Equilibrium Play and Best Response to (Stated) Beliefs in Constant Sum Games*

Pedro Rey Biel†

Abstract

We report experimental results on one-shot two person 3x3 constant sum games played by non-economists without previous experience in the laboratory. Although strategically our games are very similar to previous experiments in which game theory predictions fail dramatically, 80% of actions taken in our experiment coincided with the prediction of the unique Nash equilibrium in pure strategies and 73% of actions were best responses to elicited beliefs. We argue how social preferences, presentation effects and belief elicitation procedures may influence how subjects play in simple but non trivial games and explain the differences we observe with respect to previous work.

First draft: May 11, 2004
This Version: December 5, 2006
JEL codes: C72; C91; D81
Keywords: Experiments, Constant Sum Games, Stated Beliefs

*I am indebted to Miguel A. Ballester, Ken Binmore, Tilman Börgers, Dirk Engelmann, Ángel Hernando-Veciana, Steffen Huck and Nagore Iriberri for extensive comments. I am also grateful to Julio González-Díaz, Antonio Guarino, Georg Weizsäcker and to participants in the European Meeting of the Econometric Society (Vienna), Simposio de Análisis Económico (Navarra), ENTER Jamboree (Mannheim), UAB and UCL for comments and to Martin Bøg, Laura Diego, Tom Rutter and Chris Tomlinson for laboratory assistance. Experimental Funding was provided by the ESRC Centre for Evolutionary Learning and Social Evolution (ELSE) and Ministerio de Ciencia y Tecnología (BEC2003-01131). Financial support from Generalitat de Catalunya (2005SGR-00836) and Barcelona Economics-XREA is gratefully acknowledged.

†Pedro Rey-Biel. Universitat Autònoma de Barcelona. Department d ´Economia i d´Historia Econòmica. 08193, Bellaterra. Barcelona (Spain). Tel: (00 34) 935812113. E-mail: Pedro.rey@uab.es.

1 Introduction

A substantial portion of the experimental literature shows that game-theoretical predictions do not work well in the laboratory, even when the games played are very simple.[1] This is particularly true when subjects play games for the first time without previous experience. However, first time behaviour is crucial to model a vast number of economic situations which are not repeated, and it helps to identify strategic principles that may be obscured by convergence in repeated play.[2] A natural question is to identify the class of games for which game theory predicts well when games are played for the first time and the reasons why it might fail in other games.

We aim to contribute to this question by studying play and first order beliefs in simple but non-trivial games with similarities to others for which experimental evidence is more negative. In particular, we study two-player 3x3 constant sum normal form games with unique equilibria in pure strategies and with different number of rounds of iterated deletion of (strictly) dominated strategies necessary to reach the Nash equilibrium. We show that in this class of games, game theory predicts subjects' behaviour better than in previous experiments and we discuss the relation of our results with previous work.

For simple games with unique pure strategy equilibria, experimental evidence is not conclusive. While in 2x2 repeated games equilibrium play has found substantial support (McCabe et al. (1994), Mookherjee and Sopher (1994)), in games with more than two strategies for each subject and no possibility of learning equilibrium predictions start to fail. Stahl and Wilson (1995) found equilibrium compliance rates of 68% in 3x3 games with three rounds of dominance solvability. However, Broseta, Costa-Gomes and Crawford (2001) obtain in 2x3 games with three rounds of deletion of dominated strategies to reach equilibrium or with no dominated strategies equilibrium compliance rates ranging from 11% to 28%. For 4x4, 5x5 and 6x6 repeated games, the evidence is even more negative (Brown and Rosenthal (1990), Rapapport and Boebel (1992), Mookherjee and Sopher (1997)). Thus, choosing 3x3 games with different numbers of rounds of iterated deletion of dominated strategies we may find reasons why game theory loses its predictive power when some characteristics of the games, like the number of actions subjects can make, are changed.

Our results are surprisingly different from Costa-Gomes and Weizsäcker (2004), who found low rates of compliance with equilibrium predictions (35%), low frequency assigned to equilibrium beliefs by opponents and low percentage of best response behaviour in a similar experiment. Our design differs in three key aspects: 1) Our games are constant sum, 2) We elicited beliefs asking about frequencies of play, not probabilities and 3) Payoffs were represented by single-digit numbers. All three changes may be behind our results.

[1] For example, see Stahl and Wilson (1994, 1995), Kagel and Roth (1995), McKelvey and Palfrey (1995), Broseta, Costa-Gomes and Crwaford (2001), Binmore et al. (2002), Crawford (2002) and Goeree and Holt (2004).

[2] Crawford (2002) argues that by foregoing repetition as a teaching device, one-shot experiments place a heavier burden on subjects' understanding, with a premium on simplicity and clarity of design.

First, in constant sum games, game theory makes confident predictions since the Nash Equilibrium outcome coincides with Minimax (and Maximin). Binmore et al. (2001) laments how little experimental research has been done on constant sum games "being the branch of game theory with the most solid theoretical foundations". Previous research on constant sum games[3] has focused on whether subjects' frequencies of play in repeated games coincide with the probabilities with which subjects should play the one-shot mixed equilibria and the results have been negative. Here we offer reassurance on Von Neumann's (1928) Minimax Theorem for one shot games with unique equilibrium in pure strategies, although we cannot separate Minimax or Maximin from Nash reasoning.

We also choose to study constant sum games because, on a theoretical level, behaviour should not be affected by distributional and efficiency concerns. Efficiency concerns should not matter since subjects' payoffs always add up to the same amount, no matter which actions are chosen. Distributional concerns should not affect behaviour as long as subjects care more for their own payoffs than for those of others.[4] On the other hand, this seems counter-intuitive.[5] In constant sum games all strategic behaviour refers to how to distribute a pie of a given size and thus, how fair the distribution is should matter to subjects with distributional concerns. Of these preconceptions, a natural one is that, everything else equal, subjects should get equal shares. Therefore, whether it is feasible to equally split payoffs or not, may have an influence on play. Our design allows us to study these questions.

Second, when subjects are asked about first order beliefs directly, it is crucial to elicit them in a meaningful manner that subjects can understand. Kahneman and Tversky (1973), express doubts on whether subjects can quantify their beliefs and even if they are, they might find some form of processing quantitative beliefs more meaningful than others. We follow Gigerenzer (2000, 2002) in eliciting beliefs by asking about frequencies of play by a pool of subjects instead of asking about probabilities of a single action chosen by a single opponent as it is frequently done.[6] This may make beliefs more meaningful when subjects only choose once in each game.

Eliciting beliefs allows us to study the degrees of complexity with which individuals are able to play games. We explicitly designed our games to discriminate equilibrium behaviour from other models assuming subjects have different degrees of cognitive complexity. Although this is a complex issue, these models approximate subjects' sophistication to whether they are able to best response to their beliefs about opponents' play and whether they form those beliefs anticipating opponents may also be strategical. Thus, they define the first degree of depth of reasoning (L1) as best responding to believing opponents choose their actions randomly and define higher degrees of depth as best responding to believing opponents are one degree less sophisticated than themselves. We specifically design our games to obtain strong separation between these models' predictions and we find that the equilibrium prediction

[3] Rapapport and Boebel (1992), McCabe et al. (1994), Mookherjee and Sopher (1997), Walker and Wooders (2001), O'Neill (1987) and Binmore et al. (2001)

[4] Camerer et al. (1998).

[5] And in particular there is ample evidence that it is not satified by Dictator Game data.

[6] McKelvey and Page (1990), Offerman et al (1996), Costa-Gomes and Weizsäcker (2004).

clearly outperforms these models.

An alternative way to study cognitive complexity is to associate it with the number of rounds of iterated elimination of strictly dominated strategies subjects are able to perform. Our games differ in the number of rounds of iterated elimination of dominated strategies necessary to reach the equilibrium outcome. We find that subjects' equilibrium behaviour across games was not affected by this measure of complexity.

Finally, calculating best responses may be more difficult for subjects when payoffs are represented by several digit numbers or when there are conversion rate between experimental currency and real monetary payoffs from the experiment. Thus, we use one-digit numbers and a one-to-one relationship between experimental currency and payoffs, while maintaining the strategic complexity of the games and we find higher percentages of best response behaviour than in previous similar research.

The paper is organized as follows. Section 2 presents the experimental design and procedures. Section 3 contains the results and the main descriptive statistics. Section 4 explains a follow-up experiment to check the robustness of our results to sequential play. Section 5 concludes. Instructions are available through request to the author.

