Available online at www.sciencedirect.com

ScienceDirect

Games and Economic Behavior 65 (2009) 572–585

GAMES and Economic Behavior

www.elsevier.com/locate/geb

Equilibrium play and best response to (stated) beliefs in normal form games

Pedro Rey-Biel

Universitat Autònoma de Barcelona, Department d'Economia i d'Historia Econòmica, 08193 Bellaterra, Barcelona, Spain

Received 24 January 2006

Available online 18 March 2008

Dedicated to the memory of Antoni Calvó-Armengol, for his constant and crystal-clear support on this project

Abstract

We report experimental results on a series of ten one-shot two-person 3 × 3 normal form games with unique equilibrium in pure strategies played by non-economists. In contrast to previous experiments in which game theory predictions fail dramatically, a majority of actions taken coincided with the equilibrium prediction (70.2%) and were best-responses to subjects' stated beliefs (67.2%). In constant-sum games, 78% of actions taken were predicted by the equilibrium model, outperforming simple K-level reasoning models. We discuss how non-trivial game characteristics related to risk aversion, efficiency concerns and social preferences may affect the predictive value of different models in simple normal form games.

JEL classification: C72; C91; D81

Keywords: Experiments; Normal form games; Constant sum games; Stated beliefs; K-level reasoning

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 behavior is crucial to model a vast number of economic situations which are not repeated, and

E-mail address: Pedro.rey@uab.es.

[1] See Kagel and Roth (1995), Crawford (2002) and Camerer (2003).

doi:10.1016/j.geb.2008.03.003

it helps to identify strategic principles that may be obscured by convergence in repeated play. 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 mostly negative. In particular, we study two-player 3 × 3 normal form games with unique equilibria in pure strategies. We find high percentages of equilibrium actions (70.2%) and best-responses to (reasonably accurate) stated beliefs (67.2%).

These results contrast with previous experiments in which simple cognitive models explain data better than the equilibrium prediction.[2] Most closely related to our paper, we replicate the previous experiment by Costa-Gomes and Weizsäcker (in press) who found low rates of compliance with equilibrium predictions (35%), low frequency assigned to equilibrium beliefs (32.2%) and low percentages of best response behavior (50.5%) in a very similar experiment with variable sum games. We here discuss some reasons why in our experiment the predictions of Game Theory may be more successful and in particular, we compare our results in constant sum (CS) and variable sum (VS) games.

First, in CS games the Nash Equilibrium outcome (*NE*) coincides with the Maxmin (*Mm*) and Minmax (*mM*) outcome. Thus, subjects do not necessarily need to think strategically in order to obtain the equilibrium outcome but only choose their safest strategy, i.e., the one that guarantees the maximum of the minimum possible payoffs irrespective of what opponents choose. Therefore, comparing otherwise similar CS and VS games (in which the *NE*, *Mm* and *mM* prediction do not coincide) may help us identify whether subjects use some strategic thinking. We find higher compliance of actions with the *NE* prediction in CS games (78%), although *NE* still outperforms *Mm* and *mM* in VS games (61.9% vs. 55.5% and 39.5% respectively). Additionally, stated beliefs were reasonably accurate in both CS and VS games and a majority of choices were best responses to subjects' stated beliefs in both treatments (69.3% in CS, 66.1% in VS). We take these results as indication that subjects may perform some strategic thinking and that risk aversion is not a main driving force of our results.[3]

Second, since our experimental design includes first-order belief elicitation we are able to further study strategic thinking.[4] We explicitly designed our games to obtain strong separation between the *NE* prediction and other previously successful models assuming different degrees of cognitive complexity (K-level reasoning). These models approximate subjects' cognitive sophistication to whether they best respond to their beliefs about opponents' play and whether they form their beliefs anticipating opponents may also be strategic. Thus, the first degree of depth of reasoning (*L1*) is defined as best responding to uniform beliefs. Higher degrees are defined as best responding to increasing degrees of sophistication by opponents. Although the *L1* model best predicts subjects' actions in the similar games by Costa-Gomes et al. (2001) and Costa-Gomes and Weizsäcker (in press), we find that the *NE* prediction outperforms most of these models in our CS games (78% for *NE* vs. 48% for *L1*), while it does not perform particularly worse in the VS ones (61.9% for *NE* vs. 56% for *L1*). Still the *D1* model, assuming best responses to uniform beliefs over the set of strictly undominated strategies, is a good predictor of actions in our games

[2] See Stahl and Wilson (1994, 1995), McKelvey and Palfrey (1995), Costa-Gomes et al. (2001), Weizsäcker (2003), Goeree and Holt (2004), Costa-Gomes and Costa-Gomes and Weizsäcker (in press) and Ivanov (2006).

