2013-10-10

The Social Egoist[*]

Anne Boschini[‡], Astri Muren[○] and Mats Persson[§]

ABSTRACT:

People cooperate more in one-shot interactions than can be explained by standard textbook preferences. We discuss a set of non-standard preferences that can accommodate such behavior. They are social, in the sense of incorporating the payoffs of other persons; they are also norm-based, in the sense of taking into account the behavior of other persons. We show theoretically that, with such preferences, a Nash equilibrium with a strictly positive cooperation rate can exist. We use experimental data on within-subject decisions to show that such preferences are empirically plausible. The data show that, in addition to the well-known types (egoist, altruist, reciprocator), there is an important group: the social egoist. Such individuals care for people who have cooperated, but ignore people who have broken the implicit cooperation norm in society. The social egoists, who turn out to be different from "conditional cooperators", account for one third of the observations in our experiment.

JEL Codes: C91, D03, D64

Keywords: social norms, prisoner's dilemma, hawk-dove game, egoism, altruism, reciprocity, conditional cooperation

[*] We thank Magnus Johannesson, Topi Miettinen and Jörgen Weibull for constructive comments and suggestions.

[‡] Department of Economics, Stockholm University, S-106 91 Stockholm, Sweden. E-mail: anne.boschini@ne.su.se.

[○] Department of Economics, Stockholm University, S-106 91 Stockholm, Sweden. E-mail: astri.muren@ne.su.se.

[§] Institute for International Economic Studies, Stockholm University, S-106 91 Stockholm, Sweden. E-mail: mp@iies.su.se.

1. Introduction

Social norms may explain why people cooperate in everyday interactions, thereby helping society to function smoothly. But for norms to be effective they have to be enforced, and norm-breakers punished. Thus, if an internal enforcement mechanism (like a bad conscience) is not enough to prevent bad behavior, external punishment is needed. The literature has pointed to the need for and existence of "strong reciprocators" (Gintis, 2000) or "punishers" (Fehr and Fischbacher, 2004) who take on this duty. However, being a punisher is an arduous task that often involves a cost to the person in question, and it is therefore not surprising that people seem to prefer to avoid playing that role (Dreber at al., 2008, Ule et al. 2009, Rand et al., 2009).[1] In fact, recent studies suggest that when given the choice between punishing norm-breakers of rewarding norm-adherers, many people prefer the latter (Almenberg et al., 2011, Nikiforakis and Mitchell, 2013).

In this paper we formulate a simple model of external norm-enforcement, in the form of heterogeneous norm-based social preferences. By norm-based, we mean that the parameters in the utility function depend on whether the counterpart has adhered to a social norm or not. In the model, people differ with respect to these preference parameters.[2] We investigate theoretically under what conditions an equilibrium with complete or partial cooperation may exist. The question then is what types of social preferences – norm-based or not – we observe in reality. To do this, we conduct an experiment using two games: the Prisoner's dilemma (*PD*) game and the Hawk-Dove (*HD*) game. We use within-subject analysis of decisions in these games (specifically, the second stages of sequential *PD* and *HD* games) to distinguish between those who defect in order to punish and those who defect in order increase their own payoff.

We find that most of the participants in our experiment can be categorized in terms of four main behavioral types. Two of these types do not show norm-based behavior, namely the *egoists* (23 percent) and the *altruists* or social welfare maximizers (around 10 percent). The preferences of these types do not depend on whether the counterpart acted in adherence with the cooperation norm or not. The other two types have norm-based preferences and attach a positive value to the

[1] In some situations, men are less inclined than women to play the role of punishers, see Boschini et al. (2011).
[2] The notion that a person's utility weight on the other person's payoff is conditioned on the latter's behavior has previously been labeled "conditional altruism" by Palfrey and Rosenthal (1988).

payoffs of another player when that player adhered to the norm. One of these types is the *reciprocator*, who values the payoff of the defector negatively. In our sample, 20 percent belong to this type. The other type with norm-based preferences is one which we call the *social egoist*. This type assigns a positive value to the payoff of a norm-abider, and a zero value to that of a defector. The social egoist is important in our data, accounting for as many as one third of the observations. To investigate whether this type has anything to do with a behavioral type well-known from the literature, namely, the conditional cooperator, we included a public-goods game in our experiment; it turned out that the conditional cooperators were evenly spread across the preference types described above.

It is standard in models of social norms to include a desire for norm-adherence in the utility function (see Lindbeck *et al.*, 1999, Lopéz-Pérez, 2008, Andreoni and Bernheim, 2009, and Bénabou and Tirole, 2011). In this paper we choose the alternative route of formulating an external mechanism for norm-enforcement. In our model, an individual's preferences over the outcomes of self and others depend on the other party's adherence to a cooperation norm. An individual's decision whether to cooperate or not depends on these preferences, in combination with what is known about the behavior of others. Our approach allows for heterogeneity of behavior, generated by the distribution of the social preference parameters.

Other scholars have emphasized that people might care about the intentions of others. This has given rise to the field of psychological game theory (Geneakoplos et al., 1989, Rabin, 1993, Dufwenberg and Kirchsteiger, 2004, Falk and Fischbacher, 2006). Such an approach has an intuitive appeal, but requires a model of how players infer intentions from the counterpart's actions. Levine (1998) has suggested a simpler route: to assume that people consider and have preferences over the counterpart's *type*.[3] Of course, type is not always observable. Even in the simplistic case of one-shot interaction with an anonymous counterpart, which is what we deal with in this paper, it would require an assumption of Bayesian updating of some *a priori* beliefs. We consider instead the simplest conceivable set-up, namely, the second stage of a game with an anonymous counterpart where people's social preferences depend on the only commonly observable factor: the counterpart's behavior.

[3] See Ellingsen and Johannesson (2008) for a more general model where agents care about the opponent's type.

We demonstrate that almost all the variation in behavior in our data can be described as generated by the distribution of parameters of our assumed utility function. Our approach to heterogeneity is in this sense different from other studies that rely on a more fundamental separation into behavioral types. For example, Andreoni and Miller (2002) allow for three kinds of preferences (Selfish, Leontief and Perfect substitutes), as does Erlei (2008), while Bowles and Gintis (2004) allow for three types (Reciprocators, Selfish and Cooperators) defined by their behavior rather than their utility functions (i.e., Cooperators always cooperate, etc.).

The paper proceeds as follows. In section 2 we present the basic model of preferences and derive conditions for the existence of equilibrium. We then give a few illustrative examples of preferences that lead to particularly simple equilibrium configurations. In section 4 we discuss the strategy for testing norm-based social preferences against experimental data, and in section 5 we describe the details of such an experiment. In section 6 we present the results. Thereafter, we test whether our social egoists can be identified with the "conditional cooperators" known from earlier literature. Section 8 concludes.

2. Norm-based social preferences

By social preferences we mean that the utility function of individual i has not only i's own consumption as an argument, but also the consumption of another individual: $u_i = u(x_i, x_j)$. Such preferences could be of the standard, altruistic kind, where $u(x_i, x_j)$ is increasing in both arguments, and encompass the consumption of the other individual as an ordinary "consumption good". But it could also involve more complex forms, to be discussed below. Throughout this paper, we will abstract from risk aversion, and thus postulate linear utility functions. Social preferences will be represented by

$$u_i(x_i, x_j) = \begin{cases} x_i + \alpha x_j & if\ player\ j\ has\ conformed\ to\ the\ norm, \\ x_i + \beta x_j & if\ player\ j\ has\ deviated\ from\ the\ norm. \end{cases} \tag{1}$$

Such preferences encompass the notion of a cooperation norm: if the other individual is a norm-breaker (i.e., does not cooperate), I frown upon his behavior by applying the weight β rather than α to his payoff. Presumably, $\beta < \alpha$, but this is of course an empirical question. We will use the term “norm-based social preferences” for this kind of preferences. Charness and Rabin (2002) use a similar formulation of the utility function; they also include an “inequality aversion term” in the utility function, but find that such a term does not add to the explanatory power of the model. Without denying that more complex preferences could be interesting for future work, we therefore use the stripped-down version (1). Letting α and β vary across individuals, we will show below that such preferences are broadly consistent with the behavior we observe in our experiment.

Consider the generic *PD* game, characterized by the following pay-off matrix.

$$\begin{array}{c|cc} & \text{Coop} & \text{Defect} \\ \hline \text{Coop} & B, B & D, A \\ \text{Defect} & A, D & C, C \end{array} \qquad (2)$$

with

$$A > B > C > D. \qquad (3)$$

With no loss of generality, we consider only non-negative values of the parameters *A*, *B*, *C* and *D*. In the case of traditional, egoistic preferences $\alpha = \beta = 0$, defection is the dominant strategy, with the resulting payoffs (*C*, *C*). Since this equilibrium outcome is Pareto-dominated by (*B*, *B*) social efficiency would be enhanced if a cooperation norm were present in society. In fact, there is empirical evidence that such a norm does exist (cf. Fehr and Fischbacher, 2004). We will thus interpret the phrase “has conformed to the norm” in (1) above as “has played *cooperate* in the *PD* game” – with the corresponding definition of “has deviated from the norm”.[4]

Preferences like (1) seem natural in a *sequential PD* game, i.e., a game where the player, in the second stage of the game, is informed about how the counterpart has acted in the first stage. Assume that player *j* has played “defect” and that player *i* is informed about that before making his move. Then he might dismiss *j* as a norm-breaker and attach a low coefficient β to *j*’s consumption. In a *simultaneous PD* game, where both *i* and *j* make their moves without knowing

[4] This general approach to social preferences (1) is not limited to *PD* games, i.e., gamers defined by (2)-(3). Later in this paper, we will also consider other games where one could argue that a cooperation norm applies, i.e., where preferences (1) might be valid.

the other's action, the psychological rationale for (1) is more complex and might involve regret. Assume that *i*'s parameters α and β, and his subjective probability π that the opponent will play "cooperate", are such that he maximizes expected utility by cooperating. If, after the outcome of the game has become known, it turns out that the opponent played "defect", then *i* may regret his choice. But when choosing *ex ante* whether to cooperate or defect, by maximizing expected utility, he takes account of this possibility of regret when he makes his choice.

