAPs, see Bohnet and Zeckhauser 2004, or Fetchenhauer and Dunning 2012) only allows the elicitation of a subject's switching point with respect to the *average* expected payoff, eliminating distributional concerns. However, in this paper, we propose that such concerns are an important determination of betrayal aversion. Aimone and Houser (2012) use a design very similar to ours, but do not measure intentions with the help of monetary payoffs. Instead, they concentrate on informational aspects: Subjects seem to prefer not to trust "their" co-player, but choose to be paid according to the decision of another receiver.

The contribution of our paper is as follows: First, we show how the reciprocity model of DK can be applied to explain experimentally observed behavior in trust games. Thereby, a formal definition of betrayal aversion unfolds. Second, we present a new, graphical way of eliciting beliefs and decisions, which allows us to use the strategy method (Selten 1967) in the receiver role, even when senders have a continuous strategy space. Third, as we control for beliefs and risk aversion, we have precise information on senders' selfish motives, and we can use these data to justify why reciprocity and betrayal aversion play only a very limited role for senders' decisions. However, we will also show that our model offers the possibility to explain seemingly opposed literature results by accounting for the different experimental parameters which have been used.

The rest of the paper is structured as follows: In Section 2, we adapt the DK model to make it applicable for trust games. In Section 3, we describe our experimental setting, followed by a presentation of the results in Section 4. Therein, we also test our theoretical predictions. Section 5 concludes.

2 Model and Research Hypotheses

The experimental setting we are interested in is a trust game similar to that of Berg et al. (1995): A sender is matched with an anonymous receiver and can send him any fraction of her endowment (which we denote as s_i in the following, $0 \leq s_i \leq 1$), where it is tripled by the experimenter. Afterwards, the receiver can return any amount, which is characterized by k_j ($0 \leq k_j \leq 3 \cdot s_i$). Rational selfish senders will choose $s_i = 0$ if they anticipate that rational selfish receivers will return nothing. The typical experimental finding, however, is that many subjects do send money and many receivers do return money. While it is yet unclear how to model sender behavior, receiver behavior can be explained with the help of social preferences (for example, with the help of the outcome-based theory of Fehr and Schmidt, 1999), and, more specifically, by using models of reciprocal behavior: McCabe et al. (2003) find that receiver behavior in the trust game can be better understood with the help of reciprocity (intention-based) models than inequality (outcome-based) models (see also Dunning et al. 2012). Accordingly, we assume that subjects are driven by reciprocal motives, and we adopt the reciprocity model of Dufwenberg and Kirchsteiger (2004) to predict behavior in trust games both in the sender and the receiver role.

DK propose that people want to reciprocate kindness with kindness (and unkindness with unkindness accordingly), where *i*'s kindness to *j* at a specific node *h* of the game, called $\kappa_{ij}\left(a_i(h), \left(b_{ij}(h)\right)_{j \neq i}\right)$, is measured by the surplus of monetary payoffs that *i* expects *j* to have gained by the end of a game (given her belief about his strategy $b_{ij}(h)$), if *i* departs from a certain reference strategy by choosing $a_i(h)$ from her strategy space. Accordingly, κ_{ij} will be negative if *i* is unkind to *j* (for further details, we refer to DK themselves). The belief of *i* about *j*'s kindness to herself is denoted as $\lambda_{iji}\left(b_{ij}(h)\right)$. DK assume that utility is created in two different mental accounts – one for utility from money and one for utility from reciprocity.

Thus, a utility function of person *i* consists of two terms, weighted with an exogenously given non-negative reciprocity parameter Y_{ij}:

$$U_i\left(a_i(h),\left(b_{ij}(h)\right)_{j\neq i}\right) = \pi_i\left(a_i(h),\ \left(b_{ij}(h)\right)_{j\neq i}\right) +$$

$$\sum\nolimits_{j\in \mathrm{N}\setminus\{i\}}\left(Y_{ij}\cdot\kappa_{ij}\left(a_i(h),\left(b_{ij}(h)\right)_{j\neq i}\right)\cdot\lambda_{iji}\left(b_{ij}(h)\right)\right). \qquad (1)$$

The first term π_i represents *i*'s direct monetary payoff and the second term (after weighting with Y_{ij}) reflects *i*'s reciprocal utility expressed in monetary units as well. Multiplying κ_{ij} by λ_{iji}, the model displays reciprocal preferences: If λ_{iji} is positive (negative), *i* can raise her utility by increasing (decreasing) κ_{ij}, if it is not too costly. Moreover, *i* will dislike situations where she is friendly and *j* is unfriendly (and vice versa).

This model is able to qualitatively predict receiver behavior in trust games (see Section 4.1 and Appendix A; Appendices are available from the authors upon request). However, DK propose an arbitrary definition of kindness, which is not suited for the trust game, and which we will therefore modify below. Furthermore, DK assume that the reciprocity parameter of the co-player, and thereby her strategy, are known. Such an approach cannot capture the fact that receiver behavior is diverse and that the sender faces a situation with incomplete information. Without modifications, DK are not able to make predictions for sender behavior under uncertainty. Therefore, we will use a modified and generalized version of the DK model, which has also been proposed in Breuer and Hüwe (2014) in the context of public goods games:

1. To determine (un)kindness, we use a different reference point than DK, who themselves admit that their reference point has been chosen without deep justification. DK measure kindness of *i* to *j* by comparing *j*'s material payoff with the average of the highest and the lowest possible material payoff of *j* that is compatible with *i* choosing an efficient strategy. Instead, our reference point relies on the status quo: We consider it to be kind if the co-

player is made better off compared to his situation before the game starts. Therefore, in the trust game, the sender's kindness to the receiver is $\kappa_{ij} = 3{\cdot}s_i - k_{ij}$, meaning that the sender has to determine the receiver's profit, which depends on her action, s_i, and on her belief about the receiver's reaction, k_{ij}. Accordingly, senders cannot be unkind in the trust game. The receiver's kindness is determined by the difference between the returned amount and the received amount, meaning that receivers can be both kind and unkind. This is in line with the definition of trustworthiness ($k_j > s_i$) of Berg et al. (1995), which they adopted from Coleman (1990). Accordingly, j's kindness to i is $\kappa_{ji} = k_j - s_i$. We will also justify these reference points with the help of our experimental findings in Section 4.1.

2. As suggested in DK, p. 291, as a potential modification of their original approach, the square root is used in order to have concave utility from kindness. This is consistent with the assumption that utility from direct payments is concave as well (see below). Therefore, $\kappa_{ij} = \sqrt{3{\cdot}s_i - k_{ij}}$. For the receiver, $\kappa_{ji} = \sqrt{k_j - s_i}$ if $k_j \geq s_i$, and if $k_j < s_i$, κ_{ji} is convex with $\kappa_{ji} = -\sqrt{s_i - k_j}$.

3. Not knowing the co-player, we can introduce a simplification: Instead of using the reciprocity parameter Y_{ij}, we will write Y_i in the following, meaning that Y_i is independent of j: Subjects in the lab play anonymously. Thus, they have no possibility to condition strategies on their co-players.

4. We want to capture the fact that first players have imperfect information about a second player's reaction. The situation might even be seen to be ambiguous, as probabilities of specific reactions are not known either. For modeling purposes, we will however assume that i knows the probability distribution of $\tilde{Y}_j$ in the subject pool, but she does not know with which player she is matched. Furthermore, it is assumed that all subjects will maximize their expected utility.

5. We allow for biased beliefs, meaning that the first player can systematically misestimate a second player's behavior: We assume that senders anticipate the receivers' optimizing calculus, but we allow this anticipation to be biased: This modification reflects the typical experimental finding that subjects' beliefs are distorted (compare, for example, Breuer and Hüwe, 2014). In our model, a factor ε_i is introduced, which distorts the believed return from receiver j as follows: $k_{ij} = \min\{k_j \cdot \varepsilon_i; 2 \cdot s_i\}$, with $\varepsilon_i \geq 0$. A sender with unbiased beliefs, as in DK, thus has $\varepsilon_i = 1$. When applying the model to our experimental data, we will use an upper limit $2 \cdot s_i$ for beliefs in order to avoid that very optimistic reciprocal subjects send less money than pessimistic ones: The reciprocal utility component is decreasing for $k_j \cdot \varepsilon_i > 2 \cdot s_i$ (see equation (7) below) because very high amounts returned come along with very low payoffs for the receiver, implying that the sender perceives herself to be not very kind. To shorten, we will denote i's belief as $k_j \cdot \varepsilon_i$ in the following, but keep the limitation in mind.

6. As it is important under uncertainty to account for risk aversion preferences, we do so by assuming constant relative risk aversion with respect to monetary payoffs, which is a common assumption. More specifically, the utility from money is set to

$$U(\pi) = \begin{cases} (\pi)^{1-r} & \text{if } r < 1, \\ \ln(\pi) & \text{if } r = 1, \\ -(\pi)^{1-r} & \text{if } r > 1. \end{cases} \tag{2}$$

Subjects with $r < 0$ are risk-seeking (i.e. risk aversion is negative in this case), whereas subjects with $r = 0$ are risk-neutral, and $r > 0$ implies risk aversion. In standard trust games, believed profits are $\pi_i = 1 - s_i + k_j \cdot \varepsilon_i$ in the sender role, and $\pi_j = 3 \cdot s_i - k_j$ in the receiver role.

With $r = 0$, $\varepsilon_i = 1$, and known Y_j, the model collapses into the version suggested in DK (except the fact that different reference points are used). As well, with $r = 0$, it is basically identical to the version used in Breuer and Hüwe (2014). Section 4 in this paper proves that actual subject behavior can be better explained by giving up these simplifying assumptions.

We will now present subject behavior derived from the model in order to formulate hypotheses which enable us to test the model predictions empirically.

2.1 Receiver role

Receivers are assumed to maximize the following utility function:

$$U_j = \left(3 \cdot s_i - k_j\right)^{1-r_j} \begin{cases} +Y_j \cdot \sqrt{\left(3 \cdot s_i - k_j\right) \cdot \left(k_j - s_i\right)}, \text{ if } k_j \geq s_i, \\ -Y_j \cdot \sqrt{\left(3 \cdot s_i - k_j\right) \cdot \left(s_i - k_j\right)}, \text{ if } k_j < s_i. \end{cases} \tag{3}$$

In contrast to the sender role (see below), we only allow for $r_j \in \{0; 0.5\}$ in the receiver role. We do so for a number of reasons. First of all, this enables us to investigate situations with risk neutrality regarding money, $r_j = 0$, as well as with risk aversion, i.e. $r_j > 0$. Risk-seeking behavior with respect to money is not common according to our experimental data. Moreover, the choice $r_j = 0.5$ mirrors our assumption concerning the curvature of the reciprocal utility component, and $r_j \in \{0; 0.5\}$ makes it possible to present analytical solutions which qualitatively capture results for other values $r_j > 0$ as well.

If $r_j = 0$ and $k_j \geq s_i$, U_j has a maximum at

$$k_j^* = \left(2 - \frac{1}{\sqrt{1 + Y_j^2}}\right) \cdot s_i, \tag{4}$$

meaning that k_j^* is linear in s_i (proofs for statements in Section 2 are presented in Appendix B). In contrast, a convex response function is derived if r_j is set to 0.5: For $k_j \geq s_i$, U_j then has a maximum at

$$k_j^* = 2 \cdot s_i - \sqrt{\frac{s_i}{4 \cdot Y_j^2} + \frac{1}{64 \cdot Y_j^4}} + \frac{1}{8 \cdot Y_j^2}. \qquad (5)$$

With $r_j < 0$, k_j^* would become concave in s_i, but as already mentioned, we leave this case apart. For $k_j < s_i$, (4) and (5) have no local maximum (see Appendix B again), making responses between 0 and s_i suboptimal. Thus, subjects have to decide between returning k_j^* and $k_j = 0$: If Y_j is large (small), reciprocal utility is given more (less) weight then utility from money, and k_j^* ($k_j = 0$) is chosen. In the case of $r_j = 0$, the according threshold is independent of s_i: It can be shown that the threshold value is $Y_j \approx 0.51$, meaning that receivers start to return money when they perceive being trustworthy to be at least about half as important as receiving money. In the case of $r_j = 0.5$, the threshold depends on s_i and can numerically be determined if the utility from returning nothing is compared to the utility from returning k_j^*:

$$U_j(k_j = 0) = \sqrt{3 \cdot s_i} - Y_j \cdot \sqrt{3} \cdot s_i = \sqrt{3 \cdot s_i - k_j^*} + Y_j \cdot \sqrt{(3 \cdot s_i - k_j^*) \cdot (k_j^* - s_i)} = U_j(k_j = k_j^*). \qquad (6)$$

Based on these derivations, our model makes the following predictions: Receivers either (1) never return money, (2) have a response function which is linearly increasing in s_i, or (3) have a response function which is zero at small sendings, and increasing at increasing rates for larger sendings. If receivers return money, then $s_i < k_j^* < 2 \cdot s_i$. Although receiver behavior is not our primary interest, we test whether our model is able to explain it more precisely than existing theories can, and we formulate:

> **Hypothesis 1:** Prediction mistakes in our model with respect to receiver behavior are smaller than in the original model of DK or in the Fehr-Schmidt model.

For the receiver, no uncertainty is present, and therefore risk aversion is irrelevant. However, risk aversion is typically seen as an implication of a concave utility function. Such a utility function creates wealth effects in situations without uncertainty, meaning that risk-averse (risk-seeking) senders should typically show a convex (concave) response function as a receiver. Nevertheless, with *r* being restricted to $\{0; 0.5\}$, we can determine *r* in both roles for each subject in our data, and we will find no statistical evidence for a correlation between roles. We will therefore not require *r* to be identical across both roles, which allows a more precise description of receiver behavior (or, in turn: risk aversion in the sender role). Thus, we assume that the elasticity of substitution between direct monetary utility and reciprocal utility can be separated from the elasticity of substitution between safe and risky choices. A similar separation has been proposed in the literature before with regard to the disentanglement of risk preferences and time preferences (see Epstein and Zin 1989).

2.2 Beliefs

What returns do senders expect from receivers? Fetchenhauer and Dunning (2009) are not aware of any literature support on overoptimistic beliefs. Instead, the authors mention support for either accurate or pessimistic beliefs, and find distinct pessimistic beliefs in their own experiment. However, Breuer and Hüwe (2014) find significant overoptimism in the context of public goods games. Furthermore, the literature on trust games typically reports that many subjects send money (compare Johnson and Mislin 2011). This is the case, although Camerer (2003) can summarize that “the fact that the [monetary] return to trust is around zero seems fairly robust.” We suppose that subjects are typically not aware of such low returns from trust. Instead, we propose:

> **Hypothesis 2:** Beliefs about average amounts returned by receivers are overoptimistically biased.

2.3 Sender Role

Utility for the sender is set to

$$U_i = \begin{cases} \left(1 - s_i + \tilde{k}_j^* \cdot \varepsilon_i\right)^{1-r_i}, \text{ if } r_i < 1 \\ \ln\left(1 - s_i + \tilde{k}_j^* \cdot \varepsilon_i\right) \quad , \text{ if } r_i = 1 \\ -\left(1 - s_i + \tilde{k}_j^* \cdot \varepsilon_{\mathrm{i}}\right)^{1-r_i}, \text{ if } r_i > 1 \end{cases} \begin{cases} + Y_i \cdot \sqrt{\left(3 \cdot s_{\mathrm{i}} - \tilde{k}_j^* \cdot \varepsilon_{\mathrm{i}}\right) \cdot \left(\tilde{k}_j^* \cdot \varepsilon_{\mathrm{i}} - s_i\right)}, \text{ if } \tilde{k}_j^* \cdot \varepsilon_i \geq s_i, \\ -Y_i \cdot \sqrt{\left(3 \cdot s_i - \tilde{k}_j^* \cdot \varepsilon_{\mathrm{i}}\right) \cdot \left(s_i - \tilde{k}_j^* \cdot \varepsilon_{\mathrm{i}}\right)}, \text{ if } \tilde{k}_j^* \cdot \varepsilon_{\mathrm{i}} < s_i. \end{cases} \tag{7}$$

Additionally to high utility from money, reciprocal senders gain utility from successful interactions with the responder, and vice versa, if the interaction fails, senders suffer from small utility from money and from reciprocal disutility. Having agreed on a utility function for reciprocal senders, betrayal aversion can now be formally defined: We suggest that in trust games, a reciprocal sender acts in a betrayal-averse way if she ceteris paribus sends less money than a purely selfish-oriented sender. Such a behavior is a consequence of expected reciprocal (marginal) utility being negative. This definition implies that the extent of betrayal aversion depends in particular on a subject's overoptimism parameter and on the perceived probability distribution of different receiver types. In Sections 4.4 and 4.6, we will show how these dependencies connect betrayal aversion to specific experimental settings, which can explain why findings in the literature are seemingly opposed to one another.

Although an explicit solution for optimal behavior according to equation (7) cannot be derived analytically, some general implications regarding sender behavior can nevertheless be provided. These implications will be presented in the following, with purely selfish senders being considered first and reciprocal senders being analyzed afterwards.

Obviously, the optimal selfish sending decision (denoted as $\hat{s}_i^*$ in the following) is zero if $\tilde{Y}_j = 0 \; \forall \, j$ (money will never be returned) or $\varepsilon_i \leq 0.5$ (it is believed that $\tilde{k}_j^* \cdot \varepsilon_i \geq s_i$ will never be returned). Furthermore, $\hat{s}_i^* = 1$ if only friendly receivers are believed to be in the subject pool ($\tilde{k}_j^* (s_i = 1) \cdot \varepsilon_i > 1 \; \forall \, j$). However, these are not typical parameter values. If linear and convex response functions as well as zero returns are common in the receiver pool, i's expected profit function is convex (already compare the "average model prediction function" in Fig. 3).

$\mathrm{E}_i\big(\tilde{\pi}_i(s_i)\big)$ always equals 1 for $s_i = 0$. If a receiver pool with limited size is considered and s_i is gradually increased, $\mathrm{E}_i\big(\tilde{\pi}_i(s_i)\big)$ jumps upwards whenever equation (6) is fulfilled for a single receiver in the pool. Sendings at the left limit of such a saltus can never be optimal, because switching to the right limit increases returns from at least one receiver, and increases costs in the form of higher sendings only infinitesimally. We will find that $\mathrm{E}_i\big(\tilde{\pi}_i(s_i)\big)$ is decreasing for small sendings, and increasing for larger ones. If it is believed to increase above 1, a positive yield from sending that amount of money is expected. Accordingly, risk-averse subjects will never choose "small" sendings, which are associated with $\mathrm{E}_i\big(\tilde{\pi}_i(s_i)\big) < 1$: These sendings are associated with less profits and higher risks than a sending of zero. In contrast, with sufficiently high beliefs and sufficiently low risk aversion, a "large" sending might be chosen. As the variance of payoffs increases in s_i (note that, by definition, $\widetilde{k_j^*}$ is linearly or convexly increasing in s_i with $\widetilde{k_j^*}' > 1$, or $\widetilde{k_j^*}$ equals zero), the sending will typically – if not being zero – increase for a subject that is less risk-averse. Risk-seeking subjects value the variance of returns mirror-invertedly: If $\mathrm{E}\left(\tilde{k}_j^*(s_i = 1)\cdot\varepsilon_i\right) > 1)$, $\hat{s}_i^* = 1$ will be chosen, because this maximizes both expected returns and the variance of payoffs. With lower beliefs, "small" sendings on the decreasing part of the expected profit function may be optimal, depending on the specific parameters.

The following statements hold true as well: The more money is sent, the more receivers will start to return money. Also, remember that receivers always return more than what they received. Thus, believing in higher returns – meaning with ε_i increasing –, higher sendings will become more attractive (when being matched with a reciprocal receiver), or yield the same utility as before (when being matched with a free rider). Accordingly, ceteris paribus, higher overoptimism will – if having any effect – increase the optimal sending. Furthermore, higher sendings are always associated with a larger spread of returns. Thus, we propose

Hypothesis 3: Sendings in the trust game (1) increase with more optimistic beliefs, and (2) decrease with higher risk aversion.

We now ask whether and how reciprocity considerations affect our findings presented so far. We have to state that this is parameter-dependent. Consider the following properties of the reciprocal utility component:

1. If the receiver has a linear response function, the reciprocal utility component depends on s_i in a linear way. The slope is positive if the receiver is believed to play friendly, and negative if he is believed to play unfriendly.

2. If the response function is convex, the reciprocal utility component of the sender is convex in s_i as well, starting in the origin with a slope of 0 in the friendly case and with a negative slope in the unfriendly case.

If $\tilde{Y}_j = 0 \; \forall \, j$ or $\varepsilon_i \leq 0.5$, it is also optimal for senders with both monetary and reciprocal preferences (optimal sendings in the trust game are denoted with s_i^* in the following) to send nothing, because with sendings, reciprocal utility is negative if returns smaller than s_i are expected. As well, if $\tilde{k}_j^*(s_i = 1) \cdot \varepsilon_i > 1 \; \forall \, j$, reciprocal utility is again maximal at $s_i^* = 1$. Regarding a mixed receiver pool, 1. and 2. imply that the expected utility from reciprocity is also convex, meaning that it decreases at small sendings (if one does not consider a saltus, where utility jumps upwards), and increases at large sendings if the receiver pool is believed to be sufficiently reciprocal.

If expected reciprocal utility is increasing (decreasing) at the selfishly optimal sending decision, reciprocal senders will either increase (diminish) the sending compared to the selfish solution, leave their decision unchanged if they are stuck at a saltus, or the sending will jump to $s_i^* = 0$ ($s_i^* = 1$) if Y_i is large enough and expected reciprocal utility at $s_i = 1$ is smaller (greater) than zero. A more concrete definition of "small" and "large" cannot be given: Due to the curvature of the reciprocal utility component, the distribution of returns matters. While the distribution of receivers in our subject pool will turn out to be such that positive reciprocal utility is expected if the monetary return is expected to be positive as well, this implication does not necessarily hold true (a proof can be found in Appendix B).

As already mentioned, risk-averse subjects choose a sending on the increasing part of the expected profit function if they believe in positive yields. Such a sending is always associated with increasing expected reciprocal utility, because if amounts returned are expected to be positive, they are also expected to increase. Accordingly, including reciprocal considerations, the sending should be slightly increased compared to the selfish case. However, if (the uncertain) reciprocal utility is unfavorably distributed, the absolute value of expected reciprocal utility can be negative, meaning that $s_i = 0$ can become optimal. Also, as expected reciprocal utility is convex, $s_i = 1$ can be optimal from a reciprocal point of view. Accordingly, it is possible that s_i jumps to one of these extrema. Similar considerations can be made for risk-loving subjects. Without knowing the model parameters, we cannot make a clear prediction as to whether subjects will behave betrayal-aversely or not, and therefore we build no hypothesis on this issue at this point. In contrast, we will derive a hypothesis after we have calculated the model parameters with the help of our experimental data, meaning that we make the hypothesis contingent on found receiver behavior, on belief distortions of senders, and on their risk aversion. As a consequence, Hypothesis 4 will be presented at the end of Section 4.4.

3 Experimental Design and Procedures

3.1 Experimental Tasks

Following our argumentation presented above, careful controlling for the sender's beliefs and risk aversion is essential if the influence of reciprocal motives on the sending decision is to be researched. We controlled for these factors in our experiment by using the following design (see also screenshots and the experimental instructions, which are available upon request): Subjects had to play the standard trust game, in both the sender and the receiver role. In the receiver role, the strategy method was used. Senders were equipped with 10 currency units (CU, henceforth, worth EUR 3.33 or approximately USD 4.51). Subjects also had to decide about the amount sent in the non-social investment task, which had the same (but unknown) return distribution as the trust game, because the receiver decision was drawn from the selfsame subject pool. To control for beliefs, we asked subjects to estimate the average expected amount returned by the receiver in the trust game conditional on s_i, denoted as $\mathrm{E}_i\big(\tilde{k}_j(s_i)\big)$ in the following. For that task and for inputs in the receiver role, we programmed a novel graphical interface, which will be explained in more detail below. To determine subjects' risk aversion in the sender role, we asked them to state certainty equivalents in the investment task, again conditional on the sending. The proceeding was incentive-compatible, similarly to the approach of Holt and Laury (2002). All of these tasks were incentivized. Additionally, several control questions were asked: We wanted to know demographic data (gender, age, country of birth, course of study if a student, wealth status, number of siblings), and asked questions taken from the World Values Survey that are associated with trust and risk ("Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?", "Do you think most people would try to take advantage of you if they got a chance, or would they try to be fair?", "Living in secure surroundings is important to this person; to avoid anything that might be dangerous.", "Adventure and taking risks are important to this person; to have an exciting life.").

3.2 Procedure

Subjects arrived at the lab and were randomly assigned to their places. After the participants had been welcomed, the written instructions were distributed, and subjects were given plenty of time to read them and to ask any questions. The experiment itself started with a lottery task which measured risk aversion in a situation where the probability distribution of returns was known: Subjects had to choose between safe amounts and the risky lottery, where they could win any amount between CU 0 and CU 10, each with equal probability. For each safe amount, which could take values of CU 1, 3, 4, 4.5, 4.75, 5, 5.25, 5.75, 6.5, and 8, subjects had to indicate whether they preferred the lottery or the safe amount. Consistently playing subjects chose the safe amount if it was high and the lottery if the safe amount was low; the crossover point thus determined a subject's certainty equivalent within (typically) close boundaries. Subjects were not informed about any outcomes during the experiment in order to avoid wealth effects and interdependencies across decisions.

After the lottery task and between the following stages, a total of six control questions were asked, so that we could check whether subjects had understood the experimental design. Such questions were incentivized with EUR 0.1 each (we converted EUR into CU only within the game itself). Then, subjects found out that they had been assigned to the sender role: At that stage of the experiment, subjects did not know that they would play both roles (and – at least in the "standard" treatment (see below) – they did not know about the investment task setting so far, either). Instead, they were only informed that they would be randomly matched with an anonymous co-player, one of them taking role A (which was the sender role) and the other one role B (the receiver role). We did so to ensure that decisions in the sender role were not distorted, as there is evidence from the literature that knowing about playing both roles affects subjects' decisions (Burks et al. 2003). Our formulation allowed us to assign all subjects to role A first and to role B later on, without deceiving them. Having been informed about their role, subjects had to decide about the amount to be sent. At the subsequent stage, subjects were unexpectedly told that they had to repeat the experiment in the sender role, but that this time they would not be matched with a real receiver, but would be sending the money to the computer. We explained that the computer would determine the amount returned by randomly drawing a decision from one of the human receivers in the lab. Furthermore, we explicitly made clear that compared to the last sending decision, the situation had not changed with regard to the sender's own earnings. The difference from the prior situation was the fact that now the sending decision was not affecting another person's payoffs. We were aware of the disadvantage that these explanations framed both treatments as being similar, but nevertheless included these sentences to ensure that senders understood that the distribution of returns was unknown but equal in both treatments. To avoid hedging considerations, only one of the two sending decisions was selected randomly and paid out at the end of the experiment.

As explained so far, a within-subject design was used. To check whether a between-subject design would lead to different results and to test for any sequence effects, we conducted two different treatments. The "standard" treatment has just been explained above. Our second treatment, the "inverse" treatment, differed only regarding the fact that subjects were told on their computer screens directly before being asked for their first sending decision that they would not be participating in the experiment as explained in the instructions. Instead, the investment-game version was presented to them. Since subjects had not known that a decision in the trust-game version would follow, the two treatments represent a between-subject design. The "inverse" treatment proceeded with the subjects being told unexpectedly that the sending decision would have to be repeated, this time in a setting as explained in the instructions.

In the next stage, the estimation task followed, in which subjects had to estimate, conditional on s_i, how much they expected the receiver to return, and they had to enter their estimation into a diagram (see also Fig. 1).