[3] However, Ivanov (2006) estimates that only 39% of his subjects are risk-neutral in a similar experiment with VS games.

[4] Our belief elicitation method replicates the methods previously used by Costa-Gomes and Weizsäcker (in press).

and in particular for VS games is the best one (67%).[5] We discuss differences between our games and other authors' games to conclude that there is no universal model that best predicts behavior in simple normal form games and we discuss which game characteristics are needed for *NE* to be a good predictor of subjects' choices.

Comparing our CS and VS treatments allows us also to study the effect of efficiency concerns and social preferences on subjects' play and first-order beliefs. Theoretically, in CS games behavior should not be affected by efficiency concerns. Additionally, many forms of social preferences should not be affected as long as subjects care more for their own payoffs than for those of others. On the other hand, it seems counter-intuitive that social preferences do not play a role in CS games.[6] In them, all strategic behavior 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 being equal, subjects should get equal shares. In order to further test if distributional concerns may be behind our results we designed a treatment in which equal splits or payoffs were feasible in all games and another in which they were not. We find that the feasibility of equal splits did not have an impact on actions or stated beliefs.

Finally, our design also allowed for systematic variations in the degree of strict dominance solvability of the games and on whether payoffs were represented by simple (one-digit) or complex (two-digit) numbers. We find that both issues have no particular strong effect on the predictive value of the *NE* model both for actions and first order beliefs.

The paper is organized as follows. Section 2 presents the experimental design and procedures. Section 3 contains the results. Section 4 concludes. Appendix A contains the games.

2. Experimental design and procedures

2.1. *Experimental design*

Subjects were presented with a series of ten 3 × 3 Normal Form Games with Unique Equilibrium in Pure Strategies.

Games are classified according to whether or not they are dominance solvable by iterative deletion of strategies that are (strictly) dominated by a pure strategy. Eight of the ten games offered were 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 and three rounds for the other. Games 3R and 3C are solvable with three rounds of dominance for one player 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

[5] Our games were specifically designed such that the *NE* and *L1* prediction never coincided. Although we aimed for separation between all other models, their predictions coincided in some games so general conclusions about the relative predictive power of other models must be cautious. Virtually all of the examined models, except *L1*, perform better in our set of games than in the previous studies. A possible reason is that our games included additional weak dominance relationships between strategies, which should increase compliance with the models.

[6] And in particular there is ample evidence that social preferences may play a role in dictator games.

one player and three rounds for the other. Finally, Games NDR and NDC are not dominance solvable and have no strictly dominated actions.[7,8]

We selected 3 × 3 games in which the prediction of how subjects would play would not be trivial. As in previous experiments, our games were designed in order to obtain strong separation between Nash Equilibrium choices (*NE*) and the choices predicted by the following previously partially successful models of cognitive reasoning (K-level thinking) and other plausible models. *L1* predicts best-responses against uniform beliefs on opponents' strategies. *L2* predicts best-responses to *L1* beliefs while *L3* predicts best-responses to *L2* beliefs. D1 predicts a best response against uniform beliefs over the opponents' undominated actions. Maximax (*MM*) predicts choosing the action leading to the player's highest possible payoff. Maxmin (*Mm*) and Minmax (*mM*) correspond to choosing the action leading to the maximum of the minimum payoffs (respectively the minimum of the maximum payoffs of the other player). Efficient (*Ef*) implies choosing the action profile including the highest sum of both subject roles' payoffs.

We used a 2 × 2 × 2 design with treatments created according to three criteria. The first criterion was whether the games were constant-sum or not. In the CS treatments all ten games were constant-sum, while in the VS treatments they were not. Equivalent games in both treatments were designed such that degrees of strict dominance solvability and the prediction of the models we compare were not affected. Notice that in the CS treatment *Mm* and *mM* would predict the same choice as *NE* and *Ef* would predict any choice. Thus, we abuse notation and use *Mm*, *mM* and *Ef* to refer to the unique prediction of those models in the VS treatment.

