Can we neutralize social preference in experimental games?

Michal Krawczyk, Fabrice Le Lec

Can we neutralize social preference in experimental games?

Michal Krawczyk*, Fabrice Le Lec†

April 27, 2015

Abstract: We propose an experimental method whose purpose is to remove social concerns in games. The core idea is to adapt the binary-lottery incentive scheme, so that an individual payoff is a probability to see one's preferred social allocation implemented. For a large class of social preference models, the method induces payoffs in the game that are in line with subjects' (social) preferences. We test the method in several popular experimental games, contrasting behaviors with and without our methodology. Our results suggest that a substantial part of the difference between predictions based on selfishness and observed behaviors seems driven by such preferences, since our method does induce more "selfish" behaviors. But they also indicate that a considerable share is left unexplained, perhaps giving weight to alternative explanations or other types of social concerns.

Keywords: social preferences, experimental game theory, ultimatum game, public goods game, trust game, prisoner's dilemma, dictator game
JEL classification: A13, C65, C72, D63, D03.

1 Introduction

Over the last decades, experimental game theory has gathered ample evidence that individuals tend to depart from what is theoretically expected

*University of Warsaw, 44/50 Dluga St, 00-241 Warsaw, Poland, mkrawczyk@wne.uw.edu.pl

†Université Paris-1 Panthéon Sorbonne, Centre d'Economie de la Sorbonne UMR CNRS, Maison des Sciences Economiques - 106-112 Boulevard de l'Hôpital - 75647 Paris cedex 13 - France, fabrice.le-lec@univ-paris1.fr

from self-interested rational players. In numerous situations such as bargaining and cooperation games, experimental results seem to be robustly at odds with the game-theoretical benchmark. One of the most promising ways to account for this regularity is to take into account that individuals may not be indifferent to situations or behaviors of others. To this aim, several models of social preferences have been proposed (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Levine, 1998; Cox, Friedman, and Gjerstad, 2007). In these models individual preferences depend not only on one's payoff, but also on the others', or some aspects of others', behaviors (which could for instance signal intentions). Such models hence provide, at least at the level of stylized facts, an account of this apparent departure from theoretical predictions and observed behaviors. Adopting them does not preclude the use of game-theoretic concepts; the latter do not require that payoffs in the game correspond solely to individual pecuniary payoff, although they are indeed typically operationalized using this additional assumption.

Any apparent departure from the theoretical prediction can then be attributed to either a departure from the standard game-theoretic model of rational strategic behavior or the assumption that monetary payoffs duly represent players' preferences. What experimentalists working with game-theory models would need to distinguish between the two is a mechanism to *induce "selfishness" within the game*, or, to put it differently, to have payoffs in the game represent individual preferences, even for socially oriented individuals. One implication of the social preference models is that, once the payoffs are aligned with individuals' social concerns, the generally observed discrepancy between the theoretical benchmark and behaviors should disappear, or at least be substantially reduced. In contrast, if the discrepancy remains, it would indicate that it is due to some mechanism other than the social preferences that the method controls for. In such a case, off-equilibrium behavior observed in many experiments and typically associated with concerns for others may need to be reinterpreted. The aim of this research is to address this question by setting up a methodology that aligns individuals' social preference with payoffs.

The method builds upon the binary-lottery payment scheme initially proposed by Roth and Malouf (1979). Subjects are first asked to state their preferred social allocation of money for the group of players. Then, they interact through games in which payoffs are probabilities to see their favorite allocation implemented rather than somebody else's. Given that individuals

have already taken into account their social preferences in the first stage, maximizing one's payoff means maximizing the chances to see one's socially preferred allocation finally implemented. As long as the player believes that there is some probability that others have chosen a different allocation, playing 'selfishly' appears dominant. This is true for simultaneous games as well as a class of sequential games, if social preferences are based on final outcomes or ex-ante distribution of expected payoff. The method also provides some control for interdependent preferences and reciprocity concerns, in the sense that once the choice of the first stage are made, most individuals have little opportunity to reciprocate.

Using this methodology, referred to as the selfishness induction mechanism, we had subjects play several games for which social preferences are often used to explain behavior observed in the lab: the Dictator Game (DG), Ultimatum Game (UG), Trust Game (TG), Prisoner's Dilemma (PD), and the Public Goods Game (PGG). Our results indicate that subjects' behavior was in general closer to the theoretical benchmark, when our selfishness induction mechanism was used: a statistically significant difference was found in the DG, TG and PGG. The impact seemed to be less clear in the UG and absent in the PD. Still, we find that the effect of our method, wherever present, is still far from inducing behaviors in conformity with game theoretical benchmark. Several interpretations of the remaining discrepancy can be offered: the mechanism may be imperfect and generate side effects; social preferences are of a different type from the class covered by the method; or part of the usually observed divergence in such games is not primary linked to social preferences *per se*.

The remainder of the paper is organized as follows: the next section presents the theoretical details of the selfishness induction mechanism, the third one spells out the experimental design, the fourth one exposes the results, the fifth discusses them, and section six concludes the paper.

2 The method and its theoretical properties

In this section, we first present the general idea and intuition behind the method, then expose it in detail, and eventually derive its theoretical properties for simultaneous games and a restricted class of dynamic games. In particular we discuss how it should control for outcome-based preferences, procedural preferences, as well as interdependent preferences, and can be expected to mitigate the effect of reciprocity. The general idea of the method

	Cooperate	Defect
Cooperate	R, R	S, T
Defect	T, S	P, P

Table 1: PD matrix: utilities

	Cooperate	Defect
Cooperate	40, 40	0, 80
Defect	80, 0	20, 20

Table 2: PD matrix: monetary payoffs

is the following: n players interact in a game, with an additional player (the "dummy player"), without any action to perform in the game, being also present. Before that, all players, including the dummy player, choose a social allocation within a constrained set: each allocation determines the monetary payoff of all players. Then, the n players play a game whose payoffs are probabilities to see one's chosen allocation implemented (hereafter 'probability payoffs'). The dummy player collects the remaining probability. Being uncertain about others' choices and having no reason to believe that a given player is more likely to have chosen any specific allocation than any other individual, each player's sole aim should be to maximize their own probability payoff.

2.1 A simple example

To illustrate the method, consider a researcher willing to implement a simple prisoner's dilemma as represented in Table 1. What she wants to do is to make sure that subjects' utility levels associated with particular outcomes satisfy inequalities defining the prisoner's dilemma game: $T > R > P > S$. The standard solution would be to assign what would seem appropriate monetary payoffs, e.g. those given in Table 2.

Numerous experiments have shown that a large share of subjects tend to cooperate in such games. This may be because they feel uncomfortable with earning more than other players (advantageous inequity aversion), are motivated by efficiency – so that, say, Player 1 prefers $(0, 80)$ to $(20, 20)$ –, etc. In such a case the conclusion that subjects actually depart from playing the dominant strategy (and hence end up off the Nash equilibrium) would be incorrect. Instead, the researcher proved unable to align monetary payoffs in

	Cooperate	Defect
Cooperate	.40, .40, .20	.00, .80, .20
Defect	.80, .00, .20	.20, .20, .60

Table 3: PD matrix: probability payoffs

the game with the ordering of the outcomes she had in mind. Let us denote strict preference by $\succ$. If Player 1's preference structure is, for instance, $(40, 40) \succ (80, 0) \succ (20, 20) \succ (0, 80)$ and similarly for Player 2, then the game is a coordination game and mutual cooperation is fully compatible with rationality. To be sure, one could easily come up with other monetary payoffs that would make cooperation dominated *for this particular player, provided her preference was observable* (and obviously, it is not).

Applying our method to this simple game could proceed as follows: first, an additional player is added (the dummy player) and subjects have to state their preferences over social outcomes. For instance, subjects are first asked whether they prefer, say 30 for everyone to 50 for oneself and 0 for the others. Once their choices are made (and not revealed yet) the first two individuals play the game represented in Table 3, where payoffs are *probabilities to see one's chosen allocation implemented*. The third number represents the payoff of the dummy player (in probabilities to see her chosen allocation implemented).

Suppose a prosocial player has chosen the allocation $(30, 30, 30)$, then as long as she is not certain about others' choices and has no specific information about any other player, it is in her (prosocial) interest to maximize her game payoff, that is, the probability that her chosen allocation is implemented. Indeed, assuming her subjective probability that a given player has chosen the symmetric allocation $(30, 30, 30)$ is μ, and is equal for all other players, the consequences of the game for Player 1 can be rephrased as in Table 4, setting $u(0, 50, 0) = u(0, 0, 50) = 0$ and $u(30, 30, 30) = 1$ without loss of generality. As long as $\mu < 1$, it is the dominant strategy to defect in game represented in Table 3, because both relevant expected utility levels are higher than corresponding values for cooperation as shown in Table 4. Likewise, for a relatively selfish type of Player 1, preferring $(50, 0, 0)$ over $(30, 30, 30)$, $(0, 50, 0)$ or $(0, 0, 50)$, it will trivially be optimal to maximize own probability payoff.