<<< Insert Fig. 1 about here >>>

For sendings of CU 1.67, 3.33, 5, 7.5, and 10, respectively, subjects had to position five dots in the diagram and thereby make their decisions. On the abscissa, the amount sent (and – in parentheses – the tripled amount which B would receive) was plotted. To make the dots appear, subjects first had to click into the diagram and they could shift the dots along the vertical lines afterwards. We chose a graphical input mechanism to allow subjects to easily express their beliefs and preferences consistently: In the diagram, subjects could sketch the function $\mathrm{E}_i\big(\tilde{k}_j(s_i)\big)$. Values between the dots were interpolated linearly. EUR 2 could be earned for a correct estimation of all five points; earnings were proportionally reduced by EUR 1 for an average estimation mistake of CU 1. Payments could not become negative. According to Hypothesis 2, we expected that subjects would show overoptimistic beliefs, and we provided this comparably strong incentive to be sure that results were reliable (see also Gächter and Renner 2010, who find that beliefs are expressed more precisely if they are incentivized).

<<< *Insert Fig. 2 about here* >>>

The estimations having been completed, we explained the following lottery tasks (see Fig. 2): For the same sendings as in the estimation task, subjects had to decide 10 times whether they preferred to receive a safe payment (Option 1) or to participate in a lottery (Option 2). These lotteries mirrored the investment task, and offered earnings dependent on the return decision of a responder who was randomly chosen from the subject pool. Again, we explicitly made clear that this draw would not have any monetary consequences for a person B. Both one of the five lotteries (each one reflecting a different sending), and one of the ten lines in the options tables (receive CU *z* or participate in the lottery) were randomly chosen at the end of the experiment and the payment was disbursed accordingly. Thus, the way of eliciting subjects' certainty equivalents was incentive-compatible. The safe payment option varied for each subject depending on his or her expected amount returned: The upper five safe payments were computed with CU $z = \Big[1 - s + \mathrm{E}_i\big(\tilde{k}_j\big)(s) + \Big(3{\cdot}s - \mathrm{E}_i\big(\tilde{k}_j\big)(s)\Big) \cdot x\Big] \cdot 10$, with *x* representing a fraction of 60 %, 30 %, 15 %, 5 %, and 0 % of the maximal possible markup on the expected profit (*s* is denoted without an index to indicate that the sending-dependent variable does not refer to a sending of a *specific* subject). The lower five options accounted for CU $z = \left[1 - s + \mathrm{E}_i\left(\tilde{k}_j\right)(s)\cdot x\right]\cdot 10$, with x = 95 %, 90 %, 80 %, 60 %, and 20 % reducing the expected amount returned. The reason for the subject-dependent calibration was the following: We wanted to offer small intervals between the safe payment options close to the individually estimated amount returned in order to be able to precisely measure the crossover point for typical degrees of risk aversion, even if beliefs were distorted. Information was again presented graphically, with the dark bar in Fig. 2 representing the amount that subjects would at least receive in that lottery, the line showing the individually expected amount returned, and the top of the light bar indicating the maximum possible profit for that sending.

Finally, subjects were told that they would have to take part in the experiment again, this time in the B-role, i.e. as a receiver. Inputs had to be entered graphically into a diagram similar to the one in the estimation exercise explained above. At the end of the experiment, we matched subjects, ran the lotteries, and informed subjects about their earnings.

All experiments were computerized, using the software z-Tree (Fischbacher 2007). The experiments were conducted in the computer lab of RWTH Aachen University in August and November 2013. In two sessions with a total of 58 participants, the “standard” treatment was played; another two sessions with 44 participants used the “inverse” treatment. Participants were – with a few exceptions – students from various disciplines, with the majority studying business administration or industrial engineering and management.

4 Experimental Results

Results for the receiver role are presented first, followed by an analysis of subjects’ beliefs. Afterwards, the sender behavior in the trust game and in the investment task is analyzed, and our hypotheses are tested. Descriptive statistics are presented in Table 1, and will be analyzed throughout Section 4.

<<< Insert Table 1 about here >>>

4.1 Reciprocity in the Receiver Role

Consistent with the literature, we find that only a minority of subjects (12.7 %) always behaves according to the selfish prediction and never returns money. Alternatively, 24 % (24 subjects) are counted as being free riders in Table 1 because $Y_j = 0$ is also attributed to subjects who return only very small amounts (see below). Most responders (81.4 %) return money, and never reduce “amount returned” if the sending is increased. Only a very small fraction (5.9 %) shows “other” patterns. The first two observations are in line with our model predictions. We conclude that our method of eliciting preferences graphically yields a very consistent dataset.

Our model predicts that either $k_j = 0$ or $k_j > s_i$ should be chosen, and indeed this is true in many cases: Roughly between 81 % (for $s_i = 0.17$ and $s_i = 1$) and 55 % (for $s_i = 0.33$) of the receivers’ reactions are in line with this prediction (due to the graphical interface, subjects could only enter values with an accuracy of about 0.2 CU). We suspect that our model is least precise at $s_i = 0.33$, because (only) at this sending, inequality-averse receivers who want to be better off than the sender must return less than what they have received (see below). Furthermore, almost all “amounts returned” (about 95 %) are smaller than $2 \cdot s_i$, which is also in line with our model.

Indeed, receivers typically answer higher sendings with either a linear or a slightly convex increase in “amount returned”. In some cases, subjects choose linear profiles with $k_j \approx s_i$ or $k_j \approx 1.5 \cdot s_i$. While such behavior is in line with our model, it cannot be predicted with the reference points which define kindness in the original model version of DK (we show a derivation of receiver behavior according to DK in Appendix A) because DK consider small sendings to be unkind and therefore predict that they are answered with no returns. Furthermore, many subjects establish equality in payments by returning $k_j \approx \max\{0; 2\cdot s - 0.5\}$. Such behavior is predicted by inequality aversion theories, such as that of Fehr and Schmidt (1999) (see Appendix C), but is captured with our convex response function (equation (5) with $Y_j = 0.39$) quite precisely as well.

To exactly evaluate the fit of our model, we minimize the mean squared error (MSE) for six equally distributed, "amounts returned" per subject (at $s_i = 0.17$, 0.33, 0.5, 0.67 (interpolated), 0.83 (interpolated), 1). Thus, either equation (4) or equation (5) is applied, and Y_j is determined in such a way that it results in the best fit. This procedure attributes $Y_j = 0$ to 24 % of all subjects, and positive reciprocity parameters to the large majority of the subject pool (compare Table 1). Doing so, returns can be predicted with an accuracy of CU 1.7 (root of the average MSE). This is slightly more precise than the original model of DK, which predicts behavior with an accuracy of CU 1.9. The difference between the MSEs is significant (Mann-Whitney-U-test, $p < 0.05$). We also calculate the fit of the Fehr-Schmidt model: While, as described above, some subjects play exactly according to their prediction, others do not, leading to an average accuracy of CU 2.2. Although the Fehr-Schmidt model is less precise than the DK model according to the root of the average MSE, a Mann-Whitney-U-test finds no significant difference to the preciseness of our model.

To summarize, with respect to the receiver role, our modifications of the DK model result only in small improvements of the accuracy. Thus, little support is found for Hypothesis 1. However, note that we modified the DK model not to describe the receiver role, but to describe behavior in the sender role, which cannot be captured by DK at all. Thus, we conclude that our modifications do not worsen predictions for the receiver role and simultaneously enable us to capture the sender role, which we will describe in more detail in Sections 4.4 and 4.5.

<<< *Insert Fig. 3 about here* >>>

We also display average model predictions for decisions in the receiver role graphically, see Fig. 3 (presented as "amount kept" by the sender + average "predicted amount returned"): The graph reveals that most convex predicted response functions jump right before or after one of the six "amounts returned" which enter the MSE calculation. This effect is driven by the optimizing process with respect to Y_j: In order to assure high (low) return predictions at high sendings without having to predict a positive (zero) "amount returned" at a lower sending, Y_j is chosen in such a way that the function jumps just before or after one of the six data points. Fig. 3 also reveals that at $s_i = 1$, actual "amounts returned" are significantly higher than predicted ones: This is due to the fact that returns from receivers who strive for equal payments are systematically underestimated at large sendings. Furthermore, as reported above, some subjects chose $k_j \geq 2 \cdot s_i$, which we cannot predict and which is especially the case at $s_i = 1$.

The line of actual "average profits" in Fig. 3 reveals that at small and medium sendings, receivers return on average less than what they received. Only if senders risk almost their entire endowment, can they expect a small profit of up to 8.1 %. This finding is consistent with previous results from the literature. One could also say: The multiplier of three in standard trust games is chosen in a such way that many senders should be unsure which sending decision is the best (as will be revealed later on, this is not only true with respect to direct monetary consequences but under reciprocal considerations as well). Next, we are interested in the issue whether senders are aware of this fact.

4.2 Beliefs Regarding "Amounts Returned"

As shown in Fig. 3, senders show significant overoptimism regarding their expected profits (according to *t*-tests, $p = 0.04$ at $s_i = 0.17$ and $p < 0.01$ for higher sendings): While CU 5.1 are returned on average over all data points, senders expect that CU 6.0 will be returned (compare

Table 1). Accordingly, Hypothesis 2 is supported, which is good news because it implies that cooperation is fostered even when senders are selfish. Furthermore, a distinct false consensus effect (Ross et al. 1977) can be found (see Table 2).

<<< Insert Table 2 about here >>>

Regression (1) in Table 2 shows that the mistake that senders make when estimating the average "amount returned" depends on their own deviation from average behavior in the receiver role. We control for s to capture the tendency that overoptimism is higher for larger sendings. These variables can explain 19 % of the variance of subjects' beliefs. The false consensus effect thus explains the literature finding that the "reciprocal [subject types] trust [i.e., send] more" (see Altmann et al. 2008): At least to a substantial degree, this is only indirectly true: The own social preferences increase beliefs, which in turn raise sendings, as will be shown later on.

Despite the fact that the belief elicitation was incentivized, subjects may want to justify their sendings by stating adapted beliefs. We test this by determining the influence of $|s - s_i|$, which is the absolute difference between the sending to which i's belief refers and i's own sending. This variable allows the measuring of whether overoptimism is more pronounced if returns have to be estimated which correspond to sendings being close to one's own sending. Indeed, according to regression (2) of Table 2, we do find such an effect. This result also holds if a dummy variable is used instead of $|s - s_i|$, see Regression (3): The dummy "own sending" takes the value 1 if i's own sending is equal to the sending that the belief refers to, and takes the value 0 otherwise. Regression (4) shows results if both aspects of regressions (1) and (2) are combined.

4.3 Risk Aversion in the Lottery Tasks and Decision Consistency

As explained in Section 3.2, in the lottery tasks, it is reasonable to "sell" the lottery for a high price, but "keep" it if the price is too low. Only 9 % (9 subjects) did not fill out all tables consistently; these subjects will be excluded from the data set in the following. To arrive at exact certainty equivalents, we assume that the crossover point between the choice of the safe amount and the lottery is the average of the lowest chosen and the highest non-chosen safe amount. If subjects never (always) chose the safe amount, certainty equivalents of the lottery were computed using x = 80 % of the maximum possible markup on the expected "amount returned" (x = 10 % of the expected "amount returned"). On average, subjects discount the expected "amounts returned" by 12 %, which is a finding almost independent of s. Thus, across all data points, subjects believe in total uncertain payoffs (payoffs consist of "amount kept" + expected "amount returned") of CU 10.53, which is on average considered to be as valuable as a certain payoff of CU 9.98. Interestingly, in the introductory lottery task with known probabilities, the average discount is only 3 %, showing that ambiguity has – as typically reported in the literature – a utility decreasing impact.

By choosing crossover points in the lottery tasks, it was possible to play inconsistently compared to one's decision in the investment task: It is rational to send that amount of money in the investment task (denoted as $\hat{s}_\mathrm{i}$ in the following) which also yields the highest certainty equivalent in the lottery tasks (denoted as $\hat{s}_i^*$ in the following). As subjects had to make their sending decision without knowing about the lottery tasks in detail, and as this relationship might not be obvious to subjects, consistent play could not be taken for granted. We find that, on average, the absolute difference between $\hat{s}_i$ and $\hat{s}_i^*$ is CU 3.4. We also analyze the loss of certainty equivalent CUs resulting from these differences. Again, if $\hat{s}_i$ lies between two data points, we interpolate the corresponding certainty equivalent linearly. For 31 % of the subject pool, $\hat{s}_i$ equals $\hat{s}_i^*$. Many of these subjects play $\hat{s}_i = 0$ or $\hat{s}_i = 1$. Another 46 % display certainty equivalent differences between $\hat{s}_i$ and $\hat{s}_i^*$ of between CU 0 and CU 2 (which is up to 20 % of the endowment). The remaining 23 % are classified as playing inconsistently, as their loss in certainty equivalents is greater than CU 2.

For most subjects, $\hat{s}_i$ is too low compared to $\hat{s}_i^*$, the average value of $\hat{s}_i^* - \hat{s}_i$ is CU 1.5. Of course, the opposed point of view may be true as well: Certainty equivalents as elicited in the lottery tasks may be systematically too high, compared to sendings in the investment task. We suppose that the second view is more plausible, because (1) if subjects are explicitly requested to determine prices for lotteries, the endowment effect may let subjects claim to "sell" their lotteries only at high prices and (2) wishful thinking or a misunderstanding of the determination of the lotteries' "selling" prices may induce subjects to demand high prices. These distortions are relevant for the determination of the highest certainty equivalent, because they depend on s (with no sendings, the certainty equivalent is fixed at CU 10, while higher sendings lead to a greater spread of possible outcomes, thus rendering the determination of certainty equivalents more prone to mistakes).

As explained, for rational decision makers, $\hat{s}_i$ should be equal to $\hat{s}_i^*$. Instead, we only find a correlation between $\hat{s}_i$ and $\hat{s}_i^*$ of 0.41. Both decisions are distinctively correlated, but the correlation is low given that there is no rational reason for a deviation. The result can be explained by remembering that (1) there is a bias of stating too high certainty equivalents in the lotteries and (2) utility from money in the investment task depends on $\hat{s}$ in a u-shaped form: In the extreme case, i's utility function has two maxima with $U_i(\hat{s}_i = 0) = U_i(\hat{s}_i = 1)$. If these subjects (rationally) randomize, sendings cannot be predicted at all. Indeed, eight of our predictions err by the whole strategy space of CU 10, which accounts for 25 % of the total prediction error. Furthermore, we are not the only ones who find that subjects to some degree behave inconsistently in experiments, compare for example Erner et al. (2013).

4.4 Modelling of Sendings in the Trust Game, and Betrayal Aversion Hypothesis

In this section, we specify how our model can be applied to determine sender behavior in the trust game. Furthermore, as pointed out in Section 2.3, we analyze the distribution of receiver behavior, of beliefs, and of risk aversion in our subject pool to formulate a hypothesis with respect to the role of betrayal aversion.

To determine sender behavior, utility from money must be correctly weighted against utility from reciprocity. We have already pointed out that modeling risk-averse behavior implies wealth effects even in certain situations. This may affect trade-offs between utility from money and utility from reciprocity. The risk coefficient *r* could be computed from the risk discount (premium) subjects express when stating certainty equivalents, but it can also be derived from the curvature of the response function in the receiver role. Furthermore, we mentioned that – opposed to the theoretical prediction – both methods do not yield consistent results. Therefore, when modeling sender behavior, we have to choose between two possible risk aversion coefficients. At the five data points, we proceed as follows:

$$U_i = {CE_{i,s}}^{1-r} \begin{cases} +\,Y \cdot \sqrt{\left(3 \cdot s_{\mathrm{i}} - \tilde{k}_j^{*} \cdot \varepsilon_{\mathrm{i}}\right) \cdot \left(\tilde{k}_j^{*} \cdot \varepsilon_{\mathrm{i}} - s_i\right)}, \text{ if } \tilde{k}_j^{*} \cdot \varepsilon_i \geq s_i, \\ -Y \cdot \sqrt{\left(3 \cdot s_i - \tilde{k}_j^{*} \cdot \varepsilon_{\mathrm{i}}\right) \cdot \left(s_i - \tilde{k}_j^{*} \cdot \varepsilon_{\mathrm{i}}\right)}, \text{ if } \tilde{k}_j^{*} \cdot \varepsilon_{\mathrm{i}} < s_i. \end{cases} \tag{8}$$

with *r* being *i*’s coefficient in the responder role ($r \in \{0; 0.5\}$), and $CE_{i,s}$ being *i*’s certainty equivalents elicited in the lottery tasks (which capture risk aversion in the sender role). Modeling of the amounts returned has already been explained in Section 4.1. We calibrate ε_i by predicting *i*’s five elicited beliefs with the help of optimized values for ε_i and modeled receiver behavior. Again, we optimize by minimizing average squared errors of all five belief predictions. For sendings between the five data points, $CE_{i,s}$ is chosen again as a linear interpolation (the linear interpolation is a simplification, because it, for example, does not capture jumps in the belief function, but still we think that by this procedure, subjects’ preferences are measured in a sufficiently precise way). Thus, $CE_{i,s}$ reflects i's substitution considerations between safe and risky choices as elicited in the lottery tasks. In contrast, r derived from the receiver role in combination with Y (also determined in the receiver role) is used to weight utility from money against utility from reciprocity. This r will, with $r = 0$ or $r = 0.5$, typically take less extreme values than if we would derive it from the lottery tasks, and is therefore the appropriate one: As we have already pointed out, wealth effects should only play a minor role for weighting money against reciprocity.

Doing so, we find that in 88 % of all cases (for 82 subjects), the reciprocal utility component does not change the predicted sending decision. For seven subjects, predicted sendings in the trust game slightly increase by an average of CU 3.1, compared to predictions for the investment task. For two subjects, predictions jump from CU 0 to CU 10. In contrast, only in two cases the sending decision should be reduced (by CU 3.3 on average). Thus, while 82 subjects should make the same decision in the trust game and in the investment task, nine subjects should increase their sending, and only two should reduce it. We propose

> **Hypothesis 4**: The effect of betrayal aversion is very limited, meaning that sendings in the trust game do not differ much from sendings in the investment task. If at all, the effect will be positive instead of negative, meaning that sendings in the trust game are higher than in the investment task.

The intuition behind Hypothesis 4 is as follows: Often, there is no tradeoff between the monetary and the reciprocal utility components. Therefore, $s_i = 0$ (if beliefs are pessimistic and senders are risk-averse) or $s_i = 1$ (if beliefs are unbiased or overoptimistic and senders are not too risk-averse) is optimal due to both monetary and reciprocal considerations. For example, if $r_i < 0$ and $\varepsilon_i > 1.01$, $s_i^* = 1$ is optimal both with monetary and reciprocal considerations. However, in cases where reciprocal gains must be balanced against monetary utility losses due to risk aversion, our model can predict increased sendings in a social setting. Also note that only experimental data from the belief elicitation stage, from the lottery tasks, and from receiver behavior are used to derive Hypothesis 4. Sending decisions in both the trust game and the investment task have not entered our calculations yet. This is important, as differences between these decisions will be tested with Hypothesis 4.

4.5 Determinants of Sender Behavior

In this section, we will test Hypotheses 3 and 4: What drives sending behavior in the trust game, and how does it differ from behavior in the investment task? First of all, a Kolmogorov-Smirnov test cannot reject the hypothesis that sendings (the same is true for the investment task and for differences in sendings between both games) are equally distributed between the "standard" and the "inverse" treatment ($p > 0.9$). In the "standard" treatment, senders had to make their sending in the social environment first, and then the non-social setting was introduced. In the "inverse" treatment, the sequence was the other way round. As we find that such framing does not influence the sending decision, both treatments are merged in the following analyses. We will first analyze our experimental results statistically, and then comment on the accuracy of the corresponding model predictions.

<<< Insert Table 3 about here >>>

With regression (1) in Table 3, we test whether "amount sent" in the trust game depends on a subject's average "amount returned" in the receiver role, on her average "belief", and on her average "certainty equivalent" (we always use the average over the five data points). Regression (1) first of all reveals that sendings are mainly influenced by "avg. certainty equivalent". Furthermore, we have already shown that, due to the false consensus effect, social preferences as measured in the receiver role influence beliefs. Beliefs of course influence certainty equivalents (both are correlated with $\rho = 0.62$; however, as the highest value for the variance inflation factor is 2.1 in all of our regressions, there are no problems of multicollinearity throughout the empirical part of this paper), and thereby predict sendings. As there is noise in the data, “avg. amount returned” and “avg. belief” as part of this line of thought retain some influence (all three corresponding regression coefficients are positive in regression (1); $p = 0.085$ for “avg. amount returned”, $p = 0.120$ for “avg. belief”). In addition, there might be a direct link between sender and receiver behavior that explains the weak significance of “avg. amount returned”, meaning that reciprocal subjects ceteris paribus might send (to some degree) more than selfish ones.

Regarding the influence of risk aversion on sendings, we add that univariate regressions which test the relevance of risk measures, such as the average difference between “certainty equivalent” and “belief”, or the certainty equivalent in the introductory lottery with known probabilities, only show an insignificant influence on s_i. The same result is found in Ashraf et al. (2006), Houser et al. (2010), and Eckel and Wilson (2004). Accordingly, the latter conclude that there is little evidence for considering trust to be a “risky decision“. However, in a multivariate regression containing “avg. belief”, “avg. certainty equivalent” is significant (compare model (1) again), suggesting that the risk discount, defined as the average difference between “belief” and “certainty equivalent”, may be significant as well if it replaces “avg. certainty equivalent” in model (1). In fact, this is the case. Eckel and Wilson (2004) find this effect as well. Therefore, while the effect of risk aversion on the sending may be fairly small, it nevertheless exists, as should intuitively be the case. Accordingly, we state that Hypothesis 3 is supported.

Models (2) to (5) serve to test Hypothesis 4. Interestingly, the coefficients shown in model (1) become insignificant and substantially smaller in model (2), where the sending decision in the investment task is regarded as well. Now, $\hat{s}_i$ is the only significant independent variable, which, on its own, explains s_i^* with $R^2 = 0.57$, see model (3). In contrast to model (1), model (2) implies that the sending decision in the trust game does not (directly) depend on subjects’ social preferences. The result of model (2) is verified with the help of regression models (4) and (5), where influences on the differences between both sending decisions are analyzed: “Avg. amount returned” on its own cannot explain the difference, see model (4). As well, trust-related control questions have no additional explanatory power, see model (5). In that model, the answer to “Trust” (“Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people?”) was coded with 2 for “Most people can be trusted”, with 1 for “No answer” / “Don’t know”, and with 0 for “You can never be too careful”; answers on “Exploitation” (“Do you think most people would try to take advantage of you if they got a chance, or would they try to be fair?”) were given on a scale from 1 (take advantage) to 10 (try to be fair)); Sex was coded as 1 for male, and 0 for female.

Additionally, some further descriptive statistics are informative: On average, CU 5.0 are sent in the trust game, and only slightly less (CU 4.7) is sent in the investment task, see also Table 1. Most subjects send exactly the same amount in both treatments (61 %). In this group, there are 10 (out of 11) free riders who therefore behave consistently. From the remaining subjects, 23 % show an absolute difference between both decisions of less than or equal to CU 2. We also asked subjects at the end of the game to explain in writing why they had differentiated between both sending decisions (if they had). Many of them, and not only selfish ones, mentioned that they “did not care if the co-player was a real person or a computer (because the return to oneself was the same)”. Some mentioned that they “did not want to harm Person B”, or “played fair”, and therefore sent more in the trust game. Only very few subjects responded somehow consistently to the idea of betrayal aversion, for example, by stating that “the computer can be trusted more than human beings, […that] person B might be greedy, […and that] you can send more money to the computer [than to the human co-player] without having a guilty conscience.” (Answers paraphrased, originally in German).

We compute model predictions for the trust game as described in the previous section. Compared to actual behavior, the predictions err by CU 3.2 on average. This is slightly more precise than predicting decisions in the investment task with the help of certainty equivalents from the lottery task, where the estimation mistake was CU 3.4 (the mistake of CU 3.2 corresponds to a correlation of actual and predicted decisions in the trust game of 0.46). Again – like utility from money, reciprocal utility is typically U-shaped – there are predictions which differ from the actual decision by 10 CU (9 cases), which has a maximal effect on the accuracy of the sending prediction, but typically only very little impact on the accuracy of predicted utility levels. Thus, inaccurate sending predictions in both the investment task and the trust game may not be due to a wrong model or a lack of understanding of subjects, but may be caused by an irrelevance of the sending choice (however, we once again point out that this conclusion is only true for standard trust games, as different believed responder behavior and different risk aversion, for example caused by different multipliers or by framing, can make sendings more or less attractive). This result makes it difficult to test hypotheses with respect to betrayal aversion, which might explain that the literature has not agreed on the effect of betrayal aversion so far. Unfortunately, it is also not possible to compare the accurateness of our predictions with that of other models, because to our best knowledge, we are the first who define betrayal aversion and test this definition with experimental data.

Based on this analysis, we state that most subjects have reciprocal preferences, but that they are almost not relevant in the sender role. If at all, they are in disfavor of the theory of betrayal aversion. Accordingly, Hypothesis 4 can be supported.

4.6 Prediction of Sender Behavior with Literature Data

As already mentioned in Section 2.3, we propose that it depends on the experimental setting whether reciprocal preferences have a positive or a negative effect on sendings. For those studies in the literature, which compare social with non-social settings and report sufficient data, we test whether sender behavior can be explained with the help of our model.

In the design of Aimone and Houser (2012), the sender makes a binary decision between “keeping” and “sending”, and the receiver either answers with “returning a bit” or with “returning half of the amount”. As the multiplier is 6 in their setting, we propose senders to have the following utility function:

$$U_i = \begin{cases} \left(1 - s_i + \tilde{k}_j^*\right)^{(1-r_i)} & \text{, if } r_i < 1 \\ \ln\left(1 - s_i + \tilde{k}_j^*\right) & \text{, if } r_i = 1 \\ -\left(1 - s_i + \tilde{k}_j^*\right)^{(1-r_i)} & \text{, if } r_i > 1 \end{cases} \begin{cases} +Y_i \cdot \sqrt{\left(6 \cdot s_i - \tilde{k}_j^*\right) \cdot \left(\tilde{k}_j^* - s_i\right)}, \text{ if } \tilde{k}_j^* \geq s_i, \\ -Y_i \cdot \sqrt{\left(6 \cdot s_i - \tilde{k}_j^*\right) \cdot \left(s_i - \tilde{k}_j^*\right)}, \text{ if } \tilde{k}_j^* < s_i, \end{cases} \tag{9}$$

with USD 5 being 100 % of the endowment, $s_i \in \{0;1\}$, and $k_j \in \{0.4; 3\}$. If the sender chooses $s_i = 1$ and if the receiver returns little, reciprocal utility is equal to −1.83; in the cooperation case it is 2.45. If senders have unbiased (slightly overoptimistic) beliefs, and expect 66 % (not more than 57 %) of receivers to be free riders, the expected reciprocal utility is equal to –0.39 (below zero), meaning that senders can increase their reciprocal utility by choosing to send nothing. Nevertheless, due to the high multiplier, senders can expect to earn an attractive yield of 28 % of their endowment. Accordingly, many senders will cooperate, but fewer subjects will cooperate in a social setting than in a non-social one, some of them thus exhibiting betrayal-averse behavior. Moreover, if senders are allowed to choose the setting, most subjects will opt for the non-social one (and cooperate) and some subjects (our model predicts: overoptimistic ones with a positive expected reciprocal utility (which we are not able to verify)) will choose the social setting (and cooperate). The results of Aimone and Houser (2012) precisely confirm our predictions.