The second criterion was whether an equal split of payoffs was feasible in all ten games. In the F treatments an equal split of payoffs was feasible following a particular combination of subjects' actions, different for each game. In the U treatments the same combination of strategies leads to an unequal split of payoffs. "Equivalent" games in both treatments were designed respecting (strict) dominance solvability and the models' predictions. Additional requirements were that the equal split in the F treatments was never the result of both subjects choosing the *NE* strategy and that it changed whether the Row or Column player earned more than the equal split of payoffs in the U treatments.

A final criterion was whether payoffs were represented by "simple" or "complex"-digit numbers. In the 1D treatments all payoffs were represented by numbers in the interval $[1, 11]$, while in the 2D treatments the same numbers were multiplied by 7, possibly making calculations of best-responses a more difficult task.[9]

Table 1 shows the number of rounds of dominance solvability for each game and subject role, the prediction of each of the models we compare and the action profile which was changed by the equal split in the F treatments (*Eq*).

[7] In the CS treatments, games 1R, 2R, 2C, 3R and NDC have additional weakly dominated actions, while in the VS treatments this occurs only in games 2C and 3R. Weak dominance solvability may have had an effect on the percentage of equilibrium actions chosen and stated beliefs, although in our small sample size we did not find statistical differences. In games 2C and 3R, deletion of weakly dominated strategies would reduce the number of steps which are needed to solve the game, relative to deleting only strictly dominated strategies.

[8] Strategy L of game 2C is strictly dominated by a continuum of mixed strategies but not by any pure strategy.

[9] Huck and Weizsäcker (1999) discuss the effects of similar manipulations in payoffs.

Table 1
Games by rounds of dominance and models' predictions

Game	Dominance	*NE*	*L1*	*L2*	*L3*	*D1*	*MM*	*Eq*	*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
NDR	No	M-C	M-R	D-R	D-L	M-R	U-R	D-R	M-L	M-R	U-L
NDC	No	M-R	D-R	M-C	U-C	D-R	D-L	U-C	U-R	D-R	D-L

2.2. Experimental procedures

The experiment was carried out with pen and paper in the ELSE laboratory during April 2004 and November 2006. Subjects were recruited by E-mail using the ELSE database, which mainly consists of UCL undergraduate and graduate students. We only recruited subjects who had neither taken courses in Game Theory nor Economics and who had no previous experience in game experiments.

We performed eight sessions, one per each combination of treatments, and each 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. No subject was aware of their role as all games were presented from the point of view of row players.

Upon arrival, subjects were randomly assigned seats and were asked to read some preliminary instructions. Then, subjects were required to pass an Understanding Test. No subject failed the test. Finally, the experiment started. Ten games were presented in random and different order to each subject to control for (possible) non-feedback learning. Subjects first read the instructions about how to choose their actions, and then played all ten 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 all actions were chosen before beliefs had been mentioned.[10]

For each game, subjects were randomly and anonymously paired with a different participant. Subjects had no feedback on other subjects' actions.

Subjects were paid for both tasks according to one randomly selected game. Actions were rewarded exactly by the amount of pounds indicated by the number selected combining their action and their matched participant's action in this game.[11] Stated beliefs were paid according

[10] Additional sessions with CS1D games were run eliciting beliefs immediately before playing each game. We did not find statistical differences in actions nor in reported beliefs when grouping the data with respect to other treatments. These tests are not very powerful given the sample size. In any case, eliciting actions before beliefs may be a strong test, not a weaker one, for equilibrium play. While Costa-Gomes and Weizsäcker (in press) do not find statistical differences in the order of tasks, Ivanov (2006) does find some differences.

[11] This number was divided by 7 in the 2D treatments.

to a Quadratic Scoring Rule (QSR) which rewarded the accuracy of predictions.[12] 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) in order to reduce the incentive for risk averse subjects to not take best response actions aiming to average payoffs.[13]

Subjects were paid the sum of a £5 fixed fee, plus their earnings for choosing actions and stating beliefs. Average payments were £12.87. Each session lasted one hour and subjects were allocated forty minutes to perform both tasks.