	Cooperate	Defect
Cooperate	$.40 + .60\mu$	$.00 + 1.00\mu$
Defect	$.80 + .20\mu$	$.20 + .80\mu$

Table 4: PD game final expected utility for Player 1

2.2 Definition of the method

More generally, suppose the interaction is a n-player game. Individuals are divided into groups of $n+1$, including one dummy player. The procedure is then divided into two stages: First, subjects have to choose, from a predefined set, a social allocation of the general form $(x_1, x_2, ..., x_{n+1})$ with x_k being the monetary payoff of the k^{th} player. These "Stage 1" choices, denoted by θ_k, $k = 1 \dots n+1$ are not revealed to other players. In Stage 2, subjects first learn their roles—dummy or active player. The active ones interact through the game, whose payoff matrix is composed of probabilities to see one's allocation chosen in Stage 1 implemented. Probability payoff of subject i will be denoted by π_i. The $n+1^{st}$ subject cannot affect the outcome of the game and collects the remaining probability, that is: $\pi_{n+1} = 1 - \sum_{k=1}^{n} \pi_k$. Formally, we will refer to Stage 0 as the stage where 'nature" randomly determines each player's preferences on the set of available social allocations, and to the overall incomplete information dynamic game, composed of Stage 0,1, and 2 as the "extended game". The Stage 2 interaction, in contrast, will be referred to as the "stage game".

In our implementation of the mechanism, the social allocations available to any player i in Stage 1 are profiles of length $n+1$: $A = (a, a, a, a, ..., a)$ and $B_i = (0, 0, ..., b, ..., 0)$ with b being the i^{th} component. We set the following constraints on a and b, to ensure that A is efficient (eq. 1) while B_i is preferred from the self-interest viewpoint (eq. 2):

$$(n+1)a > b \tag{1}$$

$$b > a > 0 \tag{2}$$

This has the following implication: by any measure of fairness, at least as found in the literature, A is deemed fairer than B. That includes inequity aversion and maximin (A is a perfectly equal situation), efficiency and altruism.[1] So from the viewpoint of fairness, A dominates B_i. A choice of

[1]We leave aside social concerns unrelated to fairness, such as spitefulness and competitiveness, because, first, they tend to get limited empirical support and second, by nature

A points at some prosocial motive, whereas B_i is picked if self-interest prevails. Ideally, b should be chosen by the experimenter such that the difference with respect to a is not completely negligible (so that a substantial minority chooses B_i, and that the possibility to be matched with a B-player is real), but not too tempting either (to ensure a substantial share of A-players, i.e., individuals choosing A in Stage 1, which, as will become apparent soon, are the most interesting cases to study here).

2.3 Theoretical properties

Let $\succ_i$ and $\succsim_i$ denote strict and weak preference for a given player i. In line with prevailing experimental evidence, we assume that for any player i, $B_i \succ_i B_j$, j being another player. We can thus distinguish two types of players, picked by "nature" in Stage 0: *A-players* are those who prefer A in Stage 1 ($A \succ_i B_i$) whereas *B-players* are mostly motivated by self-interest ($B_i \succ_i A$).[2] Given the anonymous context of the experiment, we also set $B_k \sim_i B_j$ for i, j, k different, denoting these equivalent allocations as B_{-i}. The perceived prevalence of types will affect subjects' behaviors in Stage 2: we refer to player i's beliefs about player j's type a μ_{ij}, that is player i believes with probability μ_{ij} that j has chosen A and with probability $1-\mu_{ij}$ that he has chosen B_j.

We rely on the assumption of *uniform prior beliefs*: at the onset of Stage 2, a given player i has the same priors for all the other players, given that they are anonymous.[3] Formally, we have $\mu_{ij} = \mu_{ik}$ for all $j \neq k$. When player i is under consideration only and there is no ambiguity, we denote this uniform prior about others as μ.

2.3.1 Outcome-based social preferences

The outcome of the game is a lottery $L(\pi, \theta) = (\pi_1, \theta_1; \ldots; \pi_{n+1}, \theta_{n+1})$. Now considering player i, the belief about the outcome of the game for i is a

they cannot be distinguished from selfishness in the games we test here, except in the Ultimatum Game, for the second player, for whom rejecting an unfair offer could be due to sufficiently strong competitive preferences.

[2]Given that choices are not revealed to other players and that there is a positive probability to see Stage 1's decision implemented, we expect participants to choose their preferred allocation. We leave aside the rather unlikely case of individuals perfectly indifferent between the two allocations, but that does not affect the argument.

[3]This also corresponds to experimental conditions where matching is made randomly, substantiating a frequentist view of the situation: in the group of possible opponents, a proportion μ_i are self-interested and the opponents are randomly picked.

lottery L_i given by, in the absence of any belief updating:

$$L_i(\pi_i) = (\pi_i, \theta_i; \mu(1-\pi_i), A; (1-\mu)(1-\pi_i), B_{-i}) \tag{3}$$

By stochastic dominance, if $\pi_i > \pi_i'$ then the corresponding lotteries L_i and L_i' are such that: $L_i \succ_i L_i'$, since $\theta_i \succsim_i A$ and $\theta_i \succ_i B_{-i}$. This implies the following proposition:

Proposition 1 (Outcome-based preferences in games with no belief updating) *An expected utility maximizing*[4] *individual with outcome-based social preferences, holding symmetric prior beliefs about the others' types, has her preference represented by the probability payoffs of the game.*

Hence, the method ensures strict monotonicity with one's probability payoffs. For example, in a simultaneous game, if a strategy is dominant in probability payoffs, then it is also a dominant strategy in the extended game.

While this proposition obviously covers all simultaneous games, some sequential games involve belief updating. If Player 2 modifies her μ_{21} in light of the first player's action, the latter, anticipating this, may send a costly signal. Then, a separating equilibrium may arise (both players' behavior depends on their types), a deviation from the selfish benchmark. Nevertheless, for some games of interest, equilibria of the extended game correspond to the equilibria of the stage game provided appropriate solution concept is applied – here we use the perfect equilibrium (Selten, 1975). In particular, the selfishness induction mechanism works for all two-player complete-information games, in which no player making a decision at a penultimate node ("last decision", leading directly to one of the endnodes of the game) can affect the sum of payoffs for Players 1 and 2 (one well known example is the Trust Game).

Proposition 2 (Outcome-based preferences in sequential games) *For a two-player game, in which at any penultimate node the sum of obtainable payoffs for Players 1 and 2 is constant, all perfect equilibria of the extended game correspond to the perfect equilibria of the Stage 2 game.*

In this class of games the last decision merely concerns distribution of the surplus accrued between Players 1 and 2. Intuitively, because one cannot be sure that the other player chose A, the best thing one can do is to grab as many probability points for oneself as possible. And because such selfish

[4]This hypothesis can be weakened to players having risk preferences satisfying first order stochastic dominance. Yet, for the sake of conciseness and clarity, we assume maximization of expected utility in the exposition.

behavior is always expected, there is no reason to behave non-selfishly at earlier nodes (see the On-line Supplement for the proof).

Proposition 2 does not imply that non-selfish equilibria will typically exist in games in which the last player to choose can affect the dummy's payoff. Yet, the conditions for a separating equilibrium to exist are quite stringent (see the On-line Supplement for the formal elaboration in the two-player case).

It may be overly demanding though to assume that players adhere to the equilibrium (especially in our experimental conditions with no repetition nor feedback). Two alternative approaches can be considered. First, a conservative approach based on no belief updating may be legitimate: given that players cannot know for sure who chose what in Stage 1, they may reason it does not matter who (among other players) gets the lottery tickets. In this case, the same result as for simultaneous games hold. The second approach is to have an agnostic view of the question, and observe subjects' play to determine how relevant is the possibility of a separating equilibrium. There seems to be very little overall difference in the behavior of A and B-players (see the Result section below for details) suggesting that separation is unlikely to occur. That is to say, given the empirical distribution of the population of players, the optimal strategy for a pro-social player would be to play in accordance with game-theoretic solutions.

It is worth noting that the plain binary-lottery payment scheme (Roth and Malouf, 1979; Berg, Rietz, and Dickhaut, 2008) in which players' payoffs are in terms of probabilities to win a fixed prize (and zero otherwise), does not guarantee that players have their preferences represented by the probability payoffs of the game. In the absence of a player that gets the remaining probabilities, there is an additional outcome of the game, namely that no one gets the prize: $(0)_{1\leq k\leq n}$. Social motives as ubiquitous as altruism or efficiency (Charness and Rabin, 2002; Engelmann and Strobel, 2004; Cox, Friedman, and Gjerstad, 2007; Levine, 1998) are then not controlled for. For individuals motivated by these types of social concerns, having another player winning may be preferred to the situation where no one wins. Cooperation can be a rationalizable strategy in the Prisonner's Dilemma for instance.

2.3.2 Preferences based on ex ante comparisons

Until recently, the vast majority of models of social preferences have focused on riskless situations. Over the last years evidence has been accumu-

lated, however, that people seem not only to care about the monetary consequences (ex post allocation) but also about the distribution of risk among individuals (*ex ante* allocation) – see Bolton, Brandts, and Ockenfels 2005; Krawczyk and Le Lec 2010; Brock, Lange, and Ozbay 2013 and the models by Trautmann 2009 and Krawczyk 2011 for instance. One approach involves assuming that social preference is defined in terms of expected, rather than final payoffs within a game. Such preference does not lead to deviations from "selfish" behavior in Stage 2 of our method, at least for A-players. The intuition is that such social motives push subjects in the same direction as self-interest, that is towards more chances to have A implemented. That is because in any case, A-players are already "behind" in expected terms so that motives such as inequity aversion or maximin would favour the player's self-interest. The same is true for efficiency or altruism: The most efficient or altruistic allocation is A, so increasing the probability of seeing it implemented increase efficiency or altruism in terms of expected payoff. Formally, one can show the following (see On-line Supplement):

Proposition 3 *If an A-player has maximin, altruistic, efficiency driven or inequality-averse social preferences in terms of expected payoffs, then her preferences are monotonic with her probability payoff.*

In sum, for A-players, the usual prosocial motives push in the same direction as self-interest, that is to increase the player's probability payoff.