2 Experimental Design and Procedures

2.1 Experimental Design

Subjects were presented with a series of ten 3x3 Constant Sum Normal Form Games with Unique Equilibrium in Pure Strategies. For each of the ten games, they were asked to perform two tasks: they had to choose an action (between "U", "M" or "D") and they had to report how many of the players on the other subjects' role they thought would play each of the three actions available ("L", "C" and "R").

We constructed a 2x2 design according to two criteria. The first criterion was the order in which subjects had to perform the two tasks. In treatments BABAF and BABAU subjects were asked for each game, first to state their Beliefs (B) and then to chose an Action (A), after which, they moved on to the next game. In treatments ABF and ABU subjects first chose an action in the ten games, without knowing what the second task would consist of, and then, after answers for all actions were collected, they were presented again with the ten same games and asked to state their beliefs about opponents' play. Comparing the BABA and AB treatments allows us to study whether eliciting beliefs before playing the games influences behaviour.

The second criterion was whether an equal split of payoffs was feasible in each of the games. As the games were constant sum, the sum of payoffs both subjects could earn was always the same and equal to £12, no matter the strategies chosen by both players. In treatments BABAF and ABF an equal split of payoffs was feasible in one of the cells of all the games subjects played. In treatments BABAU and ABU payoffs in all games were substituted in this cell by an unequal split, such that one subject would get a payoff of £7

and the other a payoff of £5. For example, in Game 4R below, payoffs when Row subjects chose M and Columns subjects chose L were £6 for both subjects in the F treatments, while they were £5 for Row subjects and £7 for Column subjects in the U treatments. The location of the cell and the changes in payoffs from the F to the U treatments were such that it never affected neither the predictions of the six behavioural models we study nor the degrees of strict dominance solvability, such that it changed whether the Row or Column player earned more than the equal split of payoffs in the U treatments and such that subjects would get higher payoffs in this cell in some games (lower in others) than in the Nash equilibrium outcome. The cell in which the equal split was feasible never coincided with the Nash equilibrium outcome. Comparing the F and U treatments allows us to study whether the feasibility of an exact equal split influenced behaviour, which may be an indication of whether subjects' distributional concerns have an effect in constant sum games.

Game 4R (U Treatment)

		Column		
		L	C	R
Row	U	4,8	2,10	1,11
	M	**5,7**	11,1	4,8
	D	7,5	8,4	10,2

Game 4R (F Treatment)

		Column		
		L	C	R
Row	U	4,8	2,10	1,11
	M	**6,6**	11,1	4,8
	D	7,5	8,4	10,2

2.2 Experimental Procedures

The experiment was carried out with pen and paper in the ELSE laboratory during April 2004. Subjects were recruited by E-mail using the ELSE database, which consists of UCL undergraduate and graduate students. As we are interested in behaviour played without previous experience, we only recruited subjects without previous experience in game experiments and whose field of study indicated that they would not be familiar with Game Theory nor Economics.

We performed four sessions (one per treatment) with twenty subjects. In each session, ten subjects were randomly assigned "Row" roles in all ten games, while the other ten subjects were assigned "Column" roles. However, no subject was aware of their role (nor other subjects' roles) as games were presented to all players from the point of view of row players.

Upon arrival, subjects were randomly assigned seats and were asked to read some preliminary instructions, which described a strategic decision situation and the 3x3 payoff matrix associated with its normal form representation. Then subjects were required to pass an Understanding Test where they had to demonstrate that they knew how to map players' actions in a game to outcomes, and outcomes to players' payoffs. Subjects were told that those who failed the test would act as "assistants" in the experiment. No subject failed the test.

The experiment consisted of ten games which were presented in random and different order to each subject to control for (possible) non-feedback learning.[7] In the BABA treatments

[7] Costa-Gomes and Weizsäcker (04) conclude that there was no learning across games in their very similar setting.

subjects first read the instructions on stating first order beliefs and choosing actions and how they would be rewarded for these two tasks. Then subjects stated beliefs and chose actions for all ten games with no feedback. Subjects stated beliefs by writing down how many of the 10 subjects in the opponents' role they believed would chose each of their three possible actions in each game. In the AB treatments, subjects first read the instructions about how to choose their actions, and then played those games (Part I). After Part I, answer sheets were collected and subjects read the instructions on beliefs. Next, they stated their beliefs for all 10 games (Part II). This procedure guaranteed that in the AB treatments, actions were chosen before, beliefs had been mentioned. Finally, all answer sheets were collected.

For each game subjects were randomly and anonymously paired with a different participant. Subjects never learned who their matched participant in each game was, neither the action which was taken by their matched participant or any other participant in any game.

Subjects were paid according to their answers in both tasks as follows. At the end of each session, a number from 1 to 10 was selected from a bingo urn. This number indicated for which of the 10 games all subjects would be paid for both tasks.[8] Actions were rewarded according to the strategies chose by each pair of matched participants in the particular game selected. Stated beliefs were paid according to a Quadratic Scoring Rule (QSR) which rewarded accuracy prediction.[9] The QSR was designed such that subjects could earn comparatively less money with their belief statements than with their action choices (Maximum of £2 and £11 respectively). Had payoffs for both tasks been similar, risk averse subjects may be induced to take actions that were not best responses to their stated beliefs in the aim to average payoffs.

Subjects were paid the sum of a £5 fixed fee, plus their earnings for choosing actions and stating beliefs. Average payments were £12.78 (around $20 at the time). Each session lasted one hour and subjects were allocated forty minutes to perform both tasks.

2.3 The Games

We classify our games according to whether they are dominance solvable or not. Eight of our games are dominance solvable. Games 1R and 1C are dominance solvable with one round of dominance to reach the equilibrium for one of the players (Row in 1R, Column in 1C) and two rounds of dominance for the other player. Games 2R and 2C are solvable with two rounds for one player (Row in 2R, Column in 2C) and three rounds for the other. Games 3R and 3C

[8] We paid subjects for one random game instead of for an aggregated measure of their answers in all 10 games to be able to maintain the one to one relationship between outcomes and payoffs. Avoiding conversion rates may help clarifying incentives, which may be particularly important in experiments in which beliefs on other subjects' behavior are elicited.

[9] When subjects are asked to predict the frequencies of play of a finite population of subjects, QSRs are not necessarily incentive compatible as subjects' average expectation of play of each action might not necessarily be equal to one of the possible empirical distributions over the finite set of opponents' actions. In any case, expected payoff maximizers can do no better by stating different beliefs than their true beliefs and given our results we think the problem is minor. For a discussion on QSRs see Offerman, Sonnemans and Schram (1996), Offerman and Sonnemans (2001) and Selten (1998). The particular QSR we used, along with an intuitive explanation for subjects highlighting that understanding the maths of the rule was not essential, can be found in the Instructions.

are solvable with three rounds of dominance for one player (Row in 3R, Column in 3C) and two for the other, although the first deletion of strictly dominated strategies is simultaneous for both players. Games 4R and 4C are solvable with four rounds for one player (Row in 4R, Column in 4C) and three rounds for the other. Finally, Games NR and NC are not dominance solvable and have no strictly dominated actions.[10] In the U treatments, Games 1R, 2R, 2C and 3R had additional weakly dominated strategies, apart from the strictly dominated ones.

Game 1R

	L	C	R
U	3,9	4,8	5,7
M	5,7	7,5	7,5
D	9,3	9,3	8,4

Game 1C

	L	C	R
U	10,2	2,10	1,11
M	9,3	8,4	2,10
D	7,5	4,8	3,9

Game 2R

	L	C	R
U	5,7	5,7	4,8
M	2,10	11,1	3,9
D	1,11	10,2	3,9

Game 2C

	L	C	R
U	11,1	4,8	7,5
M	4,8	4,8	1,11
D	7,5	5,7	7,5

Game 3R

	L	C	R
U	5,7	4,8	5,7
M	3,9	1,11	4,8
D	3,9	3,9	11,1

Game 3C

	L	C	R
U	9,3	1,11	8,4
M	10,2	10,2	9,3
D	8,4	11,1	7,5

Game 4R

	L	C	R
U	4,8	2,10	1,11
M	5,7	11,1	4,8
D	7,5	8,4	10,2

Game 4C

	L	C	R
U	7,5	8,4	9,3
M	5,7	11,1	9,3
D	3,9	1,11	10,2

Game NR

	L	C	R
U	8,4	5,7	1,11
M	5,7	5,7	5,7
D	2,10	5,7	7,5

Game NC

	L	C	R
U	1,11	7,5	3,9
M	4,8	4,8	4,8
D	8,4	2,10	3,9

We selected 3x3 games in which the prediction of how subjects would play would not be trivial. Accordingly, we designed the games such that we were able to discriminate Nash Equilibrium choices[11] from the choices predicted by five other models that have proven to be at least partially successful in previous studies on depths of reasoning.[12] These models

[10] Out of the six possible types of 3x3 constant sum games with unique pure strategy equilibria, we covered all but one possible case according to their degree of strict dominance solvability. The remaining case has a dominated strategy for one of the subjects and it is not dominance solvable.

