University of Warsaw
Faculty of Economic Sciences

Working Papers

No. 6/2012 (72)

Michał Krawczyk
Fabrice Le Lec

Testing game theory without the social preference confound

Warsaw 2012

University of Warsaw
Faculty of Economic Sciences

Testing game theory without the social preference confound

Michał Krawczyk
University of Warsaw
Faculty of Economic Sciences
e-mail: mkrawczyk@wne.uw.edu.pl

Fabrice Le Lec
Catholic University of Lille
Lille Economie & Management
UMR CNRS 8179
e-mail: fabrice.lelec@icl-lille.fr

Abstract

Abstract: We propose an experimental method whose purpose is to induce selfish behavior in games for a broad class of social preferences. It provides a theoretical framework for testing game theoretical predictions by confronting subjects with a commonly known payoff matrix actually representing their preferences. The paper describes the empirical tests of this method based on the comparison of results from several popular experimental games played with and without our methodology. Apart from it being a test of validity of the method, our experiment helps answer the question of how useful social preferences could be in explaining commonly observed deviations from selfish rationality. Results suggest that our method does induce more selfish behaviors: a substantial part of the difference between predictions based on selfishness and observed behaviors seems indeed driven by such preferences. But they also indicate that a considerable share is left untouched, perhaps giving weight to alternative explanations.

Keywords:
social preference, experimental game theory, ultimatum game, public goods game, trust game, prisoner's dilemma, dictator game

JEL:
A13, C65, C72, D63, D03

Acknowledgements:
Marcin Malawski has made helpful comments on the manuscript and Eliza Subotowicz has proofread it. All the remaining errors are ours.

1 Introduction

Several years ago, Gary Becker proposed his now-famous Rotten Kid Theorem showing how one can induce one's offspring to behave less selfishly. What experimentalists working with game-theoretic models would need, however, is roughly the reverse. Indeed, one of the core assumptions of most game theoretic developments is that the payoff matrix represents the preferences of players. Testing the accuracy of game theory as a model of human behavior requires therefore the ability to control for this auxiliary assumption. It is often convenient to assume that monetary payoffs represent these preferences although it is in no way implied by the core of the theory (Binmore, 2005). Ample experimental evidence has been indeed gathered over the years indicating that the assumption that individuals are purely selfish is unlikely to hold, and models of social preferences have been proposed to account for that (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Levine, 1998; Cox, Friedman, and Gjerstad, 2007). The resulting difficulty for game-theoretic experiments has been long recognized (see e.g. section 2 of Goeree and Holt (2001)). One immediate consequence is that in the presence of such social preferences, it is impossible to experimentally test game theoretic concepts *per se* (such as backward induction or Nash equilibrium): what is theoretically required indeed is not only that payoffs represent preferences for individuals, but also common knowledge of this fact among the players. In a well-known theoretical article, Kreps, Milgrom, Roberts, and Wilson (1982) show that even a low level of uncertainty about other players' preferences allows to maintain cooperation in most rounds of a finitely repeated Prisoner's Dilemma game.

It is therefore difficult to disentangle apparent violations of game-theoretic solutions from plain departure from common knowledge of selfishness. Simply eliciting social preferences in a separate task without revealing it is unlikely to solve the problem since subjects will not know others' preferences, and so common knowledge cannot possibly hold. Eliciting preferences just to reveal them to all other players raises obvious strategic issues, possibly making the social preference elicitation task unreliable. We propose here an experimental method aligning, under reasonable assumptions, the values shown in the game matrix with agent's preferences and thus allowing to test for the convergence of game theoretic concepts and actual play by human subjects. In addition to its usefulness as an experimental procedure, our methodology also provides a direct test for various developments in behavioral game theory, in particular models of social preferences.

The main feature of the method is to ask subjects to state their preferred social allocation of money within a group of players, including themselves, and then have them play games where the payoffs are probabilities to see their own (rather than somebody else's) allocation implemented in the end. For any outcome-based utility model, this implies that maximizing payoffs (that is playing 'selfishly') is a weakly dominant strategy as long as the player believes that there is some probability that the opponent has chosen a different allocation from hers. The mechanism is also transparent in terms of others' preferences in the sense that all subjects face the same situation, and then it should be immediately clear what others' preferences are.

Using this methodology, 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 of common knowledge of selfish rationality when the 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, if present, is still far from inducing behaviors in conformity with game-theoretical solutions. Although only part of the deviation between experimental behavior and selfishness-based game theory may be understood in terms of outcome-based other-regarding preference, the same can be said of some alternative explanations, such as reciprocity and insufficient learning.

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

At the beginning of the experiment, subjects are told they would be divided into groups of n, whereby $n-1$ is the (largest) group size in the actual experimental games in which the experimenter seeks to induce selfishness. In Stage 1 of the experiment, each subject is asked to choose between two allocations: Allocation A giving the same payoff to every group member and Allocation B giving a somewhat higher amount to herself and a low payoff (*e.g.* zero) to all other group members. These numbers should be calibrated

to the subject pool at hand in such a way that a substantial majority of subjects go for the fair Allocation A, yet subjects perceive it to be quite likely that any subject they interact with could be of the selfish type. For instance, for $n = 5$, $A = (40, 40, 40, 40, 40)$, $B_1 = (55, 0, 0, 0, 0)$, $B_2 = (0, 55, 0, 0, 0)$ etc. would be appropriate, since all existing models of social preferences point to A (inequity aversion, maximin, revealed altruism, kindness within reciprocity-based models and efficiency). None of these choices are revealed to any other subject.