Assume that individual *i* is randomly paired with an unknown opponent to play a simultaneous *PD* game, and that *i* attaches the probability π to the event that the opponent plays "cooperate". The expected utility for *i* of playing "cooperate" then is $\pi(B+\alpha B)+(1-\pi)(D+\beta A)$ while the expected utility of playing "defect" is $\pi(A+\alpha D)+(1-\pi)(C+\beta C)$. Individual *i* will play "cooperate" if

$$\pi(B+\alpha B)+(1-\pi)(D+\beta A)\geq\pi(A+\alpha D)+(1-\pi)(C+\beta C). \qquad (4)$$

Inequality (4) shows the necessary and sufficient condition for a player characterized by preference parameters (α,β) to cooperate in a *PD* game. For another player, with a different pair (α,β), inequality (4) may be reversed; he will thus defect. We will now consider two aspects of the model: one where each individual is characterized by an arbitrary parameter triplet (α,β,π) and one where π is the same for all individuals and consistent with a Nash equilibrium.

Assume that each individual is characterized by a vector (α,β,π). We might think of $\alpha\geq 0$ and $\beta\leq\alpha$, but those restrictions are not necessary for our model; there might in principle be individuals with other configurations of α and β. For π, however, there is a natural restriction: $\pi\in[0,1]$. We can now state

Proposition 1: (i) Other things equal, a higher α makes an individual more prone to cooperate.
(ii) Other things equal, a higher β makes an individual more prone to cooperate.

Proof: These properties follow trivially from (4) and the fact that $B>D$ and $A>C$.

In a *Nash equilibrium*, inequality (4) holds for a fraction π of the population. To make the notation more compact, we define

$$\left.\begin{array}{l} a \equiv \pi(B-D) \\ b \equiv (1-\pi)(A-C) \\ c \equiv (1-\pi)(C-D)+\pi(A-B) \end{array}\right\} \tag{5}$$

Inequality (4) can thus be written $a\alpha + b\beta \geq c$. Denote the joint cumulative distribution of α and β by $F(\alpha,\beta)$ with the corresponding density $f(\alpha,\beta)$. The fraction of the population playing "cooperate" (i.e., the fraction of the population for which inequality (4) is satisfied) is given by

$$\pi = \iint_{a\alpha+b\beta\geq c} f(\alpha,\beta)\, d\alpha d\beta. \tag{6}$$

Equation (6) is a non-linear equation in π (remember that *a*, *b* and *c* are functions of π). A Nash equilibrium (*NE*) is a π that satisfies (6). We have

Proposition 2: There exists at least one *NE*.

Proof: Since π is a probability (or a fraction), it is defined on the compact set $[0,1]$. Since the right-hand side of (6) is continuous in π we can invoke the fixed-point theorem saying that a continuous mapping of a compact set into itself has at least one fixed point. Thus there is at least one *NE*. *Q. E. D.*

An *NE* may imply full cooperation (i.e., $\pi = 1$), an interior cooperation rate ($1 > \pi > 0$) or non-cooperation ($\pi = 0$). Let us start with full cooperation. Substituting $\pi = 1$ into (6) yields

$$1 = \iint_{\alpha\geq\frac{A-B}{B-D}} f(\alpha,\beta)\, d\alpha d\beta.$$

From this equation we have:

Corollary 2a: Full cooperation is an *NE* if and only if the lower support of the α distribution $\alpha_{lower} > \frac{A-B}{B-D}$.

Here, the existence of a full-cooperation *NE* does not preclude other equilibra. Even if $\alpha_{lower} > \frac{A-B}{B-D}$, we might have interior equilibria $0 < \pi < 1$, and even a non-cooperation equilibrium $\pi = 0$.

For the case of non-cooperation, we substitute $\pi = 0$ into (6). From this equation we see that the following holds.

Corollary 2b: Non-cooperation is an *NE* if and only if the upper support of the β distribution $\beta_{upper} < \frac{C-D}{A-C}$.

Note that although the corollary states a necessary and sufficient condition for a non-cooperation *NE* to exist, it does not preclude other equilibria. That is, even if $\beta_{upper} < \frac{C-D}{A-C}$, there might exist an interior equilibrium $0 < \pi < 1$, and even a full-cooperation equilibrium $\pi = 1$ (if the condition in *Corollary 2a* is satisfied). However, if $\beta_{upper} \geq \frac{C-D}{A-C}$ *all* equilibria will display at least some positive degree of cooperation.

Assume now that the condition in *Corollary 2a* is not satisfied, i.e., that $\alpha_{lower} < \frac{A-B}{B-D}$. By *Proposition 2*, there is at least one equilibrium. In that equilibrium $\pi < 1$ by *Corollary 2a*. Assume further that $\beta_{upper} > \frac{C-D}{A-C}$. By *Corollary 2b* this means that $0 < \pi$. We have thus proved

Corollary 2c: A sufficient condition for an interior solution $0 < \pi < 1$ is that $\alpha_{lower} < \frac{A-B}{B-D}$ and $\beta_{upper} > \frac{C-D}{A-C}$.

3. Some simple examples

Assume that α and β are independently and uniformly distributed variables: $\alpha \sim U[0,1]$ and $\beta \sim U[0,1]$.[5] With this joint distribution, we see that the conditions for Corollary 1c are satisfied for any *PD* game with *A* > *B* > *C* > *D*. We would therefore expect no corner solution $\pi = 1$ or $\pi = 0$, but only interior solution(s) $0 < \pi < 1$. Let's see whether this is the case.

With a uniform, joint distribution, equation (6) reads

$$\pi = \iint\limits_{a\alpha + b\beta \geq c} f(\alpha, \beta)\, d\alpha d\beta = 1 - \frac{c}{a} + \frac{b}{2a}.$$

Making use of the definition of *a*, *b* and *c* in (5), this equation can be written

$$\pi^2(B-D) = \pi((B-D) - (1-\pi)(C-D) - \pi(A-B) + (1-\pi)\frac{A-C}{2}. \tag{7}$$

This is a second-order equation in π that yields the *NE*. For a numerical example, we assume that the pay-off matrix (1) is defined by

$$A = 600,\ B = 500,\ C = 100,\ D = 0. \tag{8}$$

The second-order equation (7) then has the solutions

$$\pi = \begin{cases} -0.352 \\ \ \ 0.852 \end{cases}$$

In this case, one might interpret the smaller root as $\pi = 0$, i.e., a non-cooperative equilibrium. This is however not correct. If nobody cooperates, the opposite of inequality (4) must hold for all values of β when $\pi = 0$. That is, we must have that $\beta < \frac{C-D}{A-C} = 0.2$, which is not the case since β is uniformly distributed on $[0,1]$. Thus we cannot interpret the smaller root -0.352 as a non-

[5] One might wish to impose the restriction that $\alpha \geq \beta$, but for simplicity, we do not impose any such restriction. We thereby allow for some individuals behaving according to the "prodigal son" parable, treating a norm-breaker better than a norm-abider.

cooperative equilibrium $\pi = 0$, which should not come as a surprise since the assumed distributions of α and β do in fact satisfy the conditions in *Corollary 2c*. Thus, there is only one *NE* in this case: $\pi = 0.852$.

If one doesn't like the "prodigal son property" of the above example (i.e., that α could be smaller than β) one might instead prefer the assumption that $\alpha \sim U[0, 1]$ and $\beta \sim U[-1, 0]$. With this distribution, (6) can be written

$$\pi = \iint\limits_{a\alpha + b\beta \geq c} f(\alpha, \beta)\, d\alpha d\beta = 1 - \frac{c}{a} - \frac{b}{2a}.$$

By definitions (5) and (8), this yields a second-order equation $\pi^2 - 1.5\pi + 0.7 = 0$ with the roots

$$\pi = \begin{cases} 0.75 + 0.371i \\ 0.75 - 0.371i. \end{cases}$$

Thus there are no real roots in the interval (0, 1). By *Proposition 2* and the three corollaries, there is only one *NE*: the non-cooperative one with $\pi = 0$. Thus, assuming that β was all non-positive implied completely non-cooperative behavior.

If we instead assume that $\alpha \sim U[0, 1]$ and $\beta \sim U[-0.5, 0.5]$, the *NE* has a positive level of cooperation. We have

$$\pi = \iint\limits_{a\alpha + b\beta \geq c} f(\alpha, \beta)\, d\alpha d\beta = 1 - \frac{c}{a}.$$

With the payoff matrix (6), this yields the second-order equation $5\pi^2 - 5\pi + 100 = 0$ with the solutions

$$\pi = \begin{cases} 0.276 \\ 0.724, \end{cases}$$

i.e., two well-behaved, interior equilibria. We also see that since the distributions $\alpha \sim U[0, 1]$ and $\beta \sim U[-0.5, 0.5]$ satisfy the conditions for *Corollarium 2c*, there can be no corner solutions $\pi = 0$ or $\pi = 1$.