[11] Simply referred as "Equilibrium", from here onwards.

[12] Stahl and Wilson (1994, 1995), McKelvey and Palfrey (1995), Broseta, Costa-Gomes and Crawford (2001), Costa-Gomes and Weizsäcker (2004), Weizsäcker (2003) and Goeree and Holt (2004).

are named L1, L2, L3, D1 and Maximax. L1 predicts that each subjects' action is a best response against the belief that the opponent is playing each action with equal probability. L2 predicts a best response against the belief that the opponent is playing according to L1 and L3 predicts a best response to the believing the opponent plays according to L2. D1 predicts a best response against a uniform belief over the opponents' undominated actions. Maximax (MM) predicts the action that is part of the action profile leading to the player's highest possible payoff in the game.[13]

Table I below shows the number of rounds of dominance solvability for each game and subject role, the prediction of each of the six models we compare and the action profile which was changed by the equal split in the F treatments.

Table I: Games by Rounds of Dominance and Models' Predictions

Game	*Dominance*	*Nash*	*L1*	*L2*	*L3*	*D1*	*MM*	*Equal*	*Mm*	*mM*	*Ef*
1R	(1,2)	D-R	D-L	D-R	D-R	D-R	D-L	M-L	D-C	M-R	D-R
1C	(2,1)	D-R	M-R	D-R	D-R	D-R	U-R	D-L	M-R	D-C	U-R
2R	(2,3)	U-R	M-L	U-L	U-R	U-L	M-L	U-L	U-L	M-R	D-C
2C	(3,2)	D-C	U-C	D-C	D-C	U-C	U-R	D-L	D-L	U-C	U-L
3R	(3,2)	U-C	D-C	U-C	U-C	U-C	D-C	U-L	D-C	U-L	D-R
3C	(2,3)	M-R	M-C	D-R	M-R	M-R	D-C	D-R	M-C	D-L	M-R
4R	(4,3)	D-L	D-R	D-L	D-L	D-L	M-R	M-L	M-L	M-R	U-R
4C	(3,4)	U-L	M-L	U-L	U-L	M-L	M-C	M-L	M-L	M-L	M-C
NR	No	M-C	M-R	D-R	D-L	M-R	U-R	D-R	M-L	M-R	U-L
NC	No	M-R	D-R	M-C	U-C	D-R	D-L	U-C	U-R	D-R	D-L

3 Experimental Results

3.1 Main Descriptive Statistics

Table II below reports, for each of the ten games, high percentages of equilibrium actions taken (80%), of frequencies assigned to opponents' choosing equilibrium actions (58%) and of best responses to stated beliefs (73%). Frequencies were similar across games and the number of rounds of iterated dominance does not seem to affect percentages in a clear cut manner. However, in the two non dominance solvable games (NR and NC) percentages were lower than in the other games. This is particularly true for the percentage of best responses.

Below we study results in more detail.

[13] Stahl and Wilson (1994) use a more sophisticated version of these models. According to their definition, L2 is a best response to a belief distribution which assigns positive weights to a portion of the population choosing actions randomly (L0) and the remaining portion to subjects best responding to uniform beliefs (L1). The reason to define the zero-level of rationality as an equal probability to play each possible strategy, and thus define degrees of rationality from there on, remains open.

Table II: Percentages of Equilibrium Actions, Beliefs and Best Responses

Game	*Equilibrium Actions*	*Equilibrium Beliefs*	*Best Response*
1R	76.25	58.5	80
1C	75	59.375	75
2R	82.5	55.875	83.75
2C	81.25	51.125	71.25
3R	82.5	64.75	77.5
3C	86.25	63.125	77.5
4R	87.5	59.625	82.5
4C	78.75	59	80
NR	72.5	52.875	47.5
NC	73.75	51.5	55
Average	**79.625**	**57.575**	**73**

3.2 Treatment Effects

In this section we study whether different treatments had an effect on subjects' behaviour. In particular we study: 1) the effect of eliciting beliefs immediately before actions were taken and 2) the effect of equal payoff splits being feasible.

We start with the first question and first look at actions chosen. We use Fisher's Exact Probability Test (FEPT) for count data[14] which tests if differences in observed proportions of actions chosen between two treatments might be expected by chance. The null hypothesis (two-tailed) is that there is no difference in the probability of playing each strategy generating the observed proportion of play of each strategy in each treatment. We used the free software R (2003) to perform FEPTs and all other tests in this paper.

We conduct FEPTs separately for each game. We first compare subjects' aggregate actions for each player role (Row or Column) in each of the ten games between the BABA and the AB treatments (without aggregating the F and U treatments). Out of the 40 possible comparisons, we can never reject the null hypothesis that the underlying probability is the same at the 5% significance level. We then perform a stronger test by pooling the data for the F and U treatments. Again, there is no p-value smaller than 5% so we cannot reject the hypothesis that there is no effect of the order of tasks performed in the aggregate actions.

Our next step is to test if the order of tasks affected subjects' belief statements. We collapse each agents' belief statements into one of four categories: for each of the three actions all the stated beliefs that assigned more than half of the frequency to an action were classified in the same category (thus creating three categories), and the last category comprises all the beliefs that do not assign more than half of the frequency to any of the three actions opponents can take. This allows us to create a contingency table and use FEPTs to test for differences in belief statements between BABA and AB treatments.[15]

[14] Developed by Fisher (1935), Irwin (1935) and Yates (1934).

[15] This procedure was previously used by Costa-Gomes and Weizsäcker (2004).

When comparing subjects' aggregate belief statements for each player role in each of the ten games between treatments BABA and AB treatments (without aggregating the F and U treatments) we cannot reject the null hypothesis of no difference in all comparisons. When we perform a stronger test by pooling the F and U treatments we can only reject it once (p-value equal to 0.003 for Row subjects in Game NC), which may be expected by chance. Thus, we conclude the following:

Result 1 *The order in which subjects performed both tasks did not affect behaviour.*

We now study whether the feasibility of equal payoff splits had an effect on behaviour. We again use FEPTs under the null hypothesis that there was no difference across treatments in the probability of playing (or stating) the observed proportions of play (or beliefs stated) of each action.

When comparing aggregate actions between the F and the U treatments for each player role (without aggregating the BABA and the AB treatments), no p-value is smaller than 5%. When we pool the BABA and AB treatments and we compare the F and U treatments across player roles, only one out of the 20 possible p-values is smaller than 5% (p-value equal to 0.006 for Row subjects in Game 4C), which may be expected by chance. We also performed Mann-Whitney tests under the null hypothesis that the median of the distribution of games in which subjects chose the strategy containing the equal split was not different between the F and U treatments at the 5% significance level. Both when we aggregate the BABA and the AB treatments and when we do not, we could never reject the null hypothesis. Thus, we conclude that actions chosen were not affected by whether equal splits were available or not.[16]

Moving on to beliefs, we used the previous classification of beliefs and we performed FEPTs comparing same games under the F and U treatments. We obtain no p-value smaller than 0.05 for the 40 comparisons when we do not aggregate treatments with respect to the order of tasks. When we do aggregate them, only one of the 20 possible p-values is smaller than 0.05 (p-value of 0.0189 for Column subjects in Game NC), which indicates that there is no effect of the feasibility of equal splits. We also performed Mann-Whitney tests comparing the distribution of average frequencies assigned to the strategy which contained the equal payoff splits between the F and U treatments, again for each game and player role. We could never reject the null hypothesis that the median of the distribution of frequencies assigned to the strategy containing the equal split was not different at the 5% significance level, both when aggregating the BABA and AB treatments and when not. Thus, we conclude:

Result 2: *Behaviour was not affected by the feasibility of equal splits.*

Small payoff differences between the equal and unequal split might explain Result 2. It would be worthwhile to study robustness to higher payoff differences. An alternative

[16] Same results were obtained for the null hypothesis that the feasibility of equal splits did not affect the median of the distribution of the number of games in which subjects played the equilibrium action neither of the number of games in which they best responded to their stated beliefs.

explanation is that the equal split was feasible (or not) in *all* the games subjects played. As subjects were only paid for one of the games, our experiment resembles the strategy method, in which a weakening of the "equal split effect" has previously been observed (Güth et al. (2001)). Results in section 4 confirm that equal splits did not affect behaviour even in sequential games.