This control for ex ante preferences can only be achieved because individuals go through Stage 1. Indeed, with the standard binary-lottery payment scheme (possibly with an additional dummy player), some ex ante prosocial player would be willing to share chances to win the prize. Given the accumulating evidence of ex ante social concerns (Bolton, Brandts, and Ockenfels, 2005; Krawczyk and Le Lec, 2010), the use of the standard binary-lottery incentive scheme would rather tend to confirm the existence of such a dimension of social concern rather than control for social preferences in a broader sense.

2.3.3 Interdependent preferences and reciprocity

One important class of models of social preferences are based on the idea of reciprocity. Reciprocal preferences (as in Rabin 1993; Levine 1998; Charness and Rabin 2002; Dufwenberg and Kirchsteiger 2004; Falk and Fischbacher 2006; Cox, Friedman, and Sadiraj 2008 among others) postulate individuals prefer to reward benevolent interaction partners (or reversely punish malevo-

lent ones), even at a cost.[5] The validity of our methodology for reciprocal and interdependent preferences is less clear. This is due to the relative complexity of our method, which is magnified by the complexity of the reciprocity models themselves.[6] Despite that, we nurture the view that our method should at least partly mitigate the effect of reciprocity-based or interdependent preferences, at least for A-players.

This is so for two main reasons: on the one hand, players face a clear identification issue regarding the benevolence/malevolence of others in our setting; and on the other hand, A-players have little available actions to reward or punish some other subjects. Regarding the first point indeed, in the absence of information about Stage 2 choices, players can hardly identify overall benevolence/malevolence of their counterparts: acting in a prosocial way in the stage game but having chosen the B allocation in Stage 1 can hardly be interpreted as benevolent while maximizing one's probability payoffs after having chosen A in the initial allocation task may be. The second point is that even if players could overcome the first difficulty, A-players do not have much room for rewarding or punishing the other. Suppose, for the sake of the argument, that player i, who chose A in Stage 1, tends to believe that j is benevolent. By positive reciprocity, the best that player i can do to be kind to j is to increase the probability of seeing A implemented: *increasing her own payoff is as kind as increasing j's probability payoff*. Given the uncertainty about the other's type as well as the dummy player's type, the room for positive reciprocation seems small. In the opposite case, where i believes that player j chose B_j, and hence is malevolent: player i can either increase the probability of seeing A implemented (which is not a strong retaliation, but that is at least better than seeing B_j implemented) or can try to have some other B_k implemented, that is having another B-player winning. Yet, this last option would be both quite costly and would mean 'rewarding' k who has also chosen B. In sum, A-players do not have much opportunity to punish or reward others but they do have the opportunity to increase the probability to see A (their own choice) implemented.

[5]We consider here that interdependent preferences and reciprocal preferences as equivalent even though there is a clear theoretical distinction between both: reciprocal preferences depend on the benevolence/malevolence of the *action* taken by some other player, whereas for interdependent preferences, benevolence/malevolence depends on the other player's *preferences*. In our setting, there is little possibility to distinguish between both.

[6]Indeed, the latter often involve beliefs and second order beliefs, are not always clearly defined for more than two players, are not aimed to deal with probabilistic outcomes, and last but not least there does not seem yet to exist a scientific consensus on how to model various features of reciprocal preferences.

We hence expect our method to reduce interdependent and reciprocity concerns. This feature would not necessarily be obtained in an alternative experimental method to study the effect of social preferences in games that would consist of eliciting (distributive) social preferences in a separate task and compare it to observed individual behaviors in standard games (see Blanco, Engelmann, and Normann (2011) for both an implementation of this idea and the related issues).

3 Design

In order to test the validity of the method we had subjects play a number of games, often studied experimentally before, under one of two treatments: in the sessions with Selfishness Induction Treatment (SIT), the mechanism was implemented, with group size $n+1 = 5$, to allow 2-player and 4 player games, whereas the same games where played directly for money in the Control Treatment (CT).

The experiment was preceded by a hypothetical pilot concerning Stage 1 only, aimed at calibrating the allocations so that the fair option would be chosen by majority but not all subjects. The fair option always involved 40 PLN (approx. 10 euro) for each participant and unfair allocation paid 0 for others. The payoff for the chooser in the unfair allocation was manipulated between subjects, who were asked to report what they would choose and what fraction of experiment participants they thought would choose each option. It appeared that a payoff of approx. 60 PLN would give the desired low but non-negligible fraction of individuals choosing the selfish option. It was also clear that most individuals were rather pessimistic about other's choice: the predicted fraction of individuals choosing the selfish option was substantially higher than the actual one. That tends to raise the incentives in the SIT for the A-players: the perceived risk of ending up with nothing is higher. Allocation B was set at 55 PLN, after the first session where it was 70.[7]

[7]After the first experimental session in which the selfish chooser's payoff was set at 70 PLN, it was found that the proportion of subjects choosing the selfish option B was 60%, much higher than in the pilot, which may have been due to the hypothetical nature of the choice in the latter, or perhaps because the show-up fee of 5 PLN was not mentioned in the pilot, making the selfish choice look even more harming to others. Either way, because the selfishness induction mechanism was expected to work primarily for the individuals choosing the fair allocation in Stage 1, the selfish chooser's payment was reduced from 70 PLN to 55 PLN in subsequent sessions. Additionally, instructions were modified slightly, clearly stating the second stage would consist of games of decision and payoffs would not

The second stage consisted of eight rounds. The groups of five were re-matched for each round. Subjects were told that one round would be chosen at random at the end of the experiment to determine the distribution of tokens (and thus chances for each participant's choice from Stage 1 to be implemented). The payoffs in tokens were calibrated in such a way that, first, based on previous research, we could expect a substantial level of seemingly non-selfish choices and, second, that the sum for the active players would never exceed 100.

One fifth of our subjects constantly played the role of an inactive dummy player (a residual claimant of tokens). These participants were asked to make analogous, yet hypothetical decisions. The active participants were actually divided into two-person subgroups during the first seven rounds. The first round involved the Dictator Game with 40 tokens to share (such that the "fifth"–the inactive dummy player–would always get 100-40-40=20 tokens). Next, subjects played an Ultimatum Game with 50 tokens to be allocated, using Minimal Acceptable Offer strategy method (Selten, 1967). The third round was a Trust Game with an initial endowment of 15 tokens. The first mover was asked to place 0, 3, 6 ... or 15 tokens in the "second account" which were subsequently tripled and divided by the second mover, whereas tokens placed in the "first account" would be kept by the first mover. The game was played using the strategy method. Rounds 4 to 6 were analogous to rounds 1 to 3; we had half the active (non-dummy) subjects act as first movers in rounds 1-3 and second movers in rounds 4-6. The other half faced the reversed order. Round 7 was the Prisoner's Dilemma game with payoffs of 25 for both players in the case of mutual cooperation, 15 for each in the case of mutual defection and the "temptation" and "sucker's" payoffs of 30 and 10 respectively. In rounds 1 to 7, no feedback was given. The final round involved a linear four-person public goods game with six periods, endowment of 2 tokens per round and marginal per capita return of .5. It was the only round in which subjects were allowed to use decimals in their choices. As is customary in these kinds of games, participants were given immediate feedback after each period concerning the total contribution in their group only. We chose the so-called partner-matching repeated PGG given its prevalence in the literature on cooperation dilemmas (Andreoni, 1988; Fehr and Gaechter, 2000).

---

depend on skill or effort. These changes were found to indeed make a difference in the first stage, but not in the second one. All the statistical tests reported below include the first session but their results hold when discarding it, unless explicitly said otherwise.

Throughout the experiment, participants could consult the printed general instructions which explained how choices in Stage 2 affected the chances of one's allocation to be implemented. Instructions for specific rounds in Stage 2 were displayed on the screen at the beginning of the relevant round. Instructions were supplemented with examples and control questions that subjects had to answer correctly in order to proceed. After the final round subjects were asked to report some demographic variables. The computer randomly selected one round and, given subjects' probability payoffs from that round of Stage 2 and their Stage 1 decisions, final monetary outcomes. These were paid out in cash, immediately after the experiment.

The CT was analogous, except that there was no Stage 1 and subjects were told that each token they earned in the selected round would be worth 2 PLN. Considering typical results described in the literature we expected the average earnings to be around 35-40 PLN, to equate in expected value earnings in SIT, assuming most people would go for the equitable distribution in Stage 1. There was obviously no need to use dummy players in this treatment.

Our design and the corresponding theoretical predictions are summarized in Table 5. The undetermined cases correspond to reciprocity preferences taken in a broad sense: a reasonable assumption though would be that the observed departure from the selfishness-based prediction should be smaller in SIT than in the control treatment. By comparing these predictions (when available) with the experimental data, it is hence possible to test how important the various types of social preferences are in explaining usually observed departures from the theory.

Participants were recruited from the subject pool of the Laboratory of Experimental Economics at the University of Warsaw using ORSEE (Greiner, 2004). Among the 134 participants about half were male. All but 9 were students, of which about half were Economics majors. Average age was 23.6. The experiment was conducted in the lab preventing communication between subjects and was computerized using LabSEE XP Software.[8] Six sessions were run in total: three CT sessions and three SIT sessions with 20 to 25 subjects per session. The experiment lasted between 70 and 90 minutes, typically a bit longer under SIT than CT. Average earnings including the show-up fee were 36.7 PLN in the CT and 36.9 PLN in the SIT. An alternative simple student job would pay 10-15 PLN per hour.

