Social Preferences in the Public Arena: Evidence from a Prisoner's Dilemma Game on a TV Show*

Michèle Belot†, V. Bhaskar‡, and Jeroen van de Ven§

April 23, 2007

Abstract

We analyze a large stakes prisoner's dilemma game played on a TV show. Players cooperate 40% of the time, demonstrating that social preferences are important; however, cooperation is significantly below the 50% threshold that is required for inequity aversion to sustain cooperation. Women cooperate significantly more than men, while players who have "earned" more of the stake cooperate less. A player's promise to cooperate is also a good predictor of his decision. Surprisingly, a player's probability of cooperation is unrelated to the opponent's characteristics or promise. We argue that inequity aversion alone cannot adequately explain these results; reputational concerns in a public setting might be more important.

JEL Classification Numbers: C72, C93, D64.

Keywords: prisoner's dilemma, social preferences, inequity aversion, cheap talk, gender differences.

*Thanks to Robin Cubitt, Erik Eyster, Simon Gachter, Sanjeev Goyal, Steffen Huck, Steve Machin, Imran Rasul, Muhamet Yildiz and seminar audiences at the Edinburgh, ESEM/EEA (Vienna), ESSET (Gerzensee), Essex, the Institute of Economic Analysis (Barcelona), Nottingham and IAREP-SABE (Paris) for comments and suggestions.

†Department of Economics, University of Essex, Wivenhoe Park, CO4 3SQ Colchester, United Kingdom. E-mail: mbelot@essex.ac.uk

‡Department of Economics, University College London, Gower St. WC1E 6BT London, United Kingdom. Email: v.bhaskar@ucl.ac.uk

§Department of Economics, University of Amsterdam, ACLE, Roetersstraat 11, 1018 WB Amsterdam, The Netherlands, E-mail: j.vandeven@uva.nl

1 Introduction

Experimental work has established that individuals do not always behave opportunistically in order to maximize their monetary payoffs – the ultimatum game, the one shot prisoner's dilemma and public goods games are all cases in point. This evidence has prompted several theories of "social preferences".[1] Experimental subjects are however anonymous, while agents are rarely anonymous in real life situations. A company manager, who is wondering whether to renege on a deal that he has informally agreed to, knows that his opportunism will be noted and possibly discussed, not only by his business partner, but also by his own colleagues or employees. How do players behave in strategic situations when their behavior is public, i.e. when it is observed by others, and affects their reputation, in a broad sense? Is behavior motivated by the concerns that seem relevant in anonymous experiments, or do different considerations become dominant?

Television game shows provide a "natural experiment" where one can explore this question since individual decisions are very public, and stakes are large. While most TV games focus on risk taking behavior and have little implications for social preferences, the game we study is classical from the point of view of the study of social preferences. Specifically, we study behavior in a prisoner's dilemma game on the show *Shafted* that was broadcast on prime time television in the Netherlands.[2] The critical decision is made at the end of the show, in the *final stage*. Each of two players has to make a decision whether to share (S) or grab (G) a sum of money, X. Each player makes his/her decision independently without knowing the choice of the other player. The monetary amounts earned, as a function of the row player's own decision and that of his opponent, are depicted in Fig. 1.

	S	G
S	$\frac{X}{2}$	0
G	X	0

Fig. 1: Money Payoffs

That is, if both players share, they each get $\frac{X}{2}$; if only one player shares, his opponent gets the entire amount; and if both players choose to grab, they both get zero. If players are concerned only with their own monetary payoffs, the game in Fig. 1 corresponds to a generalized prisoner's dilemma game, where

[1] A non-exhaustive list includes Levine (1998), Fehr and Schmidt (1999), Bolton and Ockenfels (2001) and Charness and Rabin (2002).

[2] The show was broadcast in the Netherlands as *Deelt ie 't of deelt ie 't niet?*

G corresponds to the "defect" action.[3] The amount available to share, X, depends upon the "earnings" of the players, which depend upon their success in answering trivia questions in an earlier stage in the game. The median value of X is €1,683, so that the stakes are substantial.

While this final prisoner's dilemma stage is an important focus of our analysis, the game has a richer structure. The game begins with five players accumulating "money" by answering quiz questions. Players are sequentially eliminated until only three players remain. At this point, the lead player (i.e. the player who is most successful in answering trivia questions) must choose one of the other players with whom to play the prisoner's dilemma, with the stake in this game, X, being given by the sum of the earnings of these two players. This is an important decision and it allows us to study the efficacy of the lead player in choosing a cooperative opponent. These two players can make speeches to each other, and this communication is free format, allowing us to study the informativeness of "cheap-talk" in a public environment.

This paper provides an analysis of the behavior of players in the rich and complex extensive form game. Players share about 40% of the time, and the probability of sharing depends significantly upon observable player characteristics – women are more likely to share than men, and a player who has "earned" a large share of the pot is less likely to share. We also find that the content of communication is a reliable predictor of the sharing decision – a player who makes an explicit *voluntary* promise to share is very likely to honour that promise. On the other hand, if a promise is not volunteered, but arises in response to an explicit question by the presenter of the show, players are no more likely to cooperate than otherwise. This sheds new light on the compulsions and constraints perceived by agents when they engage in what is normally considered "cheap talk".

Our empirical analysis is guided by, and seeks to test, the theory of inequity aversion (Fehr and Schmidt (1999); Bolton and Ockenfels (2000)). A key quantitative prediction we derive from the theory is the 50% rule – sharing can only emerge in equilibrium if the probability of sharing by *each* player is at least 0.5. This is true under the extremely weak assumption that individuals are weakly self regarding in their inequity aversion – that is, the social component of preferences is such that an individual weakly prefers an asymmetric allocation where she is better off, to its mirror image where her opponent is better off. Under somewhat stronger but still very reasonable assumptions, the sharing probability must be even larger, at least $\frac{2}{3}$. In our data, in only 7% of the episodes have pairs of players a predicted probability of sharing above 0.5, and for all but

[3] By a generalized prisoner's dilemma, we mean a game where defect is a weakly dominant strategy.

one pairs of players in our data, we can reject the hypothesis that the sharing probability is greater than 2/3. This implies that inequity aversion does not explain sharing as an equilibrium phenomenon. We suggest that reputational considerations, arising from the publicness of the show, may be more important.

Our second surprising finding is that characteristics which reliably predict a player's sharing behavior *do not* affect the sharing probability of his or her opponent. For example, although a woman is more likely to share than a man, she is no more likely to benefit from her opponent's sharing decision than a man is. Furthermore, a player who promises to share – a promise that our data shows to be quite credible – does not seem to induce sharing behavior by the opponent. This second finding runs counter not merely to the predictions of the theory of inequity aversion, but also to a large body of experimental evidence on reciprocity and conditional cooperation, which shows that players reward niceness and punish nastiness by their opponents. This suggests that behavior in a large stakes public environment could be motivated by very different considerations from those that motivate agents in anonymous experiments. Indeed, reputational considerations may explain this finding, since it is cheaper to acquire a reputation for niceness when the opponent is less likely to share. Our overall conclusion is that some mixture of motives – reputational concerns and inequity aversion – may best explain these empirical results.

It is worth clarifying that we do not study the effect of publicness *per se* upon the propensity of players to share. Budgetary considerations preclude an experimental treatment with comparable stakes and anonymous players or smaller audiences – we conjecture that cooperation levels are likely to be much lower than those we observe on the TV show. Rather, our focus is on deriving conclusions from theories such as inequity aversion, and testing them in the context of a game played in a public environment.

The layout of the remainder of this paper is as follows. Section 2 describes the game show and provides basic descriptive statistics on the participants and their characteristics. Section 3 sets out the alternative theoretical frameworks that guide our empirical work. Section 4 is devoted to the empirical results, while the final section discusses the related literature and concludes.

2 Description of the game and data

The focus of our paper is on the critical decision made by the players, in the final stage of the game, when they must choose whether to share or grab. However, before this decision, the game also has two other important phases:

1. Selection Stage: The game starts with five players, who accumulate earn-

ings by answering trivia questions. At the end of each of three rounds, one player is eliminated, the choice being made by the player who has earned the most in that round. The third of these rounds is particularly relevant: at the end of this round, the player with the highest earnings (the *lead player* henceforth, who has earnings E_1) has to choose one of the two remaining players $\{2, 3\}$ to play the final stage of the game. Players 2 and 3 have earnings E_2 and E_3 respectively, where players are indexed so that $E_2 \geq E_3$. If the lead player chooses player j, then he plays the final stage of the game with this player, and the other player plays no further part in the show. The total stake is given by $X = E_1 + E_j$. The *chosen player* can therefore either be the second or third player.