We use results 1 and 2 to pool the data across treatments and analyze actions and beliefs in the following sections.

3.3 Actions

Table III shows the percentage of compliance with equilibrium predictions for each game by subject role.

Table III: Percentages of Equilibrium Actions

Game	*Row*	*Column*	*All Subjects*	*Nº Rounds Iterated Dominance*
1R	80	72.5	76.25	(1,2)
1C	60	90	75	(2,1)
2R	95	70	82.5	(2,3)
2C	75	87.5	81.25	(3,2)
3R	92.5	72.5	82.5	(3,2)
3C	87.5	87.5	86.25	(2,3)
4R	87.5	87.5	87.5	(4,3)
4C	67.5	90	78.75	(3,4)
NR	92.5	52.5	72.5	No
NC	72.5	75	73.75	No
Average	**81**	**78.5**	**79.625**	

On average, subjects played equilibrium actions in 79.625% of the cases. There is no clear pattern between the number of rounds of iterated deletion of dominated strategies required to reach the equilibrium and the percentage of equilibrium actions played. For example, games 1R and 1C show a lower percentage of equilibrium actions than games 3C or 4R. We also noticed that the lowest percentage of equilibrium play occurred in the non-dominance solvable games (NR and NC). We created contingency tables with the number of subjects who played equilibrium actions in each of the games (aggregating both subject roles)[17] and performed McNemar's tests[18] under the null hypothesis that there was no statistically significant difference in the proportion of compliance with equilibrium between each pair of games. We do not find statistically significant differences between games at the 5% level. When we

[17] This creates seven categories: subjects who reach the equilibrium strategy in 1 round of iterated deletion, 2 rounds, 2 rounds with simultaneous deletion in the first round, 3 rounds, 3 rounds with simultaneous deletion in the first round, 4 rounds and non dominance solvable. Notice that not all these categories have the same number of subjects, but that the Chi-square test allows us to do this comparison.

[18] In the following, we use McNemar's to exploit the statistical power derived from having the same subjects playing across different games. When this is not fulfilled, we use Chi-square tests.

do not group subject roles we perform Chi-Square test under the same null and find some significant differences, for example between Row subjects in game 2R and NC, but no clear pattern emerges. Thus the degree of iterated dominance needed to reach equilibrium is not a straightforward measure of the proportion of equilibrium play. This result may indicate that either subjects were not reasoning in terms of iterated deletion of strictly dominated strategies or that the number of rounds of iterated deletion was not high enough to make a difference. Crawford (2004) argues that in their initial responses to games subjects "seldom play dominated strategies but usually respect at most three of four rounds of iterated dominance". Our results do not contradict this claim.

Overall we conclude:

> Result 3: *Subjects played equilibrium strategies in 80% of the cases. The number of rounds of necessary deletion of strictly dominated strategies to reach the Nash equilibrium was not a clear indicator of the percentage with which the equilibrium strategies were played.*

We now compare how well the equilibrium model predicted actions taken in comparison to other models. Table IV shows the percentage of actions taken that were predicted by the standard equilibrium model, together with the percentage rates predicted by each the other five models described in section 2.3.

Table IV: Percentage of Actions Matched by Models' Predictions

Game	*Nash*	*L1*	*L2*	*L3*	*D1*	*Maximax*
1R	76.25	51.25	76.25	76.25	76.25	51.25
1C	75	62.5	75	75	75	47.5
2R	82.5	17.5	62.5	82.5	62.5	17.5
2C	81.25	56.25	81.25	81.25	56.25	16.25
3R	82.5	38.75	82.5	82.5	82.5	38.75
3C	86.25	48.75	51.25	86.25	86.25	13.75
4R	87.5	50	87.5	87.5	87.5	12.5
4C	78.75	61.25	78.75	78.75	61.25	17.5
NR	72.5	66.25	22.5	62.5	66.25	21.25
NC	73.75	50	46.25	11.25	50	15
Average	**79.625**	**50.25**	**66.37**	**66.75**	**70.37**	**25.125**

Equilibrium outperforms the predictions of the other models in all games.[19] Although the games were intentionally constructed to highlight differences between models' predictions, it is noticeable L1 and L2, which were the most successful models in Costa-Gomes & Weizsäcker (2004), perform clearly worse across all games than Nash.[20] Of the models analyzed, the one

[19] Equilibrium also outperforms each of the other models in all games when subject roles are not pooled.

[20] Notice that L2 predicts the same outcome as Equilibrium in six games, while L1 does not predict the same outcome as Equilibrium in any game. Thus, we should not infer that L2 captures behavior better than L1. L3 coincides with Equilibrium in all but Games NR and NC, where it performs significantly worse.

that comes second in predicting the aggregate of actions is D1, with a percentage of 70.375%. D1 predicts the same action as Nash for five of the ten games. In the five games were the predictions of both models are different, Nash outperforms D1 in all games, with an overall success rate of 77.75% against 49.35%.

We now look at individual behaviour. First, the cumulative distribution function (CDF) of the percentage of subjects who played at least a certain number of games according to each models' predictions shows that while 20% of the subjects played according to the Equilibrium prediction in all ten games, at most only 1.25% of the subjects played in all ten games according to any of the other models here studied. 70% of the subjects chose at least 8 actions according to the Equilibrium model.

Second, Table V classifies subjects according to the model whose predicted action subjects chose in the highest number of games. First, there were 56 out of the 80 subjects that could be clearly classified to a model according to this criterion. i.e., who responded the highest number of times according to only one model. Of these, 69.6% of subjects were classified as "Equilibrium". There are 24 subjects who could not be classified in this manner, as there were ties between various models. Columns "Ties" and "Overall" adds up to more than 100% because we include in each model category all subjects who play according to such model the highest number of times, no mateer the ties. In any case, 87.5% of the subjects who tied between two or more models, chose the highest number of actions according to "Equilibrium" and some other model, while only 50% did it according to "D1" and some other model. In the column "Overall", we add up both the clear cases and the ties to conclude that 75% of the 80 subjects can be classified as "Equilibrium", while only 26.25% of subjects can be classified as D1. Other models show lower percentages. Finally, we show in parenthesis the average number of games in which subjects classified in each model category chose actions according to each model. Notice that this average measures the intensity with which subjects were classified with respect to each model and thus, it shows that subjects classified in each category were quite consistent with the model in which they were classified.[21]

Table V: Classification in models to which subjects respond most times

Model	*Clear Cases*	*Ties*	*Overall*
Nash	69.6 (9.05)	87.5 (7.86)	75 (8.63)
L1	5.36 (8.66)	8.33 (6.5)	6.25 (8.2)
L2	5.36 (8.66)	41.66 (7.2)	16.25 (7.53)
L3	1.78 (9)	25 (7.33)	8.75 (7.33)
D1	16.07 (8.66)	50 (8.08)	26.25 (8.34)
Maximax	1.78 (8)	8.33 (6.5)	3.75 (7)

Thus, we conclude:

[21] Had subjects chosen randomly they would have answered on average in 3.3 games according to each model and, given the structure of the games, the average intensity of subjects classified in each category would have been 5.1.

Result 4: *Equilibrium captures actions played by subjects better than the alternative models, both at the individual and aggregate levels.*

Although we cannot discard that there may be other models that capture behaviour better than those studied here and in particular models allowing for errors in subjects' actions depending on the size of the payoffs, it is clear that Nash is a good predictor of actions taken for the class of games here studied.

3.4 Stated Beliefs

Subjects expected on average that their opponents would play the Nash equilibrium action with the highest frequency in each game (58%). Frequencies assigned to equilibrium play were disperse but again Wilcoxon tests at the 5% significance level confirmed there was not a clear pattern between the number of rounds of iterated dominance and the frequency assigned to equilibrium actions by opponents.

Table VI shows the average frequency of beliefs assigned to the predictions of the six models we compare for each of the games. Although in all but game 1R the highest frequency was assigned to the prediction of the Nash model the predictive value of the six models was more similar for beliefs than for actions. Notice also that the order in which each of the models is successful is practically the same with beliefs stated as it happened with actions (although with beliefs L3 outperforms L2).