The question is whether it is possible to conceive of an experiment that could yield information about how of α and β are distributed in reality. We now proceed to this issue.

4. Experimental strategy

The aim is to empirically determine the distribution of α and β in a population. For this purpose, we design an experiment that will provide such information in a simple way. The basic *PD* game is not sufficient to provide the information we need, but by varying the information provided to the participants, and the payoff matrix, we obtain a reasonably rich picture of the distribution of α and β.

We develop the experiment along two dimensions. One dimension regards the counterpart's decision: we use the second stage of a two-stage game, where each participant is informed that she will be matched against the decision of a counterpart who has cooperated (i.e., $\pi = 1$), or defected ($\pi = 0$) in a previous game. [6] The player's response then allows us to draw conclusions about her α and β, while avoiding the confound of beliefs formed by the player about π. Previous experimental studies of sequential *PD* games are Clark and Sefton (2001) and Blanco et

[6] A sequential *PD* game has two stages. In the first stage, player 1 chooses "Cooperate" or "Defect" while knowing that player 2 will observe this choice before deciding what to do. In the second stage, player 2 is informed that the counterpart has chosen "Cooperate" or "Defect", i.e., that $\pi = 0$ or $\pi = 1$. In the first stage, the expected utility for player 1 of cooperating is $\pi_X B + (1 - \pi_1)D + \pi_X \alpha B + (1 - \pi_X)\beta A$, where π_X is the probability that player 2 will cooperate, given that player 1 has cooperated. Similarly, the expected utility for player 1 of defecting is $\pi_Y A + (1 - \pi_Y)C + \pi_Y \alpha D + (1 - \pi_Y)\beta C$, where π_Y is the probability that player 2 will cooperate, given that player 1 has defected. Let us denote the utility parameters of player 2 by α_j and β_j, with the distribution functions F_α and F_β, respectively. We then have that

$$\pi_X = Prob\big(B + \alpha_j B > A + \alpha_j D\big) = 1 - F_\alpha\left(\frac{A - B}{B - D}\right),$$
$$\pi_Y = Prob\big(D + \beta_j A > C + \beta_j C\big) = 1 - F_\beta\left(\frac{C - D}{A - C}\right).$$

Substituting these expressions for π_X and π_Y into the expressions for expected utility above, player 1 can choose whether to cooperate or defect in the first step. He cooperates if

$$\alpha(\pi_X B - \pi_Y D) + \beta\big((1 - \pi_X)A - (1 - \pi_Y)C\big) > \pi_Y A - \pi_X B + (1 - \pi_Y)C - (1 - \pi_X)D - (1 - \pi_1)D$$

In the experiment, we only let the subjects play the second step; thereby, we did not have to assume that the subjects had any clear idea of the distribution functions F_α and F_β in our one-shot game.

al. (2011). The former utilize the whole sequential game, and the latter use the strategy method to elicit responses for both information nodes in the second stage. Our approach is similar to the strategy method in that it disconnects any possible relation between first-stage and second-stage players, but our players are matched with actual counterparts for each decision.

The other dimension regards the *pay-off matrix*. We use two sets of pay-offs, one of them a *PD* game, where the pay-offs (2) satisfy $A > B > C > D$, the other a Hawk-Dove game (*HD*) where $A > B > D > C$. By a suitable choice of these pay-offs, we can determine upper and lower limits for the parameters in each player's utility function: $\alpha_{lower} \leq \alpha \leq \alpha_{upper}$ and $\beta_{lower} \leq \beta \leq \beta_{upper}$.

For the $PD_{\pi=1}$ game, where the player is informed that the counterpart has played "cooperate", we substitute $\pi = 1$ into (4); thus the player chooses to cooperate if $B + \alpha B \geq A + \alpha D$. A subject who cooperates in this particular game thus has a parameter α that satisfies

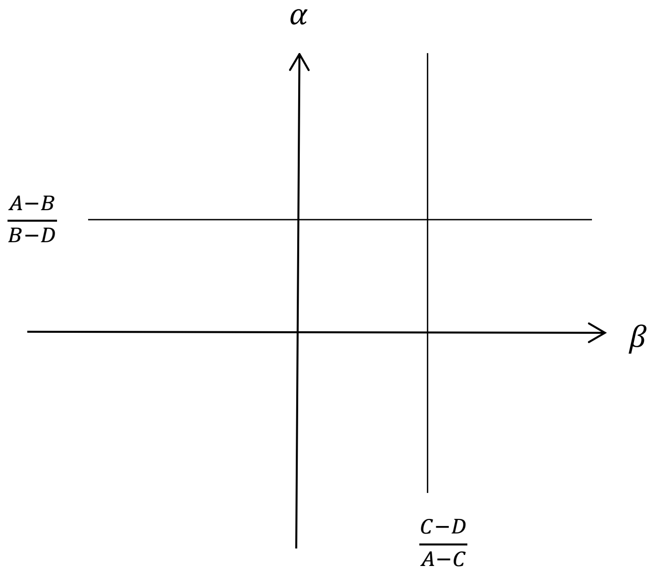
$$\alpha \geq \frac{A - B}{B - D}. \tag{9}$$

For the $PD_{\pi=0}$ game, where the player is informed that the counterpart has played "defect", we substitute $\pi = 0$ into (4). Here, the player chooses to cooperate if $D + \beta A \geq C + \beta C$. A subject who plays defect in the $PD_{\pi=0}$ game thus has a parameter β that satisfies

$$\beta \geq \frac{C - D}{A - C}. \tag{10}$$

By seeing how the subjects played in these two second-stage *PD* games we can thus enter them in (α, β) space as depicted in Figure 1. Because of the inequality $A > B > C > D$ characterizing the *PD* game, the straight lines representing the cut-offs (9) and (10) must intersect in the positive quadrant:

Figure 1: Regions in (α, β) space for a *PD* game.



Thus, the cut-offs (9) and (10) yield some information about an individual's location in (α, β) space. A *PD* game with $A > B > C > D$ does not, however, tell us whether negative values of β are common in the population. If so, there would be individuals willing to sacrifice money to inflict punishment upon norm-breakers (see, for instance, Gintis, 2000, Bowles and Gintis, 2004, and Fehr and Fischbacher, 2004). By changing D so that $A > B > D > C$, we can shift the vertical $\frac{C-D}{A-C}$ line to the left of the origin, thereby making it possible to assess the number of individuals in the experiment who must have negative values of β. This new game becomes the classic Hawk-Dove game (*HD*), first formalized in Maynard Smith and Price (1973). Such a change of D also shifts the horizontal $\frac{A-B}{B-D}$ line in Figure 1.[7]

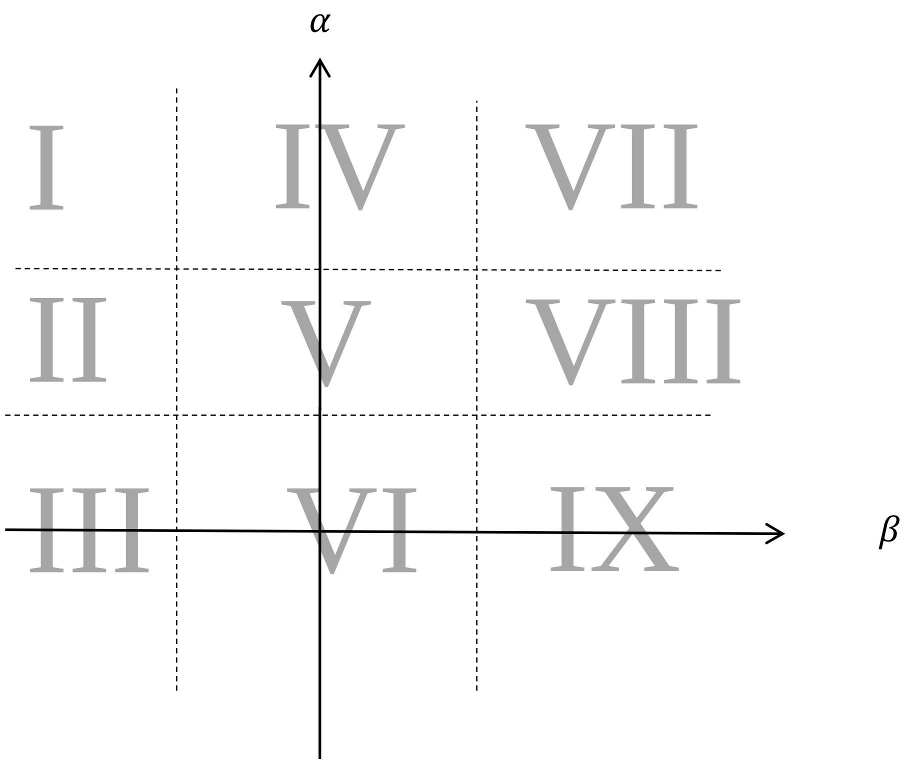
Some further comment is warranted for the game defined by $A > B > D > C$. Even if it is not a *PD* game, the notion of cooperation *vs.* defection is meaningful. The reason is that we choose the values of *A*, *B*, *C* and *D* such that the outcome ("Cooperate", "Cooperate") maximizes total pay-off, and one may thus argue that there is a cooperation norm in that game, too. Therefore, norm-based preferences (1) are relevant. In the two-stage form, where the subject is informed that the

[7] Another potentially interesting question is how many of the individuals have negative values of α (i.e., who want to harm people who have been nice to them). This would entail a change in *A* or B to shift the horizontal line $\frac{A-B}{B-D}$ line below the origin.

counterpart has defected, replying with a “Defect” implies punishing the norm-breaker. Compared to the *PD* game, the cost of punishing is positive (while it is negative in the *PD* game). This means that one has to have a lower value of one’s β in order to punish in the *HD* game.