2. Communication Stage: the two players make speeches to each other in sequence, the chosen player speaking first, and the lead player speaking second. This communication is free format. Sometimes the chosen player may respond to the lead player's speech and in some episodes, the compere of the show may ask one or both players if they will share. At the end of the communication stage, the game moves to the final stage, where the share-grab decisions are made.[4]

We use data from all the 69 episodes of the show, which were aired in the spring of 2002.[5] The total prize at stake, X, varies between €380 and €26,600, with a median value of €1,683. These are considerable sums given that the median monthly income in the Netherlands is roughly €1,200.[6] One criticism of TV show data is that participants may not be a representative sample of the population. While this criticism has some validity, our data suggests that participants have a diverse background in terms of education and occupation, compared to the pool of undergraduate students that are usually the subjects in laboratory experiments. Thus the game show particpants are possibly more representative of the overall population than undergraduate students – see Gertner (1993) and Harrison and List (2004) for a discussion.

Table 1 summarizes basic descriptive characteristics of the players reaching the last selection stage[7]. The average player is 34 years old, with no significant

[4] Before these speeches, the players also declare their "intention" to the TV audience. That is, they simultaneously make "intended choices", from $\{S, G\}$. These intentions are non-binding and are only observed by the TV audience, not by their opponent nor by the compere.

[5] Essentially the same game has been produced in the UK and in Australia, but we were unable to obtain videotapes of these for analysis. The US game show, *Friend or Foe*, has an identical final stage, but the overall extensive form is quite different (see List (2004, 2006), Kalist (2005) and Oberholzer-Gee et. al. (2004). We defer a more detailed comparision of the two games to the concluding section.

[6] This is the median disposable income of a full-time employed person in 2000 (CBS, 2006).

[7] We have analyzed all the selection stages; however, since these are less critical we do not report these findings in order to economize on space.

Table 1: Individual characteristics of players

	lead player	chosen player	eliminated player
N obs.	69	69	69
Mean age	33.9	32.1	35.6
Share women	22%	52%	49%

Table 2: Distribution of outcomes and stakes

Outcome	Frequency	Median stake (€)
S,S	19%	3,090
G,S	48%	1,533
G,G	33%	1,850

difference between ages according to player rank. Women constitute about one-half of players 2 and 3, but only 22% of the first-ranked players, due to the fact that they tend to answer fewer questions than men.[8]

In the final stage of the game, 43% of the players choose to share. Table 2 shows the distribution of joint decisions and median stakes. 19% of the episodes end with a joint outcome "S-S", one-third with "G-G" and 48% with "S-G". Since 0.19 is almost exactly equal to $0.43^2 (= 0.185)$, the decisions of the players are clearly independent, a surprising finding in view of the fact that the game is played in public and each player observes the other's characteristics. We return to this issue in greater detail later.

Cooperative behavior seems to be related with a player's demographic characteristics and his/her performance in the game (see Table 3). Men are 20% less likely to share than women. Lead players are also almost 20% less likely to share than chosen players. Both statistics are highly correlated, since 78% of the lead players are men. Finally, the cooperation rate is slightly higher when stakes are larger (above the median value).

3 Theories of Social Preferences

We now consider how various types of non-pecuniary motivations can influence behavior in the game form. The benchmark case is where agents are concerned

[8] Men answer on average 2.1 questions per series of 10 questions, while women answer 1.7. The frequency of correct answers is roughly the same across gender (78% for men against 75% for women). Answers translate into earnings in a way that depends upon how players choose to "invest". Men also earn more because they invest more aggressively.

Table 3: Individual characteristics and cooperation

	Percentage sharing (%)
Overall	43
Male	36
Female	55
Final rank = 1	33
Final rank = 2	52
Young (age ≤ 34)	41
Old (age > 34)	44
Opponent shares	44
Opponent grabs	42
Prize [€0, €1,683[	41
Prize [€1,683, € 26,600]	44

only with their own monetary payoffs. In this case, G is a weakly dominant strategy, and the unique Nash equilibrium in weakly undominated strategies is (G, G). The game also has asymmetric Nash equilibria, where one player plays S and his opponent plays G, but in these, the S-player plays a weakly dominated strategy – if this player is even slightly uncertain about the other player's intentions (as seems likely in the context), she should play G. So if players have only monetary incentives, the straightforward prediction is that they will all choose to grab. Since this is clearly rejected by the data, we now consider models where players have non-pecuniary motivations.

3.1 Inequity Aversion

Individuals generally make positive contributions in public goods experiments, which suggests that they are altruistic, i.e. they derive utility from the monetary payoffs of others.[9] On the other hand, responder behavior in the ultimatum game suggests that they can be spiteful. Inequity aversion has been proposed as a parsimonious explanation of these and other experiments.[10] Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) propose closely related versions of this theory. A common feature is that a player is assumed to incur disutility if payoffs are unequal. It is also assumed that the total payoff is additively separable in the monetary and "social payoff", i.e. the disutility due to inequity aversion.[11] Let $\tilde{\beta}_i(X)$ be the disutility suffered by player i at the allocation $(X, 0)$, i.e. when i gets X and his opponent gets zero. Similarly, let $\tilde{\alpha}_i(X)$ be

[9] Ledyard (1995) provides an excellent summary.

[10] Levine (1998) proposes an alternative theory, where some individuals are altruistic and others are spiteful; furthermore, the utility that individual i derives from j's monetary payoff also depends on whether j is altruistic or spiteful.

[11] One may allow for risk aversion towards money earnings but we follow the literature and assume linearity for simplicity. In any case, this does not seem to modify the analysis significantly. For example, the 50% rule we derive does not require risk neutrality.

the disutility suffered at the allocation $(0, X)$, i.e. when i gets zero and his opponent gets X (see Figure 2). This is very similar to the model of Fehr and Schmidt (1999).[12]

The intensity of inequity aversion is unlikely to be common knowledge. We therefore assume that there is incomplete information about the preferences of the players. To that purpose, we consider a Bayesian game where for each player i, the vector $(\tilde{\alpha}_i(X), \tilde{\beta}_i(X))$ has distribution $\tilde{F}_i(.)$ on some compact support which is subset of the positive quadrant. That is, nature draws a payoff realization for each player, where these draws are independent, and player i observes his own payoff realization but not that of his opponent. The structure of the game is common knowledge between the players.

	Share (S)	Grab (G)
Share (S)	$\frac{X}{2}$	$-\tilde{\alpha}_i(X)$
Grab (G)	$X - \tilde{\beta}_i(X)$	0

Fig.2 : Payoffs with inequity aversion

We say that inequity averse preferences are *weakly self regarding* if for any type of player, $\tilde{\alpha}_i(X) \geq \tilde{\beta}_i(X)$, i.e. the social component of preferences ranks $(X, 0)$ weakly above $(0, X)$. This assumption, that inferiority aversion exceeds superiority aversion, is widely adopted. Loewenstein, Thompson and Bazerman (1989) provide empirical support for this assumption. Our first result requires only that preferences are weakly self regarding.[13]

To obtain stronger results, we invoke additional assumptions, which are still fairly weak. To reduce the payoff uncertainty to one-dimension, let us assume that $\tilde{\alpha}_i(X) = \gamma\beta_i X$ and $\tilde{\beta}_i(X) = \beta_i X$, where $\gamma \geq 1$ is a fixed constant and β_i is a random variable. In this case, a player's type is one-dimensional. Let β_i be distributed by a cdf $F_i(.)$ with full support on $[0, \bar{\beta}]$. Define $\beta_i^x = \beta : F_i(\beta) = x$. So $\beta_i^{0.5}$ denotes the median value of β_i. Assume that $\beta_i^{0.5} \leq 1$, so that the median type weakly prefers the allocation $(X, 0)$ to the allocation $(0, 0)$. Fehr and Schmidt (1999) make the much stronger assumption that $\beta_i < 1$ for *every* type of player.

Our first proposition makes quantitative predictions regarding the extent of cooperation in any equilibrium with positive cooperation.

[12] We are assuming that the social preferences of a player are defined over a pair of allocations, her own and that of her partner in the final stage. There are differences between the models of Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) —- in Fehr and Schmidt, agents dislike inequality vis-a-vis every other individual, whereas in Bolton and Ockenfels the reference point is the average of payoffs of all agents. However, in the context of two-player games, these differences are not consequential.

[13] Bolton and Ockenfels (2000) assume symmetry in inequity aversion, which is consistent with weakly self regarding preferences but not strictly self regarding preferences. They avoid assuming the latter mainly because it plays no role in their proofs (see p. 180).

Proposition 1 *If preferences are inequity averse and weakly self regarding, then in any equilibrium where any player shares with positive probability, each player must share with probability no less than 0.5. If additionally the types of players are one-dimensional and the median type of any player weakly prefers the allocation* $(X, 0)$ *to* $(0, 0)$*, then in any equilibrium with positive sharing, each player must share with probability no less than* $\frac{2}{3}$.

Proof. Let μ be the probability assigned by i to the event that j shares. Suppose that i has no concern for money, and is only concerned about inequity aversion. If $\tilde{\alpha}_i(X) = \tilde{\beta}_i(X)$, then i will be indifferent between S and G if $\mu = 0.5$. If i also assigns positive weight to his monetary payoff or $\tilde{\alpha}_i(X) > \tilde{\beta}_i(X)$, then the required value of μ is strictly larger than 0.5.