Table VI: Average Frequency of Stated Beliefs on Models' Predictions

Game	*Nash*	*L1*	*L2*	*L3*	*D1*	*Maximax*
1R	58.5	61.25	58.5	58.5	58.5	61.25
1C	59.38	51.38	59.38	59.38	59.38	42.38
2R	55.88	33	56.13	55.88	56.13	33
2C	51.13	49.75	51.13	51.13	49.75	33.63
3R	64.75	44.38	64.75	64.75	64.75	44.38
3C	63.13	49.75	36.75	63.13	63.13	23.38
4R	59.63	49.63	59.63	59.63	59.63	30.63
4C	59	54.25	59	59	54.25	31.25
NR	52.88	49	29.63	19.13	49	28
NC	51.5	39.88	40.75	22.13	39.88	26.38
Average	**57.58**	**48.23**	**51.56**	**51.26**	**55.44**	**35.73**

We thus conclude:

Result 5: *While Equilibrium still captures belief statements better than the other models studied here, differences with other models are smaller than with actions.*

Subjects believed equilibrium actions would be played with lower frequency that they were actually played. As observed in previous experiments[22] stated beliefs were conservative, in the sense that the empirical distribution of beliefs was flatter than the distribution of actions played. The percentage of belief statements that assigned frequency one to all ten opponents playing one particular strategy was 11.125%. However, tendency to conservatism does not mean that subjects assigned equal frequency to their opponents playing each of their three available actions. The percentage of uniform belief statements[23] is only 5.875%, much lower, in fact, than the percentage of belief statements that assigned zero frequency to at least one of the opponents' actions (42%). Costa-Gomes and Weizsäcker (2004) claim that the higher percentage of zero-belief statements than uniform beliefs is a reason to discard that conservatism may be caused by risk aversion. They argue that since QSRs punish large mispredictions, risk averse subjects would avoid losses by making roughly uniform belief statements, which subjects did not make in most of the cases. However, notice that even a highly risk averse subject would state zero beliefs to two of his opponents' actions if he was sufficiently certain about the actions that all opponents would take in a particular game.

The average mean square error deviation of stated beliefs was 23.17 out of a feasible range of [0,200]. We assess the accuracy of belief statements in the aggregate by looking at whether subjects predicted the average "structure of frequencies" correctly. We define correct structure of beliefs as subjects assigning highest frequency to the actions which were played with highest frequency and assigning lowest average frequency to the actions which were played with lowest frequency. Table VII compares, for each game and subject role, the average frequency with which each of the three actions was played by subjects, with the average percentage of stated frequency assigned by the opponents to those same actions. It is noticeable that for all but three comparisons, aggregate average beliefs get the "structure of frequencies" played correctly.[24]

[22] Huck and Weizsäcker (2001), Costa-Gomes and Weizsäcker (2004)

[23] Defined as statements that assigned frequency of 3 to two actions and 4 to the other one.

[24] The difference between frequencies assigned in those three games was, however, very small. These games are indicated in Table VII with a star (*). The double star (**) in game 4C indicates that the order of beliefs with which the second and the third actions were played was inverted.

Table VII: Structure of Aggregate Beliefs and Actions

Game	*Row Actions*			*Column Beliefs*		
	U action	M action	D action	U belief	M belief	D belief
1R	0	20	80	3	15.25	81.75
1C	5	35	60	16	34	50
2R	95	5	0	66.5	20.25	13.25
2C	25	0	75	42.25	12.75	45
3R	92.5	2.5	5	66.25	8.25	25.5
3C	0	85	15	6.25	73.25	20.5
4R	0	12.5	87.5	5.5	28.25	66.25
4C	67.5	32.5	0	50.25	40.75	9
NR	2.5	92.5	5	18.25	60.25	21.5
NC	2.5	72.5	25	16.25	53.5	30.25
	Column Actions			*Row Beliefs*		
1R	22.5	5	72.5	40.75	24	35.25*
1C	2.5	7.5	90	11.5	19.75	68.75
2R	30	0	70	45.75	9	45.25*
2C	5	87.5	7.5	11.75	57.25	31
3R	15	72.5	12.5	27.25	63.25	9.5
3C	0	12.5	87.5	20.25	26.25	53
4R	87.5	0	12.5	53	14	33
4C	90	2.5	7.5	67.75	21.75	10.5**
NR	7.5	52.5	40	16.75	45.5	37.75
NC	5	20	75	22.5	28	49.5

However, when looking at each subject individually, the patterns of aggregate behaviour do not translate well into individual behaviour across games. While 75% of the subjects assigned highest frequency in six or more games to the action that was played with highest frequency, only 8.75% of the subjects did, at the same time, assigned the lowest frequency to the action that was played with lowest frequency in those six or more games, and thus, answered with the same structure of frequencies of beliefs as their opponents played. 28.25% of belief statements assigned the same frequency to the two actions that were not believed to be played with highest frequency.

We conclude:

> Result 6: *Subjects were good at predicting the actions that were played with highest frequency by their opponents, although stated beliefs tended to be "conservative".*

3.5 Best Response of Actions to Stated Beliefs

We finally check for consistency by analyzing whether actions chosen were best replies to stated beliefs. We define best replying behaviour as choosing the action that gives the highest expected payoff given the distribution of beliefs stated. According to this definition, best

replying implies that subjects' utilities only depend on own monetary payoffs and that subjects are risk neutral. Results below show that a majority of subjects satisfied this definition.

First, as it would be obvious from previous results, subjects clearly best responded to their stated beliefs more often than they would have had they chosen their actions randomly. Kolmogorov-Smirnoff Goodness of Fit Tests comparing the empirical CDFs to the CDF implied by random behaviour gives p-values of virtually zero. Table VIII shows the percentage of best responses by game and player role. Overall, subjects best responded to their stated beliefs in 73.375% of the cases, higher than the 50% observed in Costa-Gomes and Weizsäcker (2004).

Thus, we conclude:

> Result 7: *Subjects best responded to their stated beliefs a high number of times (73% of the cases).*

By comparing the percentage of best responses across games for all subjects using McNemar's test (5% significance level), we again observe the familiar pattern that the number of rounds of iterated dominance does affect in a clear way the percentage of best replies. However the percentage of best responses was significantly lower in the two non-dominance solvable games (NR and NC) than in some of the other games. Notice that in these two games at least one of the subjects could obtain the same payoff no matter the action chosen by its rival which, as we discuss below may be important. At an individual level, 70% of subjects best responded to their stated beliefs in seven or more games.

Table VIII: Percentage of best Responses to Stated Beliefs

Game	*Row Subjects*	*Column Subjects*	*All Subjects*
1R	80	77.5	78.75
1C	60	90	75
2R	90	88.5	83.75
2C	80	80	80
3R	80	72.5	76.25
3C	77.5	80	78.75
4R	82.5	85	83.75
4C	57.5	87.5	72.5
NR	60	40	50
NC	60	50	55
Average	**72.75**	**74**	**73.375**

Although the proportion of non-best response behaviour is not insignificant, it is small. We look into the nature of non-best response behaviour by calculating how much subjects lost for not best responding to their stated beliefs. We use the monetary losses subjects made when non-best responding to their stated beliefs as a proxy for how important it was for them.

