האוניברסיטה העברית בירושלים
THE HEBREW UNIVERSITY OF JERUSALEM

PARITY, SYMPATHY, AND RECIPROCITY

by

WERNER GÜTH and MENAHEM E. YAARI

Discussion Paper # 354 **April 2004**

מרכז לחקר הרציונליות

CENTER FOR THE STUDY OF RATIONALITY

Feldman Building, Givat-Ram, 91904 Jerusalem, Israel
PHONE: [972]-2-6584135 FAX: [972]-2-6513681
E-MAIL: ratio@math.huji.ac.il
URL: http://www.ratio.huji.ac.il/

PARITY, SYMPATHY, AND RECIPROCITY[1]

Werner Güth

and

Menahem E. Yaari

1. Introduction

In this paper, we consider the results of an experiment in which subjects had systematically deviated from the pursuit and maximization of personal gain. We hypothesize that these departures from self-seeking behaviour are due to one or more of the following factors. (a) *Parity* (also known as *Inequality Aversion*): In choosing among actions, individuals may be attempting to promote equality of outcomes, even at the cost of some reduction in unilateral personal gain. (b) *Sympathy* (also known as *Altruism*): In choosing among actions, individuals may be taking into account not only their own unilateral gains (or losses) but also the gains (or losses) of others. (c) *Reciprocity*: In choosing among actions, individuals may be motivated, to one extent or another, by a desire to apply *measure for measure*, i.e., to reward kindness and unkindness in like manner.

Significant departures of decision makers from purely self-seeking behaviour have been noted, and studied, for a very long time (see Dawes and Thaler [1988]). Experimental investigations of such departures date back at least 50 years (see Sally [1995] for a fairly recent survey).

Parity ("inequality aversion") has been recognized as a possible motivational factor in several experimental settings (see, e.g., Bolton [1991] Fehr and Schmidt [1999], Bolton and Ockenfels [2000]). In some cases, a concern for parity, or equity, is assumed to exist only when the decision maker's position is *inferior*, relative to that of his/her opponent ("one-sided inequality aversion"). In other cases, this concern is assumed to operate symmetrically, regardless of who is holding the advantage – oneself or one's opponent.

Sympathy (the capacity to "feel with one's neighbour as with oneself") has enjoyed prominence since antiquity. David Hume thought that sympathy was at the basis of the virtues that characterize any civilized society. Experimentally, sympathy (sometimes

[1] This essay was written, largely, in the early 1990's. It is based on experimental evidence gathered in the late 1980's. The second author, whose hesitancy had prevented the essay's publication until now, wishes, by presenting it herewith, to pay a debt of affection and gratitude to the first author.

labeled "altruism") has often been used to account for cooperative behaviour that cannot be rationalized on the basis of self-interest alone (see, e.g., Dawes and Thaler [1988], Cooper, DeJong, Forsythe and Ross [1996]).

Finally, reciprocity ("measure for measure") has also long been identified as a force that motivates behaviour, and it has been treated as such in several recent theories of human interaction (see, e.g., Rabin [1993], Fehr, Gechter and Kirchsteiger [1997], Bolton and Ockenfels [2000]).

2. Background

Consider two sums of money, say a dollars and ε dollars, and think of a being quite a bit larger than ε . The following is a 2-player game-form in which the outcomes are dollar payments to be received by the two players.

	C		D	
C	a	a	$-\varepsilon$	$a+\varepsilon$
D	$a+\varepsilon$	$-\varepsilon$	0	0

Figure 1

If each player's evaluation of outcomes is strictly in accordance with his/her dollar earnings (the higher, the better) then this game-form becomes a simple run-of-the-mill Prisoner's Dilemma. We shall refer to preferences of this type (determined solely, and monotonically, by one's own monetary earnings) as **self seeking**. Experimental interactions can easily be designed, where subjects actually play out this game-form, and this has been done many times (for a survey, see, e.g., Roth [1995], Section IIIA). When this is done, one finds that, as the ratio a/ε increases, subjects tend to opt, systematically and with increasing frequency, for the "cooperative" strategy pair (C, C) . Under self seeking preferences, C is a strongly dominated strategy, so observing the strategy pair (C, C) would seem to indicate that behaviour is either irrational or non-self-seeking.