For the experimental structure of Fetchenhauer and Dunning (2009 and 2012), we compute a reciprocal utility of 1.41 if the receiver cooperates and –2 if he defects. Therefore, when the probability of meeting a cooperator is higher than 59 %, we predict more frequent cooperation in social settings than in non-social ones, and vice versa. Fetchenhauer and Dunning (2009) elicit an average belief in cooperative outcomes of only 45 %, meaning that an average subject should have a slight tendency to behave betrayal-aversely. In contrast to our prediction, the authors find that many senders (64 %) cooperate, although, given their beliefs and compared to their decisions in the non-social lottery, only 30 % are assumed to cooperate because of monetary motives. Thus, the social setting increases cooperation rates. While these average numbers look like evidence against the betrayal aversion hypothesis, the results may support the contrary on the subject level: Due to the false consensus effect, many cooperative senders may believe in a cooperation level of greater than 59 %. These subjects should cooperate in the social setting to maximize their reciprocal utility, and they may not invest their stake in the lottery if they are risk-averse. In contrast, subjects with low beliefs should neither send money in the lottery nor in the trust game. However, as individual belief data are not reported in the paper, we are not able to test this prediction.

In Fetchenhauer and Dunning (2012), senders are informed about the probability of being matched with a cooperator, allowing the clear prediction that fewer subjects should send money in the trust game than in the investment task in the low (46 %) probability treatment, and more subjects should send money in the high (80 %) probability treatment. While the authors do not find a significant difference in the 80-%-treatment, the effect in the 46-%-treatment is significant, and it is opposed to our prediction: Subjects send money in the investment task less frequently than in the trust game. We explain our contradicting prediction as follows: Some senders may have altruistic preferences (see also Cox 2004, Ben-Ner and Halldorsson 2010, or Sapienza et al. 2013), which especially surface in the low probability treatment, because in the non-social setting, participating rates are of course very low if senders know that a negative return is to be expected (only 28.6 % participate, compared to 54.3 % in the trust game). Also note that, like us, Fetchenhauer and Dunning (2012) do not endow the receiver at the beginning of the experiment, which is generally associated with higher sendings due to social concerns such as distress and guilt (see Johnson and Mislin 2011). In settings where both parties are endowed equally (Bohnet and Zeckhauser 2004; Bohnet et al. 2008; Hong and Bohnet 2007), which increases the “threat” for the sender to end up with a smaller payoff than the receiver, or in settings which offer the possibility of not knowing that one has been betrayed (Aimone and Houser 2012), betrayal-averse behavior can be found again.

Bohnet and Zeckhauser (2004) ask subjects for their minimum acceptable probabilities of getting money returned when they send money in a social and in a non-social setting. In their experiment the multiplier is 2, $s_i \in \{0;1\}$, and $k_j \in \{0.8; 1.5\}$. In our model, this results in negative reciprocal utility of –0.49 in the case of defection, and positive utility of 0.5 in the case of cooperation. If senders have unbiased (slightly overoptimistic) beliefs and expect 29 % (up to 49 %) of receivers to cooperate, expected reciprocal utility is negative with –0.20 (below zero). Again, subjects are predicted to behave in a betrayal-averse way. Indeed, on average, higher minimum acceptable probabilities are chosen in the social setting than in the non-reciprocal ones. A similar fit can be established for the results presented in Bohnet et al. (2008) and in Hong and Bohnet (2007).

Similarly to us, Houser et al. (2010) triple the amount sent and use a continuous strategy space. Consistent with our argumentation, sendings increase if senders are informed about a probable return distribution, which reduces ambiguity (compare their treatment Trust-2 with Trust-1). As well, their results confirm our finding that sendings in the social setting are insignificantly higher than in the non-social one (Trust-2 vs. Risk-1). However, as mentioned in Section 1, Houser et al. (2010) only controlled for risk and social preferences, but not for beliefs. In addition, and opposed to us, they refrained from any theoretical analysis of betrayal-averse behavior.

Summarizing, we find that our theory is in line with most of the results reported in the literature. This is especially noteworthy, as the literature results seem to contradict each other. Our model implies that the occurrence of betrayal aversion is situation dependent because the believed *distribution* of returns matters. This implication resolves the contradictions mentioned above to a large extent. However, a systematic proof of this assumption has not been provided so far. Furthermore, from the literature review, the question arises as to how preferences for outcomes can be disentangled from preferences for intentions. This is important, as betrayal aversion is a consequence of intentions, not of outcomes. The question whether the receiver is endowed equally to the sender before the game starts or not addresses distributional concerns. In contrast, this experimental design question does not affect intentional concerns, because the endowment is not part of the players' kindness functions. Thus, from Johnson and Mislin (2011) we know that the endowment matters for the sender's decision, but it is unclear so far how the experimental results, which the literature explains with the help of betrayal aversion, are in fact driven by such distributional concerns. Again, more research is needed to investigate this in carefully controlled experimental settings and to describe the results with a model which captures both distributional and intentional effects.

5 Conclusion

In this paper, we proposed a modified version of the Dufwenberg and Kirchsteiger (2004) reciprocity model, which is able to predict behavior in trust games in the receiver as well as in the sender role. In the receiver role, no uncertainty is present and the receiver knows how friendly the sender is. Accordingly, reciprocal receivers will simply answer kindness with kindness, and return more money than is sent to them. In the sender role, the decision is more difficult, because friendly sendings can backfire: If the receiver is selfish and keeps the money, reciprocal senders will suffer twice, as the money is lost, and trust has been betrayed. On the other hand, in the case of a successful interaction, reciprocal senders gain utility from money as well as from reciprocal utility. In total, expected reciprocal utility is small in our experimental setting. Moreover, as both utility from money and utility from reciprocity depend on the sender's belief in a comparable way, receiver behavior only confirms the selfish decision in most cases, meaning that senders behave identically in a trust game compared to a non-social lottery offering the same returns. If the tradeoff between both utility components matters, reciprocal preferences increase sendings in trust games compared to non-reciprocal settings, implying that these subjects do not act betrayal-aversely, but should better be described as reciprocity-seeking. However, we also showed that in other trust game experiments reported in the literature, reciprocal preferences can indeed have a negative effect on sendings.

Our findings imply that cooperation in trust-game-like situations is not fostered very much by appealing to reciprocal motives, which, for example, explains why social peer-to-peer lending only leads a niche existence compared to classical, non-social investments into bank accounts. In contrast, inducing high beliefs in receivers' returns will generate cooperative and therefore welfare increasing outcomes. We find that people typically have overoptimistic expectations, and Orbell and Dawes (1991) argue that such a bias may have evolved because it can be evolutionary advantageous within certain cooperative dilemmas. In turn, given that economic interactions are apparently built on biased beliefs in trustworthy behavior, our economy may be more vulnerable to changes in people's perceptions than classical economists might have thought.

We predict the occurrence of betrayal-averse behavior to be strongly parameter-dependent, and more research is needed to clarify whether this is indeed the case. Connected to this question, in order to identify the role of betrayal aversion more precisely, additional research has to be done to disentangle non-reciprocal social sending motives, such as inequality aversion (which may depend on the initial endowment of both players), efficiency-maximizing preferences (which may depend on the multiplier), or altruism (which can be influenced by using framing), from those motives which stem from reciprocal considerations.

References

Aimone, J. A., and Houser, D. (2012). What You Don’t Know Won’t Hurt You: A Laboratory Analysis of Betrayal Aversion. *Experimental Economics*, 15, 571-588.

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

Ashraf, N., Bohnet, I., and Piankov, N. (2006). Decomposing Trust and Trustworthiness. *Experimental Economics*, 9, 193-208.

Ben-Ner, A., and Halldorsson, F. (2010). Trusting and Trustworthiness: What Are They, How to Measure Them, and What Affects Them. *Journal of Economic Psychology*, 31, 64-79.

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

Bohnet I., Greig, F., Herrmann, B., and Zeckhauser, R. (2008). Betrayal Aversion: Evidence from Brazil, China, Oman, Switzerland, Turkey, and the United States. *American Economic Review*, 98, 294-310.

Bohnet, I., and Zeckhauser, R. (2004). Trust, Risk and Betrayal. *Journal of Economic Behavior and Organization*, 55, 467-484.

Breuer, W., and Hüwe, A. (2014). Explaining Individual Contributions in Public Goods Games Using (only) Reciprocity and Overoptimism. Working Paper.

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

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

Coleman, J. (1990). Foundations of Social Theory. *The Belknap Press of Harvard University Press*.

Cox, J. C. (2004). How to Identify Trust and Reciprocity. *Games and Economic Behavior*, 46, 260-281.

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

Dunning, D., Fetchenhauer, D., and Schlösser, T. M. (2012). Trust as a Social and Emotional Act: Noneconomic Considerations in Trust Behavior. *Journal of Economic Psychology*, 33, 686-694.

Eckel, C. C., and Wilson, R. K. (2004). Is Trust a Risky Decision? *Journal of Economic Behavior and Organization,* 55, 447-465.

Epstein, L. G., and Zin, S. E. (1989). Substitution, Risk Aversion, and the Temporal Behavior of Consumption and Asset Returns: A Theoretical Framework. *Econometrica*, 57, 937-969.

Erner, C., Klos, A. and Langer, T. (2013). Can Prospect Theory Be Used to Predict an Investor’s Willingness to Pay? *Journal of Banking and Finance*, 37, 1960-1973.

Fehr, E. (2009). On the Economics and Biology of Trust. *Journal of the European Economic Association*, 7(2-3), 235-266.

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

Fetchenhauer, D., and Dunning, D. (2009). Do People Trust too Much or too Little? *Journal of Economic Psychology,* 30, 263-276.

Fetchenhauer, D., and Dunning, D. (2012). Betrayal Aversion versus Principled Trustfulness – How to Explain Risk Avoidance and Risky Choices in Trust Games. *Journal of Economic Behavior and Organization*, 81, 534-541.

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

Gächter, S., and Renner, E. (2010). The Effects of (Incentivized) Belief Elicitation in Public Goods Experiments. *Experimental Economics*, 13, 364-377.

Holt, C. A., and Laury, S. K. (2002). Risk Aversion and Incentive Effects. *The American Economic Review*, 92, 1644-1655.

Hong, K., and Bohnet, I. (2007). Status and Distrust: The Relevance of Inequality and Betrayal Aversion. *Journal Economic Psychology,* 28, 197-213.

Houser, D., Schunk, D., and Winter, J. (2010). Distinguishing Trust from Risk: An Anatomy of the Investment Game. *Journal of Economic Behavior and Organization*, 74, 72-81.

Johnson, N. D. and Mislin, A. A. (2011). Trust Games: A Meta-Analysis. *Journal of Economic Psychology*, 32, 865-889.

Kazuhiro, A. (2009). Defining Trust Using Expected Utility Theory. *Hitotsubashi Journal of Economics*, 50, 99-118.

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

Orbell, J., and Dawes, R. M. (1991). A “Cognitive Miser” Theory of Cooperators’ Advantage. *The American Political Science Review*, 85, 515-528.

Ross, L., Greene, D., and House, P. (1977). The “False Consensus Effect”: An Egocentric Bias in Social Perception and Attribution Processes. *Journal of Experimental Social Psychology*, 13, 279-301.

Sapienza, P., Toldra-Simats, A., and Zingales, L. (2013). Understanding Trust. *The Economic Journal*, 123, 1313-1332.

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

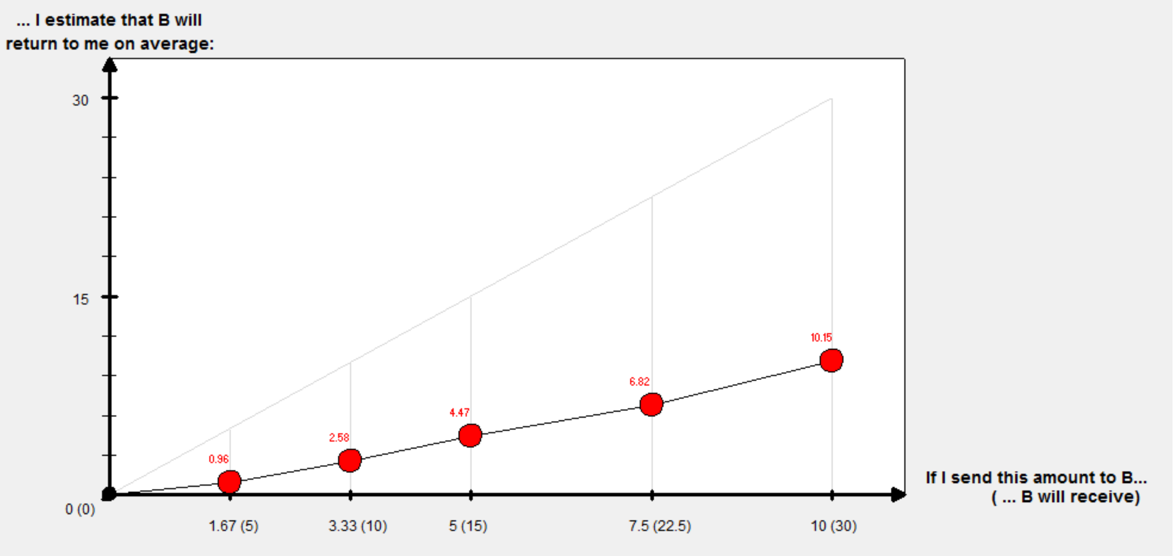
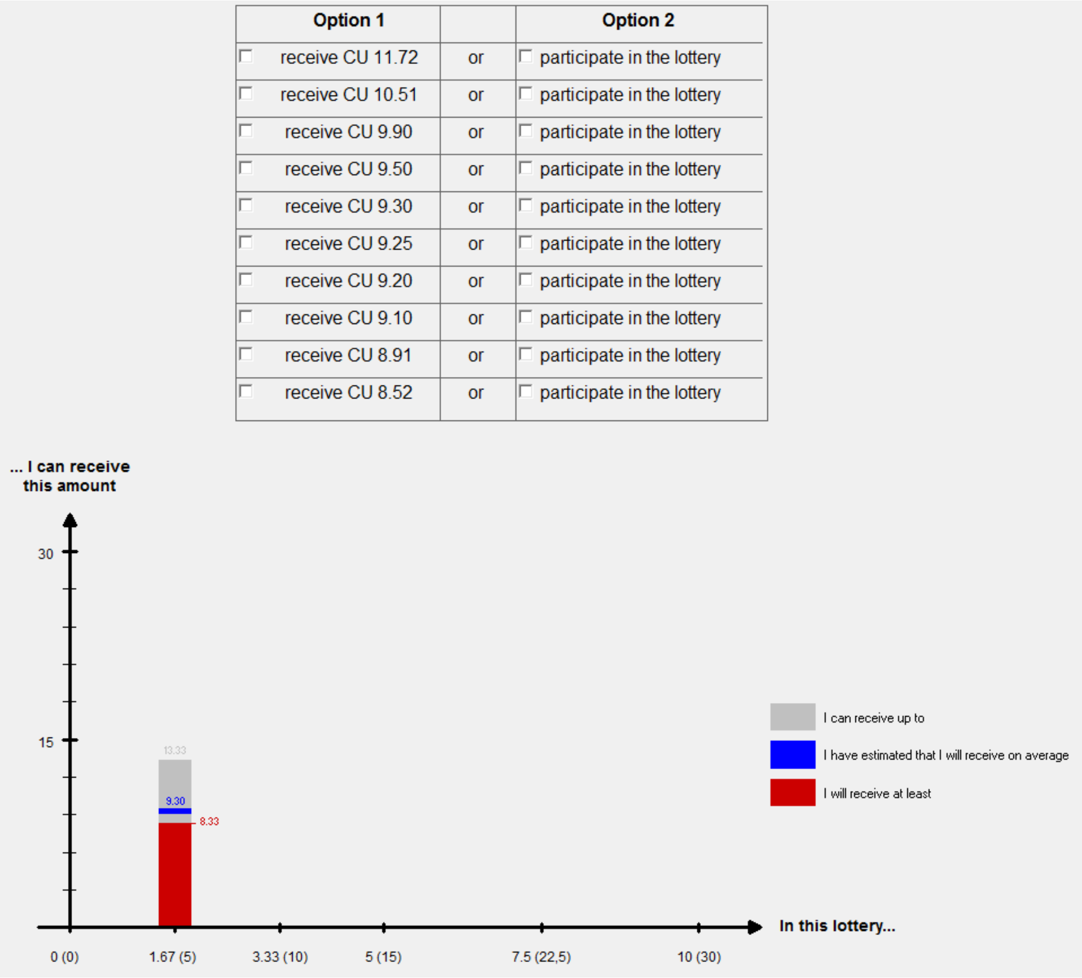


Fig. 1: Graphical belief interface



Option 1		Option 2
□ receive CU 11.72	or	□ participate in the lottery
□ receive CU 10.51	or	□ participate in the lottery
□ receive CU 9.90	or	□ participate in the lottery
□ receive CU 9.50	or	□ participate in the lottery
□ receive CU 9.30	or	□ participate in the lottery
□ receive CU 9.25	or	□ participate in the lottery
□ receive CU 9.20	or	□ participate in the lottery
□ receive CU 9.10	or	□ participate in the lottery
□ receive CU 8.91	or	□ participate in the lottery
□ receive CU 8.52	or	□ participate in the lottery

Fig. 2: Graphical interface for the lottery at s = CU 1.67

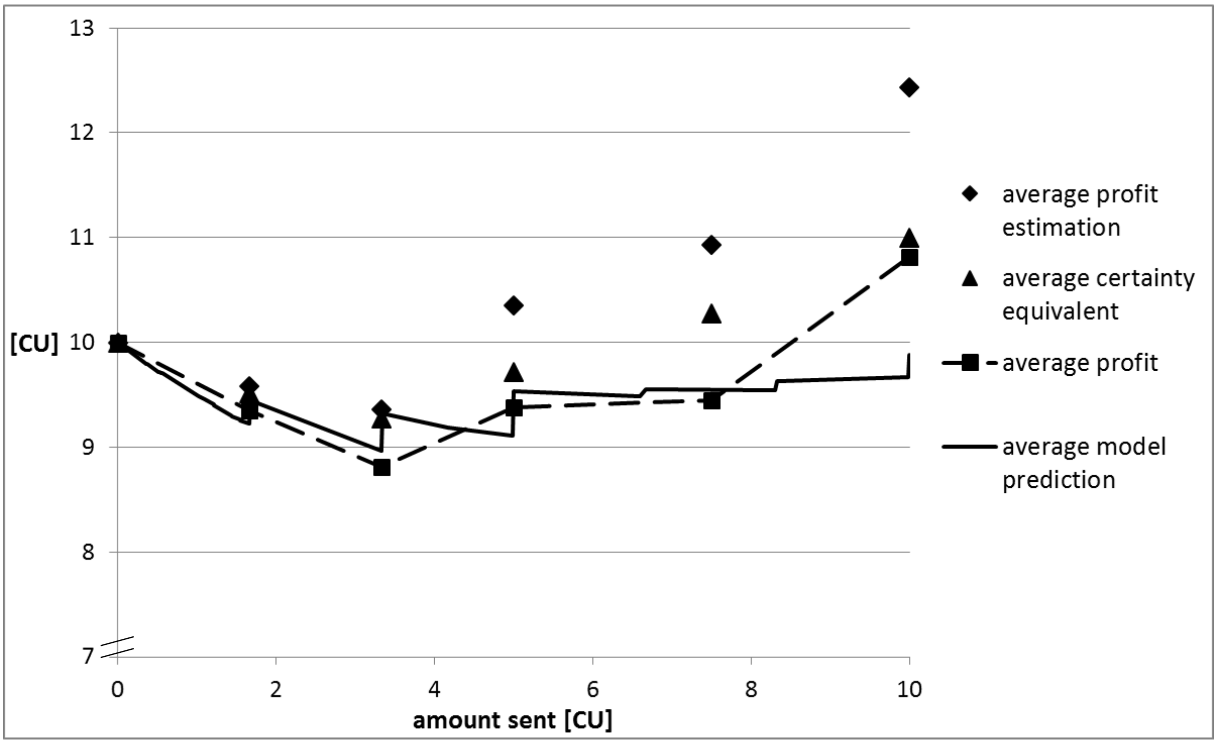


Fig. 3: Average profit estimations (amount kept + average belief regarding the average amount returned), average profits (amount kept + actual average amount returned), average certainty equivalents, derived from the lottery tasks, and average model predictions of average profits (amount kept + predicted average amount returned).

Table 1 - Definitions and descriptive statistics

Variable	Meaning	Average value
s_i	*i*'s sending in the trust game	CU 4.96
$\hat{s}_i$	*i*'s sending in the investment task	CU 4.67
$E_i(\tilde{k}_j)$	*i*'s belief about *j*'s average amount returned	CU 6.03[1)]
$E_i(\tilde{\pi}_i)$	believed payoff in the sender role	CU 10.53[1)]
k_j	*j*'s amount returned	CU 5.06[1)]
ϵ_i	*i*'s overoptimism parameter	1.19[1)]
CE$_i$	*i*'s certainty equivalent of payoffs in the investment / lottery task	CU 9.98[1)]

# subjects: 102
  # reciprocal subjects: 78
  # free riders: 24

[1)] Average over all elicited data points. Thus, actual payoffs from the experiment differed from these values because they were dependent on the random assignment mechanism and on the senders' decisions.

Table 2 - Explaining beliefs[1)]

Dependent variable	Estimation mistake (CU) with respect to $E_i\left(\tilde{k}_j(s)\right)$			
	(1)	(2)	(3)	(4)
$k_i(s) - \mathrm{E}\left(\tilde{k}_j(s)\right)$	0.32*** (0.05)			0.27*** (0.05)
s	0.18*** (0.05)	0.22*** (0.05)	0.15*** (0.05)	0.21*** (0.05)
$\lvert s - s_i \rvert$		-0.35*** (0.05)		-0.25*** (0.05)
Own sending			1.71*** (0.42)	
Constant	0.01 (0.20)	1.13*** (0.26)	-0.12 (0.22)	0.81*** (0.26)
Observations	505[2)]			
R²	0.19	0.13	0.07	0.24

Significant at the 1 percent (***), 5 percent (**), 10 percent (*) level.

[1)] OLS regressions with robust standard errors in parentheses.

[2)] One subject excluded, due to a computer blackout during the belief elicitation stage.

$E_i\left(\tilde{k}_j(s)\right)$: *i*'s belief about the average amount returned, in CU.

$k_i(s) - E\left(\tilde{k}_j(s)\right)$: Difference between *i*'s amount returned in the receiver role and the average amount returned in the subject pool, in CU.

Own sending: Dummy variable being equal to 1 if *i*'s own sending equals the sending to which *i*'s belief refers, $s_i = s$, and 0 otherwise.

$\lvert s - s_i \rvert$: Absolute difference between the sending to which the belief refers and *i*'s own sending, in CU.

Table 3 - Explaining "amount sent" in the trust game[1)]

Dependent variable		s_i		$s_i - \hat{s}_i$	
	(1)	(2)	(3)	(4)	(5)
$\hat{s}_i$		0.66*** (0.10)	0.76*** (0.07)		
Avg. amount returned	0.24* (0.14)	0.10 (0.09)		0.04 (0.07)	0.04 (0.07)
Avg. belief	0.37 (0.23)	0.18 (0.17)			
Avg. certainty equivalent	0.55*** (0.18)	0.22 (0.15)			
Trust					-0.24 (0.28)
Exploitation					0.10 (0.11)
Sex					-0.27 (0.65)
Constant	-5.58** (1.51)	-2.58* (1.44)	1.40*** (0.46)	0.10 (0.48)	0.01 (0.97)
Observations	93[2)]	93[2)]	102	102	102
*R*²	0.33	0.63	0.57	0.00	0.01

Significant at the 1 percent (***), 5 percent (**), 10 percent (*) level.

[1)] OLS regressions with robust standard errors in parentheses.

[2)] Subjects excluded if not a single certainty equivalent could be determined.

Avg. amount returned: Average of "amount returned" per data point in the receiver role.

Avg. belief: Average of "expected returns" per data point.

Avg. certainty equivalent: Average of certainty equivalents per data point, derived from the lottery tasks.

Trust, Exploitation: Selected control questions from the World Values Survey, see Section 3.1.

Appendix – not for publication, only for referees' information

A: Receiver's Reaction to Sendings in the Trust Game According to Dufwenberg and Kirchsteiger (2004)

The payoff to the sender *i* at the end of the trust game is $1 - s_i + k_j$, the payoff to the receiver *j* is $3\cdot s_i - k_j$. The highest monetary payoff of *i* which *j* can establish, is to send back everything he received, which is $k_j = 3\cdot s_i$. Accordingly, $\max\{\pi_i\} = 1 - s_i + 3\cdot s_i = 1 + 2\cdot s_i$. The unkindest behavior is to send $k_j = 0$, resulting in $\min\{\pi_i\} = 1 - s_i$. Thus, the reference point which separates kind from unkind behavior is $\pi_i^{e_j} = \frac{1}{2}\cdot[(1 + 2\cdot s_i) + (1 - s_i)] = 1 + 0.5\cdot s_\mathrm{i}$, and *j*'s actual kindness to *i* is the difference between his actual strategy and his reference strategy: $\kappa_{ji} = \left(1 - s_i + k_j\right) - (1 + 0.5\cdot s_i) = k_j - 1.5\cdot s_i$.

According to the utility function proposed in Dufwenberg and Kirchsteiger (2004), *j*'s utility is

$$U_j = 3\cdot s_i - k_j + Y_{ji}\cdot\left[k_j - 1.5\cdot s_i\right]\cdot\left[\left(3\cdot s_i - k_j\right) - \frac{1}{2}\cdot\left(\left(3\cdot s_{i,\,max} - k_j\left(s_{i,\,max}\right)\right) - 0\right)\right]. \qquad \text{(A.1)}$$

Thus, *j*'s utility is the sum of utility from money (which is equal to *j*'s earnings), and the reciprocal utility, which is the product of *j*'s kindness to *i* and *i*'s kindness to *j*, weighted with the reciprocity parameter Y_{ji}. While we have determined *j*'s kindness above, *i*'s reference strategy is unclear so far because it depends on *j*'s reaction on s_i. The most unfriendly strategy of *i* is obviously to send nothing. To determine the friendliest possible strategy, we define $s_{i,\max}$, which is the sending of *i* that maximizes *i*'s perceived kindness λ_{jij}. The reaction of *j* on maximal kindness of *i* can be determined by maximizing $U_j\left(k_j\left(s_{i,\max}\right)\right)$ over k_j.

$$U_j\left(k_j\left(s_{i,\max}\right)\right) = 3\cdot s_{i,\max} - k_j + Y_{ji}\cdot\left[k - 1.5\cdot s_{i,\max}\right]\cdot\left[1.5\cdot s_{i,\max} - 0.5\cdot k_j\right] \qquad \text{(A.2)}$$

Utility maximization implies $k_j = 2.25 \cdot s_{i,\max} - \frac{1}{Y_{ji}}$. Accordingly, $\lambda_{jij,\max} = 1.5 \cdot s_{i,\max} - 0.5 \cdot \left(2.25 \cdot s_{i,\max} - \frac{1}{Y_{ji}}\right)$, which is, as $\lambda_{jij,\max}$ is strictly increasing in s_i and as the domain of s_i is restricted, implying that $s_{i,\max} = 1$.

Thus, the reference strategy of *j* has been determined and (B.1) can be specified to:

$$U_j = 3 \cdot s_i - k_j + Y_{ji} \cdot \left[k_j - 1.5 \cdot s_i\right] \cdot \left[\left(3 \cdot s_i - k_j\right) - \left(0.375 + \frac{1}{2 \cdot Y_{ji}}\right)\right]. \tag{A.3}$$

Intuitively, *i*’s believed kindness to *j* decreases in Y_{ji}, because the higher Y_{ji}, the more will *j* return (see below), which reduces *j*’s payoffs and therefore *i*’s kindness. Maximizing (A.3), one has to consider that k_j is restricted to the positive domain. Thus,

$$k_j^* = \max\left\{0;\ 2.25 \cdot s_i - 0.1875 - \frac{3}{4 \cdot Y_{ji}}\right\}. \tag{A.4}$$

Accordingly, receivers always send nothing back if sendings are small, and increase k_j in s_i by 2.25, whereas the starting point of reciprocal behavior is determined by Y_{ji}.