We proceed by calculating, for each subject, the sum of its expected loss when not best responding to their stated beliefs averaged over the ten games each subject played. We find that Row subjects lost on average £0.3037 per game and Column subjects lost on average £0.3205 per game. Given that subjects were only paid for their actions in one game, these were the average losses per subject. Next, we calculate the average maximum feasible loss had subjects have played, in all games, the action that gave them the lowest possible expected payoff, given their stated beliefs. On average, Row subjects could have lost £3.05 per game while Column Subjects could have lost £2.69 per game. Finally, we divide both numbers to calculate for each subject in each game, the percentage of the maximum loss they incurred by not best responding. Averaging over all games for each subject role we obtain that Row subjects lost on average 10.97% of the maximum losses they could have made, while Column subjects lost 15.96% of the maximum possible losses. To put things in perspective, Row subjects would have lost 40.21% of the maximum possible losses they could have made had they chosen the action that neither was a best response nor the worst response to their stated beliefs in all ten games. Column subjects would have lost 55.24% of the maximum possible losses had their chosen this action in all 10 games.[25] Therefore, we conclude:

Result 8: *As subjects best responded in most of the games, they did not lose much with respect to the maximum losses they could have made.*

Not best responding is not the only kind of mistake subjects could have made. Subjects could also err in the accuracy of their predictions of opponents' play. Although the monetary loss derived from this mistake would be minimal, as payments for stated beliefs have an upper bound of £2, a bad prediction of how opponents play, even if it was a best response to stated beliefs, could result in taking a non-optimal action, given the frequencies with which opponents really played. We address whether both types of mistakes (bad predictions and non-best response behaviour) are related, by calculating the correlation between each subjects' average mean square error of his predictions and the average percentage of maximum loss for not best responding each subject makes. We find that there is positive significant correlation between both series (Pearson's coefficient of 0.559 with a p-value of 6.8e-08).[26] This high correlation means that subjects who chose equilibrium actions, also expected a high proportion of their opponents to choose equilibrium actions, and that this prediction was right. This suggests that subjects may have believed that their opponents would choose their actions in a similar way as they did. We thus conclude:

Result 9: *Subjects who are better at predicting the frequencies of play of their opponents are also the ones who lost, on average, less for not best responding.*

[25] An alternative way of calculating the hypothetical loses is to use the real frequencies of play by the opponents instead of the stated beliefs. Given that the percentage of best response to stated beliefs is similar to the percentage of best response to "real" play by the opponents, overall percentages only slightly.

[26] We also calculated the correlation between each subject's number of best responses with the mean square error of predictions and Pearson's coefficient was, as expected, negative and significant (Pearson's coefficient: 0.55, p-value 8.6e-08).

4 Exploring Social Preferences Further: Sequential Games

We here show a replication of our experiment with sequential games which share the same payoff matrix as the previous ones and compare results. As such, we here check whether subjects play according to the unique Subgame Perfect Equilibrium Prediction. Previous experiments have yielded large and systematic deviations from subgame perfect predictions.[27] Experimental procedures were the natural extension of the previous experiment to sequential games, although due to having a more complicated strategy space, we did not elicit beliefs. Treatments differed in whether it was the first or second mover who share the payoffs of the Row or Column player of the previous experiment and again whether equal splits were feasible. Instead, subjects were asked to explain how they took their decisions once they have finished chosen them.[28]

A possible reason for differences in the outcomes of simultaneous and sequential games with the same payoff matrix may be that subjects may put greater weight on other regarding preferences in sequential games. This would seem particularly true for models of other regarding preferences that incorporate intentionality, as the sequentiallity of the games makes clear that a second player's decision is contingent on the first player's choice and therefore, the way a second mover interprets the intentions of the strategy chosen by a first mover can clearly influence the outcome of the play. Anticipating this, a first mover may carefully select his own strategy in order to make, for example, the second mover interpret his intentions in a way that may induce him to reward supposedly kind behaviour by the first mover.

There is at least one type of sequential constant sum games in which there is evidence that other regarding preferences may affect laboratory play: dictator games.[29] In them, a single subject has to allocate a fixed amount between him and another subject, with no strategic decision being taken by the receiver. When dictator games have been played in the laboratory strictly controlling for anonymity (both between subjects and with respect to the experimenter) a significant proportion of subjects does not concord with the equilibrium prediction consisting in allocating the minimum possible amount to the other player.[30] In our experiment, the situation faced by second movers is similar to the allocators' situation in dictator games. In fact, we could define second movers' strategic situation as "mini-dictator" games, since second movers do not have a continuous choice but they can only choose between three actions. There is a difference however between mini-dictator games and our games: when second movers in our games have to choose their action, they are limited by the action taken by first movers and therefore, intentionality and willingness to reward kind behaviour may affect their choices. In dictator games, there is no possible response to the allocator's strategy and thus, non equilibrium outcomes may be explained by distributional preferences by themselves, with no need of reciprocal or intentionally driven other regarding preferences. In Falk and Kosfeld (2005) second movers decide an allocation of a constant

[27] See Crawford (2002), Johnson et al. (2002) and Binmore et al. (2002).

[28] Instructions are available by request to the author.

[29] See Hoffman, McCabe and Smith (1999) and Roth (1995).

[30] See Bolton and Zwick (1995) and Bolton, Katok and Zwick (1998).

quantity between then and a first mover, after observing whether the first mover decides to restrict or not the interval in which the second mover can decide. Thus, second movers face a dictator game situation once first movers have restricted them or not. They observe that when first movers restrict second movers, they allocate less to first movers. Thus, although the subgame perfect equilibrium prediction is not fulfilled, the "intentions"[31] signalled by whether first movers restrict or not makes a difference on second movers. Fey, McKelvey and Palfrey (1996) carried out sequential constant sum centipede games in which, at the first round, payoffs are divided evenly and, as the players pass, the division gets more and more lopsided. They observe that the subgame perfect equilibrium prediction in which the first mover takes in the first round works much better than in centipede games which are not constant sum.

In terms of both subjects having an option to decide strategically, our games also resemble ultimatum games, in which non subgame perfect equilibrium outcomes are frequently observed (Güth et al. (1982)). A key difference with our games is that in ultimatum games, the second mover has the clear option to punish the first mover by rejecting his allocation and leaving both players with no payoffs. In ultimatum games, such a threat would not be credible if second movers are only concerned for own payoff maximization, but it has been observed that not only a significant proportion of second movers exercise such threat, but that this threat is credible to first movers and they rarely allocate the minimum possible amount to second movers. The most frequent explanation for such behavior is that subjects have other regarding preferences that include intentionality. Ultimatum games are not constant sum because of the possibility of rejecting offers and leaving both players with no payoffs. In the games studied in this chapter, this possibility does not exist and in fact, the maximum "punishment" a second mover can inflict on a first mover is by choosing his own payoff maximizing strategy. However, although in constant sum games there is no possibility of punishment, second movers' intentionality driven other regarding preferences could manifest themselves in second movers rewarding kind behaviour by first movers and thus, giving up some units of payoffs in favour of first movers who have taken and action interpreted as kind by second movers.

Table IX presents the main descriptive statistics for each game when grouping all treatments and subject roles. We report, for each of the ten games, the percentage of times the combination of first movers' and second movers' choices reached an Equilibrium outcome, as well as the percentage of first movers' actions taken according to Equilibrium and the percentage of second movers' actions that were best responses to their matched first mover's choice. Results are clear. On average, 91.5% of times, the Subgame Perfect Equilibrium was reached. First movers played Equilibrium 93.5% of the times, and second movers best responded to their matched first mover's choice in 94% of the times. Percentages were high and non statistically different across all games.

[31] Referred as "trust" by the authors.

Table IX: Percentage of Equilibrium Played and Best Responses

Game	*1st Eq. action*	*2nd BR*	*Eq. Played (Sequential)*	*Eq. Played (Simultaneous)*
1R	92.5	97.5	90	57
1C	92.5	92.5	92.5	54
2R	92.5	87.5	85	66.5
2C	97.5	100	97.5	65.63
3R	95	92.5	90	67
3C	90	92.5	90	74.38
4R	92.5	92.5	92.5	76.57
4C	92.5	95	92.5	60.57
NR	92.5	92.5	90	48.56
NC	97.5	97.5	95	54.375
Average	**93.5**	**94**	**91.5**	**62.48**

Table IX also shows the comparison of subgame perfect equilibrium played in the sequential games and the percentage of Nash equilibrium in the simultaneous ones. Chi-square tests for differences in proportions of equilibrium play between games with the same name confirm the null hypothesis that the proportions of play were different between the two experiments in all games with the same name at the 5% significance level. This result is important because if we believe subjects reasoned in game theoretic terms it provides evidence that subjects may be better able to backward induct in our simple sequential games than to calculate Nash equilibria in the simultaneous ones. Although second movers had the obvious advantage of observing first movers' choices, this result may be a first indication that behaviour was not affected by intentional reciprocity.