In Stage 2, $n - 1$ subjects play one or several games, while the n^{th} subject stays inactive (and could be asked, for instance, to make equivalent yet hypothetical choices). The payoffs in the experimental games are expressed in terms of lottery tickets.[1] Each ticket represents one percent chance that a given player's selected allocation, A or B, will be implemented. The game is designed in such a way that the joint payoff cannot exceed one hundred.[2] Suppose for example that $n = 3$ and in Stage 2 subject 1 earned 23 points and subject 2 earned 48 points. Then a random device implements subject 1's chosen allocation with probability 23%, subject 2's with probability 48% and subject 3's with probability 29% $(100 - 23 - 48)$.

For outcome-based preferences, i.e. preferences dependent on final allocations only, one should intuitively choose the preferred allocation in Stage 1 and then maximize the probability of getting it, that is, maximize the number of tickets. It is so since, obviously, others may have chosen one's non-preferred allocation. It is worth noting that as long as nothing is known about others' preference and thus their selected allocations, the subject does not care who gets to decide if it is not herself.

Formally, suppose that Player 1's selected allocation is $\theta_1 \in \{A, B\}$. Then, given uncertainty about others' choices, her preferences are represented by:[3]

$$U(q) = \pi_1 u_1(\theta_1) + \sum_{2 \leq k \leq n} [\pi_k(\mu_{1k} u_1(A) + (1 - \mu_{1k}) u_1(B_k)],$$

[1] In the present experiment we call them "tokens" instead because "lottery tickets" would sound odd in the control treatment, where their value was fixed.

[2] This may be seen as a limitation of the method. However, we are not aware of an experiment testing game-theoretic predictions that would place no limit on subjects' earnings (and we would be rather afraid to conduct one!)

[3] For the sake of simplicity we consider that the agent is an expected utility maximizer, but the same reasoning holds for any other model of preferences under risk that obeys stochastic dominance.

where $\pi = (\pi_1, \pi_2, ..., q_n)$, with $\sum_{i=1}^{n} \pi_i = 1$ is the vector of probabilities of having each player's chosen allocation implemented (the payoffs in lottery tickets), and μ_{1k} is the subjective probability that Player 1 holds about player k having chosen Allocation A.[4] Since all other players are anonymous, the initial beliefs are uniform,[5] $\mu_{12} = \mu_{13} = \ldots \mu_{1n} < 1$, the inequality being a minimalistic pessimism assumption. In a static game, there is no room for belief updating, so only equilibria consistent with selfish preferences in the stage game can arise. Consider the one-shot Prisoner's Dilemma: if played directly for money, sufficient level of altruism or preference for efficiency can make mutual cooperation the unique equilibrium, and sufficient level of (advantageous) inequality aversion will make two pure strategy equilibria possible: mutual cooperation and mutual defection. With the use of the proposed procedure, only mutual defection can happen, no matter what subjects' preference concerning distributions is, as long as subjects hold reasonably pessimistic beliefs about others' types.

One concern about the reliability of the procedure in the case of *dynamic* games is that the first player could use a strategy signaling her Stage-1 choice. Costly signaling can be rational, if the change in beliefs implied for the second player yields some benefit for the first player, i.e., make the second player allocate her a higher payoff. It turns out, however, that it is predominantly A-types that are willing to sacrifice their own payoffs to increase payoffs of others who are revealed to be of type A. Thus, as a result of the costly signal, it is the selfish type of the other player that gets relatively higher payoffs, so that the signal does not pay in the end.[6]

The issue is more complex for reciprocity-based preference; existing models add a significant layer of complexity both in one-shot (Rabin, 1993) and sequential games (Dufwenberg and Kirchsteiger, 2004; Charness and Rabin, 2002) and are possibly intractable in the overall incomplete information environment implied by our method. However, the same line of reasoning seems to hold and render it very unlikely that rational players with reciprocity-based preferences would depart from the stage 2 game-theoretical solution.

[4]For the sake of conciseness, we will call individuals preferring Allocation A "type A individuals" or simply "A-types" and similarly for those choosing B.

[5]It is not strictly speaking required that prior beliefs are uniform for every other player, but since players are matched randomly and anonymously, this must be the case in our experimental situation.

[6]A rigorous presentation of this argument is provided in the Appendix, based on three innocuous assumptions: reasonable pessimism, common priors regarding anonymous subjects, and no separating equilibrium without incentives.

Indeed, it may appear that assigning tickets to another player would be generally considered a benevolent action, possibly triggering positive reciprocity, but it is unclear why a second player of type A should reciprocate: first, she could end up assigning tickets to the first player who turns out to be of type B and whose apparently kind action was actually a subterfuge to hide a previously not so kind one; second, she may reason that she has already reciprocated by choosing action A so that keeping all the tickets for herself is not in any way meaner than to give some of them to the first player. As a consequence, reciprocity as a motive does not seem to be a very convincing driver of significant departure from the theoretical solution of the second stage game.