[8]Developed by Robert Borowski, see http://www.labsee.com.

Games	Description	Outcome-based and Ex ante (A-types)	Reciprocal and interdep. pref.
DG	$x_1 \in [0, 40]$ $\pi_1(x_1) = 40 - x_1$ $\pi_2(x_1) = x_1$	$x_1 = 0$	$x_1 = 0$
UG	$x_1 \in [0, 50]$ $x_2 \in \{a, r\}$ $\pi_1(x_1, a) = x_1$ $\pi_1(x_1, r) = 0$ $\pi_2(x_1, a) = 50 - x_1$ $\pi_2(x_1, r) = 0$	either $x_1 = 49$ or $x_1 = 50$ and $x_2 = a$ (no signal.) or separating eq.	undetermined
TG	$x_1 \in [0, 15]$ $x_2 \in [0, 3 \times x_1]$ $\pi_1(x_1, x_2) = 15 - x_1 + x_2$ $\pi_2(x_1, x_2) = 3 \times x_1 - x_2$	$x_1 = 0$ $x_2 = 0$	$x_1 = 0$
PD	$x_1, x_2 \in \{c, d\}$ $\pi_1(c, c) = \pi_2(c, c) = 25$ $\pi_1(c, d) = \pi_2(d, c) = 10$ $\pi_1(d, c) = \pi_2(c, d) = 30$ $\pi_1(d, d) = \pi_2(d, d) = 15$	$x_1 = d$ $x_2 = d$	$x_1 = d$ $x_2 = d$
PGG (repeated)	$x \in [0, 2]^4$ $\pi_i(x) = 2 - x_i + \frac{1.5}{4} \sum_j x_j$	either $x_i = 0$ (no signal.) or separating eq.	undetermined

Table 5: Experimental games and predictions in SIT

4 Results

4.1 Control treatment and Stage 1 allocation choices

To assess the impact of our proposed experimental manipulation, it is helpful to establish that our subjects' behavior in the control treatment did not deviate from typical findings in the literature. On top of that, since our selfishness induction mechanism is expected to make a difference primarily for subjects that are non-selfish in the first place, choices in Stage 1 will be summarized too.

The first precondition seems to hold: in most games, our results replicate typical findings in the experimental literature. In the DG, decision makers gave away 34% of the pie, the equal split and complete selfishness being frequent choices. The offers were somewhat higher in the UG (42% of the available sum), with a mode at the equal split (40% of offers); the average

minimal acceptable offer was 26% of the pie, with a median of 30%. In the TG, subjects passed 52% of the available sum and second movers on average would send back 20% of the available money, this percentage increasing with the money passed by the first player from 13 % for 3 tokens sent to 32 % for 15. In the PD, subjects cooperate 42% of the time; in the PGG, subjects start contributing exactly half of the endowment on average and cooperation goes down over time. Overall, our subjects seem to follow the general trends observed in previous studies on these games – see *e.g.* Camerer (2003) for a survey – perhaps with a very slight skew towards prosociality, which is fortunate from our perspective, as it makes the theoretical difference between SIT and CT more salient.

Regarding Stage 1 choices in SIT, we observe that, except for the first session, the fraction of subjects choosing the equal allocation was, as expected, quite high: 63% of subjects in the SIT treatment chose the A allocation. This is fortunate as it provides numerous valid observations but also because the proportion of B-choosers is far from negligible (a little more than one third): the probability to face a selfish player increasing the incentives to behave "selfishly" in the extended games.[9]

To study the results, we present two types of analyses: first, we run standard non-parametric tests of the treatment effect. Then, we run an OLS regression with the strategy chosen in given game as the dependent variable and dummy variables for the treatment, the order of choice tasks and various socio-demographic features (gender, income[10], a dummy for Economics major) as independent variables.[11] In order to have clear comparisons between the games, we use exactly the same explanatory variables in all cases: we keep all the variables that seem to play a significant role in at least one game. We also conduct a logit analysis on the occurrence of selfish play (defined as

[9]As mentioned before, in a non-incentivized pilot session subjects were asked about their beliefs regarding others' choices in such an allocation task, and the percentage given of B-choosers was greater than the actual percentage, suggesting that subjects would rather overestimate others' selfishness than underestimate it. Note also that for session 1, the proportion was 60 % of B-players when $b = 70$ to be compared with a proportion of 25 % for other sessions, when $b = 55$.

[10]Note that income is a categorical variable coded as several dummies, since in the post-experimental questionnaire only general ranges were offered to subjects in order to take into account the general reluctance to state personal income in European countries. We also ran analyses of variance with the same variables to take this into account, and obtained similar results.

[11]The only exception is the PD, where response variable is binary, for which we used a logit model with the same independent variables.

being less than 10% away from the game-theoretical prediction, these 10% being scaled on the length of the strategy set). Based on these statistics, we observe a general tendency to be more selfish in SIT than in CT.

4.2 Games with treatment effects

In the **Dictator Game**, our mechanism induced more selfish behavior: the subjects give on average 11.27 tokens (28.17% of the total of 40 tokens), which is 1.97 tokens less than in the control treatment (W=2125, p=.04 in a one-sided Mann-Whitney test). The proportion of 0 offers in the SIT was 27%, compared with 15% in the control treatment, whereas the equal split was chosen by 25% and 34% respectively. These results are displayed in Fig. 1. It suggests that the main difference is the shift away from the (almost)

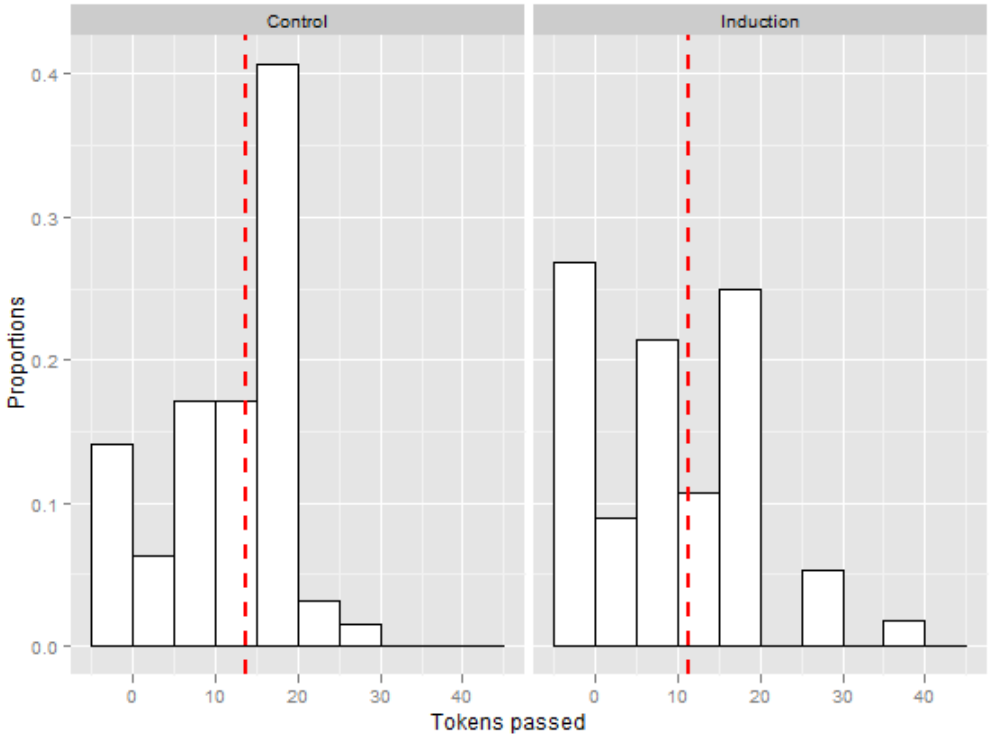

Figure 1: *Distribution of offers in the DG*

equal split and towards more complete selfishness: the former is modal in CT, while the latter is in SIT. If we classify subjects as selfish when they gave less than 10% of the available amount, a χ^2 test of treatment effect turns out weakly significant (2.9, p=.087). A linear regression confirms these

findings, and its results are summarized in table 6: Even when controlling for the possibility of a slightly different composition of subject samples in both treatments, the treatment effect is (weakly) significant and substantial.

	OLS DG	OLS TG passed	OLS TG sent-back (15)	OLS PGG (round 1)	Logit PD	OLS UG offers	OLS UG MAO	OLS Selfishness Index
Constant	12.09**	5.16*	14.41**	3.85***	−0.13	21.24***	18.26***	0.60***
	(4.88)	(2.88)	(5.61)	(1.36)	(0.40)	(4.00)	(5.65)	(0.09)
Induction	−3.39*	−1.61	−1.72	−0.96*	0.22	−0.74	2.58	0.06**
	(1.77)	(1.05)	(2.04)	(0.50)	(0.45)	(1.45)	(2.05)	(0.03)
Order	−4.67***	−1.19	−3.45**	0.25	0.08	−4.04***	−3.97**	0.06**
	(1.50)	(0.88)	(1.72)	(0.42)	(0.39)	(1.23)	(1.74)	(0.03)
Dummy Econ.	−4.16***	−2.25**	−6.10***	−0.57	0.94**	−0.35	−5.05***	0.12***
	(1.54)	(0.90)	(1.76)	(0.43)	(0.39)	(1.26)	(1.78)	(0.03)
Session 1	1.50	−0.24	0.84	0.48	−0.90	−1.39	−1.16	−0.06
	(2.35)	(1.38)	(2.70)	(0.66)	(0.59)	(1.92)	(2.72)	(0.04)
Gender	1.61	2.08**	1.38	0.27	0.03	2.86**	2.11	−0.05*
	(1.54)	(0.91)	(1.77)	(0.43)	(0.39)	(1.26)	(1.78)	(0.03)
Adj. R^2	0.13	0.09	0.14	0.00		0.11	0.04	0.19
F (p)	2.45 (.007)	1.96 (.04)	2.61 (.004)	.96 (.49)		2.19 (.02)	1.46 (.16)	3.26 (< .001)
AIC (BIC)					167.00 (183.72)			
Log Lik. (Deviance)					-77.50 (155.00)			
Num. obs.	120	120	120	120	120	120	120	120

*=significant at the .10 level, **= significant at the .05 level, ***= significant at the .01 level. Control Variables (not displayed): Income. Order=1 when subjects played the DG, UG, TG as first player first, then as second player.