B: Model Properties

1) Proof that equation (4) is correct.

For $k_j \geq s_i$:

$$\frac{dU_j}{dk_j} = \frac{Y_j \cdot (2 \cdot s_i - k_j)}{\sqrt{3 \cdot s_i - k_j} \cdot \sqrt{k_j - s_i}} - 1 = 0 \tag{B.1}$$

$$\Rightarrow Y_j^2 \cdot \left(2 \cdot s_i - k_j\right)^2 = \left(3 \cdot s_i - k_j\right) \cdot \left(k_j - s_i\right)$$

$$\Leftrightarrow 4 \cdot s_i^2 \cdot Y_j^2 + 3 \cdot s_i^2 = k_j \cdot \left(4 \cdot s_i + 4 \cdot s_i \cdot Y_j^2\right) - k_j^2 \cdot \left(1 + Y_j^2\right)$$

$$\Leftrightarrow k_j = \frac{2\cdot s_i + 2\cdot s_i\cdot Y_j^2}{1 + Y_j^2} \overset{(+)}{-} \sqrt{\left(\frac{2\cdot s_i + 2\cdot s_i\cdot Y_j^2}{1+Y_j^2}\right)^2 - \frac{4\cdot s_i^2\cdot Y_j^2 + 3\cdot s_i^2}{1+Y_j^2}}$$

$$= s_i \cdot \left(2 \overset{(+)}{-} \sqrt{4\cdot\frac{1 + Y_j^2}{1 + Y_j^2} - \frac{4\cdot Y_j^2 + 3}{1 + Y_j^2}}\right)$$

$$= \left(2 \overset{(+)}{-} \frac{1}{\sqrt{1 + Y_j^2}}\right) \cdot s_i. \tag{B.2}$$

Inserting (B.2) into (B.1) gives

$$\frac{\mathrm{d}U_j}{\mathrm{d}k_j} = \frac{Y_j \cdot s_i \cdot \left(2 - \left(2 \overset{(+)}{-} \frac{1}{\sqrt{1 + Y_j^2}}\right)\right)}{\sqrt{3\cdot s_i - k_j} \cdot \sqrt{k_j - s_i}} - 1,$$

which does not equal zero for $k_j = \left(2 + \frac{1}{\sqrt{1 + Y_j^2}}\right) \cdot s_i$. Accordingly, $k_j = \left(2 - \frac{1}{\sqrt{1 + Y_j^2}}\right) \cdot s_i$ is the only valid solution for equation (B.1). As $U_j{}'\left(k_j \to s_i\right) \to +\infty > 0$ and as $U_j{}'\left(k_j \to 3\cdot s_i\right) \to -\infty < 0$, one can conclude that $\left(2 - \frac{1}{\sqrt{1 + Y_j^2}}\right) \cdot s_i$ refers to a maximum.

For $k_j < s_i$:

$$\frac{\mathrm{d}U_j}{\mathrm{d}k_j} = \frac{Y_j \cdot (2\cdot s_i - k_j)}{\sqrt{3\cdot s_i - k_j} \cdot \sqrt{s_i - k_j}} - 1 = 0.$$

Similarly, it can be shown that extremum candidates are given by

$$k_j = \left(2 \overset{(+)}{-} \frac{1}{\sqrt{1 - Y_j^2}}\right) \cdot s_i.$$

In this case, only $k_j = \left(2 - \frac{1}{\sqrt{1-Y_j{}^2}}\right) \cdot s_i$ can be a valid solution. The extremum exists and k_j is positive only for $Y_j < \frac{\sqrt{3}}{2}$. As $U_j{}'\left(k_j \to s_i\right) \to +\infty > 0$, and as $U_j{}'\left(k_j = 0\right) = \frac{Y_j \cdot 2}{\sqrt{3}} - 1 > 0$ if $Y_j < \frac{\sqrt{3}}{2}$, the extremum must be a minimum. Accordingly, a receiver will never return $0 < k_j < s_i$.

2) Proof that equation (5) is correct.

For $k_j \geq s_i$:

$$\frac{\mathrm{d}U_j}{\mathrm{d}k_j} = \frac{1}{\sqrt{3 \cdot s_i - k_j}} \cdot \left(\frac{Y_j \cdot (2 \cdot s_i - k_j)}{\sqrt{k_j - s_i}} - \frac{1}{2}\right) = 0$$

$$\Rightarrow 2 \cdot Y_j \cdot \left(2 \cdot s_i - \mathrm{k_j}\right) = \sqrt{k_j - s_i}$$

$$\Leftrightarrow 4 \cdot Y_j^2 \cdot k_j^2 - k_j \cdot \left(16 \cdot s_i \cdot Y_j^2 + 1\right) = - s_i - 16 \cdot Y_j^2 \cdot s_i^2$$

$$\Leftrightarrow k_j = \frac{16 \cdot s_i \cdot Y_j^2 + 1}{8 \cdot Y_j^2} \genfrac{}{}{0pt}{}{(+)}{-} \sqrt{\left(\frac{16 \cdot s_i \cdot Y_j^2 + 1}{8 \cdot Y_j^2}\right)^2 - \frac{s_i + 16 \cdot Y_j^2 \cdot s_i^2}{4 \cdot Y_j^2}}$$

$$= 2 \cdot s_i + \frac{1}{8 \cdot Y_j^2} \genfrac{}{}{0pt}{}{(+)}{-} \sqrt{\frac{s_i}{4 \cdot Y_j^2} + \frac{1}{64 \cdot Y_j^4}}. \tag{B.3}$$

As above, $2 \cdot s_i + \frac{1}{8 \cdot Y_j^2} + \sqrt{\frac{s_i}{4 \cdot Y_j^2} + \frac{1}{64 \cdot Y_j^4}}$ is not a valid solution and $2 \cdot s_i + \frac{1}{8 \cdot Y_j^2} - \sqrt{\frac{s_i}{4 \cdot Y_j^2} + \frac{1}{64 \cdot Y_j^4}}$ refers to a maximum.

Similarly, for $k_j < s_i$, the extremum can be shown to correspond to

$$k_j = 2 \cdot s_i - \frac{1}{8 \cdot Y_j^2} - \sqrt{\frac{1}{64 \cdot Y_j^4} - \frac{s_i}{4 \cdot Y_j^2}}.$$

Again, this defines a minimum. Accordingly, a receiver will never return $0 < k_j < s_i$.

3) Proof that $Y_j \approx 0.5073$ is the threshold which separates defection from cooperation for receivers with $r_j = 0$:

The receiver’s utility, as defined in equation (3) in the paper, is obviously maximal at $k_j = 0$ if Y_j is low. As well, it is obviously maximal at k_j^*, as defined in equation (4) if Y_j is large. Responders will be indifferent between these decisions if

$$U_j\left(k_j = 0\right) = 3\cdot s_i - Y_j\cdot\sqrt{3\cdot s_i^2} = \left(3\cdot s_i - \left(2 - \frac{1}{\sqrt{1 + Y_j^2}}\right)\cdot s_i\right) +$$

$$Y_j\cdot\sqrt{\left(3\cdot s_i - \left(2 - \frac{1}{\sqrt{1 + Y_j^2}}\right)\cdot s_i\right)\cdot\left(\left(2 - \frac{1}{\sqrt{1 + Y_j^2}}\right)\cdot s_i - s_i\right)} = U_j\left(k_j = k_j^*\right)$$

$$\Leftrightarrow 2 - \frac{1}{\sqrt{1 + Y_j^2}} = Y_j\cdot\sqrt{3} + Y_j\cdot\sqrt{\left(3 - \left(2 - \frac{1}{\sqrt{1 + Y_j^2}}\right)\right)\cdot\left(\left(2 - \frac{1}{\sqrt{1 + Y_j^2}}\right) - 1\right)}$$

$$\Leftrightarrow 2 - \frac{1}{\sqrt{1 + Y_j^2}} = Y_j\cdot\sqrt{3} + Y_j\cdot\sqrt{\left(1 + \frac{1}{\sqrt{1 + Y_j^2}}\right)\cdot\left(1 - \frac{1}{\sqrt{1 + Y_j^2}}\right)}$$

$$\Leftrightarrow 2\cdot\left(1 + Y_j^2\right) - \sqrt{1 + Y_j^2} = Y_j\cdot\sqrt{3}\cdot\left(1 + Y_j^2\right) + Y_j^2\cdot\sqrt{1 + Y_j^2}$$

$$\Leftrightarrow \left(-1 - Y_j^2\right)\cdot\sqrt{1 + Y_j^2} = Y_j\cdot\sqrt{3}\cdot\left(1 + Y_j^2\right) - 2\cdot\left(1 + Y_j^2\right)$$

$$\Leftrightarrow \left(1 + Y_j^2\right) = Y_j^2\cdot 3 - 4\cdot Y_j\cdot\sqrt{3} + 4$$

$$\Leftrightarrow Y_j^2 - 2\cdot Y_j\cdot\sqrt{3} + 1.5 = 0$$

$$\Leftrightarrow Y_j = \sqrt{3} - \sqrt{1.5} \approx 0.5073. \qquad \text{(B.4)}$$

4) Proof that expected reciprocal utility can be both negative if $\mathrm{E}(\pi_i) > 1$ and positive if $\mathrm{E}(\pi_i) < 1$.

We consider an extreme case with only two receiver types in the following. Subjects (with overoptimistic or unbiased beliefs) believe receivers either to be kind (the expected amount returned is denoted with $c{\cdot}s$ in the following, the proportion of kind receivers is denoted with p), or to return nothing. If all kind receivers return the same amount, $c{\cdot}s$ (unkind receivers return nothing), and if $\mathrm{E}(\pi_i) = 1$ is considered, the following equation must hold:

$$\mathrm{E}(\pi_i) = p \cdot (1 - s + c{\cdot}s) + (1 - p) \cdot (1 - s + 0) = 1$$

$$\Leftrightarrow c \cdot p \cdot s - s = 0$$

$$\Leftrightarrow p = \frac{1}{c}. \tag{B.5}$$

Accordingly, expected reciprocal utility is equal to

$$= s \cdot \left[\frac{1}{c} \cdot \sqrt{(3 - c) \cdot (c - 1)} - \left(1 - \frac{1}{c}\right) \cdot \sqrt{3}\right], \tag{B.6}$$

which is smaller than zero for $p < \frac{2}{3}$, respectively c > 1.5. Consider a receiver pool where reciprocal utility is distinctly negative because p is distinctly smaller than $\frac{2}{3}$, but entails one receiver who returns slightly more than $c{\cdot}s$. In that case, expected reciprocal utility will be still smaller than zero, but payoffs will be positive on average.

In turn, for $p > \frac{2}{3}$, expected reciprocal utility will be positive in this example. It will remain positive if one receiver returns slightly less than $c{\cdot}s$, meaning that expected payoffs will be negative.

Also note that one *cannot* conclude that reciprocal utility is always positive if the proportion of kind receivers is larger than $\frac{2}{3}$ and $\mathrm{E}(\pi_i) > 1$: Note that due to the concavity of positive reciprocal utility, a pool of kind receivers who equally return c is the most favorable distribution. Having a more diverse pool of kind receivers, expected reciprocal utility can become negative even if more than $\frac{2}{3}$ of all receivers return money.

C: Receivers' Reactions to Sendings According to Fehr and Schmidt (1999)

For the trust game, the utility function of an inequality-averse receiver, according to Fehr and Schmidt (1999), is the following:

$$U_j(k_j) = 3 \cdot s_i - k_j - \alpha_j \cdot \max\{1 - s_i + k_j - (3 \cdot s_i + k_j); 0\} - \beta_j \cdot \max\{3 \cdot s_i + k_j - (1 - s_i + k_j); 0\}$$

$$= 3 \cdot s_i - k_j - \alpha_j \cdot \max\{1 - 4 \cdot s_i + 2 \cdot k_j; 0\} - \beta_j \cdot \max\{4 \cdot s_i - 1 - 2 \cdot k_j; 0\}, \qquad \text{(C.1)}$$

with $\beta_j \leq \alpha_j$ and $\beta_j < 1$: Receivers prefer equal payoffs, but they also prefer to have more than the sender over having less. Receivers also face a tradeoff between inequality and higher payoffs for themselves.

If $\beta_j < 0.5$, $k_j = 0$ is always the best reply because receivers weight utility from money more strongly than disutility from inequality. Money will only be returned if $\beta_j > 0.5$ (the receiver is sufficiently inequality-averse), $k_j \leq 2 \cdot s_i - 0{,}5$, and $s_i \geq 0.25$ (the receiver does not want to earn less than the sender). In this case, (B.1) simplifies to

$$U_j(k_j) = 3 \cdot s_i - k_j - \beta_j \cdot (4 \cdot s_i - 1 - 2 \cdot k_j), \qquad \text{(C.2)}$$

which is maximal at the corner solution $k_j^* = 2 \cdot s_i - 0.5$. Thus, receivers with $\beta_j > 0.5$ will establish equality, receivers with $\beta_j < 0.5$ will keep the whole sending, and receivers with $\beta_j = 0.5$ are indifferent between returning nothing and returning k_j^*.

Instructions For and Screenshots Of the Experiment

In the following, we give an English translation of the instructions which were handed out to the subjects. As well, we show the most important extracts from the experiment, which was conducted in German originally.

Guidelines For the Experiment

Please read the following instructions **carefully**. If you have a question, call out your seat number. Please also read the instructions carefully which are provided during the experiment.

General:

- The experiment consists of the following components: An **interactive experiment**, an **estimation exercise,** and **selection decisions**.
- In most cases, currency units **(CU)** will be used instead of euros. This will allow you to calculate with round sums. You can convert CU into euros at any time: CU 3 are worth EUR 1.
- Experimental results as well as your payout will not be revealed to you before the end of the experiment. To determine your payout, some of your decisions will be randomly drawn. As this will be done at the end of the experiment, **any of your decisions may be relevant for your payout**.
- During the experiment, we will sometimes ask you test questions. We do so to ensure that you have understood the experiment. For each correct answer, you can earn 10 euro-cents.
- Please use a period instead of a comma and also enter values without the currency unit. For example: enter “3.5” and not “CU 3,5”.

Details:

Interactive Experiment

- The experiment will be conducted with two players (called A-role and B-role). Your co-player **will be drawn randomly and anonymously** by the computer.
- The interactive experiment will proceed as follows: **Player A gets CU 10, player B gets nothing**. A can **keep** the CU 10, or **send any portion of it to player B**. The computer will **triple** the amount sent. B receives this triple amount and can **keep** all of it, or return **any portion of it to A** (the amount returned will not be increased). The interactive experiment is then over.

- Two **random** examples:
  - A keeps the CU 10. In that case, B will get 0 CU. Thus, B cannot return any money. In the end, A will be paid CU 10, and B will get CU 0.
  - A sends the CU 10 to B. B will therefore get CU 30 (CU 10 · 3 = CU 30). B decides to return CU 0 to A. In the end, B will be paid CU 30 and A will get CU 0. If B returns everything to A, B gets nothing and A gets CU 30.
- Your input in the A-role:
  As player A, you enter into an input box how many CU you want to send to player B.
- Your input in the B-role:
  If you are player B, you will not be informed of the amounts that A has sent until the experiment ends. Therefore, you have to define a return amount for each possible amount sent by A. At the end of the experiment, from your return amounts, the one that corresponds with A’s amount sent will be selected. You must enter your return amounts into a diagram. Now take a look at **diagram 1 on the additional sheet which was handed out to you**:

- This diagram will be displayed to you in the B-role. When you click on the vertical lines in the diagram, **red dots** will appear which you can **move up and down with the help of the computer mouse. This is how you set your decisions.**

- The horizontal x-axis shows the amounts which A could send to you (you would receive the triple amount). If A sends CU 0 to you, you cannot return any money. The more money A sends to you (move to the right on the horizontal x-axis), the more you can return (move upwards on the vertical y-axis). If A sends you CU 10 (on the x-axis to the far right) you will receive CU 30, and you can return any amount between CU 0 and 30.
- A can also chose to send an amount somewhere between the labeled values on the x-axis, e.g. CU 6. In such cases, the computer will calculate your decision with the help of connecting lines, which will be plotted between the red dots later on.

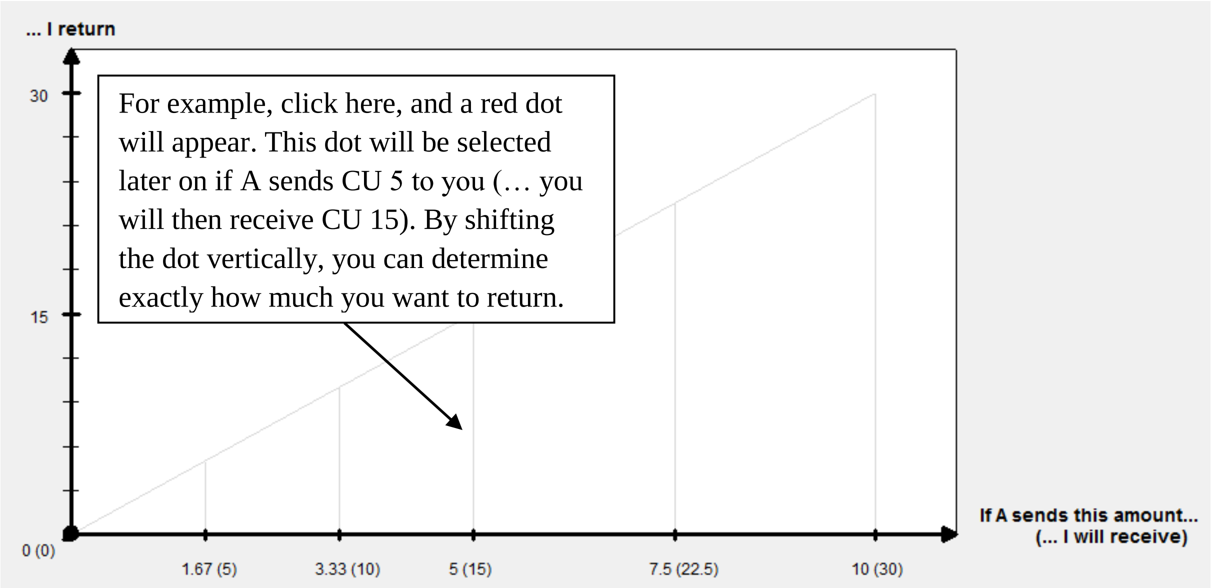
Estimation Exercise

- Within the experiment, you will do an estimation exercise. **The more correctly you estimate, the more you will earn.** The estimation exercise consists of several individual estimations. Your earnings will be calculated on the basis of your average estimation error. In total, you can earn **up to 2 euros.**
- For each CU that you have misestimated by, your earnings will be reduced by 1 euro. (If you are out by more than CU 2, you will earn EUR 0: Then, nothing will be deducted from your previously earned money.) For example: Your average estimation error is CU 1.5. In that case, you will earn EUR 0.5.

Selection Decisions

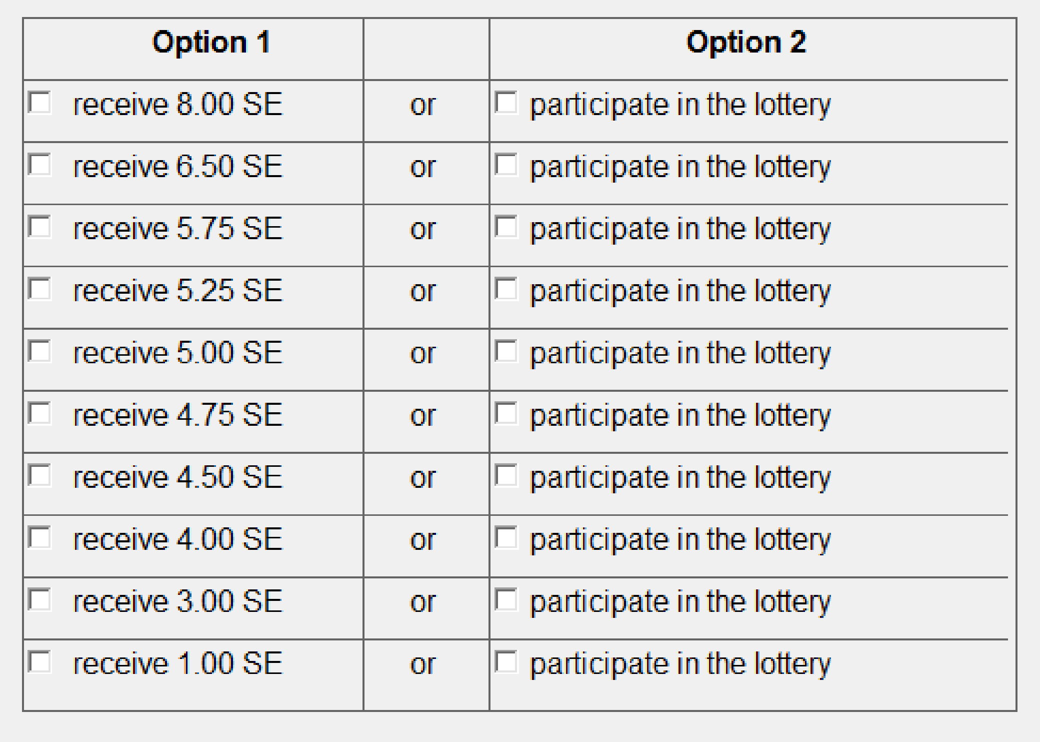
- During the experiment we will ask you several times whether you want to **take part in a lottery**. Instead of participating in the lottery, you can choose to receive a safe amount of money. Please now take a look at **graphic 2 on the additional sheet.**
  - In the presented lottery, you can win any amount between CU 0 and 10 with the same probability.
  - For each row you have to decide whether you want to take the safe amount, which is displayed in the left-hand column of the table, or whether you want to take part in the lottery. Example: Consider the third row in the table. Here, you have to decide whether you want to get CU 5.75 or to take part in the lottery, where you can earn something between CU 0 and 10. Of course, in the upper rows of the table, Option 1 is particularly attractive, whereas in the lower rows, Option 2 becomes more attractive.
  - At the end, the computer will **randomly choose a row from the table**. **Only your decision in this row will be paid out**. Any single decision in the lottery tasks can, therefore, be the only payout relevant one.
  - Apart from the lottery presented to you on the additional sheet, you will take part in several other lotteries. In those lotteries, the exact chances of winning will be unknown to you, but you will be able to estimate them approximately. Further information will be provided to you during the experiment. Here, too, only one of your decisions will be selected randomly at the end and paid out.

Diagram 1 (Role B)



For example, click here, and a red dot will appear. This dot will be selected later on if A sends CU 5 to you (… you will then receive CU 15). By shifting the dot vertically, you can determine exactly how much you want to return.

Graphic 2



Option 1		Option 2
□ receive 8.00 SE	or	□ participate in the lottery
□ receive 6.50 SE	or	□ participate in the lottery
□ receive 5.75 SE	or	□ participate in the lottery
□ receive 5.25 SE	or	□ participate in the lottery
□ receive 5.00 SE	or	□ participate in the lottery
□ receive 4.75 SE	or	□ participate in the lottery
□ receive 4.50 SE	or	□ participate in the lottery
□ receive 4.00 SE	or	□ participate in the lottery
□ receive 3.00 SE	or	□ participate in the lottery
□ receive 1.00 SE	or	□ participate in the lottery

Add. 2: Course of the Experiment

Control Questions

We asked for details of sex, age, student, course of studies if student, number of siblings, country of birth, and wealth status.

Questions associated with trust and risk attitudes:

Generally speaking, would you say that most people can be trusted or that you need to be very careful in dealing with people? Possible answers: Most people can be trusted / You can never be too careful when dealing with others / do not know; refused

Do you think most people would try to take advantage of you if they got a chance, or would they try to be fair? Answers: Scale from 1 to 10.

Now I will briefly describe a fictive person. Would you please indicate in the following on a scale whether that person is very much like you (1) or not at all like you (10)?

Living in secure surroundings is important to this person; to avoid anything that might be dangerous.

Adventure and taking risks are important to this person; to have an exciting life.

[…]

Test Questions

How much will A earn in total if A sends CU 5 and B determines in his diagram that CU 15 will be sent back?

How much will B be paid out if A sends CU 10 and B returns CU 30?

How much will B be paid out instead if B does not return CU 30, but CU 0 (A is still sending CU 10)?

How much will A be paid out in this case?

Explanation of the Investment Task (Only in the "Inverse" Treatment)

Thank you. You have now been assigned to **role A**.

Please now note that you will take part in a modified version of the experiment. Unlike the version explained in the instructions, no **real** Person B will be matched with you. Instead, the computer will receive the (triple) amount sent, and perhaps send a portion of it back. How will the computer decide which amount to return? **There are subjects in this room who are in the B-role and who are interacting with subjects in the A-role. The computer will randomly choose a return decision of one of these B-persons.** With regard to your own pay-outs, the situation has therefore not changed from the version explained in the instructions. The only difference now is that other persons' pay-outs will not be affected by your decision.

First Input Stage, Role A

In the "standard" treatment (input for the trust game):

You have been assigned to role A for the interactive experiment!

Please make your input now. You have CU 10. You can keep this money or you can send any portion of it to B. The amount you send will be tripled. B can keep this triple amount or return any portion of it to you.

I want to send to B □ CU.

In the "inverse" treatment (input for the investment task):

This stage was identical to the second input stage, role A, in the "standard" treatment, see below.

Explanation of Second Input Stage

In the "standard" treatment:

Thank you. In the next stage you will take part in the experiment in the A-role again. Unlike the last stage, this time **no real person B** will be matched with you. Instead, the computer will receive the (triple) amount sent and perhaps send a portion of it back. How will the computer decide which amount to return? **The computer will randomly choose an answer from a real person B in this room.** If you only consider your own payments, the situation therefore has not changed compared to the last stage. The difference is that now payments to another person **are not affected** by your decision.

In the "inverse" treatment:

Thank you. In the next stage you will take part in the experiment in the A-role again. Unlike the last stage, this time **a real person B** will be matched with you. The situation now is as explained in the instructions: You send an amount to a real, randomly drawn person B, the amount is tripled, person B can then return any portion of it. With regard to your own payments, the situation has therefore not changed compared from the last stage. The only difference now is that pay-outs to other persons **will be affected** by your decision.

Second Input Stage, Role A

In the "standard" treatment (input for the investment task):

You have CU 10. What amount do you want to send to the computer? The computer will randomly select one answer from a B-person in this room and return the corresponding amount to you.

I want to send an amount of CU □ to the computer.

In the "inverse" treatment (input for the trust game):

The stage was identical to the first input stage, role A, in the "standard" treatment, see above.

Explanation and Test Questions, Belief Stage

Thank you. At the next stage, you will have to estimate how much the Bs will return to you on average. The more correct your estimations are, the more you will earn. You must enter your estimations into a diagram, as was explained to you in the instructions handed out previously. Thus, at the next stage, click into the diagram several times and slide each of the red dots to the level of the expected return.

Two examples:

Consider the position 5 (15) on the x-axis. Setting that red dot as high as possible, at 15, will have the following implication: You expect that **all the Bs, without exception**, will return everything (CU 15), if they are sent CU 5. If you take the dot to the right of this position, at 7.5 (22.5), and set it at a level of CU 0.1, you estimate that on average the Bs will only return CU 0.1 if CU 7.5 are sent to them. (These examples have been chosen randomly and may therefore be unrealistic!)

Please answer the following test questions. Note: Instead of a comma you have to use a period.

Assume that CU 7.5 were sent to B. If (for whatever reason) you believe that half of the Bs will return the whole amount (CU 22.5), and the other half nothing: At which level will you have to set the red dot at the position 7.5 (22.5) in the next stage?