One may wonder whether the menu of distributions in Stage 1 is optimal. For outcome-based models, the main criterion is to maximize the utility difference between player's preferred (and thus selected) distribution and what she expects another player to choose, so that effective stakes are possibly high. This, however, is easier said than done. One could for instance consider giving an option to choose from a continuum, *i.e.* freely allocate a fixed amount among all the players, thus maximizing the utility of the selected distribution given research budget constraints. The downside of this solution is that it would seriously complicate the design, making the mechanism more difficult to understand for the subjects. Additionally, subjects could then possibly expect that it is less likely than with our A/B menu that they end up with nothing, thereby reducing the aforementioned utility difference. Such a solution would also be a source of problems if subjects try to guess and roughly match others' distribution choices and then possibly behave non-selfishly in Stage 2. While this is also a concern in our design, it takes more pessimism on the part of the subjects to behave in such a way.

To summarize, under common knowledge of outcome-based preference, the mechanism will induce complete selfishness in the games of Stage 2, be they simultaneous or sequential. Reciprocity might interfere with our method of selfishness induction, even though it is far from clear what the notion of positive reciprocation exactly means in our context. To the extent that the motivations to behave "non-selfishly" are either obscure or far-fetched, if not simply irrational, we expect the subjects to recognize the predominance of selfish actions and they will expect others to recognize it as well, establishing something very close to common knowledge of selfishness.

3 Design

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

The experiment was preceded by hypothetical pilots concerning stage I 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 the predicted fraction of individuals choosing the selfish option was substantially higher than the actual one, which clearly contributes to increasing the "stakes" in the SIT—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]

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

[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 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.

non-selfish choices and, second, that the sum for the active players would never exceed 100.

In line with the concept describe above, on fifth of our subjects constantly played the role of an inactive dummy player, residual claimant of tokens. These participants were asked to make analogous, yet hypothetical decisions and no feedback was given. The active participants were actually divided into two-person sub-groups 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 the Ultimatum Game using Minimal Acceptable Offer strategy method (Selten, 1967). The third round was a Trust Game with 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 through 6 were analogous to rounds 1 through 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 facing 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. 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 this kind 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).

Throughout the experiment participants could consult the printed general instructions, explaining 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 subjects were told which round it was and what their resulting payoffs were, including the

5 PLN show-up fee. 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, similarly to the 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.

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 in the SIT. An alternative simple student job would pay 10-15 PLN per hour.

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

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

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 in the second round, this percentage increasing with the money passed by the first player. 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 quite high, reaching three quarters for the entire sample. That has the convenient feature of providing numerous valid observations from our subjects' sample but also of having a proportion of B-choosers that is not negligible, so that the probability to face a selfish player is not minute, a pattern necessary for our preference induction mechanism to work.[9]

We present two types of analyses: first, we run standard non-parametric tests of the treatment effect on the central tendency shown in the data. Then, we run an OLS regression with the strategy chosen in one 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 homogeneous comparisons for the different games, we use exactly the same explanatory variables in all cases: we kept all the variables that seemed to play a significant role in at least one game. We have also conducted a logit analysis on the occurrence of selfish play (defined as being less than 10 % away from the game-theoretical prediction, these 10 % being scaled on the

[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.

[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.

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 A shift towards more selfish play...

In the **Dictator Game**, our mechanism induced more selfish behavior: the subjects give on average 11.27 PLN (28.17% of the total of 40 PLN), which is 17.5% 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) equal

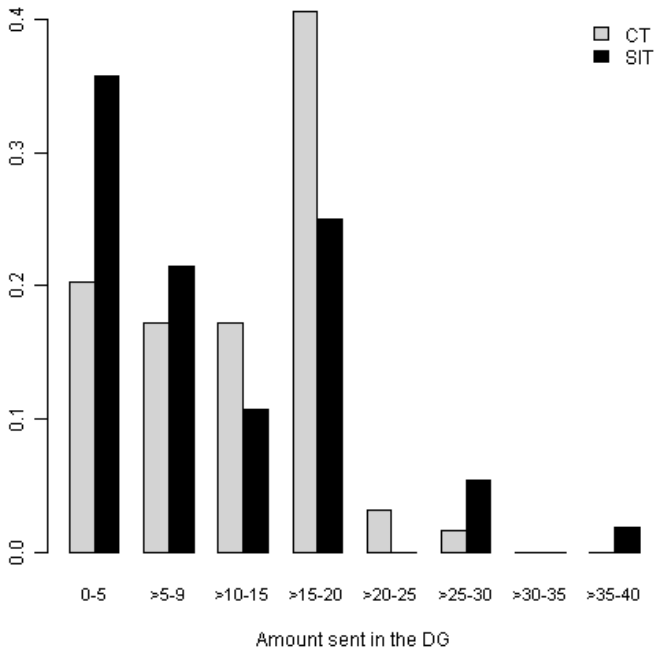

Figure 1: *Distribution of offers in the DG*

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 1: 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.

Amount passed in DG	Coef. Est.	SE	t value	p-value
SIT	-3.391*	1.774	-1.912	.059
Order	-4.668***	1.500	-3.111	.002
Econ major	-4.162***	1.535	-2.711	.008

*= significant at the .10 level, **= significant at the .05 level, ***= significant at the .01 level, n=119, control variables (not displayed): session 1, gender and income.