The cut-off values of α and β that are given by inequalities (9) and (10) define, together with the two pay-off matrices of the *PD* and *HD* games (for numerical examples, see Section 5 below), nine regions in (α, β) space, as illustrated in Figure 2. There is a one-to-one correspondence between these nine regions and the way the individuals play the found second-stage games in our experiment. Let us consider the regions one by one.

Figure 2: The nine different types of social preferences



Consider a person who always cooperates when the counterpart cooperates (in $PD_{\pi=1}$ and $HD_{\pi=1}$), and defects when the counterpart defects (in $PD_{\pi=0}$ and $HD_{\pi=0}$). This person must belong to region I, i.e., she must have a large positive α and a negative β. Such people are “reciprocators” in the sense that they like norm-adherers and frown on norm-breakers – even to the point of paying a price in order to punish the latter. The individuals in region II are similar; they also punish norm-breakers, but they are somewhat less generous towards norm-adherers. In fact, they do cooperate in $PD_{\pi=1}$ but not in $HD_{\pi=1}$. Since a cooperating counterpart does not suffer so much if the player defects in $HD_{\pi=1}$ as in $PD_{\pi=1}$, it is more tempting to defect in $HD_{\pi=1}$, unless one does have a very large α.

Further, in region III we find “punishers”, with an α close to zero and a negative β; they are neutral towards norm-adherers, but punish norm-breakers. Their behavior looks like that of the “altruistic punishers” in Fehr and Fischbacher (2004), but note that since their βs are negative, they actually enjoy punishing norm-breakers.

A particularly interesting group of people have values of β close to zero and a strictly positive α. Such individuals are friendly towards nice people but do not care, neither negatively nor positively, about people who have misbehaved. We encounter these people in region IV and region V, and we will call them “social egoists”. In region IV we have the social egoists who cooperate in $PD_{\pi=1}$, $HD_{\pi=1}$, and $HD_{\pi=0}$, and defect in $PD_{\pi=0}$, while in region V we have the social egoists who also defect in $HD_{\pi=1}$.

The egoists are in region VI, with both α and β close to zero. The “altruists” of regions VII and VIII value the consumption of others highly and cooperate also when the counterpart has misbehaved. People in region VIII do not let their altruism be affected by any cooperation norm (they are equally benevolent towards norm-breakers and norm-adherers), while the altruists of region VII are somewhat more generous towards norm-adherers.

The personality type represented by region IX is characterized by a low α and a high β. Although a person with $\beta > \alpha$ cannot be ruled out on *a priori* grounds, such a person defies a simple intuitive label. Perhaps some kind of “Prodigal son’s dad” attitude, with more generosity towards sinners than towards those who behave well, would be appropriate.

5. Experimental design

To assess the distribution of α and β in a population, we ran an experiment at Stockholm University in February 2010. The participants were 391 students in the Introductory

Microeconomics course.[8] The students were informed that participation in the experiment was voluntary, but that answering the questions would be a useful exercise for the next lecture (which dealt with strategic interaction). Nearly everyone participated, which makes the problem of selection into the experiment a non-issue in this case. The experiment consisted of answering a questionnaire (see Appendix 2) which was distributed to everyone. Participants were anonymous, and information about, e.g., gender, was self-reported on the last page of the questionnaire.

The participants played three sets of games. The first two sets consisted of one simultaneous and two second-stage games. One set consisted of regular *PD* games; the other set consisted of *HD* games. Finally, they played one public-goods game (*PG*). In the *PD* games, we use the pay-off matrix (8):

$$A = 600,\ B = 500,\ C = 100,\ D = 0.$$

In the *HD* games we used the following matrix:

$$A = 600,\ B = 500,\ C = 100,\ D = 200. \tag{11}$$

All units were in Swedish crowns (the exchange rate at the time of the experiment was 7.25 crowns per US dollar).

The stakes were real, but for administrative and budgetary reasons each participant could not be paid in every game. Instead, 10 percent of the participants were randomly selected to be paid in one of the seven games. Average payment for those paid was 381 Swedish crowns ($\approx$ 53 US dollars) net of income tax.[9] The whole experiment took 10-15 minutes; thus expected hourly payment was $4 \cdot 38.10 = 152.40$ Swedish crowns ($\approx$ 21 US dollars). Those 10 percent who were actually paid their gains had to give their names and addresses due to the requirements of the Swedish tax authorities. This information was handled by an administrator at the Stockholm University Economics Department, who was not in any other way involved in the experiment.

[8] None of the authors of this paper was involved in teaching the course.

[9] For each person we also paid 30 percent income tax to the tax authorities. This means that for a person who received 100 crowns, we paid a gross amount of $100/0.7 \approx 143$ crowns.

One challenge when studying social norms experimentally is to distinguish norm-based explanations from other explanations of behavior. As mentioned above, Fehr and Fischbacher (2004) remove second-party revenge motives by allocating the option of costly punishment to a disinterested third-party, who acts as an external enforcer. In our experiment, the counterpart in each game was another anonymous and randomly selected student in the class. In the second-stage games, each participant was informed that he/she had been matched with the decision of a counterpart who had cooperated, or defected, in the previous simultaneous game. For each new game, the participants were explicitly informed that they would be matched with a new, randomly selected, person. There is thus no direct relation between our player and his/her counterpart, and in this sense our design resembles a third-party punishment design.[10] In order to avoid any emotional connotation, we did not use the terms "cooperate" and "defect" in the questionnaire (see Appendix) but called the actions "*X*" and "*Y*", respectively.

There were four different versions of the questionnaire, in order to control for order effects. The four types of questionnaire were the following:

Table 1: Configurations of order

No.	Order of games
1.	PD_{sim}, $PD_{\pi=1}$, $PD_{\pi=0}$, HD_{sim}, $HD_{\pi=1}$, $HD_{\pi=0}$, PG
2.	PD_{sim}, $PD_{\pi=0}$, $PD_{\pi=1}$, HD_{sim}, $HD_{\pi=0}$, $HD_{\pi=1}$, PG
3.	HD_{sim}, $HD_{\pi=1}$, $HD_{\pi=0}$, PD_{sim}, $PD_{\pi=1}$, $PD_{\pi=0}$, PG
4.	HD_{sim}, $HD_{\pi=0}$, $HD_{\pi=1}$, PD_{sim}, $PD_{\pi=0}$, $PD_{\pi=1}$, PG.

We tested for two types of order effects: whether the *PD* game or the *HD* game came first, and whether, in the second-stage games, the game where the opponent had cooperated or defected came first. There were no effects of the order of the *PD* and *HD* games. There is a weak effect (at the 5 percent level) of having the defection game first, but only in the *HD* game. Here, there was a higher level of cooperation in $HD_{\pi=1}$.

[10] We consider behavior as such, and not directed towards any particular person. This is subtly different from *indirect reciprocity*, where the individual cares about behavior, or rather the intentions revealed by behavior directed towards someone else. See, for instance, Milinski et al., 2001, Nowak and Sigmund, 2005.

Before going to the main results, we present summary statistics of cooperation rates in the experiment, see Table 2.

Table 2: Summary statistics

Game	Fraction of cooperators
PD_{sim}	$\pi = 0.527$
$PD_{\pi=1}$	$\pi = 0.611$
$PD_{\pi=0}$	$\pi = 0.184$
HD_{sim}	$\pi = 0.747$
$HD_{\pi=1}$	$\pi = 0.542$
$HD_{\pi=0}$	$\pi = 0.719$

We see that the results are quite similar in the *PD* and *HD* games, in the simultaneous as well as the $\pi = 1$ variants. This is consistent with our assumption that the cooperation norm of the *PD* game is also present in the *HD* game. The striking difference between the two games shows up in the $\pi = 0$ variants, where the cooperation rate is 71.9 percent in the *HD*, versus only 18.4 percent in the *PD*, game. We now proceed to a within-subjects analysis of behavior in the four second-stage games.

6. Results

6.1 Personality Types

What personality types emerge in our data? With the pay-off matrices (8) and (11), the inequalities (9) and (10) become

$$\alpha > \frac{1}{5} \qquad \beta > \frac{1}{5} \qquad \alpha > \frac{1}{3} \qquad \beta > -\frac{1}{5}\,.$$

These inequalities define the nine regions in (α, β) space illustrated in Figure 2. For instance, an "egoist" in region VI is a person whose choices have revealed that she has $\alpha \leq \frac{1}{5}$ and $-\frac{1}{5} < \beta \leq \frac{1}{5}$, while a "social egoist" of region V has $\frac{1}{5} < \alpha \leq \frac{1}{3}$ and $-\frac{1}{5} < \beta \leq \frac{1}{5}$.

There were 391 participants in the experiment. Of these, almost 90 percent (343 individuals) acted consistently with our model.[11] Table 3 shows the results.

Table 3: Number of persons in the data belonging to each personality type (N = 391)

Region No.	Personality type	No. of indiv's	No. of Women	No. of Men	
I	Reciprocators	54	35*	19*	
II	Reciprocators	10	6	4	
III	Punishers	22	8	14	Norm-based pref.
IV	Social egoists	83	46	37	
V	Social egoists	42	27	15	
VI	Egoists	85	37*	48*	
VII	Altruists	34	14	19	Standard pref.
VIII	Altruists	8	2	6	
IX	"Prodigal son's dad"	5	3	2	
	Unclassified	48	27	21	

Note: * denotes gender difference at the ten percent level. One of the altruists did not indicate his/her gender in the questionnaire

Two results stand out. First, norm-based preferences are very common. In the table, we have drawn a line between the personality types with norm-based (types I-V) and standard (types VI-IX) behavior.[12] The former group is the larger one, consisting of 211 subjects. Within the group of norm-based individuals, there are three types: reciprocators, punishers and social egoists. Second, the social egoists are quite numerous. In fact, types IV and V consist of 125 individuals which makes the social egoist the most prevalent personality type of all.