Table 6: Econometric estimations

A similar treatment effect is visible in the behavior of first movers in the **Trust Game**, with 6.16 tokens out of 15 being sent on average in the SIT and 7.82 in the CT, a difference of 21%. Again, the distributions are significantly different according to a Mann-Whitney test (W=2138, p=.03). The results are displayed in Fig. 2. Once again, an analysis of the occurrence of ratio-

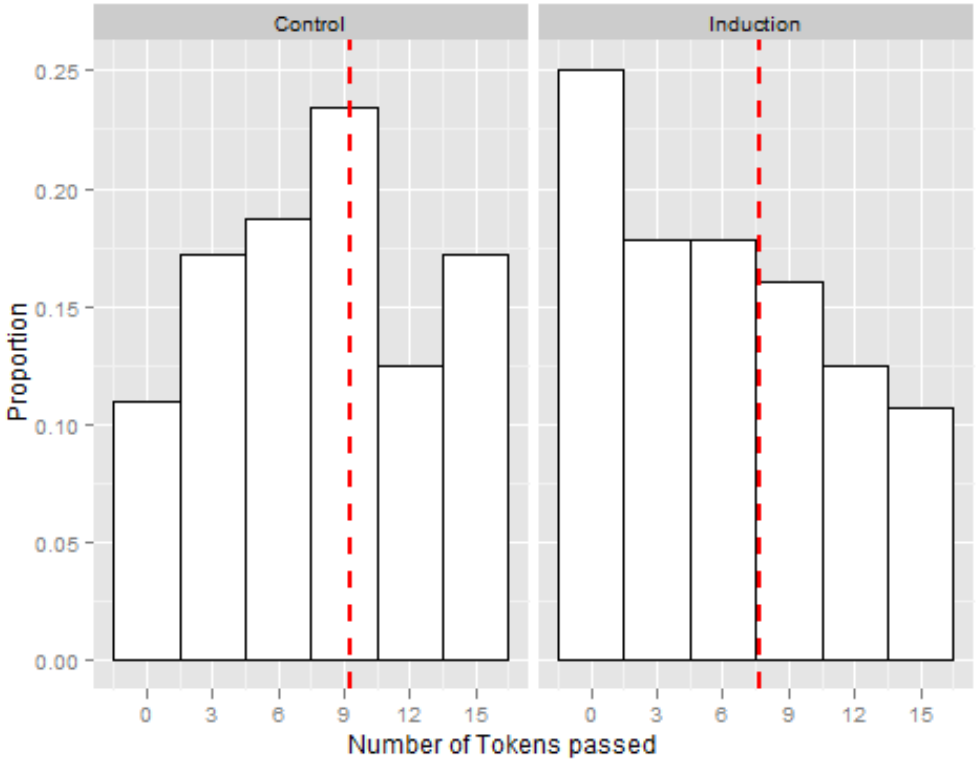

Figure 2: *Distribution of tokens sent in the TG*

nal choice (in the sense of backward induction under common knowledge of selfishness) in both treatments using a χ^2 test reveals a weakly significant difference (3.17, p=.075) between the two treatments. Also, an OLS regression controlling for various socio-demographic and experimental variables leads to a milder conclusion, with treatment being only on the verge of weak significance (with $p = .12$ for the estimated coefficient of Induction, see table 6).

In contrast, no difference is observed regarding the behavior of second players. The proportion of money sent back is displayed in Fig. 3. No statistical difference can be found between the two treatments, for any specific

amount of money sent or for an aggregate measure of planned repayment. Analyses of variance on the money sent back for 3, 6, 9, 12 and 15 tokens show that the Economics major and occasionally the order of tasks play a significant role in the money sent back, but SIT is never associated with a p-value less than .30. χ^2 tests of frequency of purely selfish behavior do not

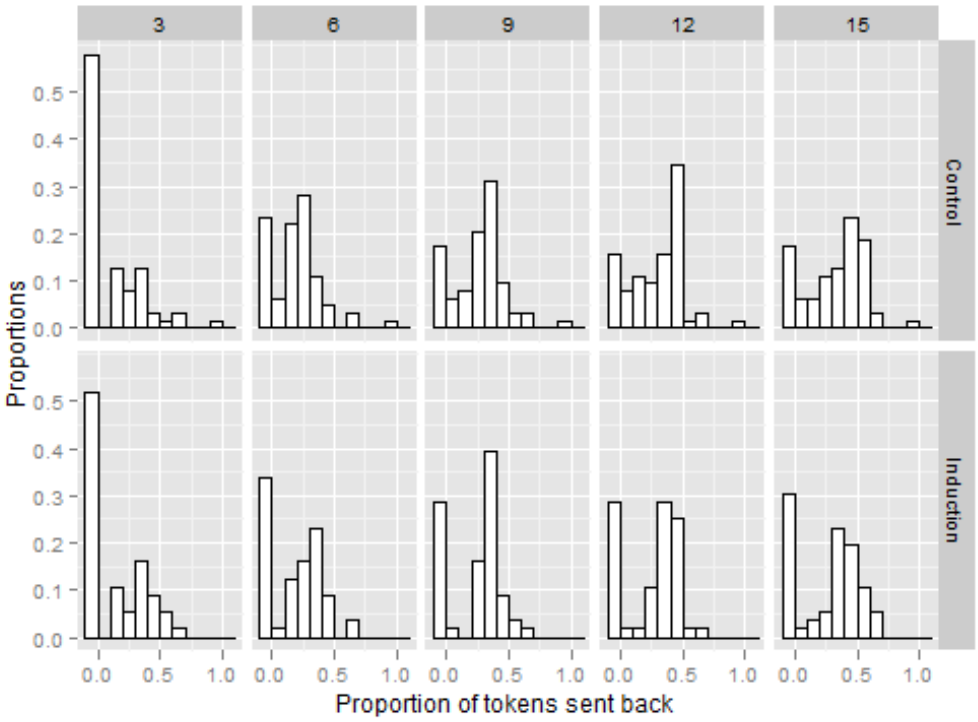

Figure 3: *Proportion of tokens sent back in the TG*

reveal any significant difference for the amount that was sent either.

In the **Public Goods Game**, we also observe a difference between the two treatments: the average contribution is 1 token (out of an endowment of 2) in the first round under CT, whereas it only reaches .70 token (35%) in SIT (see Fig. 4). The distributions of individual contributions are significantly different (one-sided Mann-Whitney, W=1379, p=.01). All further rounds are also significantly different except for the last one. To study the data in all the rounds with full statistical independence, we chose groups as appropriate level of analysis.[12] Focusing on the sum of contributions per group for each

[12]Indeed, individual behavior in the second and subsequent rounds depends critically on other members in the group in the first round, in particular because of reputation or

round, we find that the first 4 rounds yield significant results (one-sided Mann Whitney, with $p < .05$ except in the case of the third round for which $p < .10$). Considering the total contribution within a group across the 6 rounds, SIT and CT differ significantly (.46 token on average per subject and round for SIT versus .82 for CT, $p = .046$ in a one-sided Mann-Whitney test). Once again, at the individual level, we also find that the proportion of

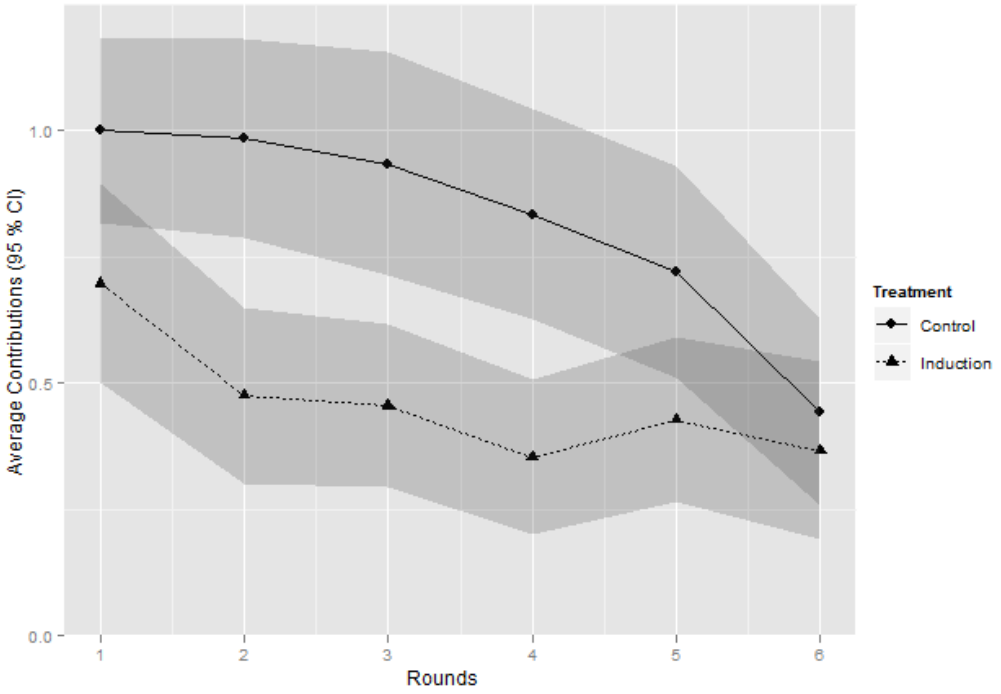

Figure 4: *Average contribution in the PGG by round*

almost purely selfish behaviors is different for both treatments: for the first round (the only one for which statistical independence holds), a χ^2 test yields a value of 3.52 (p=.06). This effect of treatment on voluntary contributions is confirmed by a linear regression with the usual control variables (Table 6).[13]

reciprocity.