Various arguments have been proposed for why people play C in the above game-form. In some notable cases (see Rabin [1993]), these arguments have been "interactive" in nature, i.e., they have to do with the player's expectation of what his/her opponent is going to do. Something like the following: "I expect my opponent to play C. Given this expectation, for me to play D would be mean. Therefore, I should play C". Conversely, a player's deliberation might go, say, as follows: "I expect my opponent to play D. Given this expectation, the problem with my playing C is not so much the loss of a measly ε

(with D , 0 is all I would get) but rather my coming out the sucker, which I abhor. For this reason, I should play D." In both cases, the underpinning of the final outcome ((C, C) in the first, (D, D) in the second) rests with "interactive" considerations, i.e., with what players expect their opponents to do. In this essay, we deliberately neutralize these interactive considerations, hoping thereby to isolate the pure effects of parity, sympathy, and reciprocity. (In the latter case, we shall consider players moving in sequence, so a minimal interactive element will be retained.) If we take the game-form in Figure 1 and "dissect" it into two separate game-forms, we get

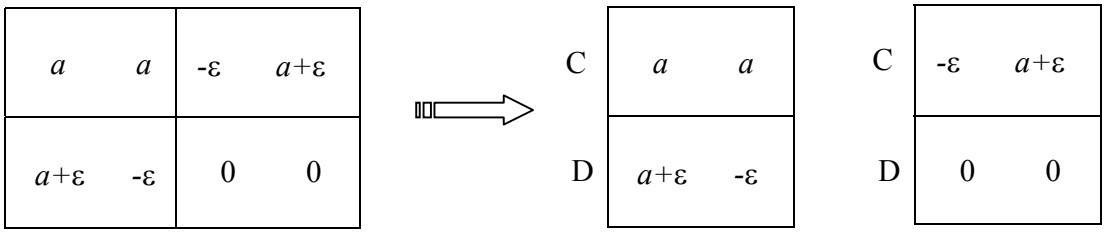

Figure 2

The two resulting game-forms, on the right, are extremely primitive. Only one of the two players has a move, and that move alone determines both players' monetary payments. Primitive encounters of this type are sometimes referred to as "Dictator Games" (see Bolton and Ockenfels [2000]). The question of what one's opponent is expected to do becomes either empty (opponent has no move) or moot (I have no move). Thus, "interactive" considerations are no longer relevant. Nevertheless, in our experimental setting, the "cooperative" move C is observed surprisingly often, with increasing frequency as the ratio a/ε increases. How can one account for this evidence, given that repeated play was completely ruled out? We claim that such an account would have to rely on the introduction on non-self-seeking elements – such as parity, sympathy and reciprocity – into the players' preferences, i.e., into the ways by which players convert game-forms into outright games.

3. The Experimental Setting

The subjects whose responses make up the evidence that will be reported herewith were 1st-year and 2nd-year students at the Hebrew University of Jerusalem. In a certain week during the academic year, experimenters were allowed to take over the last half-hour of class sessions in two large enrollment courses, namely "Introduction to Logic" and "Introduction to Constitutional Law". The students who were in attendance were invited to spend the remaining class time participating in a simple and "potentially fairly lucrative" experiment. They were told that, by the end of the period, each one of them would have earned, possibly, as much as 5 times the University's official hourly student wage, and that average earnings were in fact guaranteed to exceed twice that standard

hourly wage. In each of the classes approached in this manner, 2 or 3 students decided not to participate, with the rest staying on for the experiment. There were 248 subjects in all, divided about equally between Logic and Constitutional Law. Each subject was given a set of instructions, describing 3 tasks to be performed. To complete a task, the subject had to select an action from a "menu" consisting of a small number (usually 2) of simple available actions, and then to carry out the selected action. Physically, the actions consisted in positioning a coloured sticker in some designated space.

In each of the 3 tasks, the subject was in fact a player in a 2-player setup. The game-forms to be played were fully specified, but presumably only the players (subjects) themselves could convert these game-forms into outright games. In all the game-forms, players were completely anonymous vis-à-vis each other, i.e., a player never knew who the opposite player was. In all cases, the opposite player was described, in the subject's instructions sheet, as "another person, also participating in this experiment". And the next sentence was: "There is no way for you to know the identity of this other person, just as there is no way for this person to know your own identity". Subjects knew, broadly, that the experiment was being carried out in several large class sessions, over several days. But they did not know which courses, nor how many courses, would be involved. Thus, the only inference which a subject could draw was that the person on the opposite side was a student in one of the large-attendance courses of the Hebrew University. Moreover, it was made clear to the subjects that opponents were selected through random matchings and that, in any two tasks, a player would in all likelihood be dealing with two different randomly selected individuals on the opposite side.

In 2 out of 3 tasks (to be labeled **Task A** and **Task B**) the game-form being played was degenerate, in the sense that only one player, the subject, had a move, with the other player passively accepting the consequences. In the third task (**Task C**), interaction was genuine, albeit strategically rather primitive.