We conclude:

Result 10: *The Equilibrium prediction works well in constant sum games. When the games are played sequentially, the prediction is even more accurate.*

A second test to study if intentional reciprocity affected behaviour is to check if the feasibility of equal splits affected subjects' choices. Following the same procedures as in the previous experiment, we first use Fisher's Exact Probability Test (FEPT) for count data. We conduct FEPT separately for each game. We first compare subjects' aggregate actions for each player role (first or second movers) in each of the ten games between the Fair and Unfair treatments. Out of the 40 possible comparisons, we can never reject the null hypothesis of the underlying probability of each subject playing each of the three strategies available being equal at the 5% significance level. Table X shows that the total number of actions taken not according with Equilibrium by first movers is very similar between the Fair and Unfair treatments and of these, the number of actions that coincided with the strategy leading to the equal split ("Fair Action") is also very similar between treatments. The same happens with the number of best responses for second movers. We finally performed Mann-Whitney tests under the null hypothesis that the median of the distribution of the number of games in which first movers chose the strategy containing the equal split was not different between

the F and U treatments. We could never reject the null hypothesis at the 5% significance level.[32]

Table X: Percentage of Non Equilibrium and fair Actions

	First movers			Second Movers		
	Non Eq. Actions	Fair Actions	%	Non BR	Fair Actions	%
F treatment	32	22	68.75%	31	7	22.58%
U treatment	29	20	68.96%	29	7	24.14%

Thus, we conclude the following:

Result 11: *Behaviour was not affected by the feasibility of equal splits.*

Small payoff differences between the equal and unequal split might explain Result 11. It would be worthwhile to study robustness to higher payoff differences. An alternative explanation is that the equal split was feasible (or not) in *all* the games subjects played. As subjects were only paid for one of the games, our experiment shares characteristics with experiments carried out under the strategy method, in which a weakening of the "equal split effect" has previously been observed (Güth et al. (2001)). In any case, and admitting these caveats, our results show that there are circumstances in which subjects do not change their behaviour whether equal splits are feasible or not when deciding how to share pies of given sizes, even if one of the subjects moved previous to the other.

We finally classify the comments made by subjects in the post-experiment informal questionnaire to asses the reasons subjects claimed for their behaviour. After reading subjects' comments we created four categories. "Equilibrium" corresponds to subgame perfect equilibrium reasoning. "Maxmin" contains comments referring to "secure" or "highest of the minimum payoffs" strategies. "Fairness" corresponds to any argument in which distributional concerns were mentioned. Finally, "Other" corresponds to explanations that we were not able to classify. Second movers' answers were classified between "Best Responses", "Fairness", when they provided some argument for distributional concerns and "Not Answer" as two subjects did not fill in the voluntary questionnaire. Table XI shows the results.

Table XI: Classification of Questionnaire Answers

	First movers		*Second Movers*
Equilibrium	65%	Best response	87.5%
Maxmin	22.5%		
Fairness	5%	Fairness	7.5%
Other	7.5%	Not Answer	5%

[32] Same results were obtained for the null hypothesis that treatment effects did not affect the median of the distribution of the number of games in which first movers played the equilibrium strategy and also for the distribution of second movers' best responses to first movers' actions.

For first movers, notice that even if both "Equilibrium" and "Minimax" would lead to the same choice and ultimately they both rely in expecting the second mover to choose their payoff maximizing strategy given the first mover's choice and then maximize against it, we distinguish between both kind of explanations. In total, 87.5% of first movers' explanations were classified under one of these reasons. The criterion to separate both reasons was whether subjects' answer included a statement referring to the "maximum of the minima". For example, subject FCC2, a Medicine student in his third year, offered the following explanation:[33]

"I assumed that B participants would choose the column in which they would gain most money, so I chose the row where I would get the most if they chose their maximum strategy given my choice".

This was classified as "Equilibrium". However subject FRR10, a Russian History student in her second year claimed:

"Compared the three rows. Looked for the lowest number in each row. Then chose which one of these was highest, which is the amount I would get paid".

This was classified as "Maxmin". One of the disagreements occurred over the following statement by subject FRR9, a second year Geography student:

"*I know that the B participant will pick the column where they stand to make the most so I have to pick the row where the minimum I can get is higher than other rows".*

This statement seems to contain both reasons, although according to our criteria it was classified as "Maxmin".

In any case, what it is surprising is the small number of statements that made reference to distributional arguments. There were only two statements by first movers to distributional concerns, both of them in "Unfair" treatments, and thus, in cases where the equal split of payoffs was not feasible. These are the following:

"*Try to choose the most equal amount",* and

"Try against 'my better judgment' to be fair in my choice of row, so that a fair amount would also be allocated to B".

With respect to second movers, 87.5% of subjects claimed they chose best responses to the action taken by first movers. Here we show a couple of such answers:

"For each table, there were only three options. I chose the option that would give me most money", and,

"Based on A's selection, I made mine with the highest number reflected in the top R corner".

There were only three second movers who made reference to distributional concerns. Of these, we here reproduce the explanation given by subject FCR9, a Linguistics student in his fourth year, who seemed to hint on intentions driven reciprocity guiding his choices:

[33]Subjects referred to first movers as "A participants" and to second movers as "B participants".

"I tried to make a balance between the amount I could get and the money 'A' person could make. I rewarded as well and paid back 'A' 's decision".

Therefore, we conclude that subjects' claims are in line with the results of the experiment and, in particular the percentage of subjects who claimed to have worried for the distribution of payoffs was low (only 6.25% of the total of subjects).

5 Discussion

When surveying the experimental evidence in dominance solvable games, Camerer (2003, Chapter 5), claims that the joint hypothesis of game theoretic reasoning and preferences that value only one's own payments is easily rejected. He then claims that the interesting question is whether the rejection is due to the pure self-interest part of the joint hypothesis or to the game theoretic reasoning part or even to both. We have here designed a simple experiment in which by using a theoretically useful control for social preferences, we check if subjects play according to game theoretic predictions, and thus, this may indicate whether subjects are able to reason in game theoretic terms. Notice that this procedure does not allow us to answer whether individuals have social preferences but only helps us to identify a class of games in which whether they have social preferences or not, the equilibrium prediction is reasonably accurate. Therefore the game theoretic part of the hypothesis is not rejected in a context in which we would not expect social preferences to influence behaviour.

Our experiment was specifically designed to discriminate between the predictions of the L1, L2 and D1 models and equilibrium. We have shown that even if these models have proven more successful in predicting the outcome of simple games with very similar strategic characteristics, here their prediction is much less accurate. This seems to suggest that other game characteristics such as whether games are constant sum, which numbers are used to represent playoffs or how beliefs are elicited may influence how subjects play and belief other play games. Therefore, although we can not discard that there may exist a model that better explains behaviour in a more extensive class of games, we can not either discard the hypothesis that subjects may reason differently when these game characteristics are changed.

Constant sum games seem like a good starting point to study how subjects reason in simple games as issues like fairness and efficiency concerns seem not to affect their choices. However, this comes with the added cost of not being able to separate minimax or maximin from Nash equilibrium or subgame perfect equilibrium choices. Further research studying the class of games for which equilibrium predictions are reasonably accurate should prove promising.

6 References

Anscombe, F., Aumann, R., (1963). "A Definition of Subjective Probability". *Annals of Mathematical Statistics,* vol. 34, 199-205.

Aumann, R., (1992). "Irrationality in Game Theory" in P. DasgUta and D. Gale (eds), *Economic Analysis of Markets and Games, Essays in Honor of Frank Hahn.* MIT Press.

Beard, R., Beil, R., (1994). "Do People Rely on the Self-interested Maximization of Others? An Experimental Test", *Management Science,* 40, 252-262.

Binmore, K., (1987). "Modelling Rational Players". *Economics and Philosophy,vol.* 3, 179-214.

Binmore, K., (1998). Game Theory and The Social Contract. Volume 1. Playing Fair. The MIT Press.

Binmore, K., McCarthy, J., Ponti, G., Samuelson, L., Shaked, A., (2002). "A Backward Induction experiment". *Journal of Economic Theory,* vol. 104, 48-88.

Binmore, K., Proulx, C., Swierzbinski, J., (2001). "Does Minimax Work? An Experimental Study". *The Economic Journal,* vol. 111, No. 473, 445-464.

Brañas, P., Morales, A., (2003). "Gender and Prisoners' Dilemma". LINEEX working paper 42/03.

Broseta, B, Costa-Gomes, M., Crawford, B., (2001). "Cognition and Behavior in Normal Form Games: an Experimental Study". *Econometrica,* vol. 68, 1193-1235.

Camerer, C., (2003). "Behavioral Contract Theory. Experiments in Strategic Interaction". Princeton University Press.

Camerer, C., Weigelt, K., (1988). "Experimental Tests of a Sequential Equilibrium Reputation Model". *Econometrica,* vol. 56, 1-36.