[13]Note that the overall quality of the model is very low, but other specifications, while improving the model, yield similar estimated coefficients and significance levels for the effect of treatment. For reasons of homogeneity, we keep the same specification across the games.

Overall, in these three games, we observe substantial differences between the control treatment and the selfishness induction treatment and these differences go in the direction that was assumed, i.e. towards more selfishness. In particular more individuals seem to behave (almost) in line with game-theoretic prediction. Yet, it is quite striking that even if some effect of SIT is observable, subjects are still far from playing as predicted (that is playing as if selfish in the stage game). In the Dictator Game, whose payoff structure is quite easy to grasp, most subjects give something (two thirds give more than 5 tokens out of 40). In the Trust Game, not only first players send money but are repaid by the second player, even though the subgame perfect equilibrium is quite simple to identify. Finally, in the PGG, although we observe some difference, the across rounds contribution is still far from 0 and does not clearly tend to 0 with time. For the last round, in particular, the average contribution is almost the same for both CT and SIT.

In addition to this remaining gap between behavior and prediction in these three games, it is noteworthy that in the two other games tested, no difference at all seems to exist between those two treatments, as explored in the following subsection.

4.3 Games with no treatment effect

In the **Prisoner's Dilemma**, the proportion of subjects who chose to cooperate reached 45% under SIT vs. 42% in CT, a non-significant difference ($\chi^2 = .0074$, p=.93). A logit model (Table 6) leads to the same conclusion. Interestingly, the PD is the game where the SIT mechanism should provide the cleanest control over social preferences: no signaling is possible and the theoretical prediction corresponds to the simplest game-theoretic notion, that is the strict dominance of a strategy.

In the **Ultimatum Game** subjects offered a bit less on average in the SIT, namely 19.91 against 21.02 under CT out of 50 tokens available, but the difference in distributions does not reach any conventional significance level (Mann-Whitney's W=1953.5, p=.19). This is illustrated by Fig. 5. Responder behavior does not seem affected by treatment either: the Minimal Acceptable Offer is slightly higher in SIT than in CT (14.14 vs 13.06) and the difference is not significant (two-sided Mann-Whitney's W=1720.5, p=.71). The results are displayed in Fig. 6. For both first and second player's behavior, an OLS regression (Table 6) does not yield significant results regarding treatment, even though the order of play had a significant impact

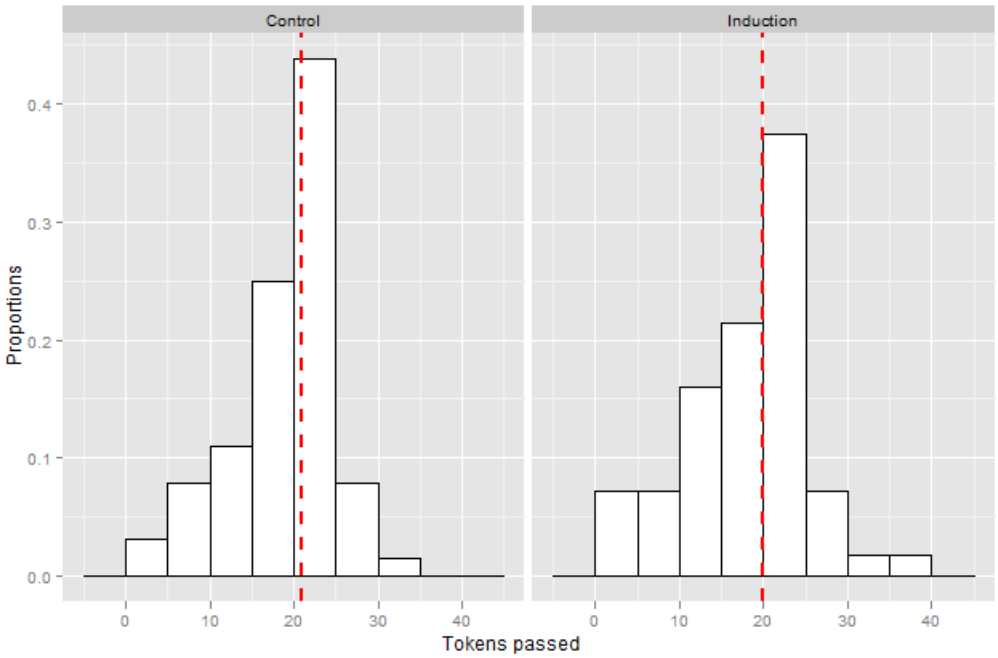

Figure 5: *Distribution of offers in the UG*

for both players, and studying economics on the second player's behavior.[14]

4.4 An overall effect

The induction procedure generates a shift towards subjects' more selfish choices in at least three different games and the opposite effect is never observed. To assess the overall impact, we computed a composite index of 'selfishness', or more specifically a measure of how distant subjects' choices are from the game theoretical solutions (Nash equilibrium or subgame perfect equilibrium for sequential games). To do so, we constructed a standardized index of relative compliance with the theoretical prediction, ranging from 0 (for largest observed deviations across treatments from the selfish equilibrium in a given game) to 1 (for equilibrium plays).[15] We then averaged the index

[14]The frequency of the subgame perfect equilibrium play cannot even be compared across treatments, since no player chose less than five tokens in the CT and only two did so in the SIT. A looser definition of "almost selfish play" does not yield any significant result in a logit analysis.

[15]An alternative would be to assign the value of 0 to the largest deviation in the strategy space (even if it was never actually played), but we think it would reduce comparability across games: for instance contributing 100% of one's endowment in the PGG appears to

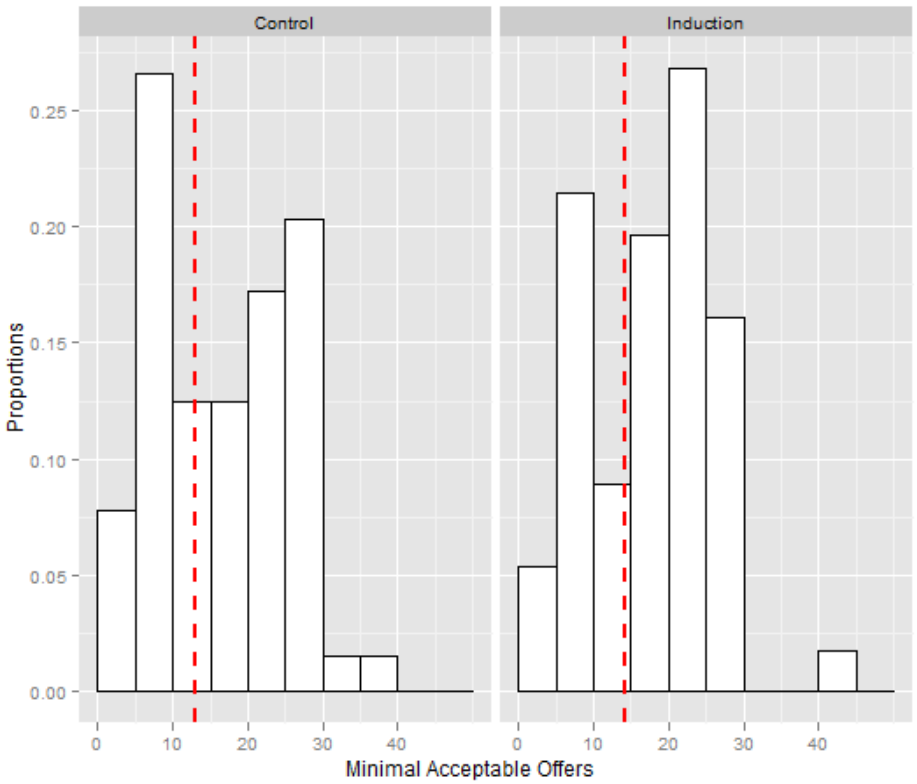

Figure 6: *Distribution of minimal acceptable offers in the UG*

over all decisions made by each individual player in the various games, and various roles in the game, such that each game played had identical weight (two roles in the same game being considered as two games). When the strategy method was used (*e.g.* TG) or when several decisions were made (*e.g* PGG) then the average index for all possible cases was taken as the index for the particular game. The distribution of the (subject-specific) values of the index is shown in Fig. 7.

A linear regression of this index with the same explanatory variables as for individual games gives the results displayed in Table 6. These results suggest indeed that our selfishness induction design had an overall effect on subjects, and that in this treatment they tended to play more in line with game theoretical solutions with selfishness assumed. This effect though was quite moderate, especially if contrasted with predictions for most games.

be less of a deviation from the prediction than keeping 0 in the UG.

