Economics Department Discussion Papers Series

ISSN 1473 – 3307

An Experiment on Forward versus Backward Induction: How Fairness and Levels of Reasoning Matter

Dieter Balkenborg and Rosemarie Nagel

Paper number 08/04

URL: http://business-school.exeter.ac.uk/economics/papers/

An Experiment on Forward versus Backward Induction: How Fairness and Levels of Reasoning Matter.*

Dieter Balkenborg† and Rosemarie Nagel‡
The University of Exeter and Universitat Pompeiu Fabra

09.05.2008

Abstract

We report the experimental results on a game with an outside option where forward induction contradicts with backward induction based on a focal, risk dominant equilibrium. The latter procedure yields the equilibrium selected by Harsanyi and Selten's (1988) theory, which is hence here in contradiction with strategic stability (Kohlberg-Mertens (1985)). We find the Harsanyi-Selten solution to be in much better agreement with our data.

Since fairness and bounded rationality seem to matter we discuss whether recent behavioral theories, in particular fairness theories and learning, might explain our findings. The fairness theories by Fehr and Schmidt (1999), Bolton and Ockenfels (2000) or Charness and Rabin (2002), when calibrated using experimental data on dictator- and ultimatum games, indeed predict that forward induction should play no role for our experiment and that the outside option should be chosen by all sufficiently selfish players. However, there is a multiplicity of "fairness equilibria", some of which seem to be rejected because they require too many levels of reasoning.

*We thank Prof. Reinhard Selten for his support in the design of the experiment and for many helpful discussions. The design of the basic game is due to him. We are grateful for the help received by those working at the Bonn Laboratory for Experimental Economics when preparing and conducting the experiment. We had several opportunities to present this work in seminars which resulted in many helpful comments and suggestions by the participants.

†(Corresponding author) Department of Economics, The University of Exeter, Streatham Court, Exeter EX2 4PU. UK, e-mail: D. G. Balkenborg at ex. ac. uk

‡Department of Economics, Universitat Pompeiu Fabra, Balmes 132, Barcelona 08008, Spain; e-mail: nagel at upf. es

We show that learning theories based on naive priors could alternatively explain our results, but not that of closely related experiments.

Abstract

JEL classification: C92, C72, Keywords: experiments, equilibrium selection, forward induction, fairness, levels of reasoning.
JEL classification: C92, C72, Keywords: experiments, equilibrium selection, forward induction, fairness, levels of reasoning.

1 Introduction

The presence of a multiplicity of Nash equilibria in many games has been a classic problem for game theory. Many theories to refine among the equilibria have been proposed in order to reduce the set solution candidates. A well known example is the theory of strategic stability by Kohlberg and Mertens (1986). An alternative approach is developed by Harsanyi and Selten (1988) who provide an equilibrium selection theory that aims to select a *unique* Nash equilibrium for every game. In this paper we discuss the results of an experiment that was conducted over 15 years ago in order to test the behavioral validity of these two theories in a game where they contradict. The experiment strongly refutes one of the theories if it is applied to the one-shot game used in our experiment and if payoffs and monetary incentives are identified. However, we always believed that elements of bounded rationality were crucial for our results.

In the recent decade many different behavioral approaches to game theory and theories of bounded rationality have been proposed. It is interesting to review our experimental results and related ones in the light of these new theories. This, besides of the overdue reporting of our experimental results, is the purpose of the current paper. We feel the task to be rewarding because the analysis leads us naturally to discuss a variety of behavioral concepts, in particular fairness theories, levels of reasoning about the rationality of players, and learning.

Our experiment is based on a version of a "Dalek"-game where the first player can choose between an outside option and the possibility to play a 2×2-game.[1] This 2×2-game has a natural focal point equilibrium, an equal division which is also risk dominant. If one accepts this focal point as *the* solution to the 2×2-subgame then backward induction implies that player 1 should take the outside option. To solve the game in this way can be justified by a strong backward induction principle, which requires that *every solution*

[1] The term "Dalek-game" was coined by Binmore (1987), Binmore (1988). It refers to a certain visual similarity between a graphic representation of the extensive form of the game and the tanks of the extraterrestials called "Daleks" in the BBC science fiction series "Dr Who".