Assume that there are 15 participants in the B-role. If CU 10 are sent and if you (for whatever reason) believe that 5 of these participants will return CU 0, 5 participants will return CU 10, and the remaining 5 participants will return CU 20: At which level will you have to set the red dot at the position 10 (30)?

Input Stage, Estimation Exercise

Compare Fig. 1 in the paper.

Explanation of Lottery Tasks

For the screenshot, compare Fig. 2 in the paper. Additionally, in the "standard" treatment, the following instructions were displayed. In the "inverse" treatment, "in role A in the second variant" was replaced by "in role A in the first variant".

Explanation:

Thank you. Now, as in the instructions explained, we ask for the Selection Decisions. Again, you have to choose between Option 1 and Option 2 in each row (look at the table on the left). If you choose Option 1, you will receive the payment which is displayed in that row. If you choose Option 2, you will participate in a lottery with an uncertain outcome. Please look at the bar chart on the left, which represents the chances of winning in the lottery.

Your chances of winning can be determined analog to the winning chances in role A in the second variant - it depends on the returns of the B-players in this room. The red bar shows your minimum payout which you will receive in this lottery; in this example it is 8.33 CU. This is the same amount you would have earned for certain in the Interactive Experiment, if you had sent 1.67 CU to B, and the computer had drawn a B-player who would have kept everything. Accordingly, in gray it is displayed how much you can maximally earn: If there are B-players in this room who return everything and if the computer draws one of these decisions, you will earn 13.33 CU. Values in between can be determined accordingly. By estimating average returns of the Bs in the last stage you have already estimated how much you will earn on average in this lottery. The blue line in the bar chart indicates this estimation. **Please consider that the bar does not contain information on the probability of single returns: If, for example, there is no B-player in this room who returns everything (nothing), it will be impossible for you in the lottery to win 13.33 CU (8.33 CU).**

For clarification: Again, the B-players are **not affected by your decisions**; only their return decisions are used to calculate your profit!

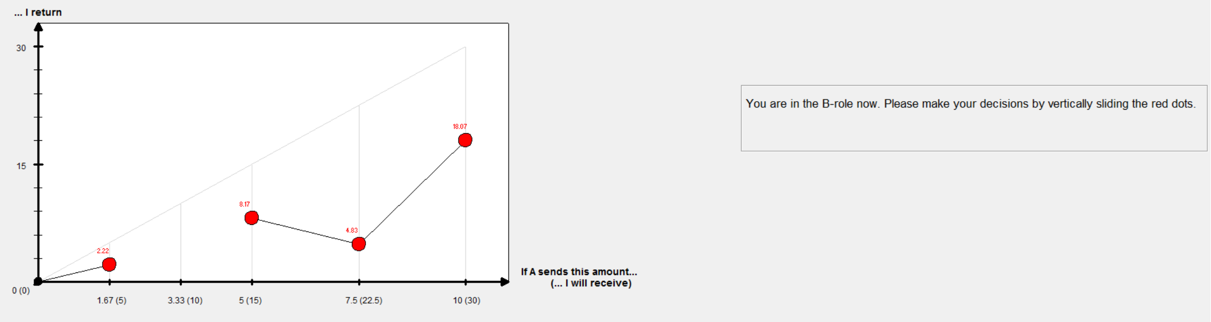
You will play analog lotteries five times in the next stages. **Only one of your decisions, meaning one of the check marks in the next five lottery stages, will be paid at the end, Thus, any of your decisions could be the only payout-relevant one!**

[…]

Explanation of the Input Stage in Role B

Thank you. Decisions in the lottery task have been completed now. At the next stage, we ask you to play the interactive experiment again, this time **in the B-role**! Other subjects in this room are playing the experiment in the A-role and one of these subjects will be randomly assigned to you and will send between CU 0 and 10 to you. In a diagram which you will recognize from the instructions handed out to you, you will have to determine what amount you want to return. You will thereby make a payout-relevant decision which is relevant for you as well as for A!

Input Stage, Role B



You are in the B-role now. Please make your decisions by vertically sliding the red dots.

Using the Carrot and the Stick? Theoretical and Experimental Insights Into Positive vs. Negative Reciprocity

by

Wolfgang Breuer[*], Anselm Hüwe[**]

November 2014

Abstract: In this paper we analyze negative reciprocity theoretically as well as experimentally. Although the reciprocity model of Dufwenberg and Kirchsteiger (2004) is often cited in the literature to justify why people punish unkind behavior, we show that this model is not able to predict punishments in ultimatum games. We therefore propose several model modifications: Amongst others, we suggest "gradual reciprocation", meaning that subjects want to reciprocate the level of the others' kindness. Moreover, we assume that perceived (un)kindness of the proposer towards the responder depends on the belief of the former regarding the punishing behavior of other responders. Accordingly, we measure this belief and find that punishments are significantly overestimated, implying fairer offers. While our model modifications lead to a correct prediction of behavior in the (convex) ultimatum game, a within-subject comparison with behavior in a trust game reveals that decisions in the second-mover roles in either game are not correlated. Robustly, we find no correlation, although our novel, graphical way of eliciting decisions from a continuous strategy space allows subjects to precisely express their preferences. Despite this result, we find a small, but significantly positive correlation between reciprocity parameters in both roles. This is possible because behavior is not only determined by a subject's reciprocal inclination, but by her individual beliefs as well.

Keywords: convex ultimatum game, overoptimism, social preferences, reciprocity

JEL classification: C72, C91, D03, D63, D83

[*] Prof. Dr. Wolfgang Breuer
Chair of Business Administration and Finance
RWTH University
Department of Finance
Templergraben 64
52056 Aachen
Germany
Phone: +49 241 8093539
Fax: +49 241 8092163
E-mail: wolfgang.breuer@bfw.rwth-aachen.de

[**] Dipl.-Ing. Anselm Hüwe
Chair of Business Administration and Finance
RWTH University
Department of Finance
Templergraben 64
52056 Aachen
Germany
Phone: +49 241 8093505
Fax: +49 241 8092163
E-mail: anselm.huewe@bfw.rwth-aachen.de

1 Introduction

Experimental research over the last decades has left little doubt that people do have social preferences, meaning that they are not only interested in their own wellbeing but also in the wellbeing of others. Researchers have made distinct progress in describing such behavior theoretically: The models of Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Dufwenberg and Kirchsteiger (2004), henceforth DK, and Falk and Fischbacher (2006) are prominent examples. However, such models face a severe limitation: They predict (some) experimental results to a very good extent, but are often not able to explain individual behavior across different games in a consistent way. While typically consistent play can, to some degree, be found if the incentive structures of the games are similar, recent experiments have shown that individual rewarding and punishing behaviors do not correlate at all (Blanco et al. 2011, henceforth BEN; Yamagishi 2012). However, a positive correlation is implied by assuming a given reciprocity parameter in reciprocal models: Gintis (2000) introduced the term "strong reciprocity" to describe non-selfish behavior, meaning that people who answer kind actions with kindness (rewarding behavior of this kind is also denoted as "positive reciprocity" in the following) are also suspected of answering unkindness with unkindness (i.e., punishing, also called "negative reciprocity"). Similarly, with the exception of Fehr and Schmidt (1999), the models mentioned above use a single parameter to capture social preferences, meaning that a preference for kindness is inevitably connected to a preference for unkindness. Fehr and Schmidt (1999) differentiate between advantageous and disadvantageous inequality, but suggest considering these parameters to be (perfectly) correlated (p. 822 and p. 864). While behavior may be heterogeneous in a pool of subjects, preferences are assumed to be stable for single persons at least over the short run, implying that individual behavior in one game would have predictive power for behavior in subsequent games.

This paper researches three questions that are connected to the idea of strong reciprocity. First, we show how punishing behavior in the ultimatum game, UG, henceforth, can be explained by assuming reciprocal preferences. We use the reciprocity model of DK, but show that major modifications are necessary to be able to describe punishments. While DK can explain games in which positive reciprocity is present (see the examples in their paper, Breuer and Hüwe, 2014a for the case of the public goods game, and Breuer and Hüwe 2014b for the case of the trust game, TG henceforth), we will show in the following that the original DK model is unable to predict punishing behavior in the UG. Given that DK are frequently cited to justify such behavior (see, among others, Bereby-Meyer and Fiks, 2013; Boarini et al., 2009; Falk et al., 2005; Falk et al., 2008; Fischbacher et al., 2013; Kamas and Preston, 2012), it is astonishing that the exact game-theoretic solutions of DK for those games have not been derived so far. Exceptions are Falk et al. (2003) and Leibbrandt and Pérez (2012), who derive precise predictions for the responder role in the UG, but avoid many of our modeling problems by not considering behavior in the proposer role.

Second, we provide additional evidence that beliefs of proposers with respect to the average behavior of responders are pessimistically biased. Given that beliefs are essential to understanding subject behavior, it is remarkable that so little attention has been paid to beliefs in UGs so far (see also Section 5 in this paper). Therefore, up to now, it has been unclear as to whether such a bias exists or not.

Third, we complement the experimental findings of BEN and Yamagishi (2012) by showing in a within-subject design that rewarding behavior of second-movers in the TG does not correlate with punishing behavior of second-movers in a convex ultimatum game, cUG henceforth (Andreoni et al., 2003). Additionally, we show that further insights may be gained by comparing these two decisions when they are modeled in a reciprocal way. We use the cUG because, in contrast to the standard UG (Güth et al., 1982), it does not force responders to make a binary choice between “accept” and “reject”. Instead, the offer can be shrunk over a continuous strategy space, allowing a precise testing of whether subjects who return *a lot* in the TG *more severely* punish low UG offers. In contrast, the standard UG can only measure whether subjects who return a lot in the TG are *more likely* to accept low UG offers, which is statistically less reliable. Furthermore, we use a graphical interface, which allows subjects to express their preferences in a very simple and distinct way.

In the cUG (as in the standard UG), a proposer *i* is matched with an anonymous responder *j* and can offer him an arbitrary portion of her endowment a_i, $0 \leq a_i \leq 1$. The modification of the cUG compared to the standard UG is that responders can “shrink the pie” to any extent, which we represent with the help of the factor m_j ($0 \leq m_j \leq 1$). Shrinking the pie means that both the offer to the responder as well as the portion that the proposer wants to keep for himself are reduced by the factor m_j. Thus, accepting (rejecting) the proposed division corresponds to $m_j = 1$ ($m_j = 0$) – the cUG entails the standard UG as a special case. In both variants, a rational selfish proposer will offer the smallest possible amount if she expects the responder to be selfish himself and not to shrink positive offers.

This paper is structured as follows: In Section 2, we show the shortcomings of DK with respect to negative reciprocity, and we propose modifications which allow the modeling of games with punishing possibilities. In addition, our hypotheses are derived. Section 3 introduces our experimental design. Section 4 presents the experimental results and proves our hypotheses. Section 5 discusses our findings and proposes avenues for future research. Section 6 concludes.

2 Theoretical Background

Why should one assume that reciprocity is the driving force behind behavior in the UG? We do so because other typically assumed social preferences, such as altruistic, welfare maximizing, or maxi-min preferences, fail to predict that most responders will reject low offers. More promising, behavior in UGs can be modeled by either assuming that subjects care about the distribution of the outcome of an interaction (see the outcome-based inequity models of Bolton and Ockenfels, 2000; Fehr and Schmidt, 1999), or that they care about intentions associated with an action (see intention-based models, such as DK). Furthermore, hybrids of these two approaches exist (for example, Falk and Fischbacher, 2006). Outcome-based models can explain UG behavior by assuming that offers which lead to unequal payments are rejected because zero but equal payments are preferred to positive but unequal ones. However, experimental results suggest that UG behavior would better be explained with the help of intentions/emotions than with a concern for distributional fairness (see BEN; Blount, 1995; Falk et al., 2003; Xiao and Houser, 2005). Interestingly, we will now show that the purely intention-based approach of DK fails to predict responder and proposer behavior. Therefore, we will make model modifications to reconcile the DK approach with the experimental results.

The reciprocal model of DK proposes that people want to answer kindness with kindness (and unkindness with unkindness, accordingly). Person *i*'s (un)kindness towards *j* at a specific node *h* of the game, called $\kappa_{ij}\left(a_i(h),\left(b_{ij}(h)\right)_{j\neq i}\right)$, is measured by the surplus (loss) of material payoffs that *i* expects *j* to have gained by the end of a game (given *i*'s belief about *j*'s strategy $b_{ij}(h)$). The surplus (loss) results if *i* departs from a certain reference strategy by choosing $a_i(h)$ from her strategy space. The belief of *i* about *j*'s kindness to herself is denoted as $\lambda_{iji}\left(b_{ij}(h),\right.$ $\left.\left(c_{ijk}(h)\right)_{k\neq j}\right)$ and may also depend on *i*'s belief about *j*'s belief about a third player *k*'s strategy, $c_{ijk}(h)$. DK assume that the utility function of person *i* consists of two terms, weighted with an exogenously given, non-negative reciprocity parameter Y_{ij}. Thereby, the first term π_i represents i's material payoff, and the second term reflects i's reciprocity utility:

$$U_i\left(a_i(h), \left(b_{ij}(h), \left(c_{ijk}(h)\right)_{k\neq j}\right)_{j\neq i}\right) = \pi_i\ \left(a_i(h),\ \left(b_{ij}(h)\right)_{j\neq i}\right) +$$

$$\textstyle\sum_{j\in \mathrm{N}\setminus\{i\}}\left(Y_{ij}\cdot \kappa_{ij}\left(a_i(h),\ \left(b_{ij}(h)\right)_{j\neq i}\right)\ \cdot\ \lambda_{iji}\left(b_{ij}(h),\ \left(c_{ijk}(h)\right)_{k\neq j}\right)\right). \qquad (1)$$

If λ_{iji} is positive, i can raise her utility by increasing κ_{ij} (if it is not too costly). In contrast, if λ_{iji} is negative, i will be unfriendly herself. Furthermore, i will dislike situations where she is friendly and j is unfriendly (and vice versa). For further details, we refer to DK themselves.

2.1 Reference Points, Efficiency, and Kindness

What do subjects perceive to be a kind or an unkind strategy in the UG? DK compute the reference point which distinguishes kindness from unkindness as "the average between the lowest and the highest material payoff of j that is compatible with i choosing an efficient strategy". A strategy is efficient if there exists no other strategy which assures a higher material payoff for some player and no lower material payoff for any player. In the UG, the only efficient strategy for the responder is that of not shrinking. Therefore, "accepting" is neither friendly nor unfriendly (called "neutral" in the following), and responders cannot be kind in the UG. In turn, it is unfriendly to shrink the pie. Accordingly, m_j = 1 is set to be the reference strategy for responders. For the proposer role, we first have to define efficiency for uncertain situations. We assume that for the efficiency determination, all possible payoffs irrespective of their probabilities are considered. In our case, all possible strategies are efficient for proposers, because selfish responders will accept each offer. Accordingly, converging to the minimal offer of a_i = 0 may result in the maximal possible payoff of 1, and converging to the maximal offer of $a_i = 1$ results in minimal payoffs for the proposer. Thus, $a_i = 0.5$ is the reference strategy for proposers. With respect to the reference points, we leave the DK theory unaltered.

DK measure kindness in absolute terms, meaning that they define kindness as the additional material payoff which is granted to the co-player: Kindness is “proportional to the size of [her] gift”. DK thereby differ from Rabin’s (1993) normalized definition of kindness. In the following, we will define kindness in the spirit of Rabin (1993) and divide the absolute (un)kindness-term by the maximal possible (un)kindness. We will discuss in Section 5 why such an approach is more suitable in the UG case.

2.2 Uncertainty About the Co-player’s Reciprocal Inclination

While the role of uncertainty for non-social decisions has extensively been researched, it has not – at least theoretically – prominently been considered for reciprocal interactions so far: DK and Falk and Fischbacher (2006) assume that the reciprocal inclinations of the co-players are common knowledge. In their papers, co-player-dependent equilibrium strategies are derived. Thereby, *i* has to consider that her kindness to *j* depends on *j*’s subsequent and initially unknown reaction: Assume that *i* considers choosing $a_i = 1$ in the UG and anticipates that *j* will react with an uncommon but existing behavior of rejecting so-called hyper-fair offers. In that case, *i*’s offer must be seen as maximally unkind. However, assume that *j* is the only responder in a large pool of subjects who accept hyper-fair offers. How kind does *i* now intend to be? Given that acceptance is most likely, $a_i = 1$ should be seen as maximally kind. We conclude that *j* should evaluate *i*’s kindness by taking into account *i*’s belief about a *typical* reaction of co-players (note that reciprocity models assume that people care about intentions, not about the outcome). Thus, *j* should evaluate *i*’s (intended) kindness independently of his actual reaction, but dependent on his belief about *i*’s belief about the expected responder reaction. If this were not the case, very reciprocal responders would reject hyper-fair offers to turn the offer into an unfriendly one (remember that “unkind”/”unkind” equilibria are preferred to “kind”/”neutral” ones), a reaction which, however, is not commonly observed among subjects typically participating in experimental research.

Under uncertainty, the second-mover's behavior is not necessarily identical to the (belief about the) first-mover's belief about the former's behavior. This adds a further degree of complexity to the model, because players do not only need to anticipate their co-players' behavior, but must also build beliefs (about beliefs…) about typical, normative behavior (for example, *i*'s belief about the average shrinkage of offers in the cUG will be denoted as m_{ik} in the following). However, this additional complexity creates a more powerful model, because it captures the fact that interactions can depend on biased information or different levels of information regarding normative behavior.

The assumption that the players' reciprocal inclination is unknown has the following additional implications: First, the reciprocity parameter Y_{ij} can be modeled as being independent of *j*, because subjects have no possibility to condition strategies on the reciprocity inclination of their co-players. Therefore, we will write Y_i instead of Y_{ij} in the following. Second, if *i* does not know *j*'s reciprocity parameter, she must build a belief about it. Therefore, we use Y_{ij} in the following to denote *i*'s belief about *j*'s reciprocity parameter, meaning that Y_{ij} is utilized with a different meaning than in DK. Third, we must specify *i*'s behavior in uncertain situations: We assume that *i* has a belief about the probability distribution of $\tilde{Y}_j$ in the subject pool, and that she will maximize her expected utility.

2.3 Reciprocal Utility and Equilibrium Justification

If proposers anticipate that responders will accept $a_i = 0.5$, a payoff of $0.5 \cdot 1$ to *j* determines *i*'s reference point of kindness towards *j*, and *i*'s intended (un)kindness towards *j* equals $\kappa_{ij} = a_i \cdot m_{ik}(a_i) - 0.5 \cdot 1$. Similarly, *i*'s belief about *j*'s intended (un)kindness towards her is $\lambda_{iji} = (1 - a_i) \cdot m_{ij}(a_i) - (1 - a_i) \cdot 1$. In the DK model (and similarly in Falk and Fischbacher

2006), i's reciprocal utility is assumed to be the product of i's (un)kindness towards j multiplied by j's (un)kindness towards i, see equation (1). As each subject can be kind (k), unkind (u), or neutral (n), six basic strategy pairs are possible. The order within these pairs is irrelevant, because the kindness terms are multiplied in the DK model. Thus, for example, $u/k = k/u$. The preference order of the strategy pairs is as follows:

$$k/k = u/u \succ k/n = n/n = u/n \succ u/k \tag{2}$$

In cooperation games, where players pay a cost so that co-players can receive higher payoffs, this preference order leads to correct predictions: If the co-player is (believed to be) unkind and does not cooperate, he is punished (because $u/u \succ n/u$, k/u). If the co-player is kind, kindness will be reciprocated ($k/k \succ n/k$, u/k). However, the UG reveals that the preference order in (2) does not always match the experimental findings: Given the order $u/u \succ n/n$, reciprocal proposers should prefer the u/u strategy pair with $a_i = 0/m_j = 0$ to the n/n one with $a_i = 0.5/m_j = 1$. However, only very few proposers offer small amounts in the UG: Second-movers may punish first-movers for unkind behavior, and first-movers may be kind in order to induce a kind reaction, but first-movers are not unkind in order to provoke an unkind reaction: This would be rather a "sadomasochistic" behavior. If such behavior were common, we would observe more small offers in the UG. Thus, we expect that unkind offers will not be made by reciprocal proposers, who typically form the majority of a subject pool, but by selfish proposers (see also Section 4.2).

In contrast to DK, we generally propose that subjects prefer to match their co-player's level of kindness:

$$k/k \succ n/k \succ u/k, \tag{3a}$$

$$n/n \succ k/n = u/n, \tag{3b}$$

$$u/u \succ n/u \succ k/u. \tag{3c}$$

The intuition of these preference orders is that the more *i*'s kindness differs from that of *j*, the less preferable her decision is. This is exactly what is found in Nicklisch and Wolff (2012), who call such behavior *gradual reciprocation* and who find this type of behavior to be by far the most common one. The following reciprocal utility function captures this notion:

$$U_i = U_{i,\pi}(\pi_i) - Y_i \cdot \left(\kappa_{ij} - \lambda_{iji}\right)^2, \tag{4}$$

and, with uncertainty, we write

$$\mathrm{E}(U_i) = \mathrm{E}\left(U_{i,\pi}(\pi_i)\right) - Y_i \cdot \mathrm{E}\left(\left(\kappa_{ij} - \lambda_{iji}\right)^2\right). \tag{5}$$

In contrast to DK, reciprocal utility is not added to utility from material payoff, $U_{i,\pi}$, by calculating the product of κ_{ij} and λ_{iji}. Instead, the (squared) difference between both kindness terms is subtracted. This assures that – holding λ_{iji} fixed – *i* prefers to reciprocate *j*'s believed level of kindness. Squaring the expression implies that reducing large kindness differences is more worthwhile than reducing small ones. Moreover, this avoids corner solutions in cases where utility from the payoff is set equal to the payoff, i.e. where risk neutrality is assumed. While risk neutrality may often be assumed, equations (4) and (5) do not rule out the possibility of modeling risk aversion or risk seeking behavior.

2.4 Responder Behavior in the cUG

According to our previous remarks, responders in the cUG are assumed to maximize the following utility function (if material utility is set equal to the material payoff):

$$U_j\left(m_j\right) = a_i \cdot m_j(a_i) - Y_j \cdot$$

$$\begin{cases} \left(\frac{(1-a_i)\cdot m_j(a_i) - (1-a_i)\cdot 1}{|(1-a_i)\cdot 0 - (1-a_i)\cdot 1|} - \frac{a_i \cdot m_{jik}(a_i) - 0.5 \cdot m_{jik}(0.5)}{\left|0 \cdot m_{jik}(0) - 0.5 \cdot m_{jik}(0.5)\right|}\right)^2, & \text{if } a_i \leq 0.5, \\ \left(\frac{(1-a_i)\cdot \mathrm{m_j}(a_i) - (1-a_i)\cdot 1}{|(1-a_i)\cdot 0 - (1-a_i)\cdot 1|} - \frac{a_i \cdot m_{jik}(a_i) - 0.5 \cdot m_{jik}(0.5)}{\left|a_{i,max} \cdot m_{jik}\left(a_{i,max}\right) - 0.5 \cdot m_{jik}(0.5)\right|}\right)^2, & \text{if } a_i > 0.5. \end{cases} \tag{6}$$

The material payoff to j is determined by i's offer, multiplied by j's shrinking rate. Additionally, if Y_j is positive, j considers a reciprocal utility component: The minuend in the squared expression shows j's kindness towards i: The numerator displays j's absolute unkindness towards i: j actually grants i a payment of $(1 - a_i) \cdot m_j(a_i)$, which must be compared to i's payoff if j plays the reference strategy of accepting the offer. Shrinking to zero determines j's maximal possible unkindness of $1 - a_i$. As mentioned before, we consider the actual unkindness relative to the absolute maximally possible unkindness. In the subtrahend, i's intended kindness towards the responders depends on their average reaction to i's offer. If i plays her reference strategy, j will believe that i believes that responders (denoted by k) will accept, because this is optimal both with respect to payoffs and with respect to reciprocal utility, $0.5 \cdot m_{jik}(0.5) = 0.5$. Offering zero is obviously maximally unkind. What is the kindest offer in the range $0.5 \leq a_i \leq 1$? First, assume that it is believed that hyper-fair offers will not be shrunk below $a_i \cdot m_{jik} > 0.5$. In that case, these offers are believed to be kind and will not be shrunk at all. In contrast, if subjects believe that hyper-fair offers will be shrunk below 0.5, these offers are believed to be unkind and will indeed be shrunk (see Appendix A that is available upon request): Reactions to hyper-fair offers are determined by self-fulfilling expectations (see also DK, p. 282, for a similar result in the case of the sequential prisoners' dilemma). The shrinking of hyper-fair offers can also be empirically observed in some societies (Henrich et al. 2001). Typically, however, societies are "stuck" in the alternative equilibrium: Hyper-fair offers are believed to be kind because it is believed that they will not be rejected because they are believed to be kind. In that case, $1 \cdot m_{jik}(1) = 1$ grants the maximal payoff to i. Equation (6) can then be simplified to

$$U_j(m_j) = a_i \cdot m_j - Y_j \cdot \left(m_j - \frac{a_i \cdot m_{jik}}{0.5}\right)^2. \tag{7}$$

Maximizing (7) over m_j yields (compare Appendix B)

$$m_j^{*} = \begin{cases} \min\left\{1;\, a_i \cdot \left(\frac{1}{2\cdot Y_j} + 2\cdot m_{jik}\right)\right\} & \text{if } Y_j > 0, \\ 1 & \text{if } Y_j = 0 \text{ and } a_i > 0. \end{cases} \tag{8}$$

If only one (representative) responder type is present in the subject pool, *j* can believe that *i* believes that the other responders will shrink like himself, $m_{jik} = m_j^{*}$. In that case,

$$\overline{m}_j^{*} = \begin{cases} \min\left\{1;\, \frac{0.5\cdot a_i}{\overline{Y}_j - 2\cdot a_i \cdot \overline{Y}_j}\right\} & \text{if } \overline{Y}_j > 0 \text{ and } a_i < 0.5, \\ 1 & \text{if } \overline{Y}_j > 0 \text{ and } a_i \geq 0.5, \\ 1 & \text{if } \overline{Y}_j = 0 \text{ and } a_i > 0. \end{cases} \tag{9}$$

<<< Insert Fig. 1 about here >>>

Equations (8) and (9) will be called response functions in the following. If reciprocal responders shrink unfair offers, shrunk offers are convex in a_i (see Fig. 1, where we have displayed two examples with $Y_j = 0.2$ and $Y_j = 2.0$). From a certain point on, each *j* accepts the offer, even if it is (slightly) below the equal split. For $a_i = 0$, each reciprocal responder will reject, and for $a_i \geq 0.5$, each responder will accept. Therefore, our model captures typically observed behavior (see Andreoni et al., 2003, and our findings in Section 4). Selfish responders will of course never shrink a pie that is larger than zero. Equations (8) and (9) can also be used to (correctly) predict behavior in the standard UG: Offers equal to or above 0.5 are always accepted, and "small" offers are rejected by reciprocal responders, with the definition of "small" depending on the responder's reciprocity inclination. To prove equation (8), we formulate Hypothesis 1:

Hypothesis 1 -responder behavior-. Responder behavior corresponds to that predicted by equation (8).