We now turn to proving the second part. Although the relative incentive to play S is not globally increasing in β, we shall show that in any equilibrium, players must follow threshold strategies, whereby a player shares if and only if β_i exceeds a critical value. Define:

$$h(\mu, \beta) = \mu\beta - (1 - \mu)\gamma\beta - 0.5\mu. \tag{1}$$

That is, $h(\mu, \beta)X$ is the payoff difference between playing S and G for a player with belief μ and payoff type β. Letting $\tilde{\mu} = \frac{\gamma}{1+\gamma}$, it is easy to verify that $h(\mu, \beta) < 0 \ \forall \mu \leq \tilde{\mu}$, so that in any equilibrium with positive sharing, the probability of sharing must exceed $\tilde{\mu}$. From equation (1) $h(\mu, \beta)$ is increasing in β for $\mu > \tilde{\mu}$. This implies that if there exists a (μ^*, β^*) such that $h(\mu^*, \beta^*) = 0$, then $h(\mu^*, \beta) \gtrless 0$ as $\beta \gtrless \beta^*$.

Fix an interior equilibrium (β_1^*, β_2^*), i.e. an equilibrium with positive sharing, and let $\mu_j^* = 1 - F(\beta_j^*)$ be the equilibrium sharing probability of player j. Since β_i^* is the type of player i that is indifferent between S and G, $h(\mu_j^*, \beta_i^*) = 0$, which yields

$$\mu_j^* = \frac{\gamma}{(1 + \gamma) - 0.5/\beta_i^*}. \tag{2}$$

From the first part of the proposition, the indifferent type must be below the median, i.e. $\beta_i^* \leq \beta_i^{0.5} \leq 1$. Since $\gamma \geq 1$, this implies $\mu_j^* \geq \frac{2}{3}$. ■

Remark 2 *The argument in the second part of the proposition iterates further. Suppose that* $\beta_i^{1/3} \leq 0.75$ *(and* $\beta_i^{0.5} \leq 1$*). This implies that* $\beta_i^* \leq 0.75$*, implying that* μ_j^* *must be greater than* 0.75*. If we assume instead that* $\beta_i^{1/3} < 0.5$ *then* $\mu_j^* > 1$ *in which case sharing cannot be supported in any equilibrium.*

Note that $\beta_i^{0.5} \leq 1$ allows a very high degree of inequity aversion, since half the players could prefer the allocation $(0, 0)$ to $(X, 0)$. The lesson we draw from

this analysis is that with reasonable degrees of inequity aversion, cooperation probabilities must be very high in order for cooperation to be sustained in equilibrium.

Remark 3 *The second part of the proposition (the two-thirds rule) can also be derived while retaining the assumption of two-dimensional types if we assume that every type prefers the allocation* $(X, 0)$ *to* $(0, 0)$.

It is worth noting that our 50-percent rule is perfectly general and applies to any symmetric binary action game where the monetary payoffs have a prisoner's dilemma structure. This follows since the only assumptions invoked in the proof are symmetry, and that inequity aversion is weakly self regarding. If inequity aversion is to explain cooperation in such a game, then the probability of cooperation must be at least 0.5 or even 2/3 if one is willing to make the additional assumptions. The key assumption for deriving the two-thirds rule is that median type of player prefers $(X, 0)$ to $(0, 0)$. This seems quite reasonable, that the median player is not so inequity averse that she prefers to throw away money. Indeed, Blanco et. al. (2006) provide empirical support – they study a dictator game and find that 97% of the players prefer the allocation $(20, 0)$ to $(0, 0)$.

Our empirical work shows that a player's observable characteristics have significant effects upon the player's probability of sharing. For example, a woman is more likely to share than a man. In a Bayesian equilibrium, these observable characteristics will also affect the *opponent's* sharing probability. In particular, a change in player i's characteristic that increases player i's equilibrium sharing probability will be associated with an increase in player $j's$ sharing probability. The intuition is that with inequity averse preferences, there is strategic complementarity in sharing decisions.

More formally, let us consider a Bayesian game where for each player $i \in \{1, 2\}$, this player's inequity aversion parameter, β, has a distribution $F(\cdot\,|\theta_i)$ that is indexed by the player's observable characteristic, θ_i. The indifference condition for player i, equation (2), can now be re-written as:

$$\mu_j^*(\theta_i, \theta_j) = \frac{\gamma}{(1+\gamma) - 0.5/\beta_i^*(\theta_i, \theta_j)}. \tag{3}$$

Consider now a change in player j's observable characteristic to θ'_j, such that in the new equilibrium, player j has a greater probability of sharing. That is, $\mu_j^*(\theta_i, \theta'_j) > \mu_j^*(\theta_i, \theta_j)$. From equation (3) it follows that $\beta_i^*(\theta_i, \theta'_j) < \beta_i^*(\theta_i, \theta_j)$. Since θ_i has not changed, this implies that $F(\beta_i^*(\theta_i, \theta'_j)) < F(\beta_i^*(\theta_i, \theta_j))$, so that player i also shares with greater probability in the new equilibrium. We have

therefore proved the following proposition.[14]

Proposition 4 *Suppose that a change in a player's characteristic from θ_i to θ_i' increases the player's equilibrium sharing probability. Then this change also increases the player's opponent's equilibrium sharing probability.*

We therefore have another testable implication of inequity aversion. Indeed, it is a significant advantage of the game show set up, where players are not anonymous, that one can specifically test for the effect of an opponent's characteristic upon a player's sharing probability. This proposition also implies that the sharing decisions of the players will be positively correlated. Note that the prediction that players are nicer to those they expect to be nice is not unique to inequity aversion, but is common to a range of different theories including that by Levine (1998), and is an essential element of theories of reciprocity (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006).

Finally, note that the size of the stake X has no implications for the probability of sharing. This is of course a knife-edge result, and small changes in modelling specification could give rise to either a small positive or small negative effect. Note also that changes in X are likely to be associated with variation in other characteristics, including the general knowledge and possibly the social preferences of the participants.

3.2 Reputational considerations

A defining characteristic of a TV show is that, unlike many experimental games where players are anonymous, the behavior of players is public. In such a context, players may be concerned with how they are perceived by the audience, and care about their reputation. Bénabou and Tirole (2006) argue that such a reputational concern can either be instrumental in nature (through its impact on future interactions), or affective (through its consequences for social esteem and shame).[15]

[14] This proposition specifies the relation between two endogenous variables. Although this is somewhat unusual, this is what is relevant for empirical analysis, since economic theory does not specify what the role of characteristics (such as gender) should be. An example of an induced increase in sharing probabilities associated with a change in characteristics from θ_j to θ_j' arises for instance if $F(\cdot \mid \theta_j')$ first order stochastically dominates $F(\cdot \mid \theta_j)$, and the equilibrium is stable.

[15] Duffy and Feltovich (2005) provide an experimental treatment where one player's action in past interactions may be oberved by his current partner. The reputation that a player seeks to develop depends on the specific game being played – e.g. it may be advantageous to have a reputation for niceness if the game is the prisoners' dilemma, but a reputation for nastiness serves you better in the game of chicken. In our context, there is only a one shot game, and players behavior therefore tells us something about the reputation that they seek to develop in the game of life.

To model reputational concerns in a simple way, let us abstract from considerations of inequity aversion, and let $\delta_i > 0$ be the non-pecuniary cost associated with playing G.[16] The payoff matrix now becomes:

	Share (S)	Grab (G)
Share (S)	$\frac{X}{2}$	0
Grab (G)	$X - \delta_i$	$-\delta_i$

Fig. 3: Payoffs with Reputational Concerns

Reputational considerations promote sharing behavior *even* if a player perceives that his opponent is unlikely to share. Indeed, since the payoff cost of S is low when the opponent is less likely to play S, a player may be more willing to play S. The cost of acquiring a reputation is low against an uncooperative opponent. Notice now that (G, G) is not a Nash equilibrium. If the stigma cost is sufficiently high for both players, i.e. if $\delta_i > 0.5X$ for $i \in \{1, 2\}$, then the unique Nash equilibrium is (S, S). If $\delta_1 < 0.5X$, then (G, S) is a Nash equilibrium, and if $\delta_2 < 0.5X$ as well, then both (G, S) and (S, G) are pure strategy Nash equilibria. Most importantly, a concern for reputation does not impose any requirement on the sharing probability in any equilibrium with positive sharing – one can construct equilibria with arbitrarily low (or high) sharing probabilities.

As with inequity aversion, one can analyze an incomplete information version of the game with reputational concerns. Assume that player i's reputation parameter δ is drawn according to a continuous cdf $\hat{F}(\cdot|\theta_i)$ that depends upon the player's observable characteristic θ_i. Consider an equilibrium with a positive probability of sharing. The indifference condition for the marginal type of player i is given by

$$0.5\mu_j^*(\theta_i, \theta_j)X = \delta_i^*(\theta_i, \theta_j). \tag{4}$$

Consider a change in player j's observable characteristic from θ_j to θ_j', such that in the new equilibrium j shares with greater probability. The above indifference condition implies that $\delta_i^*(\theta_i, \theta_j') > \delta_i^*(\theta_i, \theta_j)$, and this implies that player i shares with lower probability in the new equilibrium. We therefore have the following proposition:

Proposition 5 *Suppose that a change in a player's characteristic from θ_i to θ_i' increases the player's equilibrium sharing probability. Then this change* decreases *the player's* opponent's *equilibrium sharing probability.*

[16] A player may be keen not to look foolish, and may incur a different reputational cost if he shares while his opponent grabs. We do not deny that this consideration may play a role; however, this cannot explain sharing behavior, since it only reinforces grabbing.