(although not necessarily every Nash equilibrium) to a subgame should be extendable to a solution to the whole game. Due to a related reasoning the theory by Harsanyi and Selten (1988) also selects an equilibrium with this outcome.

In contrast, Kohlberg and Mertens (1986) discuss the strong backward induction principle (see Property (BI2) in subsection 2.6 of their paper) but reject it in favor of an alternative approach known as forward induction. The formalization of the forward induction principle in van Damme (1989) motivated our experiment. He argues that the presence of an outside option can make one of the equilibria in the 2×2-subgame focal and hence determine how the subgame is to be played. Applied to our game his principle implies that the outside option is not taken and that the play should result in an unequal division which is favorable to player 1. This equilibrium is in fact selected by many solution concepts based on the normal form representation of the game. Apart from strategic stability, the iterated dominance of weekly dominated strategies or the concept of strict equilibrium sets from evolutionary game theory (Balkenborg and Schlag (2006)) also select this solution.

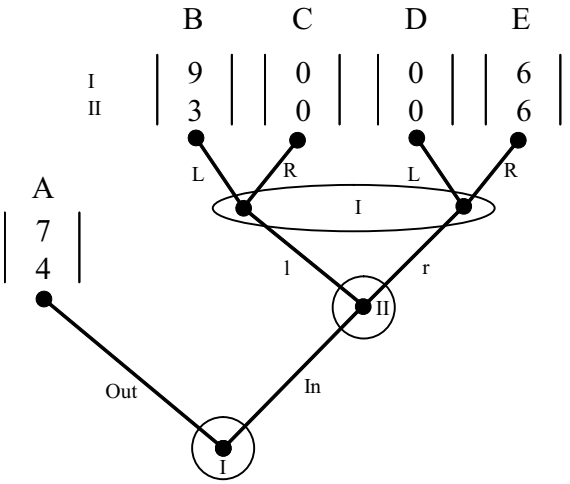

Figure 1: The Dalek Game

This unequal division is, however, virtually not observed in our data and so the forward induction outcome is refuted in favor of backward induction.

Substantial empirical evidence in the experimental literature often reject predictions based on backward induction since it requires a too complicated reasoning or is in conflict with fairness consideration. Moreover, standard game theory does not explain why we obtain strong evidence against forward induction while other authors (e.g. Cooper, DeJong, Forsythe, and Ross (1993) and Brandts and Holt (1992)) get favorable evidence in

games which are game-theoretically virtually the same. Therefore we analyze our game here in the light of alternative, descriptive theories that have been developed to explain deviations from fully rational and selfish behavior. It is hard to deny (and not unintended by our design) that fairness considerations play a role in our experiment. We start hence by applying the fairness theories by Bolton and Ockenfels (2000) and Fehr and Schmidt (1999) to our game. The important feature of these theories is that players can have varying degrees of fairness attitudes and this is reflected in their utility functions. Because the fairness attitudes of players are unknown, one has to study a game of incomplete information different from the original game. Under assumptions consistent with the experimental evidence from ultimatum games (see e.g. Kagel and Roth (1995), Camerer (2003)) the "unfair" forward induction outcome is ruled out as a Bayesian equilibrium outcome of the incomplete information fairness game. Nonetheless, the multiplicity problem of Nash equilibria is not resolved, which contrasts with most applications of fairness theories discussed in the literature.