3. Experimental results

3.1. *Treatment effects*

Table 2 reports the average percentage of Nash Equilibrium actions and best-responses to stated beliefs by subject role in each of the treatments. A first look reveals no important differences among subject roles, no differences between the 1D and 2D treatments neither between the F and U treatments. However, it seems that whether the games were constant-sum or variable sum made a difference, specially in the percentage of equilibrium actions. While overall 78.4% of actions taken in the CS treatments coincided with the equilibrium prediction, only 61.9% did so in the VS treatment. In the CS treatments, 69.4% of the actions were best-responses to stated beliefs, while in the VS treatments 64.9% of actions were best-responses.

To confirm which treatments had an effect on subjects' behavior, we use Fisher's Exact Probability Test for count data (FEPT). This test checks whether differences in observed proportions of actions (beliefs) between two treatments might be expected by chance, under the two-tailed null hypothesis of equal probability between observed proportions.

We start with actions and proceed by steps.[14] We first compare actions taken by the same subject roles in equivalent games varying in just one aspect of the treatments. For example to test

Table 2
Percentages of equilibrium actions and best-responses by treatments

		% of equilibrium actions				% of best-responses			
		1D		2D		1D		2D	
		F	U	F	U	F	U	F	U
Row	CS	78	89	74	74	65	70	64	75
	VS	58	74	65	58	65	64	67	65
Column	CS	84	76	71	81	64	67	83	67
	VS	55	64	66	55	57	61	74	66

[12] With a finite population of subjects, QSRs are not necessarily incentive compatible. In any case, expected payoff maximizers can do no better by stating different beliefs than their true beliefs and given our results, the problem is minor. For a discussion on QSRs see Offerman and Sonnemans (2001).

[13] Instructions together with the quadratic scoring rule used to reward elicited beliefs are available at http://pareto.uab.es/prey/instructionsweb.pdf.

[14] Although FEPT is specifically designed for small samples it is still not a very powerful test with only ten observations in each treatment. For example using this test, we cannot reject that the distribution $(3, 2, 5)$ in one treatment is

whether equal splits made a difference, we compare the aggregate number of actions taken by Row subjects in each pair of "equivalent" games in the CS1DF and the CS1DU treatments. We find no statistical effect of equal splits nor of the number of digits representing payoffs at the 5% significance level. We find, however, significant differences between constant and variable sum games (11 out of 80 comparisons). We use this result to collapse data on the treatment effects not being tested and perform stronger FEPTs. Following our example, we now compare the aggregate observed proportions of play by Row players in pairs of otherwise equivalent games between the CSF and the CSU treatments. We again find no differences emerging from the equal splits nor from the number of digits representing payoffs, but differences between the constant and variable sum games at the 5% significance level (8 out of 40 comparisons). Using this result, we finally collapse all data and conduct FEPTs comparing proportions of actions the CS and VS treatments by game and subject role, and we again find significant differences (5 out of 20 comparisons). We thus conclude that whether games were constant-sum or variable sum had an effect on the proportions of actions chosen, while other treatment effects were not significant.

Moving on to stated beliefs, we classify each subjects' belief statements into one of four categories: for each of the three actions all the stated beliefs that assigned (strictly) more than half of the frequency to an action were classified in the same category (thus creating three groups), and the last category comprises all beliefs which do not assign more than half of the frequency to any of the three actions opponents can take. We proceed similarly as with actions, by progressively collapsing data when no significant treatment effects appeared and using the 5% significance level. We again conclude that whether games were constant-sum or not was the only treatment effect which made a statistical difference in the observed proportions of beliefs stated.

Using these results, we aggregate data from different treatments and we study differences between the CS and VS treatments.

3.2. Actions

NE is a good predictor of subjects' actions in the CS games, while it is not as good in the VS ones. Using McNemar's tests, we find no clear pattern between the number of rounds of iterated deletion of strictly dominated strategies required to reach the equilibrium and the percentage of equilibrium actions played across games in both treatments. For example, in both treatments games 1R and 1C show a statistically significant lower percentage of equilibrium actions than games 3C or 4R at the 5% level. Crawford (2002) 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. In any case, conclusions on this result should be cautious given the power of the test used and the existence of some weakly dominated strategies in some games.[15]

We now compare how well *NE* predicted actions taken in comparison to other models. Table 3 shows the percentage of actions taken that were predicted by the equilibrium model, together with the percentage predicted by each of the other models described in Section 2.1.

significantly different than the distribution $(1, 7, 2)$ in another treatment. The power of the test increases with the number of observations and thus we proceed by steps.