Note also that the comparative statics of an increase in X is to reduce the equilibrium sharing probabilities – this can be verified readily from equation (4) – assuming that θ_1, θ_2 and the associated distributions are fixed, the only way the pair of equations can hold for the two players is if the critical thresholds for cooperation increase, thereby reducing the probability that each player shares. As before, we have the caveat that variations in X are likely to be associated with variation in other characteristics.

3.3 Communication

We now turn to the role of communication. In our game, we may assume that a player always prefers the allocation $(X, 0)$ to $(0, 0)$. This implies that he prefers that his opponent plays S, irrespective of what he himself intends to play. The structure of the game is therefore similar to that to Aumann's (1993) stag hunt game, where a player strictly prefers that his opponent hunts the stag, irrespective of the action that he intends to take. Aumann's point is that if talk is completely cheap – so that a player's messages have no direct payoff consequences for him – then one should expect communication to be completely ineffective in such games. A player will send whatever message induces his opponent to play S. Recognizing this, his opponent is unlikely to attach any credence to the message, or any promise made.[17]

The experimental literature suggests however, that subjects do not like to lie, and suffer disutility from doing so (see for instance Gneezy (2005)). Ellingsen and Johannesson (2004) formally model this in the context of a trust game, where they assume that a player suffers a cost from lying.[18] If we assume that players suffer a cost if they promise to share but do not honour the promise, then the game effectively becomes a signalling game. With inequity averse preferences, one can show that communication can be effective, in the sense that it increases the extent of cooperation that is possible in equilibrium. All types of player promise to share, but only some of them honour the promise. Since lying is costly, the set of types that share in equilibrium is larger than in the absence of communication.[19]

On the other hand, if reputational considerations are important, then promises do not increase the probability of sharing, since the structure of the game is similar to the game of chicken. If a player i thinks that her opponent has reputational concerns, it is not, prima facie, optimal for her to make a promise that

[17] Farrell and Rabin (1996) present an alternative point of view, arguing that cheap talk may allow players to coordinate on efficient equilibria, even in the context of a stag hunt type game.

[18] Miettinen (2006) extends the analysis to general normal form games, and characterizes equilibrium outcomes with lying costs.

[19] The details of this construction are available from the authors.

she will share. This is the case because such a promise may increase the opponent j's incentives to grab, since the reputational cost of playing G is constant, but the payoff gain from grabbing is increasing in the probability that i shares. Indeed, in this case, a player may want to announce that she will play G. We do not observe players making such an announcement, possibly for two reasons. First, if there is only a disutility from breaking a promise but no disutility from failing to carry out a threat, such an announcement may not be credible. More importantly, if a player announces that she will play G, this may reduce the reputational sanction on the other player. If i announces that she will play G, it becomes common knowledge, to her partner j and to the audience, that i is not a nice person, and is very unlikely to share. In this case, the reputational cost to j from playing G may become small.[20]

4 Empirical analysis

We now turn to the empirical analysis. First, we explore the relation between observable player characteristics and sharing, via a bivariate probit analysis. This leads us to consider the question of selection bias, possibly induced by the lead player's choice of a partner. We then test the 50% rule, and the role of promises. Our results also allow us to calculate the "gender-premium", i.e. the value of having a woman as an opponent in the final stage. We also draw on an experiment we ran, in which we examine whether subjects who watch DVDs of the TV show can predict player trustworthiness, in order to examine the robustness of our results and interpretations.

4.1 Cooperative characteristics

We estimate a bivariate probit model, where the dependent variables are the decisions of pairs of players in the final round (lead and chosen players). Denote by y the decision to share ($y = 1$) or not ($y = 0$) and z the corresponding latent variable such that:

[20] This discussion suggests a more sophisticated model of reputation than the simple one set out earlier, where the reputational cost from grabbing, δ_j, is an increasing function of μ_i^p, the *public* belief that player i will share. If the public belief is that player i is very likely to share, but j nevertheless chooses to grab, then the reputational sanction suffered by j will be large. An announcement by i that she will grab reduces μ_i^p and therefore the reputational cost.

$$
\begin{aligned}
y_{lead} &= 1 \text{ if } z_{lead} = X'_{lead}\beta + \varepsilon_{lead} > 0 \\
&= 0 \text{ otherwise,} \\
y_{chosen} &= 1 \text{ if } z_{chosen} = X'_{chosen}\beta + \varepsilon_{chosen} > 0 \\
&= 0 \text{ otherwise,} \\
&\quad [\varepsilon_{lead}, \varepsilon_{chosen}] \sim \text{bivariate normal } [0,0,1,1,\rho]
\end{aligned}
$$

where X is a vector of player characteristics (age, gender, relative contribution and total stake) and ρ is the correlation coefficient between the disturbances ε_{lead} and ε_{chosen}.

Table 4 reports the marginal effects of demographic characteristics on the individual's probability of sharing. While age does not have any significant effect, the gender effect is substantial.[21] A robust finding is that women are almost 20% more likely to share, confirming the simple cross-correlation presented in the previous section. We also find that players are more likely to share when the stakes are high. An increase in the size of the prize by €1,000 raises the probability of sharing by 3%.

What are the effects of the players' relative contributions to the final pot upon sharing probabilities? We would expect that a player who contributes more to the pot is less likely to share, since she may feel entitled to a larger share of the pie, and this is confirmed – an increase of the relative contribution by 10 percentage points reduces the probability of sharing by 7.1%, when evaluated at the mean. This is consistent with the findings of Cherry et. al. (2002) in the context of a dictator game – a dictator who has contributed more to the pot is significantly less likely to allocate substantial amounts to the recipient. Interestingly, we find that this extends a strategic setting of a *simultaneous move* game, where things might have been expected to be less clear cut.[22] A player who has contributed less might rationally expect her opponent to be more likely to grab, and could possibly be induced to grab. We find no evidence of this here. On the contrary, the chosen players who have contributed relatively little (that is, less than 30%) are significantly more likely to share than the chosen players who have contributed more than 30% (see Table 4, col. 2).

Finally, the estimated correlation coefficient $\widehat{\rho}$ is small (.14) and not significantly different from 0. Hence, we find no evidence of a correlation in the unobservables determining the decisions to share or grab.

[21] We experimented with the age variable in different ways to test whether the effect was non-linear but could not find any supporting evidence.

[22] We find no evidence that inequality of contributions per se affects the probability of sharing.

4.1.1 Testing for a selection bias

An important feature of our show is that the lead player can *choose* his opponent in the final. This potentially biases our estimates – arguably, the lead player tries to select the person who seems most cooperative, possibly by using private signals of trustworthiness that are not captured by public characteristics. If this is the case, the chosen player is more likely to be cooperative than a randomly selected player, giving biased estimates (we elaborate on this issue and its implications in the appendix). Lead players, on the other hand, secure their position in the final round by answering questions and investing wisely, rather than being chosen based on cooperative signals. Since we found no correlation between the performance in the game (measured by the number of questions answered and the average investment) and cooperation, there does not seem to be any reason why lead players should be less or more cooperative on average.

This asymmetry in how lead and chosen players reach the final round could possibly explain why those who contributed relatively little are more likely to cooperate – by definition, a lead player contributes more than 50%, while a chosen player contributes less than 50%. To check this possibility, we pool the sample and estimate a simple probit model. We compare the behavior of selected and lead players who contributed between 30 and 70% (Table 4, col. (3)). The estimated coefficient for the difference in behavior is close to zero and not significant, so that the contribution effect is unlikely to be driven by unobservable characteristics.

Selection bias may also affect other results, in particular the coefficients on public characteristics such as gender. If the lead player associates a public characteristic with cooperativeness, he will require a better private signal from those who do not have that characteristic than from those who do, in order to select them to play the final round. For example, if the lead player believes that women are more likely to share than men, men will only be selected if their private signal more than compensates the gender effect (*ceteris paribus*). In that case, the difference in cooperativeness between men and women playing the final round will be smaller within the sample of chosen players than in the population.