[11] The 48 individuals who could not be assigned to a specific personality type gave inconsistent answers in the sense that, for instance, $\alpha > \frac{1}{3}$ $and < \frac{1}{5}$, or that $\beta > \frac{1}{5}$ and $\beta < -\frac{1}{5}$. Note that the 48 subjects are not necessarily irrational; they may very well be rational but have more complicated preferences than the ones in (1). Also, there are of course many reasons why one might check the wrong box in an experimental questionnaire. There were no significant differences in gender composition between those 48 and the remaining 343 individuals.

[12] Whether the 5 "prodigal son's dad" individuals in region IX can be said to have standard preferences is of course a matter of judgment.

An interesting question arises, namely, the long-term survival of our nine personality types. We do not pursue evolutionary issues in this paper, but a first step in this direction is provided by the payoffs for the different types. The payoffs are given in Appendix 1 in both monetary and utility terms.

To get some indication of the extent to which our participants are aware of their personality type, we added a question at the end of the questionnaire. There, we asked them to choose one of four statements describing how they had acted. The descriptions fitted with the personality types altruist, egoist, social egoist and reciprocator. Alternatively, they could phrase their own description of their behavior. (See the questionnaire in Appendix 2.) It turned out the latter choice was popular which limits the value of this part of the data, but we do see a clear pattern for the two largest groups. Among the 122 social egoists (regions IV and V), 51 percent chose the statement "When I chose I mainly tried to: Be nice to the other if they had been nice, but otherwise consider what is best for me." Similarly, among the 84 egoists (region VI), 76 percent agreed with "When I chose I mainly tried to: Give myself as much as possible".

Finally, looking at the nine regions separately, there are few gender differences. The only ones that might be discerned are in categories I and VI: women are perhaps more likely to be reciprocal than men, while men are more likely to be egoists. However, if we combine all Reciprocators (region I plus region II), the gender differences are more significant: women are more reciprocal at the 5-percent level. If we look at all Altruists (region VII plus region VIII) men are weakly more altruistic (significance level 10 percent).

6.2 The simultaneous games

Next, we consider whether the distribution of types, derived from the second-stage sequential games, can be used to predict behavior in the simultaneous games. In the simultaneous *PD* and *HD* games, where the players are assumed to maximize expected utility, we do not know what value of π each player had in mind. One might argue that the players conceived of the "true" value as given in Table 2, or the *NE* given by equation (6). Here, however, we take an agnostic view and allow individuals to assign any value to π as long as $\pi \in [0, 1]$.

In column 3 of Table 4, we give data on the number of people in each region of Figure 2 who have played "Cooperate", and who have played "Defect", in the simultaneous *PD* game defined by the payoff matrix (8). In column (5) we give the corresponding data referring to the simultaneous *HD* game, with payoff matrix (11). Thus, for each personality type, there are both cooperators and defectors. Is this consistent with rational behavior in our model?

Table 4: Possible values of π for different personality types.

1. Region No.	2. Personality type	3. Number of Cooperate and Defect in PD_{sim}, payoff (8)	4. Choices in column 3 consistent with our model for any $\pi \in [0, 1]$	5. Number of Cooperate and Defect in HD_{sim}, payoff (9)	6. Choices in column 5 consistent with our model for any $\pi \in [0, 1]$
I	Reciprocators	40 C, 14 D	Yes	42 C, 12 D	Yes
II	Reciprocators	5 C, 5 D	Yes	6 C, 4 D	Yes
III	Punishers	4 C, 18 D	No	9 C, 13 D	Yes
IV	Social egoists	48 C, 35 D	Yes	71 C, 12 D	Yes
V	Social egoists	24 C, 18 D	Yes	27 C, 15 D	Yes
VI	Egoists	24 C, 61 D	No	62 C, 23 D	Yes
VII	Altruists	30 C, 4 D	No	29 C, 5 D	No
VIII	Altruists	6 C, 2 D	No	8 C, 0 D	Yes
IX	"Prodigal son's dad"	1 C, 4 D	Yes	2 C, 3 D	Yes
The 48 unclassified individuals		24 C, 24 D		36 C, 12 D	

Consider inequality (4). With payoff matrix (8), it can be written

$$\pi \cdot (\alpha - \beta) \geq \frac{1}{5} - \beta. \tag{12}$$

For regions I, II, IV and V, we have that $\alpha > \beta$, and thus we can write the condition for cooperation as

$$\pi \geq \frac{\frac{1}{5} - \beta}{\alpha - \beta}. \tag{13}$$

If (13) is satisfied for some combination of π, α and β, then an individual with that combination will cooperate; otherwise he/she will defect. Consider now the corner between regions I, II, IV and V, i.e., $\alpha = \frac{1}{3}$ and $\beta = -\frac{1}{5}$. For these values of α and β, inequality (13) yields $\pi \geq \frac{3}{4}$. Thus, in a small neighborhood of the point $\alpha = \frac{1}{3}$ and $\beta = -\frac{1}{5}$ some individuals will choose to cooperate and other individuals will choose to defect, depending on their beliefs about π, with their beliefs satisfying $\pi \in (0, 1)$. Thus the choices observed in regions I, II, IV and V are consistent with rational behavior, which we indicate by a "Yes" in column 4 of the table above for each of these four regions.

For region IX, we know that $\alpha \leq \beta$ and thus we write (12) as

$$\pi \leq \frac{\frac{1}{5} - \beta}{\alpha - \beta}.$$

It is easy to find combinations (for instance, $\alpha = 0.1,\ \beta = 0.3, \pi = 0.5$) such that an individual in the interior of that region is indifferent between cooperation and defection; thus both types of behavior are possible in that region, which is consistent with the 1C, 4D in Table 5. We indicate this with a "Yes" in that box of the table.

Consider now regions III and IV. If $\beta > \alpha$, (12) can be satisfied as an equality only if $\pi < 0$. For all $\pi \in [0, 1]$, everybody in those regions want to defect. Thus the data in Table A2, where some people in those regions actually cooperate, is inconsistent with rational behavior for the case $\beta > \alpha$. If instead $\beta < \alpha$, (12) can be satisfied as an equality only if $\pi > 1$. Thus, for $\pi \in [0, 1]$ everybody wants to defect. We can conclude that regardless of the relation between α and β, the presence of cooperators in regions III and IV is inconsistent with rational behavior. We indicate this with "No" in that box of the table. A similar reasoning applies to regions VII and VIII; in those regions, everybody wants to cooperate if $\pi \in [0, 1]$, which is inconsistent with us observing some defectors in those regions.

In sum, observed behavior in five of the nine regions is consistent with rational choice when the subjects played *PD* with the payoff matrix (8). In terms of numbers of individuals, there were 34 individuals out of 343, i.e., less than ten percent, who displayed an inconsistent behavior.[13]

The number of cooperators and defectors for the simultaneous *HD* game, with payoff matrix (11), is shown in column 5 of Table 4. Using the same kind of reasoning as above for finding out whether observed behavior was consistent with rationality in this game, we obtain the evaluations shown in column 6. For this game, there was only one region (No. VII) where observed behavior was inconsistent with rationality; the 5 altruists who defected should have cooperated instead. In fact, one of these five persons also was among the four altruists who defected in PD_{sim} .[14]

There were 305 persons who made choices in the simultaneous *PD* and *HD* games that were consistent with a $\pi \in [0, 1]$. In total, among the 391 participants in our experiment, the choices of 78 percent (391 – 48 – 38 = 305) were consistent with our model in a within-subject analysis of all the six (simultaneous and second-stage) *PD* and *HD* games. Of course, the remaining participants (22 percent) do not really have to be irrational; they might just have checked the wrong box in the questionnaire. It is also conceivable that a slightly more complicated model (for instance, with concave utility) might accommodate the behavior of at least some of these participants.[15]

7. Are social egoists conditional cooperators?

Our data indicate that a sizeable proportion of individuals belong to a personality type that has been somewhat disregarded in the literature: the social egoist, who is kind to norm-adherers but does not bother about norm-breakers. A reasonable question is whether social egoists *are* present in the literature, but under another name. A possible candidate is the "conditional cooperator" of

[13] We could also ask whether rational expectations, i.e., $\pi = 0.525$ (see Table 1), is consistent with observed behavior. This is a much more stringent requirement. Moreover, there is no obvious learning mechanism that would lead to an *NE* in this experiment, where who are basically unknown to one another meet once and play the games. Nevertheless, it turns out that for regions I, IV, V and IX, observed behavior in PD_{sim} is consistent with an *NE* with, $\pi = 0.525$.

[14] In the KG_{sim} game, observed behavior in all regions except VII and VIII is consistent with an *NE* with $\pi = 0.747$.

[15] For payoffs in the simultaneous games, see Appendix 1.

Ostrom (2000), who, in a number of public-goods experiments seems to reciprocate good behavior in others, while disregarding bad behavior.

Conditional cooperation is a strategy, or a type of behavior (like tit for tat), rather than a set of parameters in the utility function. Thus we do not know whether conditional cooperation (or tit for tat) is the result of utility maximization – and in that case, with what utility function. By contrast, our social egoists are defined by their preferences. The question then is: Do individuals with such parameter values behave like conditional cooperators in other types of interaction than the *PD* game?