[15] I am grateful to a referee for pointing this out.

Table 3
Percentage of actions matched by models' predictions

Treatment	Subject	*NE*	*L1*	*L2*	*L3*	*D1*	*MM*	*Eq*	*Mm*	*mM*	*Ef*
CS	Row	79	42	68	70	65	20	47	50	37	25
	Column	78	54	69	69	73	30	41	47	46	29
	Average	**78.4**	**48**	**68.5**	**69.5**	**69**	**25**	**44**	**48.5**	**41.5**	**27**
VS	Row	64	56	52	58	66	37	36	49	42	38
	Column	60	56	65	60	68	41	43	62	37	42
	Average	**61.9**	**56**	**58.5**	**59**	**67**	**39**	**39.5**	**55.5**	**39.5**	**40**

Table 3 shows that *NE* clearly outperforms the predictions of the other models for both subject roles in the CS treatment.[16] *NE* performs similarly as the other models in the VS treatments, although *D1* is the best predictor. It is noticeable that *L1* and *L2*, which were two of the most successful models in Costa-Gomes and Weizsäcker (in press), perform worse than *NE*, even in the VS treatment. Notice that *L2* predicts the same outcome as *NE* in six games, while *L1* does not predict the same outcome as *NE* in any game. Thus, we should not directly infer that *L2* captures behavior better than *L1*, when such coincidences do not occur. *L3* coincides with *NE* in all but Games NDR and NDC, where it performs significantly worse in the CS treatment (64.5% vs. 20% for the subject role whose actions differ). *D1* predicts the same action as *NE* for five of the ten games. In the five games where the predictions of both models are different, *NE* outperforms *D1* in all games in the CS treatments (67.2% vs. 29%). However, *D1* is the model which best predicts data in the VS, where it outperforms *NE* (for the games for which the predictions do not coincide, *NE* predicts 41.6% of actions while *L1* predicts 56%). Notice that all other models (*MM*, *Eq*, *Mm*, *mM* and *Ef*) are worse predictors in both treatments. Finally, notice that *Mm* and *mM* are not particularly good predictors of behavior in the VS treatments where they even perform worse than *NE*. Thus, not all the effect of the high percentage of *NE* choices in the CS treatments may be due to the fact that the predictions of *NE, Mm* and *mM* coincide in constant sum games.

We now look at individual behavior. In the CS treatment, 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 21.25% of the subjects played according to the *NE* prediction in all ten games, at most only 3.75% of the subjects played in all ten games according to any of the other models here studied. In the CS treatment 82.5% of subjects chose at least 7 actions according to *NE*. In contrast, in the VS treatment, only 41% of subjects chose *NE* action at least in 7 games.

Second, Table 4 classifies subjects according to the model whose predicted action subjects chose in the highest number of games. First, a total of 102 subjects out of 160 (50 in the CS treatment, 52 in the VS treatment) could clearly be classified according to this criterion, i.e., they responded the highest number of times according to only one model ("Clear Cases"). Of these, 62% of subjects in the CS treatment and only 13% in the VS were classified as *NE*. Of the 58 subjects whose number of predictions tied between two models, most (35% in CS, 17% in VS) tied between *NE* and some other model ("Ties"). Finally, the column "Overall" adds up both clear cases and ties to conclude that 70% of the 80 subjects in the CS treatment can be classified as *NE*, while only 24% of subjects in the VS treatments are *NE*. Notice that 34% of subjects in

[16] When data are not pooled across treatments, only in 5 out of 80 comparisons and only for one subject role the *NE* model is outperformed by other models.