Table 4: Player Characteristics & the Probability of Sharing
Probit estimates - Marginal effects

	(1)		(2)		(3)	
Age	.00	(.01)	.00	(.01)	.00	(.01)
Female	.18	(.09)**	.18	(.09)**	.18	(.09)**
Prize (x €1000)	.03	(.01)**	.03	(.01)**	.03	(.01)**
Contribution (%)	-.71	(.31)**				
Contribution < 30%			.28	(.12)**	.29	(.13)**
Contribution >70%			-.13	(.12)	-.14	(.13)
Contribution 30-70% & chosen					.00	(.10)
H_0 : Joint effect own characteristics = 0 (p-value)	.00		.00		.00	
Rho (standard error)	14	(.20)	.14	(.21)	-	
N. observations (pairs)	69		69		69	

* significant at 10% level, ** 5% level and *** 1% level;(1) and (2) are bivariate probit estimates, (3) are probit estimates; standard errors clustered by episode; the marginal effects are computed at the means of the lead player's characteristics; in col (2) the marginal effects of the contribution dummies correspond to the chosen player's equation and (<30%) and the lead player's equation respectively (>70%).

Table 5: Probability of being chosen by lead player (Probit estimates)
$P(chosen = 1|rank = 2)$

2nd male - 3rd female	.01	(.13)
2nd female - 3rd male	.06	(.11)
Age 2nd - age 3rd	-.01	(.00)*
Score 2nd - score 3rd (x € 1000)	-.16	(.08)**
Contribution 2nd - contribution 3rd (%)	.88	(.43)**
Pseudo R^2	.19	
N obs	69	

* significant at 10% level, ** 5% level and *** 1% level;
standard errors clustered by episode

We present a series of tests to investigate the presence of the potential selection bias in the sample. First, to get more insight into the selection decision, we estimate an equation for the lead player's choice, where the dependent variable is a dummy that equals one when the second player is chosen (see Table 5). Somewhat surprisingly, we find that lead players do not discriminate in favour of women. This indicates that the estimated coefficient for the gender effect on sharing is not biased. Lead players do discriminate against older players, as these are less likely to be chosen. If that is true, older players must have sent a

Table 6: Probability of sharing for lead and chosen players
Bivariate probit estimates - Marginal effects

	Lead player		Chosen player	
Age	-.01	(.01)	.01	(.01)
Female	.31	(.16)*	.14	(.13)
Contribution	-.22	(.47)	-1.18	(.46)***
Prize (x €1000)	.03	(.01)**	.05	(.03)*
H_0 : Equality of coefficients for lead and chosen players (p-value)	.19			
Rho (standard error)	.09	(.21)		
N. observations (pairs)	69		69	

* significant at 10% level, ** 5% level and *** 1% level; standard errors clustered by episode

better private signal to the lead player in order to reach the final round, and the age effect could consequently be biased upwards. Note that money and relative contributions seem to be what matters most in the selection decision. Those who bring most to the final pot are more likely to be chosen.

A second test we propose is to compare the effects of gender and age between lead and chosen players. If there is a selection bias, the age and gender effects will differ systematically between lead and selected players. The bivariate probit specification allows for different coefficients for lead and chosen players. The gender effect is larger among lead players, and less so among chosen players. The difference is not statistically significant, however. Table 6 reports the estimated coefficients separately for chosen and lead players. We cannot reject the null hypothesis that all coefficients are equal for the chosen and lead players.

Our third test is to see how the gender and age of the *excluded* player affect the sharing probability of the chosen player. If a player i is chosen above a player with a favorable public characteristic, then the value of i's private signal must be larger than in the case where i is selected in preference to a player who does not have favorable public characteristics. We find that the characteristics of the excluded player have no significant effect on the sharing decision of the chosen player (Table 7).

Finally, we ran an experiment where we asked our subjects to predict the sharing probabilities of all three players, including the eliminated player. Presumably, the private signals observed by lead players could also be noted by our subjects who were shown DVDs of the TV show. If this is the case, and if lead players do indeed observe useful private signals, we would expect chosen players to be predicted as being more cooperative than eliminated players, controlling

Table 7: Prob. that chosen player shares & characteristics of eliminated player
Probit estimates - Marginal effects

Age	.01	(.01)
Age X eliminated player	.00	(.01)
Female	.14	(.13)
Female eliminated player	.06	(.13)
Contribution 30-50%	-1.19	(.48)**
Prize (x €1000)	.05	(.04)
N. observations (pairs)	69	

* significant at 10% level, ** 5% level and *** 1% level;
standard errors clustered by episode

for observed characteristics. However, we found that there were no significant differences between the predictions for the two types of player.

Overall, our tests find no evidence of a selection bias for gender, age or contributions. Lead players do not consistently select in order to have a cooperative partner. Indeed, it seems that lead players more or less follow a social norm of choosing the second ranked player – they do so 75% of the time. They do not seem to be very strategic in selecting their partner, and in view of this, the selection biases mentioned above may not exist or not be very important.

In a companion paper on the sources of the beauty premium (Belot et al., 2007), we investigate the relationship between physical attractiveness and performance in the different stages of the show. Our striking finding is that less attractive people are significantly more likely to be eliminated, even though beauty has no effect on any aspect of performance on the show – the answering of questions, contributions or on cooperativeness – suggesting a "consumption-value" role for beauty. Since attractive players are no more likely to share than unattractive ones, this reinforces our basic conclusion in this section – selection by lead players does not systematically lead to more cooperative opponents. Thus, the selection bias does not seem to be a serious concern.

4.2 Opponent's characteristics and cooperation

Table 8 investigates how the characteristics of a player's *opponent* influences the decision to share. Somewhat surprisingly, one cannot reject the hypothesis that none of the opponent's characteristics matter for a player's decision. Furthermore, the correlation coefficient between the disturbance terms remains close to zero and not significant. Since there is no correlation between the decisions to

Table 8: Opponent's characteristics and Prob. of sharing
Bivariate probit estimates - Marginal effects

	(1)		(2)	
Opponent female	-.13	(.10)	-.15	(.09)
Both female			.15	(.37)
Opponent's age	.00	(.01)	.00	(.01)
Contribution (%)	-.78	(.34)**	-.79	(.34)**
Prize (x €1,000)	.04	(.02)**	.04	(.02)**
H_0 : Joint effect opponent characteristics = 0 (p-value)	.32		.43	
Rho (standard error)	.14	(.20)	.14	(.20)
N observations (pairs)	69		69	

* significant at 10% level, ** 5% level and *** 1% level;
standard errors clustered by episode

share of the two players, this implies that characteristics that we do not observe but that are observable to the opponent are not driving sharing behavior. This negative finding is inconsistent with a model where cooperation is solely driven by inequity aversion. In that case, a player who is, say, matched against a woman should be more likely to share. The data indicate that there is no evidence for this. This negative finding is striking in view of the large body of experimental evidence for reciprocity in behavior and conditionally cooperative preferences. In our view, the publicness of the environment and large stakes may result in very different motivations, as compared to anonymous experiments. Finally, we also found no correlation between the attractiveness of the opponent and the probability of sharing. Attractive players are not treated more favorably in the final round.

4.3 Testing the 50% Rule

We are now in a position to test a critical implication of inequity aversion set out in proposition 1 – in any equilibrium where sharing occurs with positive probability, each player shares with probability greater than 0.5. Furthermore, if we assume that $\beta_i^{0.5} \leq 1$, then the sharing probability must be greater than 2/3 for both players. The simplest test of 50% rule is as follows. Suppose that a player predicts correctly the aggregate probability of sharing, and uses this in order to estimate the sharing probability of her specific opponent, without conditioning on her opponent's characteristics. The relevant null hypothesis is a one-tailed test where the true aggregate probability, p, is weakly greater than 0.5. This null is easily rejected – our point estimate of p is 0.43, and the corresponding p-value is .04.

This simple test is subject to a caveat. It is possible that players do not

Table 9: Predicted probabilities of sharing

Characteristics	Mean	90% confidence interval	Prob($\widehat{p} \geq 0.5$)
Lead player	.34	[.30,.37]	.00
Woman lead	.43	[.35,.50]	.03
Man lead	.31	[.27,.35]	.00
Chosen player	.52	[.48,.56]	.87
Woman chosen	.61	[.56,.65]	.00
Man chosen	.43	[.38,.49]	.01

play the cooperative equilibrium, with positive sharing, all the time, but only in some fraction of the episodes. Suppose players condition their behavior on some commonly observed signal (a sunspot), and play the cooperative equilibrium (with positive sharing) for some realizations of the sunspot, but play the bad equilibrium (where both players play G with probability one) otherwise. That is, play corresponds to a correlated equilibrium of the Bayesian game. This could conceivably lead to positive sharing, but with an aggregate sharing probability which is less than 0.5. However, this explanation for low sharing implies that sharing behavior should be correlated; in particular, the probability that a player's opponent shares, conditional on the player sharing, should be greater than 0.5. However, we find that this conditional probability is 0.45 in the data, and the null hypothesis that this is greater than 0.5 is easily rejected at the 5 percent level of significance.

To explore the hypothesis further, at the level of each pair of players in our data, we use the results in Table 4 (column (1)) to predict the probabilities of sharing for each player in the final ($\widehat{p}$). We find that there are only 5 episodes where both players had a joint predicted probability of sharing higher than 0.5. We report the mean predicted probabilities in Table 9 and split the results according to observable characteristics such as gender and relative position (lead vs chosen player). We can reject that the probability of sharing is equal or larger than 0.5 for men and for lead players. Only women who are not lead players have a predicted probability of sharing well above 0.5. Overall we can reject that the probability that both final players share with a probability equal or larger than 0.5 in 64 out of 69 episodes. Furthermore, there is only one a pair such that both players have a predicted probability of sharing greater than 2/3.