We find two types of perfect Bayesian equilibria for the incomplete information game. One is a partially separating Bayesian equilibrium of the fairness model that is selected by most refinement concepts and also by Harsanyi and Selten's theory. It has the characteristic that all sufficiently fair minded types of player 1 give up the outside option in order to reach the fair outcome in the 2×2-subgame. The other is a pooling equilibrium which is less fair because in it *all* types of player 1 choose the outside option. The partially separating equilibrium requires four steps of reasoning. Experimental evidence (see e.g. Costa-Gomes and Crawford (2006); Crawford, Gneezy, and Rottenstreich (2008); Nagel (1995); Stahl and Wilson (1995); Camerer (2003), Chapter 5) suggests that already three steps demand too much of most subjects. Not surprisingly, the pooling equilibrium fits better with our data. It should be a novelty in the experimental literature that a fairer outcome does not arise because it requires too many steps of reasoning. In previous experiments fairness equilibria tended to be cognitively simple (see for instance Johnson, Camerer, Sen, and Rymon (2002) where subjects apparently replace complex backward induction reasoning by simple fairness consideration). Camerer and Fehr (2006) point out that theories based on the "economic man" may fail because economics agents may not be rational or because they may not be selfish. In our experiment standard rationality (in the sense of backward induction) seems to be restored because fairness and bounded rationality are themselves at conflict.

At this point one may argue that perhaps uncertainty about the behavior alone explains our results. If player 1 believes that player 2 takes his two choices with equal probability then it is rational for him to take the outside option. While this argument seems very convincing for our data, it does not explain why forward induction is so suc-

cessful in the data of Cooper, DeJong, Forsythe, and Ross (1993) where the 2×2-game is a symmetrical battle-of-the-sexes game with no focal point. We hence believe that fairness matters for our results.

Finally, Binmore and Samuelson (1999) argue that both equilibrium components for the Dalek game are asymptotically stable for learning processes when an inward pointing drift is added. Their considerations are important to explain the robustness of the outside option outcome in the long run. However, the theory does not discriminate between the two equilibrium components of the original game, it simply explains why the forward induction outcome does not necessarily arise from learning.[2] One may still concede that fairness consideration matter for the selection of the outside option equilibrium component. Ironically, because their theory is agnostic on the question of whether players' preferences are shaped by fairness considerations, one can combine their theory with the recent fairness theories and obtain an argument why the "superfair" equilibrium of the fairness models may not be learned.[3]

There is a large literature on experimental testing of the forward induction reasoning. The common feature of most of these experiments is that there is a two-stage game. In the first stage a player can typically choose an outside option, burn money or pay an entry fee. In the second stage there is typically a symmetric conflict, which is a pure coordination game or a battle of the sexes. Our game, in contrast, has non-symmetric components and includes fairness outcomes in the game following the outside option. Ochs (1995) gives a survey of the forward induction experimental literature. Many of the experiments he mentions support the forward induction argument, see e.g. Cooper, DeJong, Forsythe, and Ross (1992), Cooper, DeJong, Forsythe, and Ross (1993) or van Huyck, Battalio, and Beil (1993). In Cachon and Camerer (1999) outcomes are observed which are similar to those in games where forward induction applies even if forward induction does not apply in the game actually played.

However, there is also contrary evidence, similar to our experiment. Cooper, DeJong, Forsythe, and Ross (1993) obtain the forward induction solution when it coincides with a dominance argument but the same outcome is predicted when forward induction makes no prediction. Brandts and Holt (1995) show that the forward induction is only a good prediction, if it coincides with a simple dominance argument, but not without the dominance story. In Cooper, DeJong, Forsythe, and Ross (1993) and also in Huck and Müller

[2] Binmore and Samuelson (1999) do suggest that the size of the basin of attraction matters for which equilibrium is selected. Unluckily the fair equilibrium in our experiment is also the one with the bigger basin of attraction and so we cannot say what drives our result.

[3] Bolton and Ockenfels (2000) are Fehr and Schmidt (1999) are correspondingly agnostic about how subjects reach equilibrium in their rather complex incomplete information games for which the calibration of the priors is much in dispute.