Table 4
% of subjects classified by models' actions most taken and consistency

Model	Clear Cases		Ties		Overall	
	CS	VS	CS	VS	CS	VS
NE	62 (9.3)	13 (8.4)	35 (7.8)	17 (8.1)	70 (8.6)	24 (8.5)
L1	6 (9.3)	15 (8.3)	1 (7)	10 (6.1)	5 (7.6)	19 (7.3)
L2	12 (8.3)	6 (8.5)	14 (7.9)	12 (8.1)	20 (8.3)	14 (8.3)
L3	6 (7.7)	8 (7.3)	17 (7.7)	12 (8)	19 (7.7)	15 (7.7)
D1	10 (8.8)	33 (8.9)	18 (8.2)	15 (7)	22 (8.3)	34 (8.4)
MM	0	6 (6.7)	1 (3)	9 (6)	1 (3)	11 (6.2)
Eq	2 (10)	0	8 (6.8)	0	9 (7.3)	0
Mm	2 (7)	15 (8.3)	0	14 (6.3)	1 (7.2)	23 (7)
mM	0	0	4 (5.7)	3 (6)	4 (5.8)	3 (5.7)
Ef	0	4 (7)	1 (3)	9 (6.3)	1 (6.1)	10 (3)

the VS treatment are classified as *D1*, which is the model that on average best predict behavior in the VS treatment. Table 4 also shows in parenthesis the average number of games in which subjects classified in each model category chose actions according to each model. Results show that the classification was consistent.[17]

Given that the *Mm* and *mM* prediction coincides with the equilibrium prediction in constant-sum games, we can not discard that the high percentage of subjects classified as *NE* behaved non-strategically and chose Maxmin as a safe strategy. In variable sum games subjects' classification shows a high degree of heterogeneity. However, notice that similar percentages of subjects in the VS treatments can be classified as *NE* and *Mm*.

3.3. Stated beliefs

On average, subjects believed equilibrium actions would be played with higher frequency than the other two actions available to opponents. However, they believed equilibrium actions would be played with lower frequency by their opponents than they were actually played. Table 5 compares the average frequency assigned to the predictions of the *NE* model with the other ones. Subjects expected on average that their opponents would play according to the *NE* prediction with higher frequency than according to the other compared models in the CS treatment (57%), but not so much in the VS one (48.5%) where *D1* obtained highest frequency (51.5%). Overall, the predictive success of all models is more similar for beliefs than for actions.

Frequencies assigned to equilibrium play were disperse, ranging from 71.5% by column subjects in game 1R in the CS treatment, to 36.75% by row subjects in the same game and treatment. Wilcoxon tests at the 5% significance level confirmed there was not a clear pattern between the number of rounds of strict iterated dominance and the frequency assigned to equilibrium actions by opponents in both treatments.

As observed in previous experiments[18] stated beliefs were conservative, in the sense that the empirical distribution of beliefs was flatter than the distribution of actions played. The proportion of belief statements that assigned frequency one to all ten opponents playing one particular

[17] 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.

[18] See Huck and Weizsäcker (2001) and Costa-Gomes and Weizsäcker (in press).

Table 5
Average frequency of stated beliefs on models' predictions

Treatment	Subject	*NE*	*L1*	*L2*	*L3*	*D1*	*MM*	*Eq*	*Mm*	*mM*	*Ef*
CS	Row	55	50	53	51	55	39	39	43	36	30
	Column	59	47	51	52	57	33	39	44	44	29
	Average	**57**	**48.5**	**52**	**51.5**	**56**	**36**	**39**	**42**	**40**	**29.5**
VS	Row	49	50	48	49	51	44	36	51	33	40
	Column	48	50	41	44	52	44	36	43	45	40
	Average	**48.5**	**50**	**44.5**	**46.5**	**51.5**	**44**	**36**	**47**	**39**	**40**

strategy was 10.63% in the CS treatments and 8% in the VS ones. Tendency to conservatism does not mean however that subjects believed all actions by their opponents were equally probable. The percentage of uniform belief statements[19] was only of 8.25 in the CS treatments and 6.75 in the VS ones. Much lower, in fact, than the percentage of belief statements that assigned zero frequency to at least one of the opponents' actions (46.5% in CS and 43.75 in VS). This may be a first indication that stating beliefs was a meaningful task for subjects.