2.5 Proposer Behavior in the cUG

If proposers anticipate the responders' reactions, they are assumed to derive their utility as follows:

$$\mathrm{E}\big(U_i(a_i)\big) = \mathrm{E}\left(U_{i,\pi}\left((1-a_i)\ \cdot\ {m_{ij}}^{*}\right)\right)\ -\ Y_i \cdot \mathrm{E}\left(\left(\frac{a_i \cdot m_{ik}}{0.5} - {m_{ij}}^{*}\right)^{2}\right). \qquad (10)$$

i gains expected utility from material payoff and expected disutility from unequal kindness terms. The latter can be explained analogously to the kindness terms in (7). Some general comments can be made with respect to the maximum of (10): Proposers will never offer $a_i > 0.5$, because such hyper-fair offers are costly and cannot be reciprocated with kindness. Proposers who are *only* interested in reciprocity will offer the equal split, because such behavior grants the maximal reciprocal utility level of zero. As well, the zero-offer results in zero reciprocal utility if all responders are expected to reject, but selfish responders are indifferent between accepting and rejecting zero-offers, and may therefore accept. This will cause disutility for reciprocal proposers, making the equal split the preferred choice. Between a_i = 0 and a_i = 0.5, all offers cause negative reciprocal utility because they are unkind and are expected not to be shrunk to zero. Thus, reciprocal utility is only zero if both players are maximally unkind (a_i = ${m_j}^{*}$ = 0) to each other, or if they play their reference strategies (a_i = 0.5; ${m_j}^{*}$ = 1). Depending on the (perceived) distribution of the responders' reciprocity parameters, selfish risk neutral proposers will offer an unequal split, which maximizes their expected payoff. The more responders (are believed to) shrink, the fairer this split will be. Risk seeking proposers may also offer lower portions and may even prefer to offer nothing. As already mentioned, this can also be optimal from a reciprocal point of view. Risk averse proposers will offer right from this point, as the spread of the returns decreases if unfair offers are raised. Due to this argument, proposers will offer more (less), the more risk averse (risk seeking) they are (see Appendix C). Very risk averse proposers will prefer the equal split or a slightly lower offer, where that subject in the responder pool who has the highest reciprocity parameter starts to shrink. Furthermore, with randomly chosen combinations of Y_j, it can numerically be shown that right from the payoff-maximizing offer, decreasing an offer always leads to higher expected reciprocal disutility. Intuitively, lower offers only come with higher payoffs if most responders do not shrink. However, being matched with such responders, reciprocal disutility increases with decreasing offers. Accordingly, the more reciprocal a risk averse proposer is, the more she will offer. The Nash equilibrium of “offer the smallest positive unit”/“accept” will be realized if both the proposer and the responder are all selfish.

Finally, the more reciprocal a responder is (believed to be), the more he will (be believed to) shrink and proposers must compensate this with respect to payoffs by offering more.

Our model predictions are summarized in the following hypothesis:

> **Hypothesis 2 -proposer behavior-.** The more proposers expect responders to shrink, and the more risk averse proposers are and the more reciprocal proposers are, the more they will offer.

In this context, it is an interesting question as to what degree proposers are generally able to correctly anticipate responder behavior: If responders are believed to be more reciprocal than they truly are, offers should increase, and vice versa. Based on the general notion that people are typically overoptimistic (compare, for example, Breuer and Hüwe 2014a and 2014b), we predict that proposers will have favorable views of responders’ shrinking behavior.

> **Hypothesis 3 -overoptimism-.** Beliefs about expected payoffs in the proposer role are overoptimistically biased, meaning that subjects expect responders to shrink unfair offers less than it is in fact the case.

2.6 Reciprocal Consistency

Typically, social preference models are used to explain behavior in single games. However, standard economics theory assumes that preferences are given. Therefore, preferences which are found in one game, or more specifically, in one role of one game, should have predictive power for behavior in other games as well. Applying our model and assuming consistent play, subjects who are very unkind in the responder role ought to be *less* unkind in the proposer role: Responders are unkind depending on their reciprocity parameter, and subjects with high parameter values bear high disutility from unkind (accepted) offers in the proposer role. To test this implication, we propose:

Hypothesis 4 -consistency of preferences between roles-. Subjects who shrink offers more heavily in the responder role offer more money in the proposer role.

Furthermore, we will compare second mover behavior in the cUG with that in the TG. In cooperation games, such as the TG, the public goods game, or the gift-exchange game, "acting reciprocally" is equivalent to "making the co-player better off", meaning that reciprocal motives cannot easily be distinguished from altruistic, efficiency-maximizing, or maxi-min ones. Contrarily, in punishment games, such as the UG, both motives are clearly distinguishable, because reciprocal responders would shrink the pie, while altruistic, efficiency-maximizing, and maxi-min ones would not. Accordingly, if those second-movers who reward in the TG do *not* shrink in the cUG, the latter preferences are supported. In sharp contrast, if rewarding subjects also shrink, the idea of strong reciprocity is supported.

While we have already shown that shrinkage in the cUG can be explained by a subject's reciprocity inclination, we still have to prove that in the TG, subjects with higher reciprocity parameters return higher amounts to the sender. This behavior has already been predicted by the version of the DK model proposed in Breuer and Hüwe (2014b). However, as that version differs from the one presented in this paper, we will now show that such a behavior also follows from equation (5). According to (5), receivers' utility in standard trust games is as follows:

$$U_j(k_j) = U_{j,\pi}(3 \cdot s_i - k_j) - Y_j \cdot$$

$$\begin{cases} \left(\frac{\left(1 - s_i + k_j(s_i)\right) - (1 - s_i + s_i)}{|(1 - s_i + 3\cdot s_i) - (1 - s_i + s_i)|} - \frac{3\cdot s_i - k_{jik}(s_i) - 0}{|3\cdot 1 - k_{jik}(1) - 0|}\right)^2 & \text{if } s_i > 0 \text{ and } k_j \geq s_i, \\ \left(\frac{\left(1 - s_i + k_j(s_i)\right) - (1 - s_i + s_i)}{|(1 - s_i + 0\cdot s_i) - (1 - s_i + s_i)|} - \frac{3\cdot s_i - k_{jik}(s_i) - 0}{|3\cdot 1 - k_{jik}(1) - 0|}\right)^2 & \text{if } s_i > 0 \text{ and } k_j < s_i, \\ 0 & \text{if } s_i = 0, \end{cases} \tag{11}$$

with s_i being the fraction of the endowment which is sent to the receiver – where it is tripled – and k_j being the amount (measured as the portion of i's endowment) which j returns to i. In equation (11), it is assumed that the receiver's reference strategy is to return the sending such that the sender is again equipped with her initial endowment. The reference strategy for the sender is to send nothing. Thus, equation (11) is based on the model proposed in this paper, with the exception of the reference point determination: As specified in Breuer and Hüwe (2014b), the reference points defined above are more suitable for the case of the TG. If $U_{j,\pi}(\pi_i)$ is set equal to π_i, maximizing (11) with respect to k_j yields (compare Appendix D)

$$k_j^* = \begin{cases} \max\left\{0; s_i + 2\cdot s_i \cdot \left(\frac{3\cdot s_i - k_{jik}(s_i)}{3 - k_{jik}(1)} - \frac{s_i}{Y_j}\right)\right\} & \text{if } s_i > 0 \text{ and } k_j^* \geq s_i, \\ \max\left\{0; s_i + s_i \cdot \left(\frac{3\cdot s_i - k_{jik}(s_i)}{3 - k_{jik}(1)} - \frac{s_i}{2\cdot Y_j}\right)\right\} & \text{if } s_i > 0 \text{ and } k_j^* < s_i, \\ 0 & \text{if } Y_j = 0. \end{cases} \tag{12}$$

Accordingly, as we wanted to show, j returns more (above a threshold) if his reciprocal inclination is high.

While we now have shown that behavior in the UG and in the TG can be described with the help of our reciprocal model, we first of all want to assure comparability with BEN and Yamagishi (2012) and therefore formulate the following hypothesis independently from any modeling assumption. We simply ask whether deviations from selfish behavior in the second-mover roles correlate between the two games on the subject level:

> **Hypothesis 5 -consistency of preferences between games-.** Subjects who return more money in the receiver role of the trust game shrink their offers more heavily when they are in the responder role in the convex ultimatum game.

3 Experimental Design

To compare positive with negative reciprocity in a within-subject design, we let subjects play both, a trust game and a convex ultimatum game. We described the results of the TG in Breuer and Hüwe (2014b), and also refer to this paper for a detailed explanation of the experimental design (screenshots and instructions from the cUG-part of the experiment are available upon request). Before we started the first part – the TG –, we ran a lottery task to measure subjects' risk aversion. Subjects had to choose between ten different safe amounts and a risky lottery, where they could win any amount from between 0 to 10 currency units (CU, henceforth, CU 10 were worth EUR 3.33 or approximately USD 4.53), each with equal probability. For each choice, subjects had to indicate whether they preferred the lottery or the safe amount. The crossover points, where subjects switched from the safe amount to the lottery, determined their certainty equivalents. When the TG was being explained, subjects were informed that a second experiment would follow, but what kind of experiment was not disclosed. However, subjects were aware of the fact that both parts would be completely independent of each other. Except the fact that two different games were played, both experimental parts had an identical structure. Accordingly, both parts were computerized, using the software z-Tree (Fischbacher, 2007) and were conducted in the experimental lab for economic research at RWTH Aachen University. As in the TG, subjects had to play the cUG in the proposer role as well as in the responder role. Again, proposers were equipped with CU 10.

They had to decide about the offer twice: In the “social setting”, they could offer any amount to a randomly selected responder. Also, proposers had to make an offer to the computer in a “non-social setting”, which determined its own shrinking behavior by using the response function of one responder randomly selected from the pool. Accordingly, in the social setting, proposers decided about both payments to themselves and to their co-player, whereas in the non-social setting, payments to another subject did not need to be considered. As in the proceeding TG experiment, in two sessions (58 participants), the social decision had to be made first, followed by the non-social decision. In two more sessions (44 participants), the sequence was inverted in order to be able to test for sequence effects. Only one of the two proposer decisions was selected for payment at the end of the experiment to avoid hedging considerations. Subsequently, proposers were asked to estimate their expected profits (corresponding to the expected shrinking behavior), dependent on the offer. A graphical input mechanism was used for this task (see the Supplementary Material). Precise estimations were incentivized with up to EUR 2. At the end, subjects were told that they had to take part in the experiment again, this time in the responder role. At that stage, the strategy method of Selten (1967) was used, meaning that responders had to indicate their shrinking behavior for each possible offer. Again, we implemented the strategy method graphically (see the Experimental Instructions again), which enabled subjects to determine their answers conditional on a continuous proposer choice set with high accuracy: In a diagram, responders indicated their shrinking behavior for seven hypothetical offers (a_i = 0; 0.125; 0.25; 0.375; 0.5; 0.75; 1), and responses to offers between these data points were interpolated linearly (which subjects had been informed of). Information on decisions of the other players and draws of the computer from both parts were not given prior to the last stage, when subjects were informed about the course of play and about their final payments.

4 Experimental Results

4.1 Preliminary Remarks

<<< Insert Fig. 2 about here >>>

The basic results of the experiment are shown in Fig. 2: The draw-through line in this figure displays the expected payoff in the proposer role, and the expected shrinking behavior of responders, depending on a_i (again measured relative to the endowment). The vertical lines mark the standard deviation of individual shrinking behavior. The dashed line displays subjects' beliefs about the average of such behavior. Offers equal to or above $a_i = 0.5$ are almost never shrunk, leading to almost certain payoffs of about $1 – a_i$. For lower offers, responders shrink with increasing intensity, the lower the offer is. Subjects are aware of this behavior, but overestimate shrinking (see Section 4.3). The lower an offer is, the higher the standard deviation of the payoff is. Many responders accept small deviations from the equal split, leading to a payoff maximum at $a_i = 0.375$. 25 % of the proposers decided in favor of $a_i < 0.375$, but only 8 % chose $a_i \leq 0.2$. Most proposers (41 %) chose the equal split, and 6 % chose $a_i > 0.5$. In the analogue experiment of Andreoni et al. (2003), about one third of all proposers offered $a_i = 0.01$, more than 40 % chose $a_i \leq 0.2$ and about a third of all proposers offered the equal split or higher portions. Thus, substantially more extremely unfair offers were made in their experiment, which interestingly corresponds to less reciprocal responder behavior: The expected payoff for 0.01-offers is about 0.57, while it is 0.23 in our setting (to clarify: numbers in this paper denote a proportion relative to the initial endowment of CU 10).

An analysis of the accuracy of payoff predictions reveals that estimations are quite imprecise: The average absolute error for all six estimations is 0.16. If each subject had simply used their own response function to estimate average behavior, the average absolute estimation error would have been 0.17, which is only slightly worse. This comparison shows how difficult it is for subjects to anticipate responder behavior: Proposers do not make a risky, but rather an ambiguous decision. Accordingly, given that moving from $a_i = 0.5$ to $a_i = 0.375$ (1) increases profits only by 0.03 (actual value), or 0.005 (average estimation), (2) increases the spread of returns, and (3) is perceived to be unfriendly, the attractiveness of the equal split is hardly a surprise.

4.2 Hypotheses 1 and 2

We proposed that behavior in the cUG in the responder role (Hypothesis 1) as well as in the proposer role (Hypothesis 2) can be explained with the help of our reciprocal model. To test Hypothesis 1, we determine the reciprocal inclination of subjects, Y_j, by calculating that value which minimizes the sum of squared errors of model predictions at the data points $a_i = 0.125$, 0.25, 0.375, 0.5, 0.75, and 1. If $a_i = 0$ is offered, selfish responders are indifferent between all possible strategies. However, in our design, that data point also determines strategies for offers $0 < a_i \leq 0.125$, because these responses are interpolated. Therefore, selfish responders should not shrink offers of zero either. Nevertheless, we are unsure as to whether subjects considered this fact. Instead, most subjects who do not shrink any positive offer reject the zero-offer. We want to classify these subjects as being selfish and therefore omit $a_i = 0$ when calibrating the reciprocity parameter. To be able to apply equation (8), m_{jik} must be modeled. As we asked for subjects' beliefs regarding expected payoffs dependent on a_i, $m_{jik}(a_i)$ can directly be computed from that beliefs if one assumes that responders believe proposers to have the same belief as themselves. We cannot calculate $m_{jik}(a_i = 1)$, because we did not ask for that belief (i will earn nothing irrespective of m_{jik}). In that case, we assume $m_{jik}(a_i = 1) = 1$, which conforms to our model prediction and is close to the actual average value of $m_{jik}(a_i = 1) = 0.93$. However, our results will not significantly be altered if we omit $m_{jik}(a_i = 1)$ from our computations.

Finding 1: Using equation (8), the actual shrinking rate can be predicted with a median error of 7 % (root of the median of the average squared differences between actual and predicted m_j per data point of each subject). Some large errors cannot be avoided, as some responders shrink offers of $a_i \geq 0.5$. Furthermore, some subjects do not shrink in a convex, but in a concave form, leading to small prediction errors. Based on the low median prediction error of 7 %, we conclude that Hypothesis 1 can be confirmed.

We will differentiate between subjects with a reciprocity parameter of $Y_j > 0$ (reciprocal subjects), and those with $Y_j = 0$ (selfish subjects) in the following. In Fig. 1, we have displayed average actual shrinking behavior and average model predictions for the reciprocal responders. The graphs illustrate the predictive power on the aggregate level, and show that those rare cases with non-rejected zero offers by reciprocal responders and shrunk hyper-fair offers cannot be explained. Such behavior might be due to “confusion”. However, not shrinking to zero can be rationalized by assuming that kindness is not valued in a relative, but in an absolute manner, as is shown in Section 5.

Next, we will analyze behavior in the proposer role. Offers are assumed to depend on subjects’ beliefs of (expected) payoffs/shrinkage, on their degree of risk aversion, and on their reciprocal inclination. First of all, we test whether the offers are influenced by a sequence effect, as suspected in Section 3. According to Kolmogorov-Smirnov tests, neither the distribution of offers in the social treatment nor in the non-social treatment depends on the sequence (in both cases, $p > 0.9$). Accordingly, both treatments are merged for the following analyses. We test Hypothesis 2 with the help of OLS regressions, see Table 1. In Table 1, “believed profit max. offer” describes the offer where subjects expect the maximal payoff. “Avg. amount reduced” indicates the average of the three shrinking decisions for the offers $a_i = 0.125$, 0.25, and 0.375. “Lottery discount” describes the difference between the expected profit and the certainty equivalent (transferred into a percentage of the expected profit) in the introductory lottery task.

$a_i - \hat{a}_i$ denotes the difference between a subject's offer in the social and in the non-social treatment.

<<< Insert Table 1 about here >>>

Finding 2: According to regressions (1) to (4) in Table 1, subjects' beliefs, their risk aversion, and their reciprocity inclination have the expected effect on the proposer offer. However, the explanatory power is very low, and the parameters are small and only weakly significant or not significant at all. Thus, no significant support for Hypothesis 2 is found.

We will analyze reciprocal influences on the offer separately in Section 4.4. The very small influence of subjects' beliefs and risk aversion is in contrast to the findings of Breuer and Hüwe (2014b), where beliefs (and risk aversion) of the same subjects can explain sendings in the TG to a much higher degree. We explain this contradiction as follows: In the TG, payoffs depend on the sending in a u-shaped form: Low or high beliefs are often the decisive factor for either sending (almost) nothing or sending a lot. In contrast, in the cUG, the payoff function is concave: Deviating from the fair and (almost) certain outcome only results in small payoff gains, and only small deviations are profitable. Furthermore, subjects expect too high shrinking rates (see below), so that 30 % of them believe $a_i = 0.5$ to be the expected payoff maximizing offer. Moreover, we have already mentioned that beliefs are very imprecise: If subjects are aware of this, it is reasonable to offer $a_i = 0.5$, which almost half of the subject pool do. Accordingly, large influences of beliefs and risk aversion cannot be found.

4.3 Hypothesis 3

To test Hypothesis 3, we compare expected proposer profits – dependent on the offer – with the actual average profits. As Fig. 2 has already revealed, our hypothesis of overoptimism cannot be confirmed:

Finding 3: Estimations are significantly pessimistic. Thus, Hypothesis 3 must be rejected.

According to two-tailed *t*-tests, we find too low estimations for $a_i > 0$: *p*-values range from 0.000 to 0.025. At $a_i = 0$, estimations do not significantly differ from actual average payoffs ($p = 0.136$), so that we cannot confirm that those estimations are biased. Nevertheless, overall, expectations are too low: Whereas the average expected payoff over all data points is 0.41, the corresponding estimated value is 0.37, resulting in a pessimistic discount of 8.8 %.

Fig. 2 also reveals that beliefs about payoffs at $a_i = 0.5$ and $a_i = 0.75$ are (significantly) pessimistic as well. As we see no reason for suspecting that such offers are shrunk, we test whether part of the discovered bias is in fact not due to pessimism, but to non-serious, random-like inputs of some subjects. We therefore exclude 19 subjects who estimate $m_{ij} < 0.8$ at $a_i = 0.5$: The remaining subjects almost correctly predict payoffs at $a_i = 0.5$, and as well at 0.75. Still, a significant negative bias is found for these subjects at offers of $0 < a_i \leq 0.375$ ($0.016 \leq p \leq 0.029$), meaning that the interpretation of the bias as resulting from pessimism is robust. Interestingly, we now also find a negative (but insignificant) bias at $a_i = 0$.

As in the TG, a distinct false consensus effect can be observed, meaning that subjects believe others will behave like they themselves do (Ross et al., 1977): The correlation between a subject's own reduction decision and her expected shrinkage rate is, depending on a_i, between 0.65 (at $a_i = 0$) and 0.19 (at $a_i = 0.75$). As a consequence, selfish subjects have correct (for slightly unfair offers) or overoptimistic (for very low offers) beliefs, and expect the profit-maximizing offer to be $a_i = 0.125$ on average. Accordingly, many offers of selfish proposers are too low from a payoff maximizing point of view (33 % offer $a_i < 0.375$, compared to 21 % of reciprocal proposers).

<<< Insert Table 2 about here >>>

4.4 Hypotheses 4 and 5

Hypothesis 4 conjectures that in the UG, subjects who punish in the responder role are less unkind in the proposer role. Our results are as follows:

> **Finding 4:** Shrinking behavior in the responder role has only a very small effect on the offer (compare the weakly significant effect in regression (3), Table 1, or the correlation of our data displayed in the "UG offer / UG responder cell" in Table 2; furthermore, the effect is insignificant in regression (4), Table 1). In contrast, the more reciprocal subjects are as responders, the smaller the difference between the social and the non-social offer is, see regression (5), Table 1. Thus, contradictory evidence is found, and Hypothesis 4 cannot be confirmed.

Table 2 displays correlations between decisions in both roles in both games and compares them to the results of BEN and Yamagishi et al. (2012) (note that BEN use the sequential prisoners' dilemma instead of the TG, which however has a very similar incentive structure). While we will comment on most correlations in Table 2 later on, we now turn to the values displayed for the UG offer / UG responder correlation in order to explain Finding 4. The weak significance of the Pearson's correlation parameter of 0.17 (which corresponds to the weak significance of "avg. amount reduced" in regression (3)) is driven by two subjects who play the Nash equilibrium of "offering zero" as a proposer / "always accepting" in the responder role (and believe in a profit-maximizing offer of zero): If these two uncommon data values are excluded, the already low support for Hypothesis 4 from regression (3) vanishes: The coefficient of the explanatory variable becomes even lower and insignificant, and R^2 decreases as well.

Interestingly, by regressing the difference between the social and the non-social decision on the shrinking behavior (regression (5)), we find that these two variables are negatively correlated. This is astonishing: On average, the social offer is 0.06 higher than the non-social one, and one would expect this difference to be due to reciprocal proposers who differentiate between selfish and social offers. But rather, the opposite holds true. A comparison of subject types complements regression (5): The difference between the social and the non-social offer is higher for selfish subjects (0.10) than for reciprocal ones (0.04). We also find that many subjects (58 %) do not differ between either decision at all, which explains why R^2 in regression (5) is so low. We discuss this counterintuitive result in Section 5.

To test the consistency of behavior in the receiver role of the TG against the behavior in the responder role in the cUG, we regress the "average amount returned" in the TG on the "average amount reduced" in the cUG. Thus, similarly to the proceeding in the cUG, reciprocity in the TG is measured as the average over all data points of sending-dependent return decisions (see Breuer and Hüwe, 2014b). As already mentioned, according to the idea of strong reciprocity, subjects should shrink more in the cUG, the more they have returned in the TG.

> **Finding 5:** Amounts returned in the trust game and shrinkage in the convex ultimatum game are not correlated, see Table 2, ρ = 0.11 (*p* = 0.269). As well, there are 13 subjects in the pool who return nothing in the TG, and 36 subjects who do not shrink, but only 9 subjects who neither return nor shrink. Thus, Hypothesis 5 must be rejected.

Although the sign of the correlation between both games is – as expected – positive, the correlation is not significantly different from zero. We also point out that many "selfish" subjects are selfish only in one of the two games. Moreover, we test for gender differences between both games and do find a significant effect: In the TG, females return 0.56 on average, and males return 0.48 (difference significant with *p* = 0.087). In the cUG, females shrink by 0.19, and males shrink by 0.29 (*p* = 0.006).

5 Discussion

5.1 Findings 1 and 2

According to Finding 1, behavior in the responder role in the cUG can be well explained by assuming reciprocal preferences. Two systematic deviations from modeled behavior are observed: Some subjects shrink hyperfair offers, and some subjects do not completely reject zero-offers. While both deviations may be due to unconscious play, they may also reveal information about preferences. Some subjects might prefer to shrink hyper-fair offers because they believe that such offers are intended to be unkind (we did not elicit beliefs at a_i = 1, but beliefs of close-by data points are highly correlated, and high beliefs at a_i = 0.75 of these subjects thus contradict this interpretation). Alternatively, they might show an outcome orientation and prefer equal but zero payoffs to extremely (advantageously) unequal ones. Not rejecting zero-offers may infer altruistic preferences, but it could also be explained by an alternative definition of reciprocity (which would, however, be opposed to the behavior of the large majority of subjects): Note that we defined punishing and rewarding in relative terms (see Section 2.4), meaning that the unkindest offer is associated with an unkindness of 100 % and is punished as harshly as possible. Alternatively, if one uses an absolute definition as in DK, offering nothing corresponds to an absolute unkindness of 0.5, which would – according to the concept of gradual reciprocation – be punished by shrinking *i*'s payoff only by an absolute value of 0.5 as well.

Finding 2 cannot confirm the assumption that the parameters "risk aversion", "expected payoffs", and "reciprocity" influence the proposer's decision. Missing support in the proposer role may be due to the following reasons: We find that 41 % of all proposers offered the equal split, meaning that this strategy was so attractive that it hid much of the variance we were looking for. We also pointed out that the offering decision was a very ambiguous one, making data which are generated by asking for well-defined beliefs and risk aversion parameters unreliable. Nevertheless, on an aggregated level, our model predicts that – with some exceptions (more precisely: risk seeking and overoptimistic) – proposers would offer the equal split or slightly unfair splits, and that is exactly what can be observed.

5.2 Finding 3

Our third finding is that proposers' beliefs are pessimistically biased. Such a bias leads to fairer offers, because proposers fear more punishment than will actually be the case. While this result is seemingly opposed to findings of overoptimism in public goods games (Breuer and Hüwe, 2014a) and trust games (Breuer and Hüwe, 2014b), they can all be unified by arguing that subjects overestimate others' reciprocal inclination instead of their own payoffs. The reason for this bias in all games may be the false consensus effect: Most subjects are reciprocal, and they "forget" that some selfish subjects exist. When is this plausible? Johnson and Fowler (2011) argue that people are overoptimistic because this bias helps to claim contested resources. Especially in the-winner-takes-it-all situations, overoptimism is advantageous. In contrast, Orbel and Dawes (1991) propose that cooperators overestimate the willingness of others to cooperate because this can be evolutionary advantageous. Thus, overoptimism might have developed to foster social interactions. If this argument holds true for punishing situations as well, punishment games allow discrimination between these two ideas: While "the-winner-takes-it-all" argument implies that own payoffs are overestimated, social-interaction-arguments imply that punishments should be overestimated. The second view is supported by our results, and it leads to the conclusion that – given a desired level of equality – biased beliefs prevent welfare-destroying punishments. Interestingly, although myriads of papers on UGs have been published, only very few of them investigate the accurateness of beliefs, and the few papers that we are aware of report contradictory results: In the UG of Suleiman (1996), beliefs are correct in their “high delta conditions”, and pessimistic in the “low delta conditions”. Bellemare et al. (2008) find that beliefs largely depend on framing: Expected acceptance probabilities of offers are higher if one asks for the portion of responders who reject rather than for the proportion of responders who accept. We avoid this framing by preliminary asking how much one expects to earn (in the instructions, we also mentioned that expected earnings correspond to expected shrinking rates), but we cannot rule out the possibility that our design is a form of framing as well. Although Bellemare et al. (2008) do not comment on the accuracy of beliefs in their subject pool, they do report that their modeling with subjective beliefs leads to the prediction of fairer offers than modeling with correct expectations, which we take as evidence for pessimistic beliefs. In the setting of Offerman (2002), first-movers estimate the second-movers’ reaction correctly if the first-movers’ “choice” is determined by a lottery. In the setting where first-movers actually decide about their choice, they are pessimistic with respect to the second-movers’ rewarding reactions (a reaction which is not possible in the standard UG design, where the proposed division can only be accepted, not rewarded), and optimistic with respect to the punishing reactions. Perez and Kiss (2012) report that people are not systematically biased in their expectations regarding the sanctioning behavior of others. In a dictator game analyzed by Fehr and Fischbacher (2004), a third party can punish dictators who only send small amounts to the recipient, and expectations of recipients about the extent of punishing behavior among third parties are only insignificantly too high.

5.3 Findings 4 and 5

According to findings 4 and 5, neither consistent play between the roles in the cUG, nor between the second-mover roles in the TG and cUG can be found. The idea of strong reciprocity must be rejected. As we do, BEN suspect that a correlation between proposer and responder behavior in the UG (see Table 2) is due to the false consensus effect, and our regression (1) in

Table 1 does show that belief-based effects may play a role. As BEN do not measure beliefs or social vs. non-social decisions, we cannot prove that the correlation found in their experiment can completely be explained by the false consensus effect. In our design, there is evidence which even seems to support a negative relationship between roles: Comparing differences in social vs. non-social offers between both games, we find the difference to be negatively instead of positively correlated with the degree of reciprocity in the responder role. Accordingly, at least some subjects might rather be viewed as being motivated by altruism than by reciprocity: For altruists, it is consistent to offer more in the social treatment than in the non-social one, and not to shrink offers as a responder.