4. Tasks A and B: Description and Results

In Section 2, we discussed a "dissected" (or "sliced") version of the prisoner's dilemma, i.e., game-forms obtained when one considers each column of the prisoner's dilemma separately (see Figure 2). The first two tasks which our subjects were asked to perform were designed to capture the patterns of behaviour when such primitive game forms are being played out. The subjects' choice problems in Tasks **A** and **B** were as follows:

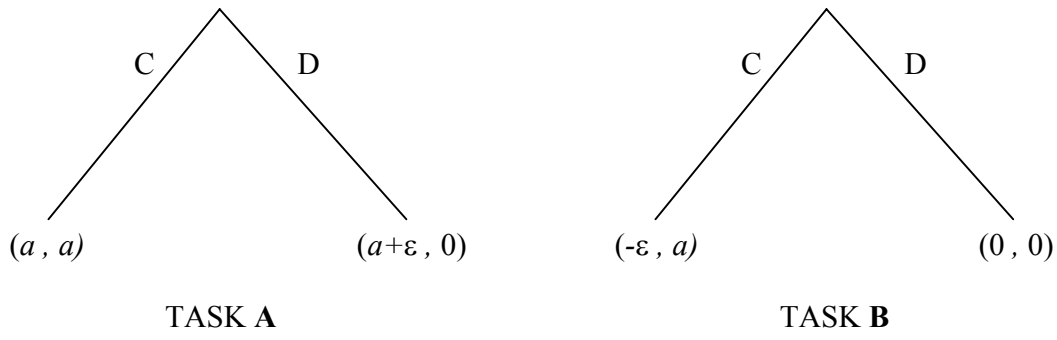

Figure 3

It will be noticed that earnings in these tasks differ somewhat from the relevant payoffs in Figure 1 ($-\varepsilon$ has been replaced by 0 in Task **A**, and $a+\varepsilon$ has been replaced by a in Task **B**). This was due to our concern not to stretch credibility by making statements like "somebody you don't know will be giving up some money". Substantively, the change is of no consequence.

The actions which are here labeled C and D were identified in the experiment in terms of the colours of two stickers, with one of the stickers to be chosen and placed in some pre-designated space. There was no reason to suppose that the stickers' colours, in themselves, affected the subjects' choices in any way. (In fact, the same colour was used to designate C in some cases and D in others.) It is therefore safe to assume that subjects judged the available actions solely by their consequences. The consequence of any given action was a pair of monetary payments, to be made by the experimenter after the completion of the experiment. All the payments, exactly as described in the subject's instructions sheet, were to be made fully and squarely, without lotteries, auctions, or other hanky-panky. The payees in all cases were, first, the subject him/herself and, second, some other unidentified person. There was no way for the subject and this other person to obtain any information on each other, save the fact that both of them were participants in the experiment. In Figure 3, the first and second components of the outcome pair are the amount to be earned by the subject and by "the other person", respectively. All payments were denominated in New Israeli Shekels (NIS).[2]

The results that were obtained for Tasks **A** and **B** are given in the following Table:

[2] At the time, the official rate of exchange was approximately 1.00 NIS = 0.50 US$.

TASK A						TASK B					
a (NIS)	ε (NIS)	C		D		*a* (NIS)	ε (NIS)	C		D	
		N	%	N	%			N	%	N	%
5	0.10	46	85	8	15	10	0.10	37	71	15	29
5	0.25	40	78	11	22	10	0.25	26	74	9	26
5	0.50	71	81	17	19	10	0.50	58	62	36	38
5	0.75	19	68	9	32	10	0.75	26	59	29	41
5	1.00	16	59	11	41	10	1.00	9	39	14	61
ALL PARAMETER VALUES		192	77	56	23	ALL PARAMETER VALUES		156	63	92	37

Table 1

5. Tasks A and B: Analysis

The evidence contained in Table 1 is obviously incompatible with subjects being strictly self-seeking. Moreover, the design of the experiment makes it virtually impossible to attribute this departure from self-seeking behaviour to considerations of repeated interaction or long-term reputation. It is therefore to be concluded that, in choosing among actions, subjects had taken into account not only payments due to be made to themselves but also payments due to be made to other individuals, about whom they know very little. Let (x, y) be the outcome of an action being contemplated by the subject, with x and y being the monetary payment levels for the subject and for the other individual, respectively. Our hypothesis was that subjects would evaluate such an outcome by calculating the quantity

$$U(x, y) = (1-\sigma)x + \sigma y - \eta|x-y|$$

Where σ and η ($0 \leq \sigma \leq 1$ and $\eta \geq 0$) are, respectively, the subject's "coefficient of *sympathy*" and "coefficient of *parity*". The coefficient of sympathy measures the subject's sensitivity to "what's happening to the other fellow" and the coefficient of parity measures the subject's sensitivity to inequality. The condition $\sigma = \eta = 0$ characterizes a purely self-seeking subject.

Notice that in Task **A**, both sympathy and parity tend to push the subject towards the cooperative action, C. That is, in Task **A**, as the values of σ and η become higher, choosing C becomes more likely. In Task **B**, on the other hand, sympathy still works in the direction of reinforcing C, while parity now works in the opposite direction, i.e. higher values of η tend to reinforce self-seeking behavior. This fact makes it possible to use our data to separate out the effects of sympathy and parity. Our hypothesis concerning the evaluation of outcomes by the subjects may in fact be rewritten as follows:

$$U(x, y) = \begin{cases} (1-\alpha)x + \alpha y & \text{if } x \geq y \\ (1-\beta)x + \beta y & \text{if } x \leq y \end{cases}$$