4.4 The Role of Promises

Before the candidates make their final decision in the last round, they get the opportunity to make a brief speech. The chosen player speaks first, and the lead player speaks after this. This speech is "cheap talk" in the sense that any

promises made are not binding and do not affect monetary payoffs. Players may say anything they please, and there is a diversity of ways in which players use this opportunity. Some players make a promise to share, while others do not – they sometimes talk about what they intend to do with the money; try to convince the other player to share, or say in general terms that "sharing is good". We code the communication into a dummy variable which equals 1 if the player makes an explicit promise to share.[23] We count 45 explicit promises in our sample (out of 138 messages).

Table 10 presents the results of bivariate estimates including a dummy for promises. We find a strong correlation between promises and cooperative behavior: Those who promise to share are *almost 50% more likely to indeed do so*. Those who will grab seem on the other hand reluctant to lie. This is in line with the experimental evidence (Ellingsen and Johannsen, 2004; Charness and Dufwenberg, 2006; Orbell et al., 1990) and suggests that making a promise is not cheap talk, since a lie may be psychologically costly for a player. Our results are also contrary to the findings of Gneezy (2005), who suggests that players are sensitive to the consequences of the lie upon their opponent. That is, player i is disinclined lie, when by doing so the opponent j is induced to take an action that has adverse monetary consequences for j. The peculiarity of our game is that a lie has *no monetary payoff consequences* for the opponent. If a player intends to play G, then her opponent's monetary payoff does not depend on his action – he gets zero irrespective of his action. Nevertheless, we find that players are reluctant to lie.

Our definition of a promise is one that is made voluntarily. In some episodes a player does not volunteer a promise, but is induced to make a promise because the compere asks them explicitly, at the end of their speech, if he/she intends to share. Out of 20 such instances, only one player did not explicitly answer "yes", and we label these yes responses as "forced" promises. We include dummies both for forced and voluntary promises in column (2) of table 10. While the effect of voluntary promises remains large and significant, we do not find any significant effect of forced promises and we can reject the hypothesis that the two coefficients are equal. We indeed find no correlation between a forced promise and observable cooperative characteristics. We conclude that forced promises do not affect behavior, while voluntary promises are an excellent predictor of sharing behavior.[24]

[23] Typical promises take the form: "I will share", "I promise to share" or "I will not let you down". The key criteria we use is that the statement includes the word "I", and a clear statement of the intention to share.

[24] In an interview, we asked the compere the reasons for asking this explicit question. He responded that this was dictated mainly by a consideration of how much time was left for the program. We find no correlation between the compere's question and the observable characteristics of players in the sample of players who not make a voluntary promise.

Table 10: Promises and cooperative behavior
Bivariate probit estimates - Marginal effects

	(1)		(2)		(3)	
Voluntary promise	.47	(.08)***	.46	(.08)***	.46	(.08)***
Forced promise			-.02	(.13)	-.01	(.13)
Voluntary promise opponent					.07	(.10)
Age	.00	(.01)	.00	(.01)	.00	(.01)
Female	.18	(.09)**	.18	(.09)**	.19	(.09)**
Contribution (%)	-.73	(.30)**	-.73	(.29)**	-.75	(.30)**
Prize (x €1,000)	.04	(.01)***	.04	(.01)***	.03	(.01)***
Chosen player						
Rho	-.02	(.25)	-.02	(.26)	-.04	(.26)
N observations (pairs)	69		69		69	

* significant at 10% level, ** 5% level and *** 1% level;
standard errors clustered by episode

Since a voluntary promise is a strong predictor of the player's behavior, we can see whether a promise influences behavior of the opponent. If players interpret promises correctly, and have inequity averse or conditionally cooperative preferences, an opponent's promise should increase a player's propensity to share. However, we find that the opponent's promise has *no effect* on a player's sharing decision – see column (3) in table 10. This is in line with the general finding that the decision to share is not influenced by opponent characteristics or even their choices. It also implies that either players do not interpret promises appropriately, or that conditionally cooperative preferences or inequity aversion do not play a major role. Finally, we cannot reject the hypothesis that promises are independent, even though players speak sequentially.

Our results raise an interesting question: why does a player who intends to choose S not make a promise to do so? Suppose that player i believes that his opponent is inequity averse; furthermore, i knows that j is aware that making a false promise is costly, at least for some types of player i. Then if i intends to play S, it is always optimal for him to make a promise, since this increases the probability that j shares. However, we find that 36% of the players who choose S do not make a promise. It seems unlikely that a player is uncertain of his own action at the communication stage – such an explanation would presume that a player resolves his uncertainty upon hearing his opponent's speech. Indeed, we find that of the lead players who play S, 39% do not make a promise – since lead players speak last, this explanation is very unlikely to apply to them. While one explanation is that some players are unaware of the role of promises – i.e. they may not realize that making a promise is a costly signal of future

intentions, an alternative explanation is that reputational considerations are more important in sustaining sharing. Recall that with reputational concerns, the relative payoff gain from sharing is decreasing in the probability that the opponent shares. The prevalence of reputational considerations, and players' recognition that their opponents may have such preferences, may provide one explanation for our results. At this point, it is worth noting that almost no player ever threatens to play G — as we have argued in section 3, such a threat may well reduce the reputational sanction on a player's opponent when she plays G, by changing the public belief about the player.

Finally, we should note that if players get disutility from lying, this changes the payoffs in the game, and may modify the predictions of proposition 1. For example, a player who has promised to share may now be induced to play S even if the probability of his opponent sharing is less than 0.5. This of course raises the question, why does such a player make a promise in the first place, if he does believe that his opponent only shares with low probability. Nevertheless, to test for robustness, we restrict attention to the sub-sample where neither player has made a promise to share. Here again we find that the probability of sharing is less than 0.5, supporting our earlier results.[25]

Lies and liars

Although a large fraction of players hold their promise, we also have some liars — indeed, 41% of the people who make a promise to share do not abide by it. Lying can only pay if it convinces the opponent to share, i.e. if the opponent is a conditional cooperator[26]. Indeed, a player who intends to grab could increase his *own* monetary payoff if the opponent believes the lie and decides to share. The higher the stakes, the higher the monetary payoff from lying. Table 11 reports probit estimates of the probability of lying (defined as an individual who makes a voluntary promise to share but chooses G) as a function of characteristics of the players. We find no clear relationship between the size of the stakes and the probability of lying. Thus, it does seem that liars do not only care about their own monetary payoff. On the other hand, we find that those who contributed relatively more to the final pot are more likely to lie, perhaps because they feel less guilty about it. The "liars" have no further distinct characteristics – gender, age or ethnicity are not correlated with lying.

[25] If we use the results of table 10 to predict the probability of sharing (including promises), we still find that in only 9 out of 69 episodes both players had a predicted probability of sharing higher than 0.5.

[26] Our results in the previous section show that promises do not affect the behaviour of the opponent, so lying seems totally useless. The fact that players lie suggests that they are unaware of this and believe that they can influence the decision of their opponent.

Table 11: Probability of lying: P(lying = 1)
Bivariate probit estimates - marginal effects

Age	.01	(.01)
Female	-.05	(.13)
Contribution (%)	.44	(.34)
Prize (x €1,000)	-.01	(.01)
Chosen player		
Rho	.05	(.25)
N observations (pairs)	69	

* significant at 10% level, ** 5% level and *** 1% level; standard errors clustered by episode

4.5 The value of a woman

From the previous sections, we know how lead players select their opponents and what the determinants of cooperative behavior really are. Combining these results gives insights into the money that is lost by not selecting the most cooperative player. We have shown that the main determinant of the lead player's decision is the monetary contribution that each player would make to the final pot. Those who would contribute more are more likely to be chosen. Most importantly, the lead player does not discriminate in favour of women.

We use our estimates of column 1 in Table 4 to predict the respective probabilities of sharing of the second and third players, $\widehat{P}_j$, $j = 2, 3$. We find that the third player has a higher predicted probability of sharing than the second player in 49 episodes, with on average a 10% higher likelihood to share. We calculated the expected value of each player, V_j ($j = 2, 3$), where V_j is equal to $\widehat{P}_j X_j$,and X_j is the final stake if player j plays the final.[27] We find that on average the second player is worth €197 more than the third player. However, there are 27 cases where the third player has a higher expected value than the second player, yet the the third player is effectively chosen in 6 of them only.

Turning to gender, table 4 shows that women are 18% more likely to share than men. There are 38 episodes where a man and a woman compete for the position in the final round. On average, a woman is worth €475 more than her male competitor and she has a higher expected value in all these 38 episodes. However, women are chosen in 20 of these 37 episodes only. In only 3 of these episodes is she the third-ranked player.