Camerer, C., Ho, T., Weigelt, K., (1998). "Iterated Dominance and Iterated Best Response in Experimental p-Beauty Contests". *The American Economic Review,* vol 88, No. 4, 947-969.

Cooper, R., DeJong, D., Forysthe, R., Ross, T., (1990). "Selection Criteria in Coordination Games: Some Experimental Results". *The American Economic Review,* vol 80, 218-234.

Cooper, D., Van Huyck, J., (2003). "Evidence on the Equivalence of the Strategic and Extensive Form Representation of Games", *Journal of Economic Theory.*

Costa-Gomes, M., Weizsäcker, G., (2004). "Stated Beliefs and Play in Normal Form Games". Harvard University. Mimeo.

Crawford, V., (2002). "Introduction to Experimental Game Theory". *Journal of Economic Theory,* vol. 104, 1-15.

Fey, M., McKelvey, R., Plafrey, T., (1996). "An Experimental Study of Constant Sum Centipede Games", *International Journal of Game Theory,* 25, 269-287.

Fisher, R., (1935). "The Logic of Inductive Inference". *The Journal of the Royal Statistical Society.* Series A, vol. 98, 39-54.

Gigerenzer, G., (2000). Adaptive Thinking. Rationality in the Real World. Oxford University Press.

Gigerenzer, G., (2002). Reckoning with Risk. Allen Lane, The Penguin Press. Penguin Books LTD.

Goeree, J., Holt, C., (1999). "Stochastic Game Theory: For Playing Games, Not Just Doing Theory", *Proceedings of the National Academy of Sciences,* 96, 10564-10567.

Goeree, J., Holt, C., (2004). "A Model of Noisy Introspection". *Games and Economic Behavior,* forthcoming.

Güth, W., Schmittberger, R., Schwarze, B., (1982). "An Experimental Analysis of Ultimatum Bargaining". *Journal of Economic Behavior and Organization,* vol. 47, 71-85.

Güth, W., Huck, S., Müller, W., (2001). "The Relevance of Equal Splits in Ultimatum Games". *Games and economic Behavior,* vol 37, No. 1, 161-169.

Huck, S., Weizsäcker, G., (2001). "Do Players Correctly Estimate what Others Do? Evidence of Conservatism in Beliefs". *Journal of Economic Behavior and Organization,* vol. 3, 367-388.

Irwin, J., (1935). "Test of Significance for Differences between Percentages Based on Small Numbers". *Metron*, 12, 83-94.

Johnson, E.J., Camerer, C., Weigelt, K., (2002). "Detecting failures of Backward Induction: Monitoring Informational Search in Sequential Bargaining". *Journal of Economic Theory,* vol. 104, 16-47.

Kagel, J., Roth, A., (1995). The Handbook of Experimental Economics. Princeton University Press.

Kahneman, D., and Tversky, A., (1973). "On the psychology of prediction." *Psychological Review*, vol.80, 237-251.

Katok, E., Sefton, M., Yavas, A., (2002). "Implementation by Iterated Best Response in Experimental P-Beauty Contests". *Journal of Economic Theory,* vol. 104, 89-103.

McCabe, K., Mukherji, A., Runkle, D., (1994). "An Experimental Study of Learning and Limited Information in Games". University of Minnesota. Discussion Paper.

McKelvey, R., Page, T., (1990). "Public and Private Information: An Experimental Study of Information Pooling". *Econometrica,* vol 58, 1321-1339.

McKelvey, R., Palfrey, T., (1995). "Quantal Response Equilibrium for Normal Form Games". *Games and Economic Behavior,* vol.10, 6-38.

Mookherjee, D., Sopher, B., (1994). "Learning Behavior in an Experimental Matching Pennies Game". *Games and Economic Behavior,* vol. 7, 62-91.

Mookherjee, D., Sopher, B., (1997). "Learning and Decision Costs in Experimental Constant Sum Games". *Games and Economic Behavior,* vol. 19, 97-132.

Nagel, R., (1995). "Unraveling in Guessing Games: An Experimental Study". *The American Economic Review,* vol. 85, No. 5, 1313-1326.

Nyarko, Y., Schotter, A., (2002). "An Experimental Study of Belief Learning Using Elicited Beliefs". *Econometrica,* vol. 70, No. 3, 971-1006.

Ochs, J., Roth, A., (1989). "An Experimental Study of Sequential Bargaining". *The American Economic Review,* vol 79, 355-384.

Offerman, T., Sonnemans, J., Schram., A., (1996). "Value Orientations, Expectations and Voluntary Contributions in Public Goods". *The Economic Journal, 106 (*July), 817-845.

Offerman, T., Sonnemans, J., (2001). "Is the Quadratic Scoring Rule Behaviorally Incentive Compatible?". University of Amsterdam. Mimeo.

O' Neill, B., (1987). "Non-metric Test of the Minimax Theory of Two-Person Zero-Sum Games". *Proceedings of the National Academy of Sciences,* vol. 84, 2106-2109.

Rapapport, A., Boebel, R., (1992). "Mixed Strategies in Strictly Competitive Games: a Further Test of the Minimax Hypothesis". *Games and Economic Behavior,* vol. 4, 261-283.

R Development Core Team, (2003). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing. Vienna, Austria. ISBN 3-9000510-00-3. URL: www.R-project.org.

Roth, A., (1995). "Bargaining Experiments" in The Handbook of Experimental Economics (Kagel, J. and Roth, A. Eds), 253-348. Princeton University Press.

Ruström, E., Wilcox, N., (2004). "Learning and Belief Elicitation: Observer Effects". University of Houston. Mimeo.

Savage, L., (1954). The Foundations of Statistics. Wiley. New York.

Stahl, D., (1993). "Evolution of Smart Players". *Games and Economic Behavior,* vol. 5, 604-617.

Stahl, D., Wilson, P., (1994). "Experimental Evidence on Players' Models of Other Players". *Journal of Economic Behavior and Organization,* vol. 25, 309-327.

Stahl, D., Wilson, P., (1995). "On Players' Models of Other Players: Theory and Experimental Evidence". *Games and Economic Behavior,* vol. 10, 218-254.

Selten, R., (1998). "Axiomatic Characterization of the Quadratic Scoring Rule". *Experimental Economics*, 1, 43-62.

Schotter, A., Weigelt, K., Wilson, C., (1994). "A Laboratory Investigation of Multiperson Rationality and Presentation Effects", *Games and economic Behavior,* 6, 445-468.

Von Neuman, J., (1928). "Zur theorie der gesellschaftsspiele". *Mathematische Annalen,* vol.100, 2950320.

Walker, M., Wooders, J., (2001). "Minimax Play at Wimbledon". *The American Economic Review,* vol. 91, 1521-1538.

Weizsäcker, G., (2003). "Ignoring the Rationality of Others: Evidence from Experimental Normal Form Games". *Games and Economic Behavior, vol. 44,145-171.*

Yates, F., (1984). "Contingency Tables Involving Small Numbers and the X^2 Test". *The Journal of the Royal Statistical Society.* SUlement 1, 217-235.

Zizzo, D., (2002). "Racing with Uncertainty: A Patent Race Experiment", *International Journal of Industrial Organization,* 20, 877-902.

Game 1R

	L	C	R
U	1,9	2,6	4,3
M	4,4	5,4	5,4
D	7,3	7,5	6,8

Game 1C

	L	C	R
U	10,2	2,10	7,11
M	7,3	6,4	7,10
D	6,6	1,7	9,8

Game 2R

	L	C	R
U	6,6	4,8	4,9
M	4,8	11,3	3,5
D	1,10	10,6	3,8

Game 2C

	L	C	R
U	11,1	1,8	7,5
M	4,8	4,8	1,11
D	6,5	5,7	2,5

Game 3R

	L	C	R
U	6,6	7,8	2,4
M	2,8	1,10	4,6
D	3,9	3,9	11,5

Game 3C

	L	C	R
U	6,1	2,8	4,2
M	7,1	10,4	9,7
D	5,1	11,3	5,5

Game 4R

	L	C	R
U	4,3	2,8	8,11
M	6,6	11,1	5,7
D	10,8	4,3	9,2

Game 4C

	L	C	R
U	5,6	2,3	7,2
M	4,4	10,3	7,2
D	2,7	1,8	8,1

Game NR

	L	C	R
U	8,6	2,6	1,11
M	4,6	7,6	3,6
D	2,7	2,5	4,4

Game NC

	L	C	R
U	3,10	5,5	3,9
M	4,9	2,9	4,9
D	9,5	3,8	2,7