where $\alpha = \sigma + \eta$ and $\beta = \sigma - \eta$. Clearly, Task **A** involves α while Task **B** involves β. Consider a subject whose sympathy and parity coefficients are given, accordingly, by $\sigma = (\alpha+\beta)/2$ and $\eta = (\alpha-\beta)/2$. In task **A**, this subject would choose the action C if $\alpha \geq \varepsilon/(a+\varepsilon)$ and, similarly, in Task **B**, she would choose C if $\beta \geq \varepsilon/(a+\varepsilon)$. Now, Table 1 contains evidence on the frequency of subjects choosing C or D in both tasks. It follows, therefore, that Table 1 contains the frequencies with which the inequalities $\alpha \geq t$ and $\beta \geq t$ are satisfied, for various values of t. In short, Table 1 contains information about the **distributions** of α and β in the population from which our subjects were drawn. We can, in fact, use Table 1 to obtain 5 points on the distribution of α and 5 points on the distribution of β. These points are drawn in the following Figure.

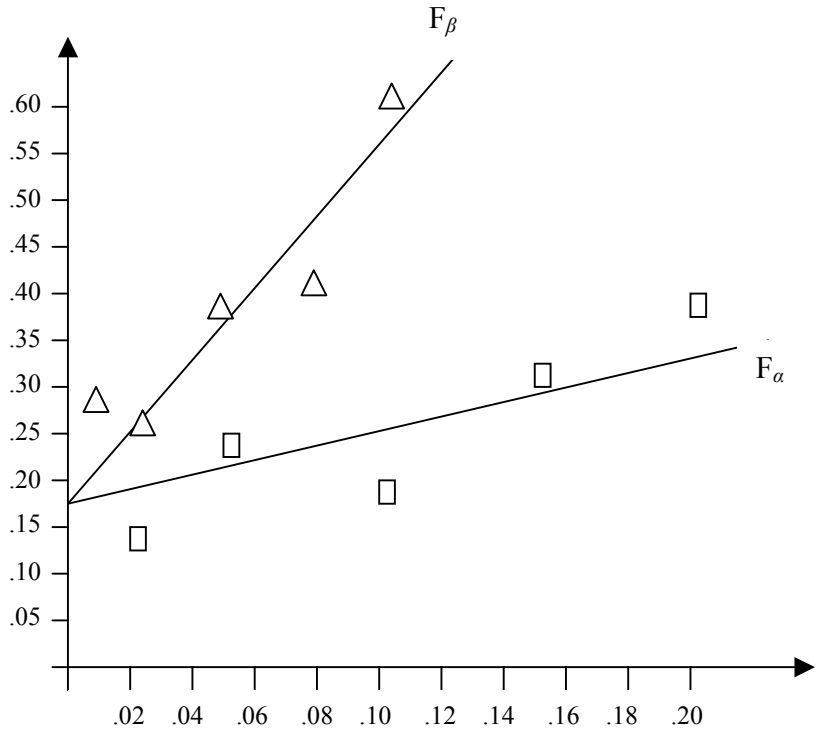

Figure 4

The points in Figure 4 can be used to estimate two distribution functions, one for α and one for β, using linear regression. The estimated distribution functions, obtained in this way, are also drawn in Figure 4.

Let the distribution functions of α and β be denoted F_α and F_β, respectively. The regression equations which have been estimated were –

$$F_\alpha(t) = c + at$$

$$F_\beta(t) = c + bt$$

The two equations were constrained to have a common intercept, c. Removal of this constraint, i.e. allowing the two equations to have separate intercepts, yields an insignificant difference between the two estimated intercepts. Theoretically, equality of the two intercepts is a consequence of our **comonotonicity** assumption (see below).

The results of estimating the regression coefficients in the above equations were as follows:

$c = 0.164$ (t-value 5.07)

$a = 1.039$ (t-value 3.48)

$b = 4.065$ (t-value 6.84)

All the coefficients are significant at the 0.01 level, and the overall fit is quite good ($R^2 = 0.87$, adjusted $R^2 = 0.83$). The estimated distribution functions for α and β (with the coefficients rounded off slightly) come out to be –

$$F_\alpha(t) = \min\{ 0.16 + t , 1\}$$

$$F_\beta(t) = \min\{ 0.16 + 4t , 1\}$$

We now proceed to use these equations in order to retrieve the distributions of the preference characteristics σ and η, recalling that $\sigma = (\alpha+\beta)/2$ and $\eta = (\alpha-\beta)/2$. In order to do this, we must say something about the degree of *dependence* among the variables being studied. Postulating independence is clearly inappropriate, because we expect sensitivity to the position of the other to go hand-in-hand (at least to some extent) with sensitivity to inequality. In the absence of an estimate for the strength of this coherence of tendencies, we postulate complete dependence. More precisely, we suppose the two tendency variables, σ and η, to be (strictly) **co-monotone**. By this we mean that the inequality

$$(\sigma_i - \sigma_j)(\eta_i - \eta_j) > 0$$

holds whenever $(\sigma_i - \sigma_j) \neq 0$, where σ_i and η_i are, respectively, the σ-value and η-value for some individual *i*, and similarly σ_j and η_j for some individual *j*. In other words, ordering the individuals according to their σ-values leads to the same ranking as ordering them according to their η-values. Another way of expressing this is to say that there exists a strictly increasing function, say f , such that $\eta = f(\sigma)$.