Table 12 reports the average difference in expected values between the second and third player, as a function of their gender. The average difference in

[27] This is the expected payoff of choosing player j if the lead player plays G in the final stage, and twice the expected payoff if he/she plays S.

Table 12: Gender mix, optimal and actual choice

Gender mix	$V_2 - V_3$	Modal optimal choice	Frequency optimal choice
Woman second, Man third	€626 ($N = 22$)	Woman	$\frac{17}{22}$
Man second Woman third	−€268 ($N = 16$)	Woman	$\frac{3}{16}$
Woman second Woman third	€184 ($N = 16$)	Second	$\frac{13}{16}$
Man second Man third	€78 ($N = 15$)	Second	$\frac{10}{15}$

expected values is €626 when the female player is in a second position. Women who are in a second position are on average more likely to share than men in a third position, i.e. the gender effect more than compensates the effect of relative contributions.

As a rule of thumb, in order to maximize expected earnings the lead player should always select the female player, and take the highest ranked one if the choice is between two players of the same gender. Table 12 shows that in around 40% of the episodes the player chosen is not the one with the highest expected value. The choice seems mostly based on the ranking of the players, i.e. the lead player chooses the second player in 75% of the cases. One possibility is that lead players simply choose the second ranked player very often, since this is the "fair" thing to do.

4.6 Can observers predict player behavior?

One possible explanation for why we do not find that a player's characteristics affect the sharing probability of his or her opponent, is that players do not correctly interpret how their opponents' characteristics influence the probability of sharing. To explore this, we ran an experiment in which subjects were shown recordings of the game show and asked to predict the choices of the players. Our goal was to examine how accurately subjects interpret observable characteristics. A detailed exposition and analysis of the results of this experiment can be found in a companion paper.

The results reveal that predictions are very noisy and that, on average, subjects are not very good at predicting. Nevertheless, we do find that students accurately interpret some basic characteristics. Firstly, subjects predict an aggregate sharing probability of 42% which is very close to what we obtain in our sample. This supports our analysis of the 50% rule, since it shows that the average subject is not significantly over-optimistic about human sharing behavior as compared to the reality of the game show. Second, subjects correctly believe

that women are more likely to share than men (although they underestimate the magnitude of the difference), and that the player who contributed most is less likely to share. On the other hand, their predictions are correlated with characteristics that are not associated with cooperativeness in the original data. For example, subjects expect older players to be more likely to share, while we find no strong relationship between age and cooperative behavior in our data. Overall, the results of this experiment provides some support for the equilibrium analysis conducted in this paper since it shows that subjects expectations are not out of line with the reality of actual behavior.

5 Concluding Comments

This paper speaks to a number of literatures, including empirical work on the prisoner's dilemma, on the nature of social preferences, on pre-play communication and on gender differences in kindness or altruism. Space considerations do not permit an exhaustive discussion of all these literatures, and we will focus on what seems most directly pertinent.

The prisoner's dilemma is probably the most famous of games, and has been the subject of innumerable experiments. Our analysis has focused on deriving and testing sharp quantitative and qualitative predictions of an influential theory – the theory of inequity aversion – in the context of this prisoner's dilemma game played on public television. These predictions also apply to *any* prisoner's dilemma with symmetric monetary payoffs, not just to the payoffs presented in Fig. 1.

The game show we analyze is somewhat similar to the US TV game show "Friend or Foe" – the final prisoner's dilemma stage is identical, but the overall extensive form is different. List (2004, 2006) and Oberholzer-Gee et al. (2004) analyze this data and find an overall sharing rate of 0.5, with cooperation dropping between the first and second seasons of the show. "Friend or Foe" has different extensive form overall as compared to Shafted, since pairs of players play as a team to accumulate prize money. This implies that is not possible to identify individual contributions to the total stake, as in our show. Players sort into teams at the beginning of the show, unlike our game where the lead player makes a selection decision towards the end, just before the prisoner's dilemma stage.

Our paper differs from this work in two respects. First, given the richer extensive form, we are able to provide empirical analysis of players' communication and the lead player's selection decision.[28] Second, we have tested

[28] "Friend or Foe" has a selection stage right at the beginning, when individuals form teams. List (2006) shows that the selection of partners at the beginning of the show is driven by

behavior against the clear quantitative and qualitative predictions of inequity aversion and other alternatives, that are derived in the present paper. We may apply these tests to the summary statistics reported in the "Friend or Foe" papers. The overall cooperation rate is not inconsistent with the 50% rule derived in proposition 1, but does seem prima facie inconsistent with the minimum probability of $\frac{2}{3}$ derived under stronger assumptions.

The literature on experimental prisoner's dilemma games is enormous – Sally (1995) provides a survey and a meta-analysis. Focusing on symmetric binary action prisoner's dilemma games where there are no serious concerns about repetition effects,[29] the cooperation rate in experiments is quite low, often in the range of 20 to 30 percent (see for instance Andreoni and Miller, 1993; Andreoni and Varian, 1999; and Cooper et al., 1996). These low rates of cooperation cannot be explained by inequity aversion, and suggest that even in a laboratory environment, other considerations are important. The cooperation rate is typically higher in experiments where subjects are not anonymous and/or can communicate (Dawes et al., 1977; Orbell et al., 1990; Ledyard, 1995; Sally, 1995, Bohnet and Frey, 1999), and indeed often well above the 50% threshold.[30] In comparison with prisoner's dilemma experiments with communication, our cooperation rate is relatively low. Since this occurs despite the publicness of the environment, it suggests that the enormity of the stakes is playing a role in inducing grabbing behavior.

From a theoretical standpoint, our main finding is that inequity aversion cannot explain cooperative behavior. Charness and Rabin (2002) and Engelmann and Strobel (2004) present experimental evidence from dictator and other games of distribution, showing that inequity aversion alone is inadequate to explain behavior. In the two-player context, these papers suggest that players are also motivated by efficiency considerations. Such a concern for efficiency could also be important in our context.

One puzzling finding is that we find many good predictors of a player's sharing decision – gender, relative contribution, and voluntary promises are all extremely informative. Nevertheless, these do not influence the opponent's

characteristics correlated with expected payoff. However, the expected payoff from a partner depends on expected earnings from answering questions as a team as well the partner's probability of sharing. The advantage of our setting is that when the lead player makes her selection decision, she observes each potential partner's realized earnings, so that uncertainty only pertains to the partner's sharing decision. We are therefore able to focus on the ability of lead players to choose "cooperative" opponents.

[29] Most laboratory experiments have some element of what subjects could construe as repetition, despite careful design. One advantage of our TV show is that subjects know for sure that it is a one shot game.

[30] The extent of cooperation depends on whether communication is written or face-to face, one-sided or two-sided (e.g. Cooper et al., 1989, 1992; Crawford, 1998). Holt (1995) claims that communication is ineffective when the incentives to defect are high.

decision in any way. This appears to be counter to the predictions of much of the experimental literature which shows that people want to be nice to the nice, and nasty to the nasty. There is a very wide range of literature establishing this phenomenon, including public goods experiments and the ultimatum game. Indeed, Rabin (1993), Levine (1998), Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) are all papers which develop theories that are motivated by this literature. We have suggested that a combination of reputational concerns and inequity aversion could explain this puzzle, but this deserves further exploration.

There is substantial experimental evidence showing that pre-play communication permits a higher level of cooperation or trust. Ellingsen and Johannesson (2004) study an experimental version of the hold-up problem with one-sided communication. In one treatment the non-investing party is allowed to make a promise, while in another treatment the investing party may make a threat. They find that pre-play communication improves efficiency when the non-investing party is allowed to communicate. Subjects are reluctant to lie, but promises are more credible than threats. Gneezy (2005) presents experimental evidence showing that the cost to a player from telling a lie depends upon the consequences upon the opponent – a lie that imposes smaller costs upon the opponent when believed is more palatable. As we have seen, in our game it can plausibly argued that lying itself imposes no monetary cost upon the opponent; nevertheless, players are reluctant to lie, and voluntary promises are surprisingly informative. Charness and Dufwenberg (2006) study a trust game, and find that pre-play communication is informative and effective. They set out a somewhat more complex model, that is inspired by the theory of psychological games. Duffy and Feltovich (2006) provide an experimental comparison of the effectiveness of communication as compared to observation of past actions in one-shot games.

Our findings also relate to the literature on gender differences in altruism and reciprocity. There does not seem to be a consensus on whether or not women are more altruistic than men – Eckel and Grossman (1999) provide a review of this literature. Croson and Buchan (1999) analyze trust games and find that women exhibit greater reciprocity, while the degree of trust exhibited does not differ according to gender. Andreoni and Vesterlund (2001) find that women are more altruistic than men when altruism is expensive, while men are more inclined to be generous when generosity is cheap. This is consistent with our results, since sharing behavior is expensive in our high stakes context. Additionally, the publicness of the context may play an important role in generating gender differences in behavior that maybe absent in anonymous experiments.