Table 1: *OLS regression on passed money in the Dictator Game*

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 approx. 21%. Again, this gap is statistically significant using a Mann-Whitney test (W=2138, p=.03). The results are displayed in Fig. 2. Once again, an analysis of the occurrence of ra-

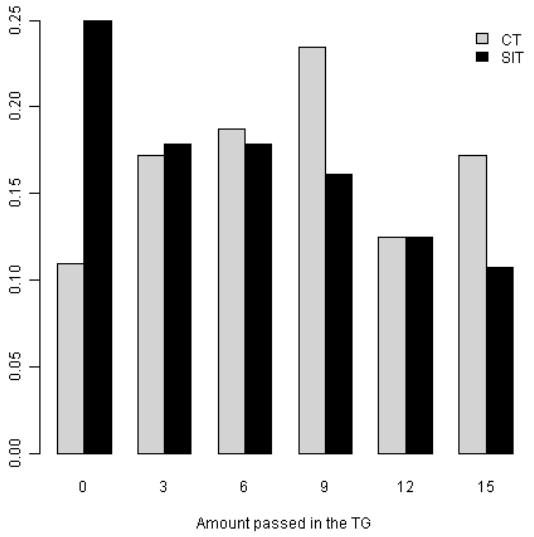

Figure 2: *Distribution of amounts of money sent in the TG*

tional 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. And 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 (Table 2).

Amount passed in TG	Coef. Est.	SE	t value	p-value
SIT	-1.611	1.045	-1.542	.126
Econ major	-2.245**	.905	-2.482	.015
Gender	2.077**	.901	2.287	.024

*=significant at the .10 level, **= significant at the .05 level, ***= significant at the .01 level, n=119, Control Variables (not displayed): Order, Session 1, Income

Table 2: *OLS regression of money sent in the TG, first player*

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 reveal any significant difference for the amount that could be sent either.

In the **Public Goods Game**, we also observe a difference between the two treatments: the average contribution is 1 PLN (out of an endowment of 2), whereas in the first round under CT, it only reaches .70 PLN (35%) in SIT (see Fig. 4). The difference is very significant (one-sided Mann-Whitney, W=1379, p=.01). All further rounds are also significantly different but 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 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 PLN on average per subject and round for SIT versus .82 for CT, p=.046 in a one-sided Mann-Whitney). Once again, at the individual

[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 reciprocity.

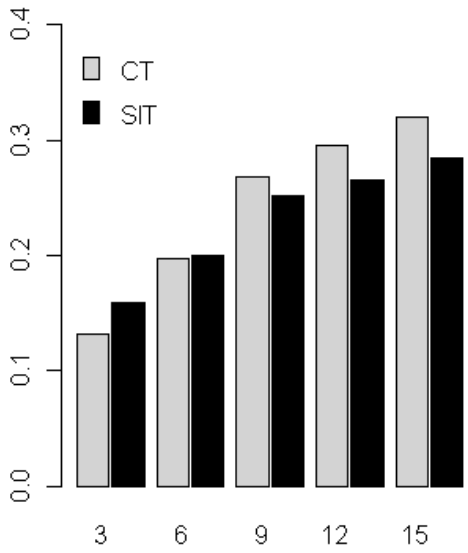

Figure 3: *Proportion of money sent back in the TG*

level, we also find that the proportion of 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 3).

Voluntary contribution in PGG (1st round)	Coef. Est.	SE	t value	p-value
SIT	-0.958*	.495	-1.936	.055
Dummy Economics and Management	-.568	.426	-1.333	.185

*=significant at the .10 level, **= significant at the .05 level, ***= significant at the .01 level, n=119, Control variables (not displayed): Order, Gender, Income, Session 1.

Table 3: *OLS regression for voluntary contributions in the PGG, first round*

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.

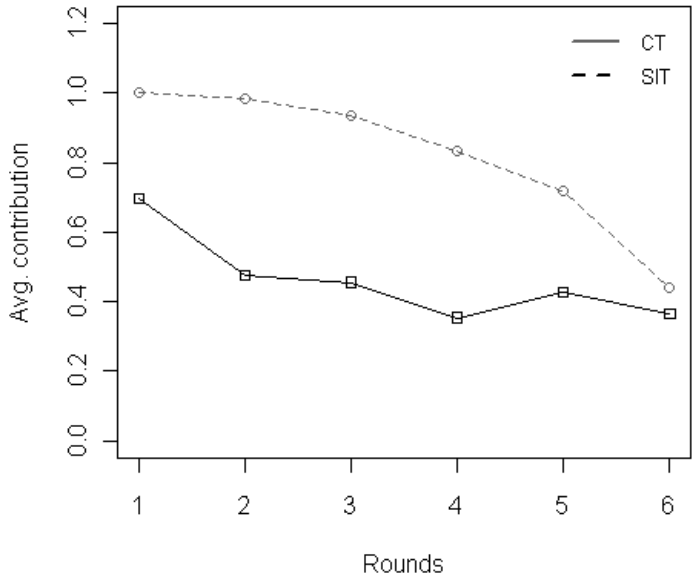

Figure 4: *Average contribution in the PGG by round*

In particular more individuals seem to behave (almost) in line with game-theoretic prediction. Yet, in the two other games tested, no such difference seems to exist between those two treatments.