To investigate this, we added a two-person public-goods game at the end of our questionnaire. Payoffs to each individual was determined by the following formula

$$\text{Own payoff} = \text{amount withheld} + \frac{(\text{own contribution} + \text{counterpart}'\text{s contr.})\cdot 1.5}{2}.$$

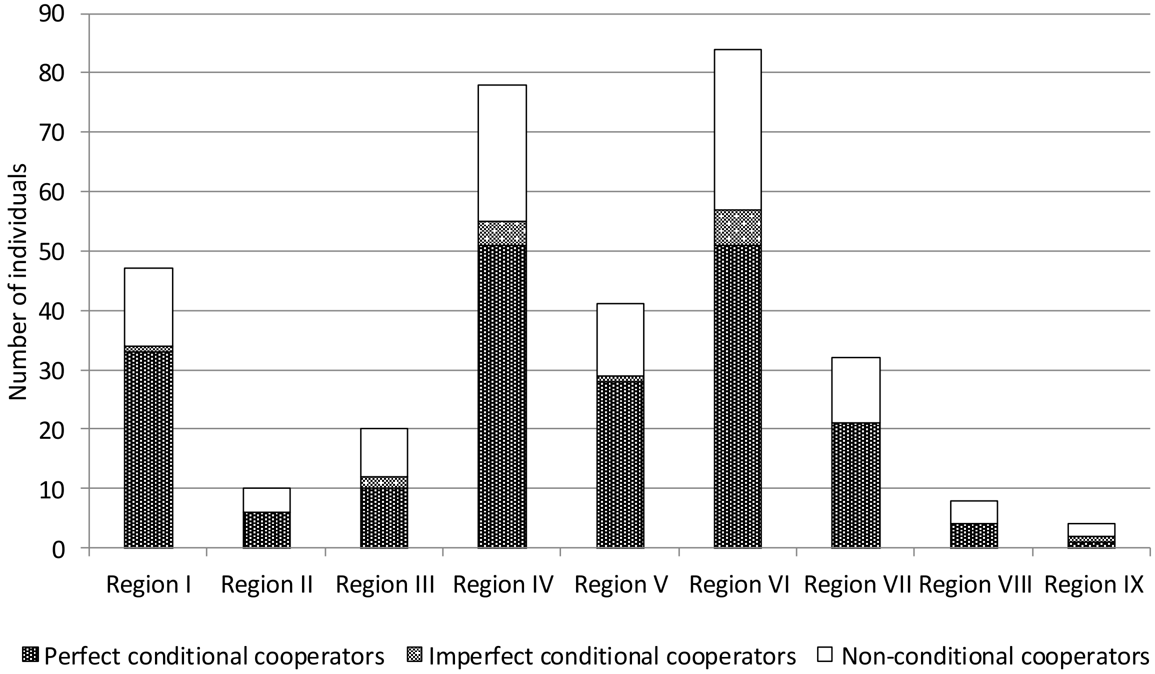
Each individual was endowed with 400 Swedish crowns, and the contributions could be 0, 100, 200, 300 or 400 crowns. We measure contributions for each possible contribution from the counterpart, using the “strategy method”.[16]

Based on the answers and using a similar classification method as Fischbacher et al. (2001), we could identify three categories of individuals: *(i)* Perfect conditional cooperators: Spearman rank correlation coefficient < 0.01, *(ii)* Imperfect conditional cooperators: 0.01 < Spearman rank correlation coefficient < 0.05, and *(iii)* Non-conditional cooperators: all others.

In Figure 3 we can see that conditional cooperators are common among social egoists (in region IV perfect conditional cooperators make up 65 percent of the individuals, and imperfect cooperators another 5 percent, and in region V the perfect cooperators make up 68 percent and the imperfect ones another 2 percent). However, they are by no means more frequent among social egoists than among other types. For instance, in region I (reciprocators) 70 + 2 percent are conditional cooperators; in region VII (altruists), 66 + 0 percent. Even among the egoists in region VI, the conditional cooperators constitute 61 + 7 percent.

[16] Thus, we asked: “Assume that the counterpart contributes *X* crowns. How much do you then contribute?”.

Figure 3: The distribution of conditional cooperators across personality types



The large share of conditional cooperators in all personality types (regions I-IX) in Figure 3 suggests that the concept of conditional cooperation explains a large part of the aggregate behavior in the public goods game. The considerable homogeneity across personality types is remarkable: of each type, two thirds of the individuals are conditional cooperators. In contrast, our results in the *PD* and *HD* games point to large individual heterogeneity in cooperative behavior. This is in line with the findings of Blanco et al. (2011) that a model can have a large predictive power at the aggregate level while its within-subjects predictive power is modest. A possible explanation, mentioned by them, is that these different games activate different behavioral norms. This is an interesting question for future research. In any case, we may conclude that the social egoist, emerging from our data as a common behavioral type, is not just another manifestation of the well-known conditional cooperator.

8. Conclusions

In this paper we have studied norm-based social preferences, and we have derived conditions for cooperation to be an equilibrium in cooperation games. It turned out that within the framework of the model, it is easy to find parameters such that an equilibrium with a positive cooperation rate emerges.

To chart the distribution of parameters in an actual population, we conducted an experiment with participants playing *PD* and *HD* games. Using within-subject analysis, the behavior of the vast majority of the participants can be explained by our notion of norm-based social preferences. We find considerable heterogeneity in social preferences, and based on this heterogeneity we categorize the participants into a number of broad personality types. Besides egoists and altruists, we identify three personality types with norm-based social preferences: reciprocators, punishers and social egoists. In fact, these three constitute more than half the population – and the social egoists alone, one third.

The social egoist is nice to those who conform to society's cooperation norm, and indifferent to those who do not. The idea of such behavior has been suggested before, by Palfrey and Rosenthal (1988) and López-Pérez (2008). But to the best of our knowledge, the prevalence of this behavioral type in a population has not been empirically assessed. When making such an empirical assessment, and finding that the social egoist is quite common, we have not limited the experimental participants to a few narrow categories of behavior. By defining the behavioral types according to parameters (α, β) in their utility functions we have allowed for the possibility that an individual behaves like one type (for instance, an egoist) in one situation, and like another type (for instance, an altruist) in another.

Many interesting questions remain to be addressed, for instance, concerning the formation of beliefs about the fraction of cooperators in a population. Also, there is the question of the emergence of norms, and the importance of repeated interaction. There is also the evolutionary aspect: will the social egoist survive in the long run? In particular, the increasing prevalence of social networking might be a suitable environment for social egoists.

References

Almenberg, Johan, Anna Dreber, Coren L. Apicella and David G. Rand (2011): “Third Party Reward and Punishment: Group Size, Efficiency and Public Goods”, in *Psychology of Punishment,* Nova Science Publishers. Eds. NM Palmetti et al.

Andreoni, James and D. Douglas Bernheim (2009). “Social Image and the 50-50 Norm: A Theoretical and Experimental Analysis of Audience Effects”, *Econometrica* 77(5), 1607-1636.

Andreoni, James and John Miller (2002). “Giving According to GARP: An Experimental Test of the Consistency of Preferences for Altruism”, *Econometrica* 70(2), 737-753.

Bénabou, Roland and Jean Tirole (2011). “Identity, Morals, and Taboos: Beliefs as Assets”, *Quarterly Journal of Economics* 126, 805-855.

Blanco, Mariana, Dirk Engelmann and Hans Theo Normann (2011). “A within-subject analysis of other-regarding preferences”, *Games and Economic Behavior* 72, 321–338.

Boschini, Anne, Astri Muren and Mats Persson (2011). “Men among men do not take norm enforcement seriously”, *Journal of Socio-Economics* 40, 523-529.

Bowles, Samuel and Herbert Gintis (2004). “The evolution of strong reciprocity: cooperation in heterogeneous populations”, *Theoretical Population Biology* 65, 17-28.

Charness, Gary, and Matthew Rabin (2002): “Understanding Social Preferences with Simple Tests”, *Quarterly Journal of Economics*, XX, 817-869.

Clark, Kenneth, and Martin Sefton (2001): “The Sequential Prisoner’s Dilemma: Evidence on Reciprocation”, *Economic Journal* 111, 51-68.

Dreber, Anna, David G. Rand, Drew Fudenberg and Martin A. Nowak (2008): “Winners don’t punish”, *Nature* 452, 348-351.

Dufwenberg, Martin and Georg Kirchsteiger (2004): “A theory of sequential reciprocity”, *Games and Economic Behavior* 47, 268-298.

Ellingsen, Tore and Magnus Johannesson (2008): “Pride and Predjudice: The Human Side of Incentive Theory”, American Economic Review 98(3), 990-1008.

Erlei, Mathias (2008). “Heterogeneous social preferences”, *Journal of Economic Behavior and Organization* 65(3–4), 436–457.

Falk, Armin and Urs Fischbacher (2006): “A theory of reciprocity”, *Games and Economic Behavior* 54(2), 293-315.

Fehr, Ernst and Urs Fischbacher (2004). “Third-party punishment and social norms”, *Evolution and Human Behavior* 25, 63-87.

Fischbacher, Urs, Simon Gächter and Ernst Fehr (2001). “Are people conditionally cooperative? Evidence from a public goods experiment”, *Economics Letters* 71, 397-404.

Geanakoplos, John, David Pearce and Ennio Stacchetti (1989): “Psychological Games and Sequential Rationality”, *Games and Economic Behavior* 1, 60-79.