Given the co-monotonicity of σ and η , we turn now to the other two ("synthetic") variables, α and β . Clearly, α and σ are also co-monotone, as are α and η . This follows immediately from the definition, $\alpha = \sigma+\eta$. As for the other variable, β , its definition is given by $\beta = \sigma-\eta$, which neither implies nor contradicts the co-monotonicity of β with the other variables. We do know, however, that there exists a real function, say g , such that $\beta = g(\alpha)$. While g is not, a-priori, an increasing function, all we need to do is to exhibit an increasing function g such that the equation $\beta = g(\alpha)$ agrees with the distributions of α and β , as estimated above. In other words, we wish to find g such that $F_\alpha(t) = F_\beta(g(t))$ holds for all t . This leads to $g(t) = t/4$, which **is** strictly increasing. We are now in a position to write the equation $\beta = \alpha/4$, which, together with the relationships $\sigma = (\alpha+\beta)/2$ and $\eta = (\alpha-\beta)/2$, can now be used to retrieve the distribution functions of σ and η , to be denoted F_σ and F_η . The result is:

$$F_\sigma(t) = \min\{\ 0.16 + 1.6\ t\ \ ,\ 1\}$$

$$F_\eta(t) = \min\{\ 0.16 + 2.7\ t\ \ ,\ 1\}$$

for $0 \leq t \leq 1$. The population from which our subjects were drawn may thus be characterized in the following way. There is a hard core (an atom) of pure self-seekers (i.e., individuals satisfying $\sigma = \eta = 0$) comprising one sixth of the population. The remaining five-sixths exhibit the properties of sympathy and parity, i.e. have positive σ's and η's, with the σ-values and η-values of these remaining five-sixths of the population being uniformly distributed. By our co-monotonicity assumption, sympathy and parity go hand-in-hand, i.e. the 2 variables are perfectly correlated. The highest possible value of σ (i.e. the upper bound of the support of σ), as estimated from our raw (**unrounded**) data, is given by 0.505, which is remarkably close to the theoretical maximum, $\sigma = ½$. ($\sigma = ½$ is the case of an agent who regards a dollar paid to oneself and a dollar paid to the other person as equally desirable – "Love Thy Neighbour as Thyself".) As for the parity coefficient, η , the highest possible value allowed by our estimated distribution is given, approximately, by $\eta = ⅓$. In other words, the **highest** degree of inequality-aversion occurs when an individual is willing to pay 1 dollar to correct a 3-dollar gap in incomes. The **average utility function** implied by our data, i.e. the utility U , as defined above, using the estimated **means** of σ and η , comes out to be –

$$U^{av}(x, y) = 0.78x + 0.22y – 0.14|x-y|$$

We know of no *a-priori* discussion of parity and sympathy that would tell us whether this representation of an average individual is "reasonable" or not. It should be remembered

also that the subjects in our experiment were informed that "the person on the opposite side" was also someone participating in the same experiment. All our estimates and results are, of course, specific to this context.

6. Task C

In Tasks **A** and **B**, subjects were dealing with degenerate game-forms, where strategic interaction was altogether absent. The game-form of Task C, in contrast, involved genuine interaction, albeit very limited in scope. The following tree diagram describes the simple setting in which the subjects found themselves:

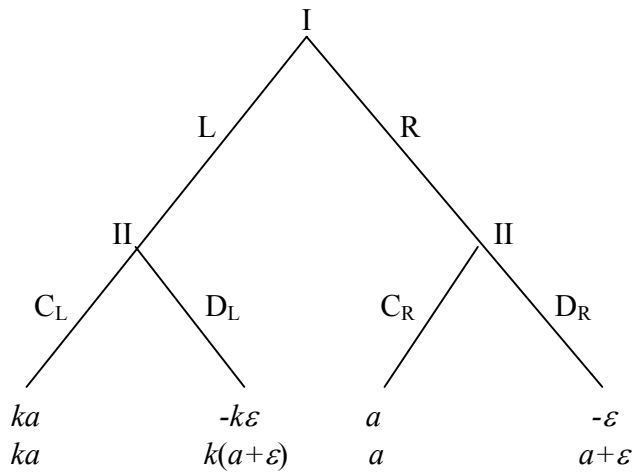

Figure 5

Each subject was assigned the role of either Opener (Player I) or Responder (Player II). Once again, subjects were informed that the person playing in the other role was also a participant in the experiment, whose identity was to remain unknown and for whom the subject's own identity would equally remain unknown. Subjects in the role of Player I were asked to perform their selected action by placing a sticker in a pre-designated space, using the colour of the sticker to indicate the action being taken (**L** or **R** in Figure 5). Subjects in the role of Player II were asked to submit **strategies**, by placing **two** coloured stickers in pre-designated spaces, each sticker indicating what the action shall be, given the colour that Player I will have selected. It was natural for the subjects to think in terms of strategies, in view of the information that the person on the other side was someone who was going to be matched to you (i.e. to the subject) at random, **after** the entire experiment had been run. It was possible, therefore, that the opening move of the opponent had not even been played, at the time of the subject's own action. Thus, specifying the selected action conditionally, for every possible opening move of the opponent, was clearly seen to be the only way.