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 4) leads to the same conclusion.

In the **Ultimatum Game** subjects offered a bit less on average in the SIT, namely 19.91 vs. 21.02 under CT out of 50 tokens available, but the difference does not reach the conventional significance level (Mann-Whitney's W=1953.5, p=.19). The distribution of offers shown in Fig. 5 reveal little difference. Responder behavior does not seem affected by treatment either: the minimal acceptable offer is slightly higher in SIT than in CT (13.06 vs 14.14) 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 does not yield significant results

Defection in PD	Estimate	SE	z value	p value
SIT	.221	.400	-.336	.737
Order	.083	.445	.496	.832
Session1	-.090	.592	-1.514	.130
Dummy Economics and Management	.940**	.394	2.384	.017
Gender	.032	.385	.083	.934

*=significant at the .10 level, **= significant at the .05 level, ***= significant at the .01 level,n=119, Income is dropped to maintain minimal number of observations in all cells

Table 4: *Logit analysis of occurrences of defection in the PD*

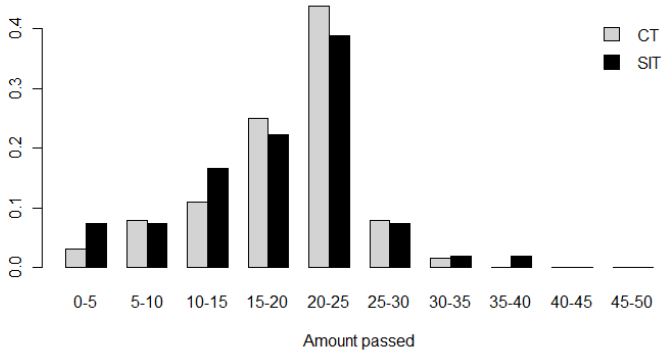

Figure 5: *Distribution of offers in the UG*

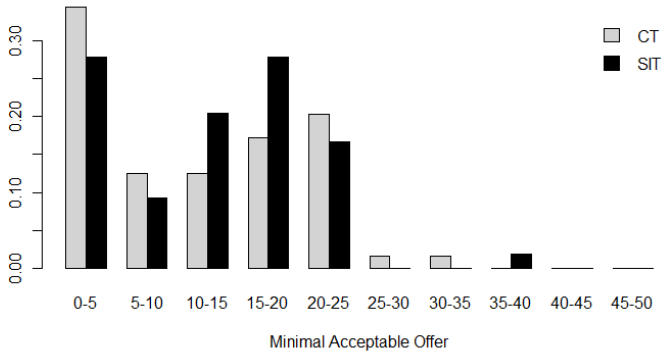

Figure 6: *Distribution of minimal acceptable offers in the UG*

regarding treatment, even though the order of play had a significant impact for both players, and studying economics on the second player's behavior.[13]

4.4 An overall effect

The induction procedure generates a shift towards subjects' more selfish choices in at least three different games (including quite a substantial one in the PGG) 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) in order to see if, overall, subjects in the SIT treatment did behave more selfishly and in line with the theory. To do so, we constructed a normalized index of relative compliance with the theoretical prediction, ranging from 0 (for largest deviations across treatments from the selfish equilibrium observed in a given game) to 1 (for equilibrium plays).[14] We then averaged the index 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 5. 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.

[13]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.

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

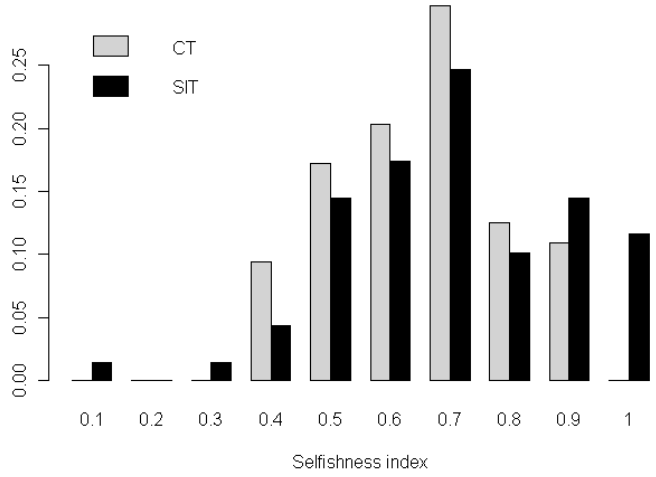

Figure 7: *Distribution of selfishness index*

Overall normalized distance to 'selfish' GT prediction	Coef. Est.	SE	t value	p-value
SIT	.073**	.033	2.185	.031
Order	.64**	.029	2.205	.029
Session1	-.062	.042	-1.469	.144
Dummy Economics and Management	.125***	.029	4.352	$< .001$

*=significant at the .10 level, **= significant at the .05 level, ***= significant at the .01 level, n=134, Control Variables (not displayed): Income and gender

Table 5: *OLS regression of selfishness index*

4.5 Departure from the theoretical predictions

Yet, the treatment effect varies a lot between games and roles: it is present in the DG, first move of the TG, first round of the PGG and (albeit not significantly so) in the first move of the UG. There is no impact in the PD, or the second stage of the TG or UG. And even if the SIT mechanism seems to make a difference, a large share of subjects still depart from the theoretical solution. Focusing only on the games where SIT implies a more selfish play, behaviors are still quite far from the game theoretical solution. 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 they contribute on average approximately one third of their 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 CT meaning that on average social preferences presumably controlled for only account for a modest share of the theoretical deviation.