Gintis, Herbert (2000): “Strong Reciprocity and Human Sociality”, *Journal of Theoretical Biology* 206, 169-179.

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

Lindbeck, Assar, Sten Nyberg and Jorgen W. Weibull (1999). Social Norms and Economic Incentives in the Welfare State, *Quarterly Journal of Economics* 114(1), 1-35.

López-Pérez, Raúl (2008). "Aversion to norm-breaking: A model", *Games and Economic Behavior* 64, 237–267.

Maynard Smith, John and George Price (1973). "The Logic of Animal Conflict", Nature 246, 15-18.

Milinski, Manfred, Dirk Semmann, Theo C. M. Bakker and Hans Jürgen Krambeck (2001): "Cooperation through indirect reciprocity: image scoring or standing strategy?" *Proceedings of the Royal Society of London. Series B:Biological Sciences* 268, 2495-2501.

Nikiforakis, Nikos and Helen Mitchell (2013): "Mixing the carrots with the sticks: third party punishment and reward", *Experimental Economics* in press.

Nowak, Martin A. and Karl Sigmund (2005): "Evolution of indirect reciprocity", *Nature* 437, 1291-1298.

Ostrom, Elinor (2000): "Collective Action and the Evolution of Social Norms", *Journal of Economic Perspectives* 14(3), 137-158.

Palfrey, Thomas R. and Howard Rosenthal (1988): "Private incentives in social dilemmas: The effects of incomplete information and altruism", Journal of Public Economics 35(3), 309-332.

Rabin, Matthew (1993): "Incorporating Fairness into Game Theory and Economics", *American Economic Review* 83(5), 1281–302.

Rand, David G., Anna Dreber, Tore Ellingsen, Drew Fudenberg and Martin A. Nowak (2009): "Positive Interactions Promote Public Cooperation", *Science* 325, 1272-1275.

Ule, Aljaž, Arthur Schram, Arno Riedl and Timothy C. Cason (2009): "Indirect Punishment and Generosity Towards Strangers", *Science* 326, 1701-1704.

Appendix 1: Individual Payoffs

In column 2 of Table A1, we report the sum of monetary payoffs in the four second-stage games, $PD_{\pi=1}$, $PD_{\pi=0}$, $HD_{\pi=1}$ and $HD_{\pi=0}$, for each personality type. We see that the egoists (type VI) are the most successful ones; over the four second-stage games, the average egoist earned 1,500 crowns in total.

Table A1: Payoffs in the four second-stage games for different personality types.

Personality type		Total monetary payoff	Total utility payoff for the average individual
I	Reciprocators $\bar{\alpha} = 0.856$ $\bar{\beta} = -0.305$	1200	1995
II	Reciprocators $\bar{\alpha} = 0.266$ $\bar{\beta} = -0.305$	1300	1425
III	Punishers $\bar{\alpha} = -0.176$ $\bar{\beta} = -0.305$	1400	1304
IV	Social egoists $\bar{\alpha} = 0.856$ $\bar{\beta} = 0$	1300	2056
V	Social egoists $\bar{\alpha} = 0.266$ $\bar{\beta} = 0$	1400	1586
VI	Egoists $\bar{\alpha} = -0.176$ $\bar{\beta} = 0$	1500	1465
VII	Altruists $\bar{\alpha} = 0.856$ $\bar{\beta} = 0.269$	1200	2379
VIII	Altruists $\bar{\alpha} = 0.266$ $\bar{\beta} = 0.269$	1300	1809
IX	“Prodigal son’s dad” $\bar{\alpha} = -0.176$ $\bar{\beta} = 0.269$	1400	1688

To calculate the utility payoffs for the nine personality types, we need information about the distributions of α and β. We get this by using the data to calibrate uniform and independent distributions (A1), as defined below.

Assume that α and β are independently distributed with distribution functions that can be characterized by two parameters *m* and *s*. Thus $Prob(\alpha \leq k_\alpha) \equiv F(k_\alpha;\ m_\alpha, s_\alpha)$ and $Prob\big(\beta \leq k_\beta\big) \equiv F\big(k_\beta;\ m_\beta, s_\beta\big)$, where $F(\cdot\,;\ m, s)$ is the distribution function.

We choose the values of k_α and k_β from (9) and (10):

$$k_\alpha = \frac{A-B}{B-D} \text{ and } \ k_\beta = \frac{C-D}{A-C}.$$

We use the cooperation rates reported in Table 2, together with the values for k_α and k_β given by the two payoff matrices (8) and (11) to obtain four equations in the four unknowns $m_\alpha, s_\alpha, m_\beta$ and s_β.

$$1 - F\left(\tfrac{1}{5}, m_\alpha, s_\alpha\right) = 0.611$$

$$1 - F\left(\tfrac{1}{5}, m_\beta, s_\beta\right) = 0.184$$

$$1 - F\left(\tfrac{1}{3}, m_\alpha, s_\alpha\right) = 0.542$$

$$1 - F\left(-\tfrac{1}{5}, m_\beta, s_\beta\right) = 0.719$$

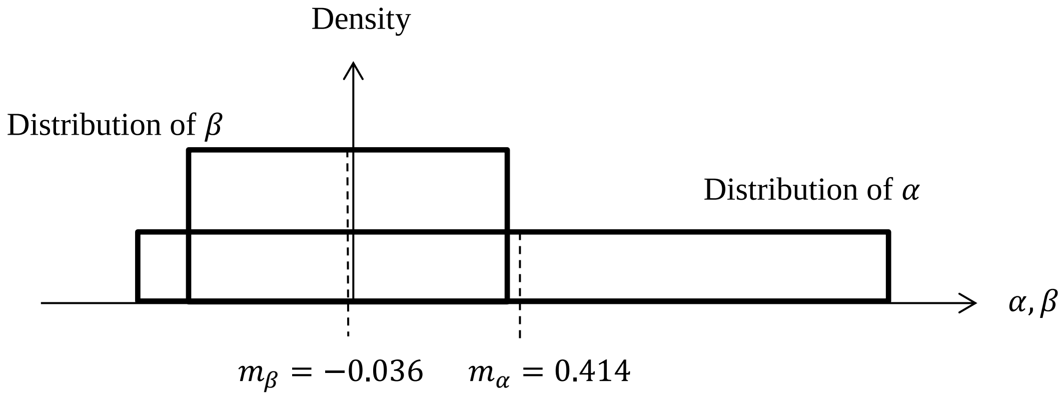
For uniform distributions, $\alpha \sim U[m_\alpha - s_\alpha, m_\alpha + s_\alpha]$ and $\beta \sim U\big[m_\beta - s_\beta, m_\beta + s_\beta\big]$, the solution is:[17]

$$\left.\begin{matrix} m_\alpha = 0.414, & s_\alpha = 0.966 \\ m_\beta = -0.036, & s_\beta = 0.374 \end{matrix}\right\} \qquad \text{(A1)}$$

These distributions of α and β are depicted in Figure A1. With these distributions, and the inequalities (9) and (10), we can calculate the average values of α and β for each of the nine personality types. We can thus compute the utility for the average individual in the nine groups as reported in the last column of Table A1. We see that altruists (VII) enjoy the highest utility, while punishers enjoy the lowest. In fact, the correlation between monetary and utility payoffs is negative and equal to -0.755.

[17] For other distributions, not characterized by two parameters only (for instance, a correlation between α and β), we will of course need more equations.

Figure A1: Calibrated distributions of α and β.



Next, we calculate the payoffs in the simultaneous games. Consider the average individual among the Reciprocators (type I). Applying inequality (4), we see that this individual will cooperate in PD_{sim} if she thinks that $\pi \geq 0.435$. We do not know what this individual thinks about π, but as a numerical illustration of how the model can be used, we apply the true cooperation rate in the population, as given in Table 2.[18] We thus set $\pi = 0.527$, which is obviously larger than the 0.435 required for cooperation. Thus the average individual of type I will cooperate in PD_{sim}. In a fraction $\pi = 0.527$ of the cases, she will meet another cooperator, and in a fraction $1 - \pi = 0.473$, she will meet a defector. Her average monetary payoff is thus 263.50 crowns, and her average utility payoff is 402.5 utils.

For HD_{sim} we also apply inequality (4) to see that the average individual of Type I will cooperate if $\pi \geq 0.664$. Applying the actual cooperation rate $\pi = 0.747$, the average Reciprocator of type I will thus cooperate. This yields an average monetary payoff of 424.10 crowns, and an average utility payoff of 697.52 utils.

Making similar calculations for all personality types, we obtain the results reported in Table A2.

[18] It would be an interesting task to design an experiment to test beliefs about π among different types.

Table A2: Payoffs in the two simultaneous games for different personality types, assuming expected π equal to the actual cooperation rates in the experiment; 0.527 for PD_{sim} and 0.747 for HD_{sim}.