Note that Player II in Task **C** is actually facing a situation which is the same as that of the decision maker in Task **A**. Indeed, Player II's choice is between a "cooperating" move, which would result in both him/herself and the person on the other side receiving equal positive payments, and a "defecting" move, which yields a somewhat greater payment to oneself but inflicts a loss on the person on the other side. This is true at both nodes where Player II has a move. The difference is that at the left-hand node the stakes are uniformly higher than at the right-hand node, by a factor $k > 1$. Thus, all that Player I's move does is to determine the size of the stakes. Moving Left, by Player I, can be thought of as "being trusting" and moving Right as "being cautious". As before, all payments were denominated in New Israeli Shekels (NIS), and specifically with values $a = 5, 10$ and $\varepsilon = 0.50, 1$. The values of the scale factor were set at $k = 2, 3$. Taking all the combinations of these parameter values yielded a 2x2x2 design, so subjects were divided at random into 8 groups of roughly equal size, with each group filling one of the cells in the 2x2x2 design. On the average, about 30 subjects were assigned to each cell, divided about equally between those in the role of Player I (Opener) and Player II (Responder).[3] Subjects who are uniformly self-seeking would always play **R** ("cautious") in the role of Player I and ($\mathbf{D_R}$, $\mathbf{D_L}$) in the role of Player II, i.e., they would play out the unique subgame perfect equilibrium of the game in which individually received money payments act as payoffs. As we already know, from Tasks **A** and **B**, subjects were not uniformly self-seeking, and this observation was reinforced by the results for Task **C**, which are summarized in the following table (numbers in parentheses are percentages):

	PARAMETERS			PLAYER I (OPENER)		PLAYER II (RESPONDER)			
	a	ε	k	R	L	C_RC_L	C_RD_L	D_RC_L	D_RD_L
1	5	0.5	2	3 (20)	12 (80)	8 (62)	1 (8)	2 (15)	2 (15)
2	5	1	2	2 (14)	12 (86)	7 (58)	1 (8)	1 (8)	3 (25)
3	5	0.5	3	4 (29)	10 (71)	9 (60)	2 (13)	2 (13)	2 (13)
4	5	1	3	4 (24)	13 (76)	7 (44)	2 (13)	2 (13)	5 (31)
5	10	0.5	2	1 (7)	14 (93)	11 (65)	1 (6)	0 (0)	5 (29)
6	10	1	2	4 (24)	13 (76)	12 (67)	0 (0)	3 (17)	3 (17)
7	10	0.5	3	2 (11)	16 (89)	14 (93)	0 (0)	0 (0)	1 (7)
8	10	1	3	2 (14)	12 (86)	8 (53)	3 (20)	0 (0)	4 (27)
OVERALL				22 (18)	102 (82)	76 (63)	10 (8)	10 (8)	25 (21)

Table 2

In our analysis of Tasks **A** and **B**, we had assumed that an individual evaluates an action that yields x Shekels to herself and y Shekels to the person on the other side by calculating the quantity

$$U(x, y) = (1-\sigma)x + \sigma y - \eta|x-y|$$

[3] In the course of running the experiment, discrepancies developed in some cases between numbers of Openers and numbers of matching Responders. In these cases, final payments were determined by randomly forming a few 2-to-1 matchings.

where the pair (σ, η) designates the individual's type. Adopting this framework also for Task **C**, we find that an individual of type (σ, η) in the role of Player II (Responder) would play $(\mathbf{C_R}, \mathbf{C_L})$ if $(1-2\sigma)/(\sigma+\eta) \leq a/\varepsilon$, and $(\mathbf{D_R}, \mathbf{D_L})$ otherwise. (Note that this condition is independent of the scale factor k.) Given the distributions of σ and η, as estimated above, and given our co-monotonicity assumption, the predicted behaviour pattern for Player II (the Responder) would be as specified in Table 3. In order to facilitate comparison with the observed data, each cell in Table 3 contains both the predicted frequency (at the lower left) and the observed frequency (at the upper right). Observed frequencies have been calculated, in each case, by pooling together the data for the 2 values of the scale parameter k.