To put it differently, using our selfishness induction method does bring subjects closer to the prediction, and so social preferences are likely to explain part of the departure from theoretically expected play, but for most part this discrepancy cannot be accounted for by preferences controlled for by our mechanism.

5 Discussion

These results raise two issues: the first one is to know determine whether the effect we measure, and its limited magnitude, is indeed related to a rather mild reaction of A-type subjects, and not linked to some unintended consequences on B-types. One could for instance raise the issue of "procedural preferences" understood as preference for equality in *expected* terms (Trautmann, 2009; Krawczyk, 2011): someone who prefers B to A in the first stage may still have some aversion to inequality 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. In any case, this would mean that only B-types do depart from selfish play in the Stage 2 game (one of the reasons why it is desirable for the experimenter to have a majority of A-types). In fact, with procedural players only, the first stage may effectively be turned into a coordination game with one payoff-dominant (all A) and the other risk-dominant (all B) equilibrium. Such B players could then explain why we do not observe perfect self-interested game theoretical behavior.[15] One way to check for that is to see whether the difference between both treatments is mostly explained by A players rather than Bs as

[15]In the same spirit, some pro-social, yet sufficient pessimism about others' choices, subjects could have chosen allocation B, but would have played in a prosocial way in the second stage, feeling a sense of guilt would have prefer not to play selfishly in the second stage. Since others are expected to choose B, the only way to balance for it on expected payoffs is to do the same and possibly play non-selfishly in Stage 2, depending on the course of the game.

well as to check whether the departure from game theoretical prediction in stage II is due to B-players.

The second issue is that such data call into question usual explanations of why subjects depart from the game-theoretical solution, at least those provided by social preferences as modeled in the literature. Before drawing strong conclusions about this, it seems sound to review which *other* phenomenon could explain this departure: we consider here reciprocal preferences and learning as alternative explanations.

5.1 Treatment effect and Stage-1 allocation choice

To deal with the first point, it is useful to know to what extent this discrepancy is driven by subjects who have chosen the B allocation in the first stage. As discussed earlier, they might experience the feeling of guilt related to this choice, which could result in the willingness to give others a chance to see their allocation (or the equal Allocation A) implemented. This would result in B subjects being less selfish in the second stage than the A 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).

These observations lead to believe that the gap between game-theoretical predictions and observed behavior is not mostly, if at all, driven by B-subjects. It suggests then that although part of the discrepancy 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 (outcome-based) social preferences: it may shed light on why, in usual experimental settings, subjects do not always play according to predictions. In particular we may consider two widely used theoretical explanations, namely learning and reciprocity-based preferences. In light of our results, they seem to fall short of explaining observed behavior.

5.2 Reciprocal preferences

Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006) as well as Charness and Rabin (2002) have proposed to explain non-selfish behavior

on the basis of preference for reciprocity: subjects seem to play nicely towards people who have been nice to them or to punish mean opponents, even if such actions are costly. Such an explanation would require specifying what being kind or unkind may mean in our setting. In particular, keeping the lottery tickets to oneself (rather than transferring them to someone else), after having selected the fair Allocation A, is not really unkind.

Putting this major issue aside, our results are still hard to account for with reciprocal preferences: in the DG for instance, where no reciprocity-driven actions can be expected, choices in SIT are still quite far from the theoretical prediction. And arguably, if reciprocity was the main driver of behavior, we should observe virtually no treatment effect not only in the UG or the PD (which is largely the case) but also in the TG or the PGG (which is not the case). So at best, reciprocity provides a partial explanation. Even when supplemented with outcome-based preferences as in Charness and Rabin (2002), it would fail to explain behaviors in the DG, but also differences between TG and UG, on the one hand, and PGG and PD, on the other hand.

5.3 Learning

One of the limitations of our experimental design is that it did not give subjects much ability to learn. It is often put forth that deviations from theoretical predictions decay as subjects become familiar with the situation and their opponents' actions (Camerer, 2003). How such an explanation could apply to a situation as simple as the DG is not so clear, but it could provide reasons why in more complicated games subjects fail to identify the rational solution. Yet, the results on the repeated PGG do not seem to support this insufficient learning hypothesis: in the last round, 18 out of 56 subjects still contribute more that 10% of their endowment and, except from round one to round two,[16] there is rather little evidence of a trend towards less cooperation. A Page's trend test on the last 5 rounds of SIT does not yield a significant result (L=2483.5, $p > .30$).