Personality type		Total monetary payoff for the average individual in PD_{sim}	Total utility payoff for the average individual in PD_{sim}	Total monetary payoff for the average individual in HD_{sim}	Total utility payoff for the average individual in HD_{sim}	Total monetary payoff (PD_{sim} + HD_{sim})	Total utility payoff (PD_{sim} + HD_{sim})
I	Reciprocators (C in *PD* & *HD*) $\bar{\alpha} = 0.856$ $\bar{\beta} = -0.305$	263.50	402.5	424.10	697.5	687.60	1100.0
II	Reciprocators (D in *PD* & *HD*) $\bar{\alpha} = 0.266$ $\bar{\beta} = -0.305$	363.50	349.1	473.50	505.5	837.00	854.6
III	Punishers (D in *PD* & *HD*) $\bar{\alpha} = -0.176$ $\bar{\beta} = -0.305$	363.50	349.1	473.50	435.0	837.00	784.1
IV	Social egoists (C in *PD* & *HD*) $\bar{\alpha} = 0.856$ $\bar{\beta} = 0$	263.50	489.1	424.10	743.8	687.60	1232.9
V	Social egoists (D in *PD* & C in *HD*) $\bar{\alpha} = 0.266$ $\bar{\beta} = 0$	363.50	363.5	424.10	523.5	787.60	887.0
VI	Egoists (D in *PD* & D in *HD*) $\bar{\alpha} = -0.176$ $\bar{\beta} = 0$	363.50	363.5	473.50	440.8	837.00	804.3
VII	Altruists (C in *PD* & C in *HD*) $\bar{\alpha} = 0.856$ $\bar{\beta} = 0.269$	263.50	565.4	424.10	784.7	687.60	1350.1
VIII	Altruists (C in *PD* & *HD*) $\bar{\alpha} = 0.266$ $\bar{\beta} = 0.269$	263.50	409.9	424.10	564.3	687.60	974.2
IX	"Prodigal son's dad" (D in *PD* & *HD*) $\bar{\alpha} = -0.176$ $\bar{\beta} = 0.269$	363.50	376.2	473.50	445.9	837.00	822.1

Appendix 2: Instructions

READ THE INSTRUCTIONS CAREFULLY

- YOU MAY GAIN SEVERAL HUNDRED KRONOR

This form has been distributed to everyone who attends today's lecture on micro theory. In order to make the procedure more interesting, Stockholm University will pay real money to the participants. The money has been made available through a research project on decision-making. Since we cannot pay everyone in the course, we will randomly select one tenth of the participants who will be paid in real money according to their decisions.

You will be completely anonymous throughout the investigation and your answers cannot be identified by classmates or teachers. Those who are randomly selected to receive money will have to provide the university administration with their names, addresses and social security numbers. The draw of winners will be based on the "lottery number" that you will choose on the last page.

(Turn the page when the
teacher gives the go-ahead)

Page 1

Questionnaire: Part 1.

The economic decision we want to investigate is the following. Imagine that you will choose one of two options, *X* or *Y*. You have a counterpart (another student in the course) who will also choose one of the options. Depending on how you choose, you will get different amounts of money.

You will choose option *X* or *Y* without knowing which option your opponent is choosing. The result of your decisions is shown in the following table:

		The counterpart chooses	
		X	*Y*
You choose	*X*	You get 500 kr and the counterpart gets 500 kr	You get 0 kr and the counterpart gets 600 kr
	Y	You get 600 kr and the counterpart gets 0 kr	You get 100 kr and the counterpart gets 100 kr

In this part of the survey, you can obtain up to 600 kr and not less than 0 kr. If you are randomly selected to get paid in this part, we will anonymously match you with another randomly chosen student in the course, and pay each person the amount resulting from your choices.

State here which option (*X* or *Y*) you choose: □

(Turn the page when the
teacher gives the go-ahead)

Page 2

Part 1 cont.

In this part of the investigation, you know that your opponent has chosen *X*. The outcome for both of you now depends on what you choose to do. Which option do you choose?

To dispense with the need to scroll back to the previous page, we repeat the table of the different outcomes for your choices:

		The counterpart chooses	
		X	*Y*
You choose	*X*	You get 500 kr and the counterpart gets 500 kr	You get 0 kr and the counterpart gets 600 kr
	Y	You get 600 kr and the counterpart gets 0 kr	You get 100 kr and the counterpart gets 100 kr

In this part of the survey, you can thus obtain between 500 and 600 kronor. If you are randomly selected to get paid in this part, we will anonymously match you with another randomly chosen student in the course who chose alternative *X* on page 1 above, and pay each person the amount resulting from your choices.

State here which option (*X* or *Y*) you choose: □

(Turn the page when the
teacher gives the go-ahead)

Page 3

Part 1 cont.

This part of the investigation resembles the previous one, but you now know that your opponent has chosen *Y*. The outcome for both of you now depends on what you choose to do. Which option do you choose?

To dispense with the need to scroll back to the previous page, we repeat the table of the different outcomes for your choices:

		The counterpart chooses	
		X	*Y*
You choose	*X*	You get 500 kr and the counterpart gets 500 kr	You get 0 kr and the counterpart gets 600 kr
	Y	You get 600 kr and the counterpart gets 0 kr	You get 100 kr and the counterpart gets 100 kr

In this part of the survey, you can thus obtain between 0 and 100 kronor. If you are randomly selected to get paid in this part, we will anonymously match you with another randomly chosen student in the course who chose alternative *Y* on page 1 above, and pay each person the amount resulting from your choices.

State here which option (*X* or *Y*) you choose: □

(Here ends part 1 of the investigation. Turn when the teacher gives the go-ahead)

Page 4

Part 2:

You will choose option *X* or *Y* without knowing which option your opponent is choosing. The result of your decisions is shown in the following table (which is a bit different from the corresponding table in Part 1 above):

		The counterpart	chooses
		X	*Y*
You choose	*X*	You get 500 kr and the counterpart gets 500 kr	You get 200 kr and the counterpart gets 600 kr
	Y	You get 600 kr and the counterpart gets 200 kr	You get 100 kr and the counterpart gets 100 kr

In this part of the survey, you can obtain up to 600 kr and not less than 100 kr. If you are randomly selected to get paid in this part, we will anonymously match you with another randomly chosen student in the course, and pay each person the amount resulting from your choices.

State here which option (*X* or *Y*) you choose: □

(Turn the page when the
teacher gives the go-ahead)

Page 5

Part 2 cont.

In this part of the investigation, you know that your opponent has chosen *X*. The outcome for both of you now depends on what you choose to do. Which option do you choose?

To dispense with the need to scroll back to the previous page, we repeat the table of the different outcomes for your choices:

		The counterpart chooses	
		X	*Y*
You choose	*X*	You get 500 kr and the counterpart gets 500 kr	You get 200 kr and the counterpart gets 600 kr
	Y	You get 600 kr and the counterpart gets 200 kr	You get 100 kr and the counterpart gets 100 kr

In this part of the survey, you can thus obtain between 500 and 600 kronor. If you are randomly selected to get paid in this part, we will anonymously match you with another randomly chosen student in the course who chose alternative *X* on page 4 above, and pay each person the amount resulting from your choices.

State here which option (*X* or *Y*) you choose: □

(Turn the page when the
teacher gives the go-ahead)

Page 6

Part 2 cont.

This part of the investigation resembles the previous one, but you now know that your opponent has chosen *Y*. The outcome for both of you now depends on what you choose to do. Which option do you choose?

To dispense with the need to scroll back to the previous page, we repeat the table of the different outcomes for your choices:

	The counterpart chooses	
	X	*Y*
You choose *X*	You get 500 kr and the counterpart gets 500 kr	You get 200 kr and the counterpart gets 600 kr
Y	You get 600 kr and the counterpart gets 200 kr	You get 100 kr and the counterpart gets 100 kr

In this part of the survey, you can thus obtain between 100 and 200 kronor. If you are randomly selected to get paid in this part, we will anonymously match you with another randomly chosen student in the course who chose alternative *Y* on page 4 above, and pay each person the amount resulting from your choices.

State here which option (*X* or *Y*) you choose: □

(Here ends part 2 of the
investigation. Turn when the
teacher gives the go-ahead)

Part 3.

You and your counterpart get 400 kronor each, that you can keep or invest in a joint pot (you can give 0, 100, 200, 300 or 400 kronor to the pot). Whatever you put in, we will add money to increase the total in the pot by 50 percent. The pot is then split equally between you.

For example, if both give 300 kronor you keep 100 kronor and there is 600 kronor in the pot. We then add 50 percent, to make 900 kronor in the pot. This is shared equally between you, so you get a total of 550 kronor each (including the 100 you saved in the beginning). If instead you give 300 kronor and the counterpart 0, you get a total of 325 and the counterpart 625. If nobody gives anything you keep your 400 kronor.

(i) If you do not know how much your counterpart gives, how much do you give? □

(ii) Assume instead that you know how much your counterpart gives. Indicate by filling out the table how much you give (0, 100, 200, 300, 400) depending on what your counterpart gives:

If your counterpart gives:	**Then you give:**
0 kronor	
100 kronor	
200 kronor	
300 kronor	
400 kronor	

If you are randomly selected to get paid in this part, we will anonymously match you with another student in the course. We will pay you the amounts resulting from your choices.

(Turn the page when the
teacher gives the go-ahead)

Page. 8

Concluding questions:

A. Indicate by circling the statement that best describes how you tried to act generally (i.e. in all games)

WHEN I CHOSE I MAINLY TRIED TO:
Give the counterpart as much as possible.
Give myself as much as possible.
Be nice to the other if they had been nice, but otherwise consider what is best for me.
Be nice to the other if the other had been nice, be mean if the other had been mean.
Other, namely (fill out yourself):

B. Indicate your age here: □

C. Circle your sex: Woman Man

D. How many terms have you studied at the university level before this term? □

Thank you for your participation!

(Turn the page when the
teacher gives the go-ahead)

Fill in the four empty boxes below. Tear off the strip at the stripe. You retain it to sign for your gains. (Make sure you enter the same information above and below the line)
Return the response form to the teacher.

	Select a number between 1 and 1000	Indicate your seminar group
2		

Fold and tear: ---

	Select a number between 1 and 1000	Indicate your seminar group
2		