PARAMETER VALUES		Frequencies of Player II's Strategies (In %)			
a	ε	C_RC_L	C_RD_L	D_RC_L	D_RD_L
5	0.50	61 75	11 0	14 0	14 25
5	1	51 67	10 0	10 0	29 33
10	0.50	79 79	3 0	0 0	18 21
10	1	60 75	10 0	8 0	22 25

Table 3

The hypothesis underlying the predicted frequencies in Table 3 is **not** the hypothesis that we had formulated at the outset for task **C**. We did not expect the framework developed for Tasks **A** and **B** to be adequate for explaining behaviour in Task **C**. Rather, we had expected that, in Task **C**, subjects in the role of Player II would display a tendency to **reciprocate**, a tendency which cannot come into play in the interaction-free settings of Tasks **A** and **B** but can (and would, we thought) come into play in Task **C**. An agent is said to engage in reciprocity if he/she tends to be kind to an opponent who is perceived to have been kind and cooperative, and to be unkind – even vindictive – to an opponent who is perceived to have been unkind and disregarding. Consider an agent in the role of Player II in Task **C**. We expected that, in this situation, at least some subjects would regard the opening move **L** of Player I as kind and trusting, and the opening move **R** of Player I as unkind and distrusting. These subjects (we thought) would then tend to reciprocate, by rewarding a trusting opponent and snubbing a distrusting one. In our setting, reciprocity would be exhibited through Player II's use of the strategy $(\mathbf{D_R}, \mathbf{C_L})$, to which one might refer as the **reciprocating strategy**. Our hypothesis was that the frequency of appearance of this response strategy, $(\mathbf{D_R}, \mathbf{C_L})$, would be positive and significantly greater than the

frequency of appearance of its mirror image, $(\mathbf{C_R}, \mathbf{D_L})$, which we had expected to be negligible. Under reciprocity, the frequencies of the two "straight" strategies, $(\mathbf{C_R}, \mathbf{C_L})$ and $(\mathbf{D_R}, \mathbf{D_L})$, would both be lower that the values predicted by the model of Tasks **A** and **B**, with both reductions together making up the frequency of the reciprocating strategy $(\mathbf{D_R}, \mathbf{C_L})$. In the event (see Table 3), subjects in the role of Player II did select the reciprocating strategy in about 8% of the cases, on average, but a similar number of subjects opted for the mirror-image strategy, $(\mathbf{C_R}, \mathbf{D_L})$, thereby displaying a pattern that seems to run contrary to reciprocity. The case for reciprocity in our data, if there is one, is very weak at best. Possibly this is due to the opening move **R** not being perceived by Responders as necessarily unkind.

Now let us consider the behaviour of Player I (the Opener). The choice as to whether to play R or L clearly hinges on Player I's assessment of who it is out there, at the opposite end of the interaction. The first hypothesis to be checked out in this respect is **Rational Expectations**. Under Rational Expectations, one assumes that the agent in the role of Player I takes the distribution of types in the population as given and picks an action optimally, given this distribution and given his/her type. One then proceeds to check whether the distribution of agents which is implicit in the observed actions of subjects who are in the role of Player I coincides with the underlying distribution, which Player I had taken as given. In the present context, Rational Expectations will have been confirmed if the distribution of types that was derived above for Tasks **A** and **B** were a fixed point of this kind. Checking the data, we find that this is not the case. Indeed, if the subjects in the role of Player I had known that the distribution of types was as described in Section 5 above, then they would all play **L** ("trusting") regardless of their own type, i.e., even a strictly self-seeking Player I would play **L**. While our data do show the rates of playing **L** to be quite high, they fall far short of 100% **L** in all cases.

Given the failure of Rational Expectations, we are led to consider the following alternative hypothesis: When contemplating the question of who it could be, sitting out there at the opposite end of the interaction, the subject's answer might be "it's probably someone more-or-less like myself". This leads us to the hypothesis that a subject of type (σ, η) in the role of Player I will act on the assumption that the opponent (Player II) is also of type (σ, η). We refer to this naïve mode of forming expectations regarding the opponent's type as the **Egomorphic Expectations** hypothesis.

It is straightforward to calculate what an agent of type (σ, η) would do, in the role of Player I, if she assumes that the opponent is of the same type. Using the distribution of types derived in Section 5 above, one can determine, for each parameter configuration, the predicted frequencies with which the 2 actions, **L** and **R**, will be selected. This calculation comes out as follows: For parameter values (a, ε, k), a fraction $(a+8\varepsilon)/(6a+12\varepsilon)$ of the population will play **R** ("cautious"), independently of the value of the scale parameter k. A comparison of these predicted frequencies with observed frequencies is displayed in the following table. (Observed frequencies reflect a pooling of responses over the 2 values of the parameter k.)

Parameter Values		Frequency (In %) of Move R ("Cautious") By Player I	
a	ε	Predicted	Observed
5	0.50	25	24
5	1	31	26
10	0.50	21	18
10	1	25	19

Table 4

We see that the move **R** ("cautious") is observed systematically less frequently than would be predicted under Egomorphic Expectations. However, when we consider the four parameter configurations separately, we find that in each case the difference between the observed frequency and the predicted frequency is not statistically significant. In this sense we are entitled to conclude that the Egomorphic Expectations hypothesis has been confirmed.