Obviously, our design does not allow us to answer this question directly and definitively; the experimental situation here is a bit different from the usual experimental designs aimed at assessing the effect of learning: the game is a repeated game with the same subjects, and the number of iterations is perhaps a bit low to draw definite conclusion. But it is worth noting that

[16]It is between rounds one and two that, as one can guess, subjects get their first information about what others tend to do in the group.

we do observe this trend for the control treatment, even with only 6 rounds, and with subjects playing within the same group. It is not clear in addition why learning in repeated games of cooperation should be less prevalent with strangers than with partners, especially when the cooperation level is already low at the very beginning. Explaining this lack of learning on the basis of reputation building seems to fall short since we do not observe (in contrast with the control treatment) the usual 'end-game effect'.

Another explanation for this relative distance from the theoretical prediction could be that subjects use their position as first player to signal their type in the first stage of the experiment. As put forth in the appendix and the second section, this signaling cannot be fully rational, but perhaps subjects do signal in a non-rational way: perhaps playing selfishly in the UG could lead the opponent to think that the first player is not a 'nice guy', and that, as a consequence, he probably chose the B allocation. In the particular case of the UG, this explanation may seem appealing and explain the lack of significant difference in the UG for the two treatments. However, as discussed earlier, it is not clear that an individual would not feel entitled to behave selfishly in the second stage having chosen A. Additionally, this explanation will not work for static games (PD and DG).

6 Conclusion

There are several ways in which our results can be useful. First, we show a way to reduce the social preference confound in experimental games. Since it comes at a cost of additional participants (dummy players) and more complex design, researchers must carefully consider whether it is worth it in each particular study at hand. Clearly, the method may be subject to further development. In particular, optimum calibration of payoffs and coupling with appropriate framing of the instructions are likely to make it more effective.

Second, 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. We propose that dropping the assumption of selfishness alone, especially by allowing outcome-based preference only, will not help us dramatically improve predictions for experimental games. Of course, this is only a putative conclusion; in particular, one would like to allow for more learning. If, however, this proposition is correct, what is then left as the main explanation of the gap between theory and experimental results? To us, it seems likely 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). Another phenomenon that could explain the data would be the social norms (Henrich, Boyd, Bowles, Camerer, Fehr, and Gintis, 2004): if norms are just principles to be followed, playing for probabilities to see one's preferred allocation implemented or to play directly for monetary units may not change behaviors much.

Finally, our findings may teach us something about behavior in organizations and thus also about how to design optimal incentive structures. Models of (outcome-based) social preference generally predict that people may be quite cooperative when their actions directly affect others' monetary payoffs. For example sales representatives may voluntarily share information helping their colleagues get better deals. When taken at face value, these very same models predict cut-throat competition between the same individuals when the stakes consist of some indivisible resource (*e.g.* power, promotion, hierarchical position). Our findings suggest, in contrast, that even though competition may be fiercer, cooperation and prosocial behaviors will be far from absent.

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.

Binmore, K. (2005): "Economic man – or straw man? A commentary on Henrich et al.," *Behavioral and Brain Science*, 28, 817–818.

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

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.

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

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

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

Goeree, J., and C. Holt (2001): "Ten little treasures of game theory and ten intuitive contradictions," *American Economic Review*, 91, 1402–1422.

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.

Kreps, D. M., P. Milgrom, J. Roberts, and R. Wilson (1982): "Rational cooperation in the finitely-repeated prisoners' dilemma," *Journal of Economic Theory*, 27, 245–252.

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

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

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.

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

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

Sugden, R. (1993): "Thinking as a team: toward an explanation of nonselfish 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.

Appendix

When a game played in Stage 2 is dynamic, an action chosen by one player could be seen as a signal concerning Stage-1 allocation. To study such an effect, consider an incomplete information sequential game. In Stage 0 nature chooses for each player i one of two types: A or B. Because subjects are not expected to make a difference between anonymous others, without loss of generality we can assume that $u_i(B_j)_{j\neq i} = 0$ and $u_i(A) = 1$ for each type of any player. We will then have $u_i(B_i) < 1$ for A-types and $u_i(B_i) > 1$ for B-types (for simplicity we assume no subject is perfectly indifferent, but it not necessary for our reasoning). Because Stage 1 choices are not revealed and players have outcome-based preference, they will always choose their preferred allocation.

Consider a two-player Stage 2 game whereby each player moves at most once, first Player 1, then Player 2,[17] possibly with some random moves of nature. Denote a typical mixed (but possibly degenerate) strategy for Player 1 by α_1 and similarly α_2 for Player 2. Player 1's payoff within Stage 2 is $\pi_1(\alpha_1, \alpha_2)$ and for Player 2 it is $\pi_2(\alpha_1, \alpha_2)$. We will show that Player 1 will try to maximize her payoff in lottery tickets, π_1, and Player 2 will try to maximize π_2, given other's strategy. More precisely, any Perfect Bayesian Nash Equilibrium of the entire game (Stages 0-II) will involve moves in Stage 2 that are consistent with the PBNE equilibrium under common knowledge of selfishness.