As in BEN and in Yamagishi et al. (2012), we find no significant correlation between second mover decisions in the TG and in the UG (see Table 2 again). Less clear evidence is reported by Kamas and Preston (2012), who classify subjects into different categories (self-interested, inequity averse, efficiency maximizing, social surplus maximizing) with the help of dictator allocation questions and find some degree of consistency between a TG and a UG. However, also in their experiment, only a minority of selfish subjects, efficiency maximizers, and social surplus maximizers accept low offers in the UG. Accordingly, their preferences are described by the categorization only to a limited extent. We also report that no correlation between positive and negative reciprocity related questions is found in a large survey (Socio-Economic Panel) conducted by the German Institute for Economic Research (Egloff et al. 2013). Surprisingly, although we can confirm these literature results, we simultaneously find a positive and significant correlation between the reciprocity parameters elicited from the second mover roles in both games with the help of our reciprocal model: *Reciprocity parameters* computed using equations (8) and (12) are significantly correlated with $\rho = 0.21$ (Spearman, $p = 0.040$), while average *decisions* are not (the Pearson coefficient of 0.11 reported in Table 2 corresponds to a Spearman rank coefficient of $\rho = 0.13$, $p = 0.196$). Apparently, a subject ranking by their reciprocity parameters can differ from a ranking by their average decisions: Neither can shrinking in the UG, respectively amounts returned in the TG, be uniquely determined by knowing Y_j (because $m_{jik}(a_i)$, respectively $k_{jik}(s_i)$, matters), nor is the reverse possible (because $m_{jik}(a_i)$ and a_i, respectively $k_{jik}(s_i)$ and s_i, matter). If $k_{jik}(s_i)$ is biased due to the false consensus effect, subjects with a high reciprocity parameter will return less in the TG than without a bias (the bias implies that sendings are perceived to be less friendly than is actually the case, see equation (12). Similarly, highly reciprocal responders in the cUG should shrink even more due to the false consensus effect (believing in high shrinking rates leads to higher perceived unkindness of low offers, see equation (8). As a consequence, even if subjects showed a constant reciprocal inclination over both games, no perfect positive correlation between the average amount reduced and the average amount returned would be predicted: The maximal possible correlation in our data set is 0.89 (Spearman rank coefficient) instead of 1, which would result if beliefs did not play a role. It is also of interest to investigate to which extent beliefs can influence the correlation at all. Decisions depend on beliefs (see equation (8)), but their influence is limited: m_{jik} is restricted to $0 \leq m_{jik} \leq 1$, meaning that – holding Y_j fix – not every m_j^* can be reached by just adapting m_{jik}. For example, free-riders will not shrink, irrespective of their beliefs. As before, we use the reciprocity parameters derived in the IG to compute shrinking rates in the cUG, but now assume such beliefs (deviating from the true ones) that minimize the correlation between the average amount returned and the average amount reduced: The correlation can be lowered to 0.32. Thus, beliefs can indeed distinctly affect decisions, but at least in our setting they cannot be the cause for a correlation of zero between two decisions.

Finding 5 is remarkable because it implies that individual behavior cannot be predicted even if one has observed the individual’s reciprocal inclination in a preceding, different situation. This opens space for future research: The question arises what individual reciprocal preferences look like if both negative and positive reciprocity are found on the aggregate level, but strong reciprocity is found on the individual level only to a very limited extent.

On a scale, possible explanations for this finding can be sorted between the following two extremes: On the one hand, one may believe that it is in principle possible to find stable individual preferences, which can precisely predict behavior in different social interactions, but that research was not yet successful. In that case, future research should aim at identifying, describing, and modeling the true motives. In addition, more research would be needed to clarify whether different motives are pursued simultaneously (by weighting them), or whether they are processed subsequently (see Fischbacher et al., 2013, who aim in that direction). For example, in the case of the responder's role in the UG, altruism and reciprocity contravene, and it is unclear how these two motives interact.

On the other hand, one may believe that subjects do not have stable preferences, or, more concretely, that they prefer to randomize between reciprocal and selfish behavior. Being unsure about which strategy should be preferred, selfish behavior may be seen as a temptation which subjects often resist but sometimes succumb to. In this context, we mention the finding of Rand et al. (2012): Subjects who reach their decision more quickly (or are forced to do so) are more cooperative, meaning that people are predisposed towards cooperation. It would be interesting to know whether people are also predisposed to punish, and only sometimes deviate from that predisposition. Especially if a second game follows a first game in direct succession, subjects may feel obliged to reciprocate in the second game if they were selfish in the first one. In turn, reciprocating in the first game may be seen as "having done their duty", therefore being free to maximize income in the second game.

Our experiment cannot clarify why subjects prefer to "sometimes free ride and sometimes distinctly reciprocate" rather than to "always reciprocate a bit". Note that almost all subjects prefer to either return nothing or to return substantial amounts in the TG, and to either accept or to heavily shrink small offers in the cUG. As well, it is an open question as to the circumstances in which such randomizing behavior is evolutionary advantageous in environments where social image and repeated, non-anonymous interactions play a role (we are only aware of Szolnoki und Perc, 2013, who find that being a strong reciprocal type is advantageous only in very narrow and unrealistic parameter regions). However, it may be worth exploring the following approach in more detail: Assume that a preference to randomize has not developed in anonymous interactions, but in situations where the reciprocal inclination of the co-player is known (because the interaction is a repeated one or because subjects have reputation). In that case, randomizing responder behavior creates uncertainty for the proposer. If proposers are risk-averse, such responder behavior will on average lead to fairer offers and result in less shrinking than with stable preferences. Again, similarly to our explanation of the pessimistic bias, randomized reactions foster social interactions. Obviously, such behavior does not emerge from the preference structure proposed in this paper: Equation (8) uniquely defines a responder's optimal behavior. Claiming to randomize in order to induce higher offers is a non-credible threat. However, assuming that a responder prefers to randomize, his uncertain reaction will increase an offer on average, compared to a situation where his reaction can be foreseen. Very risk averse (selfish) proposers, will then even offer the equal split. Nevertheless, data from our subject pool only support a very small effect: In our lottery task, the average certainty equivalent is only 2.8 % below the expected payoff, corresponding to a risk coefficient of r_i = 0.016 (we assume a utility function of the form $U_i(\pi_i) = \pi_i^{1-r_i}$ at this point). Furthermore, we assume that each responder randomizes such that his behavior exactly mirrors that of responders in our pool. With r_i = 0.016, a selfish proposer will offer $a_i^* = 0.36$. However, a risk-neutral selfish proposer, whose utility is not affected by randomized reactions, offers (almost) the same amount. This means that an average proposer almost does not take risk considerations into account. With (according to Holt and Laury 2002) a "very high" risk aversion parameter of r_i = 0.9, the optimal selfish offer increases sligthly to $a_i^* = 0.38$. As reciprocal subjects are assumed to be risk-averse with respect to reciprocal utility, uncertainty regarding the kindness of responders affects the reciprocal utility component similarly to the payoff considerations presented so far: With $Y_j = 0.05$, the optimal offer for a risk-neutral proposer is $a_i^* = 0.42$ if uncertain responder behavior is considered. In contrast, assuming that the reaction of the responder is certain and that it results in payoffs corresponding to the average payoff regarded previously, a_i^* decreases to 0.39.

Above, we presented two possible experimental outcomes: Either, one may find that individual preferences are stable, or one may find the contrary. These two possibilities can also be interpreted as not contradicting each other, as assuming unstable preferences can simply mean that we model behavior as being random as long as the underlying motives have not been completely understood. Similarly, BEN propose that the Fehr and Schmidt (1999) model is an "as if" model, which is "qualitatively able to capture different important motives in different games but that the low predictive power of the model at the individual level is driven by the low correlation of these motives within subjects" (p. 333). Accordingly, the question is that of how sensitive subjects react to certain triggers and how difficult it is to discover the underlying motives. Having social preferences which are easy influenced may be advantageous, because it allows subjects to rapidly adapt to changes in the social structure of their society.

Third, one might assume that subjects have stable and well defined reciprocal preferences, but differentiate between positive and negative reciprocity: Some subjects (according to our data: especially male subjects) punish, while others (especially female) reward. Our gender-dependent result is supported by Burnham (2007) who finds that men who reject low UG-offers have higher testosterone levels than those who accept. As well, Eckel and Grossman (2001) find that female responders are more likely to accept an offer than male ones (contrary evidence is reported by Garcia-Gallego et al. 2012, who find that women reject more than men). Thus, it is possible to differentiate between positive and negative reciprocity in a type-specific way.

Also, the correlations reported in Table 2 support this view: If situations are compared where the reciprocal decision is kinder than the selfish decision (TG offer, TG responder, UG offer) significantly positive correlations are found (with the exception of the TG offer / UG offer comparison in BEN). In contrast, if the reciprocal decision is unkinder than the selfish decision (UG responder), no correlation with the other decisions is found (with the exception of the UG offer / UG responder, which we – see above – attribute to the false consensus effect). Such an explanation also conforms to the literature finding that social preferences are consistent between comparable games: Consistency is high if subjects repeatedly play the same game (Andreoni and Miller, 2002; Volk et al., 2012), and it is still substantial across similar games (Dariel and Nikiforakis, 2014; Yamagishi et al., 2013).

6 Conclusion

In this paper we showed how outcomes of the (convex) ultimatum game can be explained with the help of a reciprocal theory based on the model of Dufwenberg and Kirchsteiger (2004) (DK). The UG gives some interesting insights into reciprocal behavior: As proposers in the UG can choose between a neutral / neutral outcome with respect to proposer / responder kindness, or can provoke an unkind / unkind outcome, experimental evidence shows that the latter outcome is not preferred to the former one. In contrast to DK, we suggest modeling reciprocity as being gradual, meaning that not only the sign of the kindness term should be reciprocated, but the magnitude as well. Furthermore, we find that a model which assumes the co-player's reciprocal inclination to be known cannot simply be applied to the typical laboratory situation, where co-players are anonymous. Instead, subjects are assumed to determine the others' kindness by anticipating the belief about typical behavior of co-players. This insight allows the derivation of self-fulfilling equilibria in UGs. To some degree, it also explains why no correlation between decisions in different games is found, although behavior may to some extent be reciprocally consistent: (Un)kindness is not necessarily (perfectly) connected to a subject's reciprocal inclination, because one's own (un)kindness depends on the perceived, individually biased belief about the (un)kindness of others.

Our experimental design reveals that a combination of four aspects explains why the equal split is the typical choice in UGs. Firstly, by offering less, payoffs can only be increased by the small amount of 0.3 currency units (the initial endowment of 10 currency units was worth EUR 3.33 or approximately USD 4.53). Secondly, such offers are risky, while the equal split almost certainly grants a payoff of 5 currency units. Thirdly, the equal split is the offer which maximizes reciprocal utility. Fourthly, while actual gains from unfair offers are already small, believed gains are even smaller: In contrast to our hypothesis, subjects overestimate the second-movers' reciprocal reactions instead of the payoffs granted to the proposers. This finding implies that fair offers are partly made because of an exaggerated fear that unfair offers are punished. Similarly to the findings in Breuer and Hüwe (2014a) and (2014b), beliefs and reciprocal preferences are found to interact. Believing in more reciprocal co-players than is actually the case, less own reciprocity is necessary to induce a socially desired outcome.

References

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

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

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

Bereby-Meyer, Y., and Fiks, S. (2013). Changes in Negative Reciprocity as a Function of Age. *Journal of Behavioral Decision Making*, 26, 397-403.

Boarini, R., Laslier, J. F., and Robin, S. (2009). Interpersonal Comparisons of Utility in Bargaining: Evidence From a Transcontinental Ultimatum Game. *Theory and Decision*, 67, 341-373.

Breuer, W., and Hüwe, A. (2014a). Explaining Individual Contributions in Public Goods Games Using (only) Reciprocity and Overoptimism. Working Paper.

Breuer, W., and Hüwe, A. (2014b). Trust, Reciprocity, and Betrayal Aversion: Theoretical and Experimental Insights. Working Paper.

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

Blount, S. (1995). When Social Outcomes Aren't Fair: The effect of causal Attributions on Preferences. *Organizational Behavior and Human Decision Processes*, 63, 131-144.

Bolton, G. E., and Ockenfels, A. (2000). ERC – A Theory of Equity, Reciprocity, and Competition. *The American Economic Review*, 90, 166-193.

Burnham, T. C. (2007). High-Testosterone Men Reject Low Ultimatum Games Offers. *Proceedings of the Royal Society: Biological Science*, 274, 2327-2330.

Dariel, A., and Nikiforakis, N. (2014). Cooperators and Reciprocators: A Within-Subject Analysis of Pro-Social Behavior. *Economic Letters*, 122, 163-166.

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

Eckel, C., and Grossman, P. (2001). Chivalry and Solidarity in Ultimatum Games. *Economic Inquiry*, 39, 171–188.

Egloff, B., Richter, D., and Schmukle, S. C. (2013). Need for Conclusive Evidence That Positive and Negative Reciprocity Are Unrelated. *Proceedings of the National Academy of Sciences*, 110, E786.

Falk, A., Fehr, E., and Fischbacher, U. (2003). On the Nature of Fair Behavior. *Economic Inquiry*, 41, 20-26.

Falk, A., Fehr, E., and Fischbacher, U. (2005). Driving Forces Behind Informal Sanctions. *Econometrica*, 73, 2017-2030.

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

Falk, A., and Fischbacher, U. (2006). A Theory of Reciprocity. *Games and Economic Behavior*, 54, 239-315.

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

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

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

Fischbacher, U., and Hertwig, B. (2013). How to Model Heterogeneity in Costly Punishment: Insights From Responders' Response Times. *Journal of Behavioral Decision Making,* 26, 462-476.

García-Gallego, A., Georgantzísa, N., and Jaramillo-Gutiérreze, A. (2012). Gender Differences in Ultimatum Games: Despite Rather Than Due to Risk Attitudes. *Journal of Economic Behavior and Organization*, 83, 42-49.

Gintis, H. (2000). Strong Reciprocity and Human Sociality. *Journal of Theoretical Biology*, 206, 169-179.

Güth, W., Schmittberger, R., and Schwarze, B. (1982). An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior and Organization,* 3, 367-388.

Henrich, J., Boyd, R., Bowles, S., Camerer, C., Fehr, E., Gintis, H., and McElreath, R. (2001). In Search of Homo Economicus: Behavior Experiments in 15 Small-Scale Societies. *American Economic Review*, 91, 73-78.

Holt, C. A., and Laury, S. (2002). Risk Aversion and Incentive Effects. *American Economic Review*, 92, 1644-1655.

Johnson, D. D. P., and Fowler, J. H. (2011). The Evolution of Overconfidence. *Nature*, 477, 317-320.

Kamas, L., and Preston, A. (2012). Distributive and Reciprocal Fairness – What Can We Learn From the Heterogeneity of Social Preferences. *Journal of Economic Psychology*, 33, 538-553.

Leibbrandt, A., and Pérez, R. L. (2012). An Exploration of Third and Second Party Punishment in Ten Simple Games. *Journal of Economic Behavior and Organization*, 84, 753-766.

Nicklisch, A., and Wolff, I. (2012). On the Nature of Reciprocity: Evidence From the Ultimatum Reciprocity Measure. *Journal of Economic Behavior and Organization*, 84, 892-905.

Offerman, T. (2002). Hurting Hurts More Than Helping Helps. *European Economic Review*, 46, 1423-1437.

Orbel, J., and Dawes, R. M. (1991). A "Cognitive Miser" Theory of Cooperators' Advantage. *The American Political Science Review,* 85, 515-528.

Pérez, R. L., and Kiss, H. J. (2012). Do People Accurately Anticipate Sanctions? *Southern Economic Journal*, 79, 300-321.

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

Rand, D. G., Greene, J. D., and Nowak, A. (2012). Spontaneous Giving and Calculated Greed. *Nature*, 489, 427-430.

Ross, L., Greene, D., and House, P. (1977). An Egocentric Bias in Social Perception and Attribution Processes. *Journal of Experimental Social Psychology*, 13, 279-301.

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, J.C.B. Mohr (Paul Siebeck).

Suleiman, R. (1996). Expectations and Fairness in a Modified Ultimatum Game. *Journal of Economic Psychology*, 17, 531-554.

Szolnoki, A., and Perc, M. (2013). Correlation of Positive and Negative Reciprocity Fails to Confer an Evolutionary Advantage: Phase Transitions to Elementary Strategies. *Physical Review X*, 1-11.

Volk, S., Thöni, C., and Ruigrok, W. (2012). Temporal Stability and Psychological Foundations of Cooperation Preferences. *Journal of Economic Behavior and Organization*, 81, 664-676.

Xiao, E., and Houser, D. (2005). Emotion Expression in Human Punishment Behavior. *Proceedings of the National Academy of Sciences*, 102, 7398-7401.

Yamagishi, T., Horita, Y., Mifune, N., Hashimoto, H., Li, Y., Shinada, M., Miura, A., Inukai, K., Takagishi, H., and Simunovic, D. (2012). Rejection of Unfair Offers in the Ultimatum Game is No Evidence of Strong Reciprocity. *Proceedings of the National Academy of Science*, 109, 20364-20368.

Yamagishi, T., Mifune, N., Li, Y., Shinada, M., Hashimoto, H., Horita, Y., Miura, A., et al. (2013). Is behavioral Pro-Sociality Game-Specific? Pro-Social Preference and Expectations of Pro-Sociality. *Organizational Behavior and Human Decision Processes*, 120, 260-271.

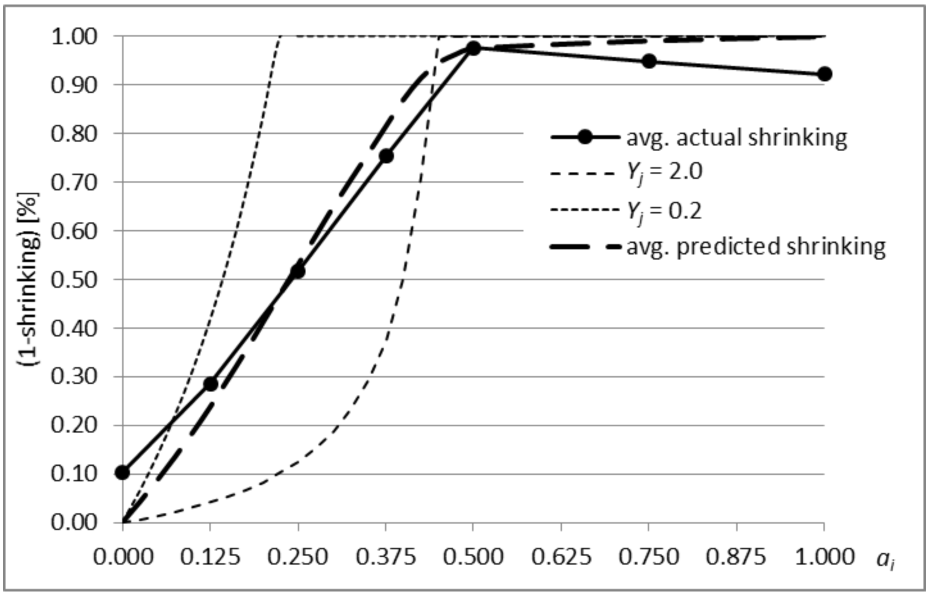


Fig. 1: Shrinking of offers with $Y_j = 0.2$ and $Y_j = 2.0$, average model predictions of shrinking behavior of reciprocal responders, and actual shrinking behavior of reciprocal responders.

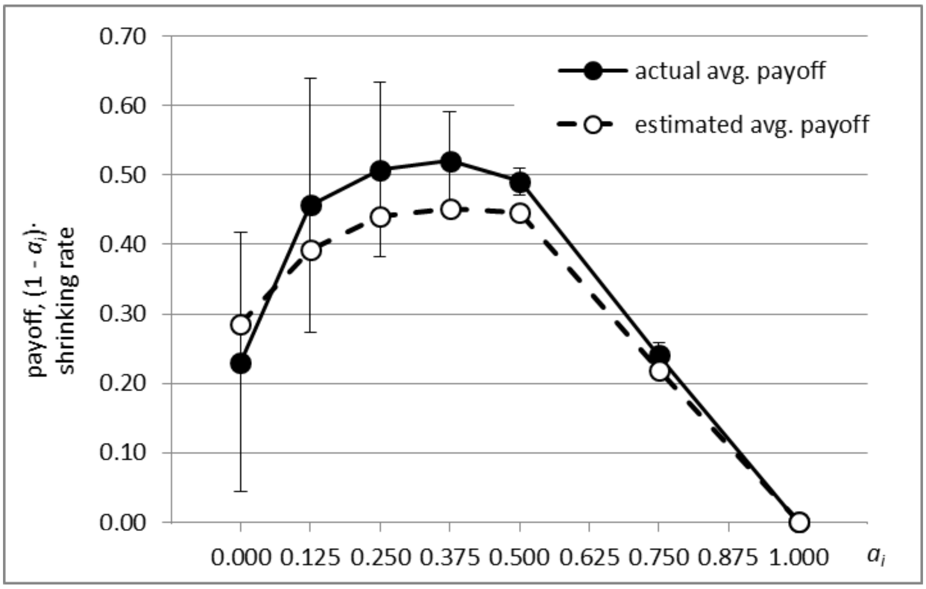


Fig. 2: Estimated and actual average profit/shrinking, depending on the offer. The vertical bars indicate the standard deviation of actual profits/shrinking behavior.

Table 1 - Offer determinants[1)]

Dependent variable			a_i		$a_i - \hat{a}_i$
	(1)	(2)	(3)	(4)	(5)
Believed profit max. offer	0.11*			0.08	
	(0.06)			(0.05)	
Lottery discount		0.07		0.08	
		(0.09)		(0.09)	
Avg. amount reduced			0.09*	0.07	-0.15***
			(0.05)	(0.05)	(0.06)
Constant	0.38***	0.41***	0.39***	0.37***	0.10***
	(0.02)	(0.01)	(0.02)	(0.03)	(0.02)
Observations	102	102	102	102	102
R²	0.04	0.02	0.03	0.07	0.05

Significant at the 1 percent (***), 5 percent (**), 10 percent (*) level.

[1)] OLS regressions with robust standard errors in parentheses.

a_i: Proposer's offer to the responder in the social treatment

Avg. belief: Expected return (average of data points at a_i = 0.125, 0.25, and 0.375)

Avg. amount reduced: Expected return (average of data points at a_i = 0.125, 0.25, and 0.375)

Table 2 - Correlations between decisions in both roles of both games[1]

	TG offer	TG responder	UG offer
UG offer	0.13 (BEN[2),3)]) 0.22**	0.49*** (BEN[3),4)]) 0.17*	
UG responder	-0.03 (BEN[2),3)]) -0.17 (YHM[5)]) 0.02	0.19 (BEN[3),4)]) -0.02 (YHM) 0.11	0.40*** (BEN[2)]) 0.17*
TG offer		0.43*** (BEN[3),6)]) 0.325***	

Significant at the 1 percent (***), 5 percent (**), 10 percent (*) level.

[1)] Pearson's ρ if not indicated otherwise. As in the rest of the paper, responder behavior in the UG (TG) is measured by the average amount reduced (returned).

[2)] BEN: Blanco et al. (2011); Spearman's ρ.

[3)] Correlations to first (second) mover decision in the sequential prisoners' dilemma instead of to the TG offer (responder).

[4)] Rank biserial correlation.

[5)] YHM: Yamagishi et al. (2012).

[6)] Phi coefficient.

Supplementary Material - For Referees' Information

Appendix

A: Optimal Shrinking if Hyper-Fair Offers Are Believed to Be Shrunk Below 0.5

For $a_i > 0.5$ and $m_{jik}(1) < 0.5$, j's utility function is

$$U_j\left(m_j\right) = a_i \cdot m_j - Y_j \cdot \left(\frac{(1-a_i)\cdot m_j - (1-a_i)}{1-a_i} - \frac{a_i \cdot m_{jik} - 0.5}{0.5 - m_{jik}(1)}\right)^2 \qquad \text{(A.1)}$$

Maximizing over m_j results in

$$U'_j\left(m_j\right) = a_i - \frac{2\cdot Y_j \cdot \left(\left(2\cdot m_{jik}(1) - 1\right)\cdot m_j - 2\cdot m_{jik}(1) + 2\cdot a_i \cdot m_{jik}\right)}{2\cdot m_{jik}(1) - 1} = 0$$

$$-a_i = \frac{-4\cdot a_i \cdot m_{jik} \cdot Y_j - 4\cdot m_{jik}(1)\cdot Y_j \cdot m_j + 4\cdot m_{jik}(1)\cdot Y_j + 2\cdot Y_j \cdot m_j}{2\cdot m_{jik}(1) - 1}$$

$$\Leftrightarrow -4\cdot m_{jik}(1)\cdot Y_j \cdot m_j + 2\cdot Y_j \cdot m_j$$

$$= 4\cdot a_i \cdot m_{jik} \cdot Y_j - 2\cdot a_i \cdot m_{jik}(1) - 4\cdot m_{jik}(1)\cdot Y_j + a_i$$

$$\Leftrightarrow m_j^{*} = \frac{a_i - 2\cdot a_i \cdot m_{jik}(1) - 4\cdot m_{jik}(1)\cdot Y_j + 4\cdot a_i \cdot m_{jik}\cdot Y_j}{2\cdot Y_j - 4\cdot m_{jik}(1)\cdot Y_j}. \qquad \text{(A.2)}$$

If only one responder type is present in the subject pool, j can believe that i believes that the other responders shrink like himself. If $a_i = 1$ is offered, $m_{jik} = m_{jik}(1) = m_j^{*}(1)$. In that case, (A.2) can be simplified to

$$m_j^{*}(1) = \frac{1 - 2\cdot m_j(1) - 4\cdot m_j^{*}(1)\cdot Y_j + 4\cdot m_j^{*}(1)\cdot Y_j}{2\cdot Y_j - 4\cdot m_j^{*}(1)\cdot Y_j}.$$

$$\Leftrightarrow m_j^{*}(1) = \frac{1}{2\cdot Y_j}. \qquad \text{(A.3)}$$

Inserting $\frac{1}{2 \cdot Y_j}$ for $m_{jik}(1)$ in (A.2) gives

$$\Leftrightarrow m_j^{\,*} = \frac{a_i + 2 \cdot Y_j - 4 \cdot a_i \cdot m_{jik} \cdot Y_j^{\,2} - a_i \cdot Y_j}{2 \cdot Y_j - 2 \cdot Y_j^{\,2}}, \tag{A.4}$$

which equals, if $m_{jik} = m_j^{\,*}$,

$$\Leftrightarrow m_j^{\,*} = \frac{a_i + 2 \cdot Y_j - a_i \cdot Y_j}{2 \cdot Y_j + 4 \cdot a_i \cdot Y_j^{\,2} - 2 \cdot Y_j^{\,2}}. \tag{A.4}$$

According to (A.4), $m_j^{\,*}$ is decreasing in a_i and in Y_j for $a_i \geq 0.5$ and converges to zero for large reciprocity parameters. Accordingly, for sufficiently high reciprocity parameters respectively offers, subjects shrink hyperfair offers.