(2005) the forward induction solution predicts well in the experiment based on the extensive form but fairs poorly when subjects are presented with the normal form game.[4] A similar problem seems to arise in Caminati, Innocenti, and Ricciuti (2006) who use games similar to ours but who work essentially with the normal form. Brandts, Cabrales, and Charness (2003) find evidence against forward induction in an industrial organization game.

Our description of the experiment and its result in Sections 2 – 4 will be brief. More details can be found in the discussion paper Balkenborg (1994). The extensive game we use in our experiment and the conflict between forward and backward induction is further explained in Section 2. Section 3 describes our experimental design and Section 4 the results. In Section 5 we discuss the implications of behavioral theories for our experiment. Section 6 concludes.

2 The Basic Game

Our experiment is based on the game in extensive form in Figure 2. It starts with a random move selecting with equal probabilities between two subgames which we refer to as the *left* and the *right subgame*.

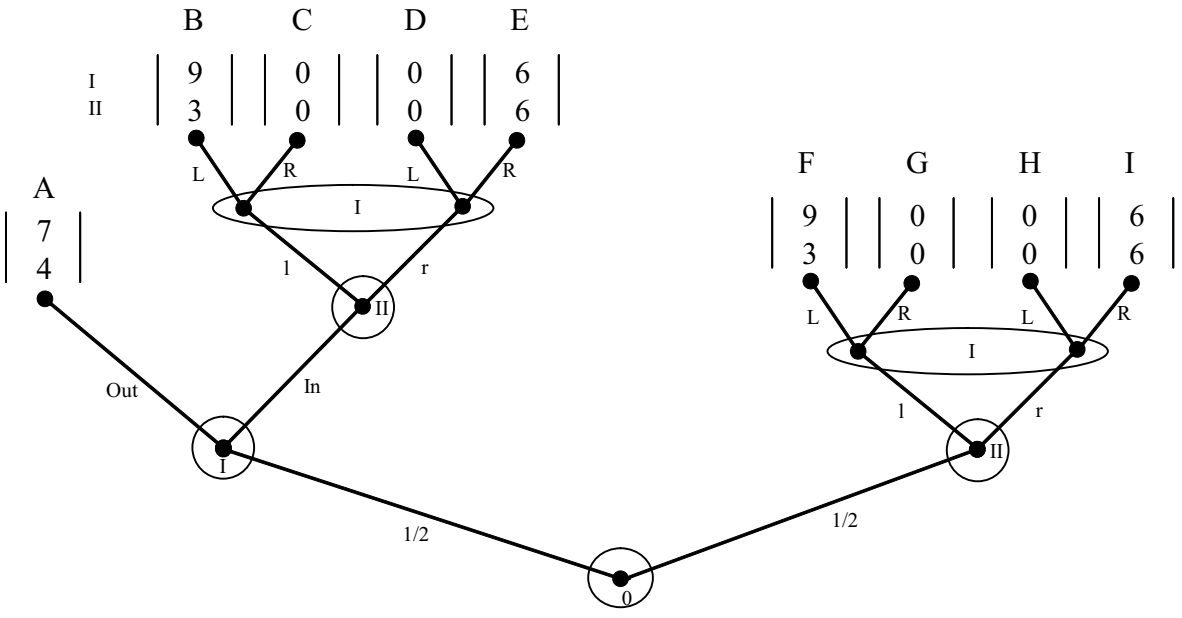

Figure 2: The extensive game used in the experiment.

[4]See also Huck and Müller (2005) for a more detailed summary on recent literature.

The *right subgame* will also be called the *right bargaining game*[5] in the following because it can be interpreted as a simultaneous-move bargaining game. Players 1 and 2 must decide simultaneously and independently between two possible agreements how to share 12 points. One of the two agreements (corresponding to outcome **I,** with both choosing right) yields an *equal division* $(6, 6)$