Ferining the notation from the main text, let $\mu_{ij}(\emptyset)$, $i, j \in \{1, 2, 3\}, i \neq j$ denotes the prior belief of player i about player j, i.e. the probability with which, at the beginning of Stage 2 (empty history ($\emptyset$)), i thinks j chose A in Stage 1. Given anonymity, common prior assumption and minimal pessimism we have $\mu_{ij}(\emptyset) = \mu_{kl}(\emptyset) = \mu < 1$ for all $i \neq j, k \neq l$. The overall expected utility of Player 1 in the supergame if she does not update the beliefs is now

$$\begin{aligned} EU_1(\alpha_1, \alpha_2) =& \pi_1(\alpha_1, \alpha_2)\max(u_1(A), u_1(B_1)) \\ &+ \pi_2(\alpha_1, \alpha_2)(\mu u_1(A) + (1-\mu)u_1(B_2)) \\ &+ [1 - \pi_1(\alpha_1, \alpha_2) - \pi_2(\alpha_1, \alpha_2)](\mu u_1(A) + (1-\mu)u_1(B_3)) \\ =& \pi_1(\alpha_1, \alpha_2)\max(1, u_1(B_1)) + (1 - \pi_1(\alpha_1, \alpha_2))\mu \end{aligned} \tag{1}$$

[17] We thus limit ourselves to the case of $n = 3$. A similar reasoning holds for more players and for more complicated games.

and similarly for Player 2. If μ is indeed never updated, then maximization of $\pi_i(\alpha_1, \alpha_2)$ by player $i = 1, 2$ follows immediately, because $\max(1, u_1(B_1)) > \mu$. Therefore, if there is non-selfish PBNE, it must involve some belief updating and thus at least partial separation of types. To simplify the reasoning, we will make a mild assumption of no separation without incentives (i.e. that when at some information set one type is perfectly indifferent between the strategy selected by the other type and another strategy, she will choose the former).

If beliefs diverge from the prior, players care about having a possibly high payoff for themselves but also about increasing the payoff of a player who, given his or her past moves, is believed to be relatively likely of type A, or decreasing it in the opposite case.

Consider any information set of Player 2, denoted I_2. As said before, if Player 2's beliefs about the type of Player 1 have not been updated, $\mu_{21}(I_2) = \mu$, then both types of Player 2 will choose the same strategy. Suppose that this is not the case and at least partial separation obtains at I_2. Denote the strategy induced by the strategy of type-A of Player 2 starting from the information set I_2 by $\alpha_2^A|I_2$ and similarly $\alpha_2^B|I_2$ for type B. Denote the (expected) payoffs they will yield to players (1,2) by $(\pi_1^A(I_2), \pi_2^A(I_2))$ and $(\pi_1^B(I_2), \pi_2^B(I_2))$ respectively.[18] If Player 2's type is A and he indeed plays $\alpha_2^A|I_2$, his expected utility at this point is $\pi_2^A + \pi_1^A\mu_{21}(I_2) + (1 - \pi_1^A - \pi_2^A)\mu$, and similarly if he plays $\alpha_2^B|I_2$. Thus, for him to choose the former rather than the latter, the following condition must be satisfied (strict inequality follows from the assumption of no separation without incentives):

$$\pi_2^A + \pi_1^A\mu_{21}(I_2) + (1 - \pi_1^A - \pi_2^A)\mu > \pi_2^B + \pi_1^B\mu_{21}(I_2) + (1 - \pi_1^B - \pi_2^B)\mu \quad (2)$$

Similarly, for type B of Player 2 to choose $\alpha_2^B|I_2$ rather than $\alpha_2^A|I_2$, it must be that

$$\pi_2^A u_2(B_2) + \pi_1^A\mu_{21}(I_2) + (1 - \pi_1^A - \pi_2^A)\mu < \pi_2^B u_2(B_2) + \pi_1^B\mu_{21}(I_2) + (1 - \pi_1^B - \pi_2^B)\mu \quad (3)$$

Subtracting the LHS of 2 from LHS of 3 we get $\pi_2^A(u_2(B_2) - 1)$ which must be smaller than the analogous difference of RHSs, which is equal to $\pi_2^B(u_2(B_2) - 1)$. It thus follows that $\pi_2^A < \pi_2^B$. Then for equation 2 to be satisfied, it must be that $(\pi_1^A - \pi_1^B)(\mu_{21}(I_2) - \mu) > 0$ (in words, strategy

[18] From now on we will skip (I_2) to simplify notation.

chosen by type A of Player 2 gives more lottery tickets to Player 1 than that chosen by type B if and only if Player 1 is more likely to be of type A than is Player 3). This is tantamount to saying that the selfish type B (of Player 2) cares relatively more about his own payoff, while the other-regarding type A cares relatively more about allocating the payoff to others who are also likely to be of type A.

Consider now the strategy of type A of Player 1, α_1^A, leading to expected payoffs (π_1^A, π_2^A) and the corresponding strategy α_1^B, leading to payoffs (π_1^B, π_2^B) Assume these strategies are different. Using the same logic as before, one can show that $\pi_1^A < \pi_1^B$ but the difference between the expected payoff of type A and type B of Player 2 is higher under α_1^A than under α_1^B. However, this would require that type A of Player 2 behave more selfishly than type B in at least some information sets which obtain more often under α_1^A than under α_1^B (i.e. sets in which Player 1 is revealed to be more likely of type A than B) and we have shown before that his incentives are exactly opposite. Thus no separation obtains, QED.

Faculty of Economic Sciences
University of Warsaw
44/50 Długa St.
00-241 Warsaw
www.wne.uw.edu.pl