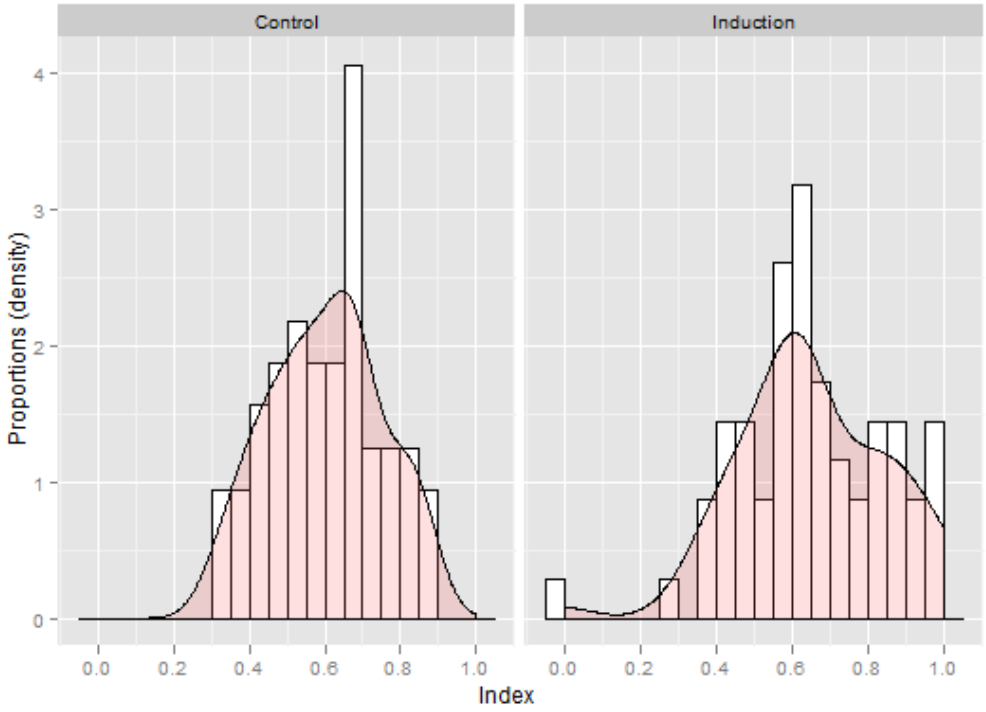

Figure 7: *Distribution of the selfishness index*

5 Discussion

On a general note, using our selfishness induction method does bring subjects closer to the prediction. Therefore, social preferences covered by our mechanism are likely to explain part of the departure from theoretically expected play. Still for a large share, the deviation cannot be accounted for by this type of preferences. First, the treatment effect varies between games and roles. Second, even for the games where the SIT mechanism seems to make a difference, a large share of subjects still depart from the theoretical prediction. In the PGG, where the effect is strong, in the first round 32 out of 56 SIT subjects contribute more than 10% of their endowment, and average contribution approximately equals one third of the endowment. In the TG, first movers pass on average nearly 40% of the endowment, and three quarters demonstrate some trust by passing a positive amount. The picture is even more striking in the DG, where subjects give on average 28% of the total sum, and only 30% of subjects give less than 10% of the total sum, with more than a fifth choosing the equal split. Overall, regarding the selfishness index, the average index is .65 for the SIT to be compared with .60 in the

CT.

Three types of explanations can be put forth: First of all, the relative weakness of the difference in the Control and Induction treatments could be driven by some experimental artifact: the SIT may have side-effects or work only on a specific subsample of subjects. Alternatively, it could be that the types of social preferences theoretically covered by the SIT method do not explain a significant part of the usual departures from predictions in experimental conditions. This could be because subjects' social preferences are of a type that does not belong to the class covered by the method; or, more radically, subjects might depart from the predictions for reasons other than their social preferences. We discuss these three views in the following subsections.

5.1 Side effects, behavior by types and signaling

One disadvantage of the method is that in increases complexity of the situation. This might affect subjects' behavior: the overall situation being more complicated, individuals may be more prone to errors, which could result in increased variance. This effect could drive deviations from predictions, particularly in the cases in which the latter are at one end of the strategy space: larger or more frequent errors would result in an apparent departure from the theoretical benchmark. We do not have much evidence for such an increase in frequency of 'unreasonable' play: 7% of subjects passed more than half the tokens in the dictator for SIT, to be compared with 4% for CT ($\chi^2 = 0.03$, $p = .85$), in the UG the figures are 9 and 10% ($\chi^2 = 0.03$, $p = .94$). For the TG, only 12% versus 6% sent back more than half the tokens in the case of an initial sending of 15 tokens by player 1 ($\chi^2 = 0.75$, $p = .39$). Overall, this gives little support to the idea that errors are larger (and more prevalent) in the case of SIT.

A possibly more promising explanation is that the aggregate picture hides major differences between B-players and A-players, with the former being the ones departing strongly from the theoretical predictions in the SIT (as is theoretically possible). One straightforward reason is "ex ante social preferences" understood as social preference applying to *expected* (monetary) payoffs (Trautmann, 2009; Krawczyk, 2011): someone who prefers B to A in the first stage may still exhibit some social concerns in expected payoffs. After having selected B and realized she is ahead of others in Stage 2, she may feel bad about having too high a chance of leaving others with nothing

(because of efficiency, altruism or inequity aversion). This would mean that only B-types do depart from selfish play in the Stage 2 games.

One way to check for that is to see whether the difference between both treatments is mostly explained by A-players rather than Bs. We can also check whether the divergence from game-theoretic predictions in Stage 2 is due to behavior of B-players. That is not what we observe: contrary to this explanation, B-players are, if anything, more 'selfish' on average in the second part of the experiment. For instance in the DG they give on average 10.18 tokens to be compared with 11.97, W=424, p=.18, and similar results are obtained for all games. Overall, the selfishness index of the B-players is higher than their A counterparts (.71 versus .60, $p < .001$ for both Student's t test and Mann-Whitney). These observations lead us to believe that the gap between game-theoretic predictions and observed behavior is not mostly driven by B-subjects.

Another promising explanation for this remaining discrepancy between actual play and the theoretical prediction can be that subjects use their position as first player to signal their type in the first stage of the experiment. Again, this would be an artifact of the method – an additional trigger of cooperative behavior that is absent in the standard, control game. Signaling could also be taking place in early rounds of the PGG. Because, as we have shown before, A types care relatively more about others' types, if such signaling was a significant factor, we would expect A types to exhibit strong conditional cooperation. In fact, panel regression analyzes do not suggest any difference between A's and B's propensity to reciprocate cooperation.

In the UG, the fact that subjects depart from the subgame-perfect equilibrium in the SIT as much as in the control treatment could possibly correspond to a separating equilibrium (that is, A-players signaling in a costly way their types). Intuitively, playing selfishly in the UG could lead the opponent to think that the first player is not a 'nice guy' and likely has chosen allocation B. If such belief updating is sufficiently strong, rejection could actually increase responder's expected payoff. Empirically, though, we observe very little evidence, if any, of such a phenomenon. A-players on average pass 18.77 tokens while B-players 20.65 ($p = .39$ for Student's t test), and the minimum acceptable offer averages 14.54 tokens for B-players versus 13.88 for A-players ($p = .79$ for Student's t test). The behaviors of the two types are virtually indistinguishable in the data, giving little support to an explanation in terms of signaling.

For the TG, the subgame perfect equilibrium is also clear (see Proposition 2) but it could be that signaling for A-players is in effect not a too costly strategy: for instance, if B-players send nothing in almost all the cases, and almost all A-players send some positive amount, then the second players would be able to distinguish both types and A-players could send back a lot in the case of a large initial contribution at almost no cost and not to do so when facing B-players. Overall, we do observe that A-players depart from equilibrium more than B's, so that some signaling may occur. Yet, based on empirical distributions of play and types in the TG, signaling for A-players is never an empirically optimal strategy: the probability to see A implemented is greatest when the first mover, of type A, sends nothing; it then decreases slowly with the tokens sent(from a probability of .69 when sending nothing to a .63 when sending 15 tokens). Moreover, for signaling to be an efficient strategy for A-players, the second players should be able to identify the types of first player based on their actions. Based on empirical distributions of actions and types, this turns out to be quite difficult: the proportion of Bs varies for non-zero amounts sent from 55.6 % to 85.7%, in a non-monotonic way. In particular, sending 9 tokens is associated with highest proportion of B-types, while 12 tokens–the lowest. The proportion of B-players among those first players who sent something is 28.6%, to be compared with 39.3% for the whole population of first players. So, there may be some attempt at signaling but most first player strategies are also followed by a significant share of Bs, rendering the identification of player 1's type very uncertain for the second player.

In sum, we find little evidence of a separating equilibrium for games where signaling could be a rational strategy (repeated PGG) and some evidence in the TG, where it is unexpected. To a certain extent, it could explain variations between games: in the UG, if the first player's move is taken as a signal of her type, then the second player would conclude that the cost of rejecting a small offer is quite low (in particular in comparison with the CT) because the dummy player is relatively more likely to be of type A than the first player. Rejecting low offers may hence represent some acceptable risk.

5.2 Alternative types of social preferences

An obvious candidate for a class of social preferences not fully covered by the method that could explain the remaining gap between behaviors and predictions is reciprocity. In section 2, we only suggested that our method would reduce, and not necessarily remove, reciprocal concerns. It is tempting to interpret the partial effect of our method as compatible with reciprocity-based

preferences. Although theoretically possible, the details of the results suggest that if relevant, reciprocity-driven behavior is only part of the explanation. Indeed, several specific features fit only weakly with such an explanation. In the DG for instance, where no reciprocity-driven actions can be expected, choices in SIT are still quite far from the theoretical prediction. In games where the role of reciprocity is expected to be strong (PD,UG,TG,PGG), we find mixed results: quite a strong effect of selfishness induction mechanism in the PGG and for the first player of the TG, and no effect for the PD, the UG and the second player of the TG.