We assess the accuracy of belief statements in the aggregate by looking at the aggregate ordering of frequencies. The percentages of games in which, on average, subjects guessed correctly which action was played with highest frequency by their opponents and at the same time which one with lowest frequency was 77.5% in the CS treatments and 66.25% in the VS ones. However, when looking at individual subjects, these patterns do not translate well into individual behavior across games. Only 32% of the total individual belief statements in the CS treatments and 30.63% in the VS ones guessed the observed ordering of frequencies played correctly. The average mean square error deviation of stated beliefs was 31.03 in the CS treatments and 28.16 in the VS ones, out of a feasible range of [0, 66.6]. On average, subjects guessed the correct ordering of frequencies in 4.3 games in the CS treatments (4.4 in VS). Thus we conclude that although stated beliefs were reasonable accurate on the aggregate, they were not so accurate at the individual level.

3.4. Best response of actions to stated beliefs

We finally check for consistency between actions and beliefs by analyzing whether actions chosen were best replies to stated beliefs. We define best replying behavior 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 best replied to their stated beliefs in both treatments.

First, subjects clearly best responded to their stated beliefs more often than they would have had they chosen their actions randomly. Chi-Square Goodness of Fit Tests comparing the empirical cumulative distribution functions (CDF) to the CDF implied by random behavior gives p-values of virtually zero. Overall, subjects best responded to their stated beliefs in 69.25% of the CS games while in 66.13% in the VS ones. Notice that these percentages are higher than the 50% observed in Costa-Gomes and Weizsäcker (in press), even in the VS treatment.[20] Second-best-responses were also more frequent (21.25% in CS, 24.75% in VS) than worst responses

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

[20] Some of the weak dominance relationships previously mentioned may be partially behind this result. Notice also that games NDR and NDC and 3C (in the VS treatments) use the same payoff vector for several outcomes.

(9.5% in CS and 9.12% in VS). We also observe that subjects were more likely to best respond when the average difference in payoffs between best response and other actions was higher (correlation coefficient of 24.77 for all treatments).

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 strict iterated dominance does not affect in a clear way the percentage of best replies. At an individual level, 63.75% of subjects best responded to their stated beliefs in seven or more games in the CS treatments, although only 55% did so in the VS treatment.

Although the proportion of non-best response behavior is not insignificant, it is small. We look into the nature of non-best response behavior by calculating how much subjects (hypothetically) lost for not best responding to their stated beliefs.

We proceed by calculating for each subject, how much they lost on average over all ten games by taking the actions they took instead of the actions that would have been their best response given the empirical distribution of their stated beliefs.[21] We find that subjects in the CS treatment lost on average £0.17 per game (£0.22 in VS). 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, subjects in the CS treatment could have lost £1.2 per game (£1.4 in VS). This means that subjects in the CS treatment lost on average 14.2% of the maximum loss they could have incurred due to not best responding (17.1% in VS). Alternatively, subjects in the CS treatment would have lost on average £0.57 had they made an uniform random choice in all ten games (£0.69 in VS). Therefore, subjects in the CS treatment lost 29% due to not best responding of what they would have lost had they randomly chosen their actions given their stated beliefs (31.9% in VS).

4. Discussion

This paper shows a set of games for which the unique Nash equilibrium may be a good predictor of subjects' choices and in which a high percentage of subjects' actions are best responses to their beliefs of opponents' play. In our games, especially when they are constant sum, the equilibrium prediction also outperforms the predictions of previously successful cognitive models based on K-level thinking, more prominently and given our design, *L1*.

Our study replicates the methods used in previous studies in which the equilibrium model has proved less successful. Thus, differences in our results may be caused by differences in our games. We specifically designed our games such that we could observe the effects of social preferences, cognitive complexity and risk aversion in play and beliefs. Thus, while maintaining many other game characteristics with respect to previous studies, we designed treatments differing in games being constant or variable sum, the feasibility of equal splits and the number of digits that represent payoffs to loom into these issues. However, our games may have also altered other game characteristics which have shown to be important for the predictive value of different models. For example, given the restrictions imposed by our treatments, our games also differed in the degree of separation allowed between the different models we compare. There may be many other sources of differences between our games and others' games which may be behind our results. For example, payoff differences between subjects or between strategies, degrees of weak

[21] Notice that this calculation does not use the empirical observed choices by opponents.

dominance solvability or the physical location in the game form representation of the strategies predicted by different models. However, notice that none of these differences define the reasoning process behind Nash equilibrium or K-level thinking. Therefore, there may be some aspect of subjects' real reasoning process which is not well captured by any of these models.