7. Relevance

Even after its recent triumphs, experimental work in economics is still held suspect, in some quarters, on grounds of relevance. The argument runs roughly as follows: Evidence regarding the behaviour of subjects in experimental settings may be of considerable interest to psychologists studying individual behaviour per se. Impeccable and robust as their findings might be, these findings can only be taken as evidence of how subjects tend to behave in the given experimental setting. In a market setting, the **very same** subjects may well behave differently, if only because in the marketplace a person's very livelihood is at stake, and this can never be the case in an experiment. Thus, experimental evidence becomes "economically irrelevant" almost by definition.

Similar skepticism can presumably be voiced in the present context: It is true that we find subjects veering off systematically from self-seeking behaviour, but in a market setting "things are *serious*", so these very subjects may well revert to being strict self-seekers. Presumably, this would render our findings economically irrelevant. Quite surprisingly – and inadvertently – we find ourselves in possession of evidence on this issue.

Let us recall some of the design features of our experiment. Subjects who had participated in the experiment were entitled to receive certain monetary payments. The exact amount due to any given subject was determined not only by this subject's own action, but also by the actions of others, who would be matched to this subject later on. Thus, payments could only be made after a delay, a couple of days following the conclusion of all the experimental sessions. This was clearly the only way to guarantee

that everyone received his/her due. Subjects were therefore instructed to write their names on their completed task sheets, and their instructions told them that they could collect their individually calculated earnings at some designated office on campus, at any time during working hours in the week following the running of the experiments.

What happened was this: By the very start of the payment week, a queue had formed outside the door of the designated payment office, of people anxious to collect their rightful earnings. After the people in the queue had been served, individuals continued turning up steadily throughout the working hours, with the pace of arrivals slackening off gradually. This process continued throughout the week. At week's end, it became apparent that only about 70% of the subjects had actually turned up to collect their earnings. It therefore became necessary to declare an extension of the payment deadline, so another week was allowed, for subjects to come and collect their money. This decision was clearly and conspicuously announced on all the appropriate bulletin boards. Most of the remaining subjects did show up within the extension period, but a small residual, consisting of about 5% of all subjects, never did turn up to collect their rightful earnings. (It was decided not to pursue these remaining subjects one by one.)

This course of events made it possible for us to explore the relationship, if any, between "taking money seriously" and a person's tendency to veer away from strictly self-seeking behaviour. Consider the view that non-self-seeking behaviour is a kind of frivolity that will tend to disappear "when things get serious". If this view is correct, then the more seriously the subject regards monetary earnings, the less likely this subject would be to depart from strictly self-serving actions. Fortunately, our observations made it possible for us to test this hypothesis in a straightforward manner. We defined two integer-valued variables, t and s , such that t_i is subject *i*'s time of arrival to collect his/her earnings, and s_i is the number of *i*'s non-self-seeking moves ("moves to the left") in Tasks **A**, **B**, and **C**. (The exact definitions of t and s are simple and unimportant.) On calculating the correlation between these two variables, we found exactly nothing: No statistical relationship whatsoever was found to exist between acting selfishly and taking money seriously. People who were very eager to meet their money were as likely to depart from self-seeking behaviour as were people who were extremely relaxed about it.

8. Conclusion

The fact that people do not always act in a purely self-interested manner has long been recognized. But a detailed analysis of such behaviour, with a view to identifying the systematic elements in it (if any) is only recently being attempted. Our intent, in the present essay, has been to try and contribute to this newly emerging research effort.

9. References

Bolton, G.E. [1991], "A Comparative Model of Bargaining: Theory and Evidence." *American Economic Review*, 81, 5, pp 1096-136.

Bolton, G.E., and A. Ockenfels [2000], "ERC: A Theory of Equity, Reciprocity, and Competition." *American Economic Review*, 90, 1, pp 166-93.

Cooper, R., D.V. DeJong, R. Forsythe, and T.W. Ross [1996], "Cooperation without Reputation: Experimental Evidence from Prisoner's Dilemma Games." *Games and Economic Behavior*, 12, pp187-218.

Dawes, R.M., and R.H. Thaler [1988], "Anomalies: Cooperation." *Journal of Economic Perspectives*, 2, 3, pp187-97.

Fehr, E., S. Gachter, and G. Kirchsteiger [1997], "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econometrica*, 65, 4, pp 833-60.

Fehr, E., and K. Schmidt [1999], "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, 114, 3, pp 817-68.

Rabin, M. [1993], "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, 83, 5, pp 1281-302.

Roth, A.E. [1995], "Introduction to Experimental Economics." In J.H. Kagel and A.E. Roth (*eds*.), **The Handbook of Experimental Economics**, Princeton: Princeton University Press.

Sally, D. [1995], "Conversation and Cooperation in Social Dilemmas." *Rationality and Society*, 7, 1, pp 58-92.