More critically perhaps, to account for our results in SIT, reciprocity would fundamentally need to apply not to final consequences (even in expected terms) but to probability payoff. For instance, interpreting the observed payback by A-types of Player 2 of the TG in terms of positive reciprocity requires that *even if the expected payoff of the first player will stay the same* increasing her probability payoff is more benevolent than not doing so. One possible interpretation is that some symbolic reciprocity comes into play: the second player accepts to bear a cost (in expected terms given that he is not sure of player 1's type) just to send an 'acknowledgment' message, which will have no material consequence to A-type of Player 1. More generally, this would mean that *reciprocity, and by extension social preferences, go deeper than usually assumed by models* because the signals are what matters rather than the ultimate consequences.

Pushing this interpretation to its extreme, our results could mean that what matters for social preferences are not mainly the consequences of a chosen actions but the actions themselves. In a way, the classical debate between consequentialism and deontologism in ethics echoes this issue: what may matter, for an action to be good, is for a large part its proximity with a normative benchmark, or its conformity with a normative rule.[16] In this sense the consequences are secondary, but prosociality would imply to adopt an action not too far from a moral standard: sharing (or splitting equally) in the DG, doing one's share in cooperation, trusting others, etc. How prosocial an action is may not primarily be measured in terms of the final consequences but within the structure of the game. In other words, the structure

[16]In philosophy, one of the most prominent advocate of a deontological approach to ethics is Kant. In his view, lying for instance is never a moral action, regardless of the consequences. In his famous example, lying is still immoral, even though telling the truth results in the death of an innocent. Empirically, several studies have put forth some aversion to lying (Gneezy, 2005; Erat and Gneezy, 2012) or cheating, that cannot be explained by consequentialist motives.

of the interaction may matter more than the consequences. This would quite straightforwardly explain that behaviors are only mildly different in SIT than in CT.

In a similar spirit, social norms (Henrich, Boyd, Bowles, Camerer, Fehr, and Gintis, 2004) may provide an explanation for our results. In general, a social norm entails a specific type of behavior; following the norm one may not choose the action optimal from the point of view of some social principle (such as efficiency or equity). If norms are just broad principles to be followed, such as "share equally", limited behavioral difference between the case of playing for probabilities to see one's preferred allocation and playing directly for monetary units is to be expected. In this view, certain situations call for certain conduct. In support of this interpretation, there is some evidence (Gneezy and Guth, 2003) that the equal split in bargaining situations applies even to asymmetric interactions, where equity would be best approached by unequal split of the available resources.

5.3 Alternatives to social preferences

An alternative explanation may stem from the fact that subjects do not conceive of strategic interactions the way game theory assumes rational players do. One possible deviation is that subjects reason collectively, as if all players were a team (Sugden, 1993). It may also well be that they use heuristics or other cognitive short-cuts (Simon, 1982; Selten, 1998). Individuals may fail to reason consequentially when the outcome is uncertain: Shafir and Tversky (1992) provide some evidence that it may be the case for the Prisoner's dilemma. In the absence of opportunities to learn in our experiment, the observed behavior may reveal that individuals tend to approach cognitively strategic interactions in a different way from the strategic rationality postulated by game theory.

It cannot be entirely excluded either that some players deemed it fair to give everyone some chance to co-decide about final allocation. However, voluminous research in social psychology tends to show that already purely symbolic (non-instrumental) "voice" given to parties involved in an interaction tends to satisfy the need for a fair procedure (Lind and Tyler, 1988). It would thus seem that once all players submitted their preferred allocation in Stage 1, it may not be necessary that each has a positive number of tokens.

6 Conclusion

Our results suggest that although part of the gap between theoretical predictions and usual experimental game results is driven by social preferences, an important share of this deviation remains to be explained, a point already put forth by Andreoni (1995) and Andreoni and Blanchard (2006). As a consequence, it raises the question of why subjects still depart from the theoretical solutions, an issue perhaps even more crucial than the explanatory power of social preferences: it may shed light on why, in usual experimental settings, subjects do not always play according to predictions.

Our results seem to indicate that only part of the discrepancy between typically observed behavior in games and standard equilibrium predictions may be "blamed" on social preferences. They suggest that dropping the assumption of selfishness alone, especially by allowing outcome-based preference only, will not help us dramatically improve predictions for experimental games. They also suggest, in line with other research, that other forms of social preferences, or alternative explanations, may form a promising avenue for future research.

References

Andreoni, J. (1988): "Why free ride?: strategies and learning in public goods experiments," *Journal of Public Economics*, 37(3), 291–304.

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

Andreoni, J., and E. Blanchard (2006): "Testing subgame perfection apart from fairness in ultimatum games," *Experimental Economics*, 9, 307–321.

Berg, J., T. A. Rietz, and J. W. Dickhaut (2008): "Handbook of Experimental Economics Results," chap. On the Performance of the Lottery Procedure for Controlling Risk Preferences. Elsevier, North-Holland.

Blanco, M., D. Engelmann, and H. T. Normann (2011): "A within-subject analysis of other-regarding preferences," *Games and Economic Behavior*, 72(2), 321–338.

Bolton, G. E., J. Brandts, and A. Ockenfels (2005): "Fair procedures: evidence from games involving lotteries," *Economic Journal*, 115, 1054–1076.

Bolton, G. E., and A. Ockenfels (2000): "ERC: a theory of equity, reciprocity and competition," *American Economic Review*, 90, 166–193.

Brock, M. J., A. Lange, and E. J. Ozbay (2013): "Dictating the Risk: Experimental Evidence on Giving in Risky Environments," *American Economic Review*, 103(1), 415–437.

Camerer, C. (2003): *Behavioral game theory. Experiments in strategic interactions*. Princeton, N.J.: Princeton University Press.

Charness, G., and M. Rabin (2002): "Understanding social preferences with simple tests," *Quartely Journal of Economics*, 117, 817–869.

Cox, J., D. Friedman, and S. Gjerstad (2007): "A tractable model of reciprocity and fairness," *Games and Economic Behavior*, 55, 17–45.

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

Dufwenberg, M., and G. Kirchsteiger (2004): "A Theory of sequential reciprocity," *Games and Economic Behavior*, 47, 268–298.

Engelmann, D., and M. Strobel (2004): "Inequality aversion, efficiency, and maximin preferences in simple distribution experiments," *American Economic Review*, 94, 857–869.

Erat, S., and U. Gneezy (2012): "White lies," *Management Science*, 58(4), 723–733.

Falk, A., and U. Fischbacher (2006): "A theory of reciprocity," *Games and Economic Behavior*, 54, 293–315.

Fehr, E., and S. Gaechter (2000): "Cooperation and punishment in public goods experiments," *American Economic Review*, 90, 980–994.

Fehr, E., and K. M. Schmidt (1999): "A theory of fairness, competition, and cooperation," *Quarterly Journal of Economics*, 114, 817–868.

Gneezy, U. (2005): "Deception: The role of consequences," *American Economic Review*, pp. 384–394.

Gneezy, U., and W. Guth (2003): "On competing rewards standards—an experimental study of ultimatum bargaining," *The Journal of Socio-Economics*, 31(6), 599–607.

Greiner, B. (2004): "An Online Recruitment System for Economic Experiments," in *Forschung und wissenschaftliches Rechnen 2003*, ed. by K. Kremer, and V. Macho. Göttingen : Ges. für Wiss. Datenverarbeitung,.

Henrich, J., R. Boyd, S. Bowles, C. Camerer, E. Fehr, and H. Gintis (2004): *Foundations of human sociality: Economic experiments and ethnographic evidence from fifteen small-scale societies*. Oxford, U.K.: Oxford University Press.

Krawczyk, M. (2011): "A model of procedural and distributive fairness," *Theory and decision*, 70, 111–128.

Krawczyk, M., and F. Le Lec (2010): "'Give me a chance!' An experiment in social decision under risk," *Experimental Economics*, 13(4), 500–511.

Levine, D. K. (1998): "Modeling altruism and spitefulness in experiments," *Review of Economic Dynamics*, 1(3), 593–622.

Lind, E. A., and T. R. Tyler (1988): *The social psychology of procedural justice*. Springer.

Rabin, M. (1993): "Incorporating fairness into game theory and economics," *American Economic Review*, 83, 1281–1302.

Roth, A., and M. Malouf (1979): "Game-Theoretic Models and the Role of Information in Bargaining," *Psychological Review*, 86(6), 574–594.

Selten, R. (1967): "Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperiments," in *Beiträge zur experimentellen Wirtschaftsforschung*, ed. by H. Sauermann. Tübingen: Mohr.

——— (1975): "Reexamination of the perfectness concept for equilibrium points in extensive games," *International journal of game theory*, 4(1), 25–55.

——— (1998): "Features of experimentally observed bounded rationality," *European Economic Review*, 42, 413–436.

Shafir, E., and A. Tversky (1992): "Thinking through uncertainty: Nonconsequential reasoning and choice," *Cognitive psychology*, 24(4), 449–474.

Simon, H. (1982): *Models of bounded rationality*. Cambridge MA: MIT Press.

Sugden, R. (1993): "Thinking as a team: toward an explanation of non-selfish behavior," *Social Philosophy and Policy*, 10, 69–89.

Trautmann, S. (2009): "A tractable model of process fairness under risk," *Journal of Economic Psychology*, 30, 803–813.