Our results, together with previous studies, indicate that there may not exist a universal simple model explaining subjects' behavior across different simple normal form games when played for the first time. It may even be possible that for different games a majority of subjects use different strategic processes. For example, differences between our constant sum and variable sum treatments may be an indication of this. We have here shown evidence that under some circumstances the unique Nash equilibrium may still be a good predictor of behavior, and that K-level models may not predict behavior as well in some simple games. Thus, we cannot discard that subjects under some circumstances may be more strategic than previously thought, since the percentage of best responses across our games is also high. Further comparative research should follow.

Acknowledgments

I am indebted to Miguel A. Ballester, Ken Binmore, Tilman Börgers, Antoni Calvó-Armengol, Dirk Engelmann, Steffen Huck, Ignacio Palacios-Huerta, David Pérez-Castrillo and two anonymous referees for extensive comments. I am also grateful to Vincent P. Crawford, Julio González-Díaz, Antonio Guarino, Ángel Hernando-Veciana, Manuel Mosquera, Nagore Iriberri, Joel Sobel, Georg Weizsäcker and to participants in the European Meeting of the Econometric Society (Vienna), ESA World Meeting (Rome), 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 Ministerio de Ciencia y Tecnología (SEJ2006-00538 and Consolider-Ingenio CSD2006-00016) and Generalitat de Catalunya (2005SGR-00836 and Barcelona Economics-Xarxa CREA) is gratefully acknowledged.

Appendix A. The games

A.1. Constant-sum games

The following ten games correspond to the CS1DU treatments. Games for the two-digit (2D) treatments are obtained by multiplying payoffs by 7. Games for the equal (F) treatments are obtained by substituting the payoffs of the "*Eq*" combination of strategies of Table 1 by 6 (42 in the 2D treatments).

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, 1	10, 2

Game NDR

	L	C	R
U	8, 4	5, 7	1, 11
M	5, 7	5, 7	5, 7
D	2, 10	5, 7	7, 52

Game NDC

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

A.2. Variable sum games

The following ten games correspond to the VS1DU treatments. Games for the two-digit (2D) treatments are obtained by multiplying payoffs by 7. Games for the equal (F) treatments are obtained by substituting the payoffs of the "*Eq*" combination of strategies of Table 1 by equal one-digit numbers (two-digits in the 2D treatments) such that the predictions of all the models considered in Table 1 are satisfied.

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 NDR

	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 NDC

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

References

Camerer, C., 2003. Behavioral Game Theory. Experiments in Strategic Interaction. Princeton Univ. Press.

Costa-Gomes, M., Weizsäcker, G., in press. Stated beliefs and play in normal form games. Rev. Econ. Stud.

Costa-Gomes, M., Crawford, B., Broseta, B., 2001. Cognition and behavior in normal form games: An experimental study. Econometrica 68, 1193–1235.

Crawford, V., 2002. Introduction to experimental game theory. J. Econ. Theory 104, 1–15.

Goeree, J., Holt, C., 2004. A model of noisy introspection. Games Econ. Behav. 46, 365–382.

Huck, S., Weizsäcker, G., 1999. Risk, complexity, and deviations from expected-value maximization. J. Econ. Psychol. 20, 699–715.

Huck, S., Weizsäcker, G., 2001. Do players correctly estimate what others do? Evidence of conservatism in beliefs. J. Econ. Behav. Organ. 3, 367–388.

Ivanov, A., 2006. Strategic play and risk aversion in one-shot normal-form games: An experimental study. Mimeo, The Ohio State University.

Kagel, J., Roth, A., 1995. The Handbook of Experimental Economics. Princeton Univ. Press.

McKelvey, R., Palfrey, T., 1995. Quantal response equilibrium for normal form games. Games Econ. Behav. 10, 6–38.

Offerman, T., Sonnemans, J., 2001. Is the quadratic scoring rule behaviorally incentive compatible? Mimeo, University of Amsterdam.

Stahl, D., Wilson, P., 1994. Experimental evidence on players' models of other players. J. Econ. Behav. Organ. 25, 309–327.

Stahl, D., Wilson, P., 1995. On players' models of other players: Theory and experimental evidence. Games Econ. Behav. 10, 218–254.

Weizsäcker, G., 2003. Ignoring the rationality of others: Evidence from experimental normal form games. Games Econ. Behav. 44, 145–171.