To conclude, we examine behavior in a prisoner's dilemma type game played

in a public setting with high stakes. We suggest that a mixture of motives, inequity aversion plus reputation, may best explain our results. Our findings suggest that behavior in a public setting may thus be motivated by different concerns from behavior in anonymous experiments, and that this difference may be economically relevant and worth exploring further. Indeed, this may be amenable to experimental investigation, since one may be able to vary the degree of publicness of individual decisions.

6 Appendix

We present the formal analysis that underlies our tests of selection bias in section 4.1.1. Suppose that if player $i \in \{2, 3\}$ plays the final stage, his or her decision can be summarized by the following decision rule, whereby i shares if and only if the variable y_i exceeds some critical level $\bar{y}$, where y_i is a linear function of three components:

$$y_i = \beta X_i + \gamma Z_i + \varepsilon_i. \tag{5}$$

X_i is a *public characteristic*; to simplify exposition, assume that it is a binary variable coding gender, with 1 denoting female.[31] Z_i is a scalar continuous *private signal* which is observed by the lead player and player i, but not by us. For simplicity we assume that Z_i is independent of X_i, and ordered so that $\gamma \geq 0$. Finally, ε_i is independent of X_i and Z_i and is observed only by player i.

Let us consider how selection by the lead player on the basis of her private signal affects the observed sharing probability as compared to random selection. Suppose that players 2 and 3 have the same public characteristic (e.g. they are both male). If the lead player chooses in order to maximize the probability that her partner shares, she will select the player with a larger value of Z_i, since they do not differ in terms of X. Since Z_i is relevant to the sharing decision, the probability that player i shares, conditional on being selected by the lead player, will be larger than if selection was purely random. This "selection-bias" will be larger the better the quality of the private signal, i.e. the larger the variance of γZ_i relative to ε_i.

Now let us consider how the manner of selection affects the difference in sharing probability between females and males. Under random selection, a male shares with a probability equal to that of the event $(\gamma Z_i + \varepsilon_i \geq \bar{y})$, while a female shares with a probability equal to that of the event $(\gamma Z_i + \varepsilon_i \geq \bar{y} - \beta)$.

[31] More generally, X_i can denote any vector of i's public characteristics – this makes no difference to the analysis.

When the lead player selects so as to maximize the sharing probability, this difference between males and females is reduced. Essentially, when players 2 and 3 are of different sex, females are selected in preference to males even if they have relatively low values of the private signal. Thus the average sharing probability of females is reduced while that of males is enhanced.

References

[1] Andreoni, J. and J. Miller (1993), Rational Cooperation in the Finitely Repeated Prisoner's Dilemma, *Economic Journal* 103, 570-585.

[2] Andreoni, J. and H. Varian (1999), Preplay Contracting in the Prisoner's Dilemma, *Proceedings of the National Academy of Sciences USA* 96, 10933-10938.

[3] Andreoni, J. and L. Vesterlund (2001), Which is the Fair Sex? Gender Differences in Altruism, *Quarterly Journal of Economics* 116, 293-312.

[4] Aumann, R. (1993), Nash Equilibria are not Self-Enforcing, In: Gabszewicz, J., Thisse, J.-F., Wolsey, L. (Eds.), *Economic Decision Making: Games, Econometrics and Optimization,* Elsevier, Amsterdam, 201–206.

[5] Bénabou, R. and J. Tirole (2006), Incentives and Prosocial Behavior, *American Economic Review*, forthcoming.

[6] Blanco, M., D. Engelmann and H. Normann (2006), A Within-Subject Analysis of Other-Regarding Preferences, mimeo.

[7] Bohnet, I. and B. Frey (1999), The Sound of Silence in Prisoner's Dilemma and Dictator games, *Journal of Economic Behavior and Organization* 38, 43-57.

[8] Bolton, G. and A. Ockenfels (2000), ERC: A Theory of Equity, Reciprocity, and Competition, *American Economic Review* 90(1), 166-193.

[9] Charness, G. and M. Dufwenberg (2006), Promises & Partnership, *Econometrica,* 74, 1579-1601.

[10] Charness, G. and M. Rabin (2002), Understanding Social Preferences with Simple Tests, *Quarterly Journal of Economics* 117(3), 817-869.

[11] Cherry, T., P. Frykblom and J. Shogren (2002), Hardnosing the Dictator, *American Economic Review* 92(4), 1218-1221.

[12] Cooper, R., V. Douglas, R. DeJong, R. Forsythe and T. Ross (1996), Cooperation without Reputation: Experimental Evidence from Prisoner's Dilemma Games, *Games and Economic Behavior* 12, 187-218.

[13] Cooper, R., V. Douglas, R. DeJong, R. Forsythe and T. Ross (1989), Communication in the Battle of the Sexes Game: Some Experimental Results, *Rand Journal of Economics* 20(4), 568-587.

[14] Cooper, R., V. Douglas, R. DeJong, R. Forsythe and T. Ross (1992), Communication in Coordination Games, *Quarterly Journal of Economics* 107(2), 739-771.

[15] Crawford V. (1998), A Survey of Experiments on Communication via Cheap Talk, *Journal of Economic Theory* 78, 286-298.

[16] Croson, R. and N. Buchan, (1999), Gender and Culture: International Experimental Evidence from Trust Games, *American Economic Review* 89(2), 386-391.

[17] Dawes, R., J. McTavish, and H. Shaklee (1977), Behavior, Communication, and Assumptions about other People's Behavior in a Commons Dilemma Situation, *Journal of Personality and Social Psychology* 35(1), 1-11.

[18] Duffy, J. and N. Feltovich (2006), Words, Deeds and Lies: Strategic Behavior in Games with Multiple Signals, *Review of Economic Studies*, 73, 669-688.

[19] Dufwenberg, M. and G. Kirchsteiger (2004), A theory of sequential reciprocity, *Games and Economic Behavior* 47, 268-298.

[20] Eckel, C. and P. Grossman, (1999), Differences in Economic Decisions of Men and Women, in C. Plott and V. Smith (eds.) *Handbook of Experimental Results,* Amsterdam: Elsevier.

[21] Ellingsen, T. and M. Johannesson (2004), Promises, Threats and Fairness, *Economic Journal* 114(495), 397-420.

[22] Engelmann, D. and M. Strobel (2004), Inequality Aversion, Efficiency, and Maximin Preferences in Simple Distribution Experiments, *American Economic Review* 94(4), 857-869.

[23] Falk, A. and U. Fischbacher (2006), A theory of reciprocity, *Games and Economic Behavior* 54, 293-315.

[24] Farrell, J. and M. Rabin (1996), Cheap Talk, *Journal of Economic Perspectives* 10(3), 103-118.

[25] Fehr, E. and K. Schmidt (1999), A Theory of Fairness, Competition, and Cooperation, *Quarterly Journal of Economics* 114(3), 817-868.

[26] Gertner, R. (1993), Game Shows and Economic Behavior: Risk-Taking on 'Card Sharks', *Quarterly Journal of Economics* 108(2), 507-22.

[27] Gneezy, U. (2005), Deception: The Role of Consequences, *American Economic Review* 90(1), 384-394.

[28] Harrison, G. and J. List (2004), Field Experiments, *Journal of Economic Literature* 27, 1009-1055.

[29] Holt C. (1995), Industrial Organization: a Survey of Laboratory Research, in Kagel, J. and A. Roth (eds.), *Handbook of Experimental Economics*, Princeton UP.

[30] Kalist, D. (2004), Data from the Television Game Show Friend or Foe?, *Journal of Education Statistics* 12(3).

[31] Ledyard, J. (1995), Public Goods: A Survey of Experimental Research, in Kagel, J. and A. Roth (eds.), *Handbook of Experimental Economics*, Princeton UP.

[32] Levine, D. (1998), Modeling Altruism and Spitefulness in Experiments, *Review of Economic Dynamics* 1(3), 593-622.

[33] Levitt, S. (2004), Testing Theories of Discrimination: Evidence from Weakest Link, *Journal of Law and Economics* XLVII, 431-452.

[34] List, J. (2004), Young, Selfish and Male: Field Evidence of Social Preferences, *Economic Journal* 114, 121-49.

[35] List, J. (2006), Friend or Foe? A Natural Experiment of the Prisoner's Dilemma, *NBER Working Paper* W12097.

[36] Miettinen, T. (2006), Promises and Lies: A Theory of Pre-Play Negotiations, mimeo, University College London.

[37] Loewenstein, G., L. Thompson and M. Bazerman (1989), Social utility and decision making in interpersonal contexts, *Journal of Personality and Social Psychology* 57, 426-441

[38] Oberholzer-Gee, F., J. Waldfogel and M. White (2004), Friend or Foe? Coordination, Cooperation, and Learning in High-Stakes Games, mimeo, Harvard Business School.

[39] Orbell, J., R. Dawes and A. van de Kragt (1990), The Limits of Multilateral Promising, *Ethics* 100, 616-627.

[40] Rabin, M. (1993), Incorporating Fairness into Game Theory and Economics, *American Economic Review* 83(5), 1281-1302.

[41] Sally, D. (1995), Conversation and Cooperation in Social Dilemmas: A Meta-Analysis of experiments from 1958 to 1992, Cornell University, 1995.