B: Proof That Equations (8) and (9) Are Correct

Proof of Equation (8):

$$U_j(m_j) = a_i \cdot m_j - Y_j \cdot \left(m_j - \frac{a_i \cdot m_{jik}}{0.5}\right)^2$$

$$U_j'(m_j) = a_i - 2 \cdot Y_j \cdot \left(m_j - \frac{a_i \cdot m_{jik}}{0.5}\right) = 0$$

$$\Leftrightarrow a_i + 4 \cdot Y_j \cdot a_{\mathrm{i}} \cdot m_{jik} - 2 \cdot Y_j \cdot m_j = 0$$

$$\Leftrightarrow m_j = \frac{0.5 \cdot a_i + 2 \cdot Y_j \cdot a_i \cdot m_{jik}}{Y_j} = a_i \cdot \left(\frac{1}{2 \cdot Y_j} + 2 \cdot m_{jik}\right) \tag{B.1}$$

Proof of Equation (9):

$m_j^{\,*} = \frac{0.5 \cdot a_i + 2 \cdot a_i \cdot m_{jik} \cdot Y_j}{Y_j}$, see equation (8) in the paper.

If m_{jik} is replaced by $m_j^{\,*}$ the following holds:

$$\Rightarrow m_j{}^* \cdot Y_j \ - 2{\cdot}a_i \cdot m_j{}^* \cdot Y_j = 0.5{\cdot}a_i$$

$$\Leftrightarrow m_j{}^* = \frac{0.5{\cdot}a_i}{Y_j - 2{\cdot}a_i \cdot Y_j} \tag{B.2}$$

C: Proof That More Risk Averse Proposers Will Offer More Than Less Risk Averse Proposers

A proposer's uncertain payoff is defined as

$$\pi_i = (1 - a_i) \cdot \tilde{m}_j{}^*. \tag{C.1}$$

If the responder is believed to shrink, i's payoff can be computed using equation (8):

$$\tilde{\pi}_i = (1 - a_i) \cdot a_i \cdot \left(\frac{1}{2{\cdot}\tilde{Y}_j} + 2{\cdot}m_{jik}\right). \tag{C.2}$$

The derivative of (C.2) with respect to a_i is

$$\frac{\partial \tilde{\pi}_i}{\partial a_i} = (1 - 2 \cdot a_i) \cdot \left(\frac{1}{2{\cdot}\tilde{Y}_j} + 2{\cdot}\frac{\partial m_{jik}}{\partial a_i}\right), \tag{C.3}$$

which is positive for $a_i < 0.5$ if we assume that $\frac{\partial m_{jik}}{\partial a_i} > 0$. Thus, by lowering a_i, a proposer is always worse off if she is matched with one of the ι responders who shrink, meaning that the payoff decreases in all cases except the best possible ones (which are: being matched with one of the $J - \iota$ responders who do not shrink). We denote

$$U^{'}(\pi_i, j \leq \iota) = \frac{\partial U}{\partial \pi_i} U(\pi_i, j \leq \iota) = \frac{\frac{\partial U}{\partial a_i} U\left((1{-}a_i) \cdot m_j{}^*\right)}{(1{-}2{\cdot}a_i) \cdot \left(\frac{1}{2\cdot Y_j} + 2{\cdot}\frac{\partial m_{jik}}{\partial a_i}\right)} \tag{C.4}$$

and

$$U^{'}(\pi_i, j > \iota) = \frac{\partial U}{\partial \pi_i} U(\pi_i, j > \iota) = -\frac{\partial U}{\partial a_i}(U(1 - a_i) \cdot 1) \tag{C.5}$$

Thus, increasing a_i increases π_i if j shrinks, and decreases π_i if j accepts. If i chooses optimally (assume that it is optimal for i not to choose a corner solution), changing the offer must not result in higher utility. Thus, marginal expected utility from increased payoffs in the cases of "accepting" must equal absolute marginal expected disutility from reduced payoffs in the cases of "shrinking" (in the following, remind that it is equally probable for i to be matched with any of the j responders):

$$\sum_{j=1}^{j=\iota} U_{\text{small RA}}{}'\left(\pi_{i,\text{ small RA}^*, j}\right)$$

$$\overset{!}{=} \sum_{j=\iota+1}^{j=J} U_{\text{small RA}}{}'\left(\pi_{i,\text{ small RA}^*, J}\right) = (J-\iota) \cdot U_{\text{small RA}}{}'\left(\pi_{i,\text{ small RA}^*, J}\right)$$

$$\Leftrightarrow \frac{\sum_{j=1}^{j=\iota} U_{\text{small RA}}{}'\left(\pi_{i,\text{ small RA}^*, j}\right)}{(J-\iota) \cdot U_{\text{small RA}}{}'\left(\pi_{i,\text{ small RA}^*, J}\right)} = 1. \tag{C.6}$$

Thereby, we assume that i's utility function is differentiable at all $\pi_{i,\text{ small RA}^*}$. U_{s} (U_{l}) denotes the utility function of a risk averse proposer whose risk aversion is comparably small (large). Risk aversion (RA) is defined by $-\frac{U''(\pi_i)}{U'(\pi_i)}$, with $U'(\pi_i) > 0$ and $U''(\pi_i) < 0$. Using the definition of RA, we can state that the following is true:

$$-\frac{U_{\text{l}}''(\pi_i)}{U_{\text{l}}'(\pi_i)} > -\frac{U_{\text{s}}''(\pi_i)}{U_{\text{s}}'(\pi_i)} \; \forall \; \pi_i \tag{C.7}$$

$$\Leftrightarrow -U_{\text{l}}''(\pi_i) > -\frac{U_{\text{l}}'(\pi_i)}{U_{\text{s}}'(\pi_i)} \cdot U_{\text{s}}''(\pi_i) \tag{C.8}$$

With respect to the right part of (C.8), we can say that:

$$-\frac{U_{\text{l}}'(\pi_i)}{U_{\text{s}}'(\pi_i)} \cdot U_{\text{s}}''(\pi_i) > -\frac{U_{\text{l}}'(\pi_{i,J})}{U_{\text{s}}'(\pi_{i,J})} \cdot U_{\text{s}}''(\pi_i) \; \forall \; \pi_i < \pi_{i,J}, \text{ if } \left(\frac{U_{\text{l}}'(\pi_i)}{U_{\text{s}}'(\pi_i)}\right)' < 0. \tag{C.9}$$

The condition in (C.9) is fulfilled, as it can be concluded from (C.7):

$$-\frac{U_\mathrm{l}''(\pi_i)}{U_\mathrm{l}'(\pi_i)} > -\frac{U_\mathrm{s}''(\pi_i)}{U_\mathrm{s}'(\pi_i)}$$

$$\Leftrightarrow \frac{U_\mathrm{l}''(\pi_i)}{U_\mathrm{l}'(\pi_i)} - \frac{U_\mathrm{s}''(\pi_i)}{U_\mathrm{s}'(\pi_i)} < 0$$

$$\Leftrightarrow \frac{U_\mathrm{l}''(\pi_i) \cdot U_\mathrm{s}'(\pi_i)}{U_\mathrm{s}'(\pi_i)^2} - \frac{U_\mathrm{s}''(\pi_i) \cdot U_\mathrm{l}'(\pi_i)}{U_\mathrm{s}'(\pi_i)^2} = \left(\frac{U_\mathrm{l}'(\pi_i)}{U_\mathrm{s}'(\pi_i)}\right)' < 0$$

Thus, from (C.8) and (C.9), it follows that

$$\Rightarrow -U_\mathrm{l}''(\pi_i) > -\frac{U_\mathrm{l}'(\pi_{i,J})}{U_\mathrm{s}'(\pi_{i,J})} \cdot U_\mathrm{s}''(\pi_i) \; \forall \, \pi_i < \pi_{i,J} \tag{C.10}$$

From (C.10), we conclude:

$$-\int_{\hat{\pi}_i}^{\pi_{i,J}} U_\mathrm{l}''(\pi_i)\mathrm{d}\pi_i > -\frac{U_\mathrm{l}'(\pi_{i,J})}{U_\mathrm{s}'(\pi_{i,J})} \cdot \int_{\hat{\pi}_i}^{\pi_{i,J}} U_\mathrm{s}''(\pi_i)\mathrm{d}\pi_i \, , \forall \, \pi_i < \pi_{i,J}$$

$$\Leftrightarrow U_\mathrm{l}'(\pi_{i,J}) - \int_{\hat{\pi}_i}^{\pi_{i,J}} U_\mathrm{l}''(\pi_i)\mathrm{d}\pi_i > U_\mathrm{l}'(\pi_{i,J}) - \frac{U_\mathrm{l}'(\pi_{i,J})}{U_\mathrm{s}'(\pi_{i,J})} \cdot \int_{\hat{\pi}_i}^{\pi_{i,J}} U_\mathrm{s}''(\pi_i)\mathrm{d}\pi_i = U_\mathrm{s}'(\pi_{i,J}) \cdot \frac{U_\mathrm{l}'(\pi_{i,J})}{U_\mathrm{s}'(\pi_{i,J})} -$$

$$\frac{U_\mathrm{l}'(\pi_{i,J})}{U_\mathrm{s}'(\pi_{i,J})} \cdot \int_{\hat{\pi}_i}^{\pi_{i,J}} U_\mathrm{s}''(\pi_i)\,\mathrm{d}\pi_i$$

$$\Leftrightarrow U_\mathrm{l}'(\hat{\pi}_i) > \frac{U_\mathrm{l}'(\pi_{i,J})}{U_\mathrm{s}'(\pi_{i,J})} \cdot U_\mathrm{s}'(\hat{\pi}_i)$$

$$\Leftrightarrow \frac{U_\mathrm{l}'(\hat{\pi}_i)}{U_\mathrm{l}'(\pi_{i,J})} > \frac{U_\mathrm{s}'(\hat{\pi}_i)}{U_\mathrm{s}'(\pi_{i,J})} \tag{C.11}$$

$\pi_{i,J}$ denotes that payoff which is received if being matched with a responder who does not shrink, $\hat{\pi}_i$ denotes those payoffs which are received if being matched with responders who shrink. Accordingly, (C.11) holds for all $j \leq \iota$, meaning that it also holds for the sum over all $j \leq \iota$. Now assume that a_i^* has been chosen such that the expected utility of a proposer with *small* RA is maximized, corresponding to payoffs of $\pi_{i,\,\text{small RA}^*,j}$. In this case, it follows from (C.11) that

$$\frac{\sum_{j=1}^{j=\iota} U_{\mathrm{l}}'\left(\pi_{i,\,\text{small RA}^*,\,j}\right)}{U_{\mathrm{l}}'\left(\pi_{i,\,\text{small RA}^*,\,J}\right)} > \frac{\sum_{j=1}^{j=\iota} U_{\mathrm{s}}'\left(\pi_{i,\,\text{small RA}^*,\,j}\right)}{U_{\mathrm{s}}'\left(\pi_{i,\,\text{small RA}^*,\,J}\right)},$$

$$\Leftrightarrow \frac{\sum_{j=1}^{j=\iota} U_{\text{large RA}}'\left(\pi_{i,\,\text{small RA}^*,\,j}\right)}{(J-\iota)\cdot U_{\text{large RA}}'\left(\pi_{i,\,\text{small RA},\,J^*}\right)} > \frac{\sum_{j=1}^{j=\iota} U_{\text{small RA}}'\left(\pi_{i,\,\text{small RA}^*,\,j}\right)}{(J-\iota)\cdot U_{\text{small RA}}'\left(\pi_{i,\,\text{small RA}^*,\,J}\right)} = 1,$$

$$\Leftrightarrow \sum_{j=1}^{j=\iota} U_{\mathrm{l}}'\left(\pi_{i,\,\text{small RA},\,j^*}\right) > (J-\iota)\cdot U_{\mathrm{l}}'\left(\pi_{i,\,\text{small RA},\,J^*}\right). \tag{C.12}$$

(C.12) implies that a proposer with a large RA can increase her utility by raising the offer compared to a proposer with a small RA. That is what we wanted to show.

Now assume that i's utility function is not differentiable at $\pi_{i,\,\text{small RA}}{}^*$. In this case, there is at least one responder in the pool for whom $\pi_{i,j}(m_j)$ is not differentiable at $\pi_{i,\,\text{small RA}}{}^*$, because that responder starts to shrink exactly at this point. Thus, it may be possible that proposers with slightly different degrees of risk aversion are stuck at this kink, meaning that a marginal increase of risk aversion may not result in a change of this offer. However, with risk aversion being sufficiently different, or with the proportion of responders who start to shrink at the same offer being sufficiently small, such exceptions can be neglected.

D: Proof That Equation (12) Is Correct

$$U_j(k_j) = U_j(3\cdot s_i - k_j) - Y_j \cdot \begin{cases} \left(\dfrac{-s_i + k_j}{2\cdot s_i} - \dfrac{3\cdot s_i - k_{jik}(s_i)}{3 - k_{jik}(1)}\right)^2, & \text{if } s_i > 0 \text{ and } k_j \geq s_i, \\ \left(\dfrac{-s_i + k_j}{s_i} - \dfrac{3\cdot s_i - k_{jik}(s_i)}{3 - k_{jik}(1)}\right)^2, & \text{if } s_i > 0 \text{ and } k_j < s_i, \\ 0, & \text{if } s_i = 0. \end{cases}$$

If $s_i > 0$ and $k_j^* \geq s_i$:

$$U_j'(k_j) = -1 - 2\cdot Y_j \cdot \left(\frac{-s_i + k_j}{2\cdot s_i} - \frac{3\cdot s_i - k_{jik}(s_i)}{3 - k_{jik}(1)}\right)\cdot\frac{1}{2\cdot s_i} = 0$$

$$- Y_j \cdot \left(\frac{- s_i + k_j}{2{\cdot}s_i}\right) + Y_j \cdot \left(\frac{3{\cdot}s_i - k_{jik}(s_i)}{3 - k_{jik}(1)}\right) = s_i$$

$$k_j^{\;*} = s_i + 2{\cdot}s_i \cdot \left(\frac{3{\cdot}s_i - k_{jik}(s_i)}{3 - k_{jik}(1)} - \frac{s_i}{Y_j}\right) \tag{D.1}$$

If $s_i > 0$ and $k_j^{\;*} < s_i$:

$$k_j^{\;*} = s_i + s_i \cdot \left(\frac{3{\cdot}s_i - k_{jik}(s_i)}{3 - k_{jik}(1)} - \frac{s_i}{2{\cdot}Y_j}\right) \tag{D.2}$$

We still have to prove that $3 - k_{jik}(1)$ is indeed the maximal payoff i can grant to j. Thus, we have to show that $s_i = 1$ is maximally kind. Assume that only one responder type is present, $k_{jik} = k_j$. For $s_i = 1$, $k_{jik}(1) = 3 - \frac{2}{Y_j}$ according to (D.1). Based on this and on $k_{jik}(s_i\,) = k_j$, $k_j^{\;*}$ can be recalculated:

$$k_j^{\;*} = s_i + 2 \cdot s_i \cdot \left(\frac{3 \cdot s_i - k_{jik}(s_i)}{3 - \left(3 - \frac{2}{Y_j}\right)} - \frac{s_i}{Y_j}\right)$$

$$\Rightarrow k_j^{\;*} = s_i + 3{\cdot}s_i^{\,2} - k_j^{\;*} \cdot Y_j \cdot s_i - \frac{2{\cdot}s_i^{\,2}}{Y_j}$$

$$\Leftrightarrow k_j^{\;*} = 3{\cdot}s_i - \frac{2{\cdot}s_i^{\,2} + 2{\cdot}Y_j \cdot s_i}{Y_j \cdot (s_i \cdot Y_j + 1)}, \tag{D.3}$$

Differentiating (D.3) with respect to s_i gives

$$\frac{\partial k_j^{\;*}}{\partial s_i} = 3 - \frac{4{\cdot}s_i + 2{\cdot}Y_j}{Y_j{\cdot}(s_i{\cdot}Y_j{+}1)} + \frac{2{\cdot}s_i^{\,2} + 2{\cdot}Y_j \cdot s_i}{(s_i{\cdot}Y_j{+}1)^2} = 3 - \frac{4{\cdot}s_i + Y_j \cdot (2{\cdot}s_i^{\,2} + 2)}{Y_j{\cdot}(s_i{\cdot}Y_j{+}1)^2} \tag{D.4}$$

Remind that the sendings to j are tripled by the experimenter. Thus, as $\frac{\partial k_j^{\;*}}{\partial s_i} < 3$, j will himself make better off if the sending is increased, meaning that $s_i = 1$ is maximally kind. As we used

(D.1), our proof is only valid for $k_j^* \geq s_i$. However, $s_i = 1$ will also be maximally kind if j returns less than he received: By definition, in this case, j's payoff increases in s_i with > 2.

Experimental Instructions and Course of the Experiment

In the following, we give an English translation of the instructions which were handed out to the subjects. We also show the most important extracts from the experiment, which was conducted in German originally.

Add. 1: Guidelines For the Second Part of the Experiment

Please read the following instructions **carefully**. If you have a question, call out your seat number. Please also read the instructions carefully which are provided during the experiment.

General:

You will participate in an interactive experiment again, **which differs from the first one**. In the following, the amounts will not be tripled and nothing can be sent back!

Details:

Interactive Experiment

- Again, the interactive experiment will be conducted with two players (called player A and player B). Your co-player **will be drawn randomly and anonymously** by the computer.

- The interactive experiment will proceed as follows: **Player A gets CU 10, player B gets nothing**. **A** can **offer any portion** of it **to player B**. Second, player B decides if he will accept the proposed distribution without any changes. **If B accepts**, **players A and B get the proposed payoffs**. B can also decide to change the payouts. However, B can **only reduce the payoffs**, and he can reduce them for both players only **in the same proportion**. The interactive experiment is then over.

- Two **random** examples:

− A proposes to keep the CU 10 and to offer CU 0 to B. B does not change the payoffs. In the end, A will be paid CU 10, and B will get CU 0. However, if B decides to reduce the payout by 100 %, both player A and player B will receive CU 0.

− A proposes to keep 2.5 and offers CU 7.5 to B. B does not agree and reduces the payoffs of A to CU 2.17 (thus by 13 %), the payoff of B is **therefore reduced by 13 %** to CU 6.53 as well. Accordingly, A receives CU 2.17, B receives CU 6.53.

- Your input in the A-role:
  As player A, you enter into an input box how many CU you want to offer to player B.
- Your input in the B-role:
  If you are player B, you will not be informed of the amounts that A has offered to you until the experiment ends. Therefore, you have to define your answer for every possible amount offered by player A. Again, you must enter your answers into a diagram. Now take a look at the **diagram on the additional sheet which was handed out to you:**
- As in the first experiment, you must **click into the diagram**, to let **black dots appear**, which you can **move up and down. This is how you set your decisions.** Exemplarily, one

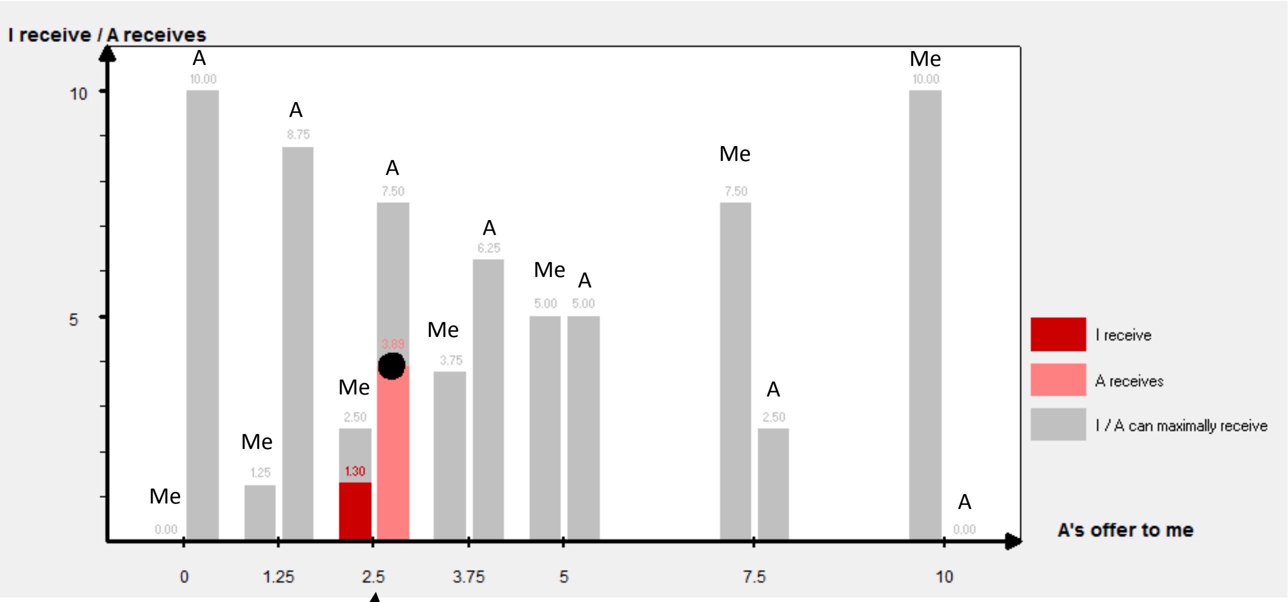
dot is already set into the diagram. The dot indicates how your decision affects the payoffs to you (dark-red, left bar), and to player A (light-red, right bar).

- The horizontal x-axis shows how much player A could offer to you. The size of the grey bars indicate how much you and player A will earn if you **do not reduce** the payoffs. To do so, move the black dot on the vertical y-axis **to the top**. **Reduce** the payouts by moving the black dot **downwards**. Thereby, both your payout and player A's payout is reduced to an equal proportion. The position of the black dot in proportion to the size of the grey bar indicates how much you reduced the payouts. Thereby the top position indicates "no reduction" while the bottom position (on the x-axis) indicates an "entire reduction". Now take a look at the case in which player A offers CU 2.5 to you (pair of bars, that is labeled with 2.5 on the x-axis): There, A keeps CU 10-2.5=7.5. If you decide to reduce the payouts by 48 % in this ("arbitrary") example, you will receive CU 1.30 instead of CU 2.5. The payout of player A will be reduced by 48 % as well A will receive CU 3.89 instead of CU 7.5. At every bar you can make your decision in an analogue way.
- Again, A can chose to offer an amount somewhere between the labeled values on the x-axis (e.g. CU 6). In such cases, the computer calculates your decision with the help of connecting lines, which will be plotted between the black dots later on.

Estimation Exercise and Selection Decisions *[Note: "Selection decisions" refers to lottery tasks (see Breuer and Hüwe, 2014b), which were also asked in this experimental part, but which are not discussed in the paper due to length restrictions]*

Again, you will have to do an estimation exercise and make selection decisions, which are similar to those in the first part of the experiment in their structure. Annotation: Perhaps, you noticed in the first part of the experiment that the fifth line of option 1 in the selection decisions always corresponded to your estimation of the expected payoff. Hence, option 1 was individually tailored to you. This will no longer be the case in the following experimental part such that you **cannot** raise the option-1-amounts by estimating high expected payoffs. Accordingly, it is **optimal for you to estimate as precisely as possible.**

Diagram for the Second Part of the Experiment (Role B)



An arbitrary example: If A offers CU 2.5 to me, A keeps CU 7.5. I reduce this amount to CU 3.89. Accordingly, my payoff is also reduced by 48 % to CU 1.30.

Add. 2: Course of the Experiment

Also compare the supplementary material of Breuer and Hüwe (2014b), where the control questions are displayed and where the course of the experiment is described (both experimental parts had the same structure)!

Test Questions

How much can A maximally earn if A offers CU 0 to B?

At which level must B place the black dot if B wants to earn as much as possible and is offered CU 2.5?

How much will A earn in that case?

How much will A earn if B halves an offer of CU 2.5?

First Input Stage, Role A

You have been assigned to role A for the interactive experiment!

Please make your input now. You decide how much you want to offer to B. Player B is a person in this room who is randomly assigned to you.

I want to offer to B an amount of CU □ (accordingly, I will keep CU 10 minus this amount if B does not reduce the amounts).

Explanation and Test Questions, Belief Stage

Thank you. At this stage, you will again have to estimate by how much players in this room in the B-role will **reduce** the offers **on average**. Put differently: How much will you earn in the A-role on average (depending on the offer)? In this estimation task, you can earn up to EUR 2.

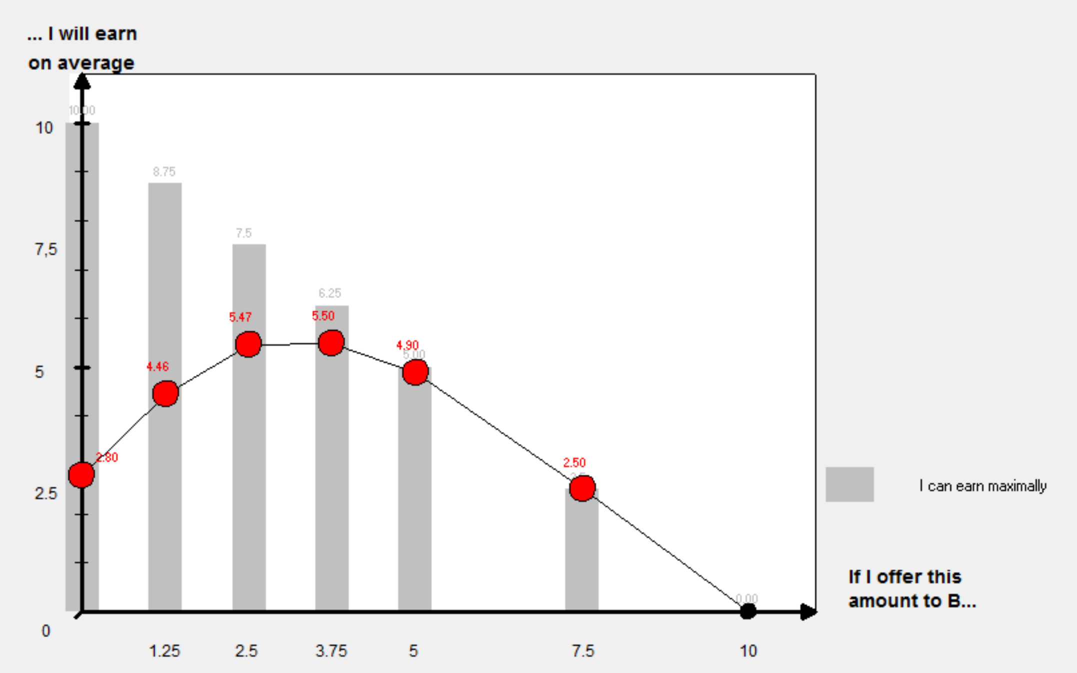
Look at the diagram displayed below. Setting a red dot at the left-most bar at CU 10 will have the following implication: You expect that **no B, without exception**, will reduce the payoffs if CU 0 is offered to him. Now look at the value 1.25 on the x-axis. There, A offers CU 1.25 to B and wants to keep CU 8.75. For example, you may believe that 50 % of the Bs will not reduce at all, and 50 % will reduce to zero. Accordingly, you should estimate 50 % * 8.75 + 50 % * 0 = CU 4.38, and you should mark this value with a red dot.

Please answer the following test question.

Look at an offer of CU 7.5 to B, meaning that you would like to keep CU 2.5 At which position do you have to mark the bar with a red dot if you believe that half of the Bs do not reduce the offer at all, and the other half reduces to CU 0?

Now click on “Next” to proceed to the estimation stage.

Input Stage, Estimation Exercise



[…]

Explanation of the Input Stage in Role B

Thank you. Decisions in the lottery task have been completed now. At the next stage, we ask you to play the interactive experiment again, this time **in the B-role**! Other subjects in this room are playing the experiment in the A-role and one of these subjects will be randomly assigned to you and will offer between CU 0 and 10 to you. In a diagram which you will recognize from the instructions handed out to you, you will have to determine to what degree you want to reduce the payoffs. Thereby, you will make a payout-relevant decision which is relevant for you as well as for A!

Input Stage, Role B

Compare the diagram displayed in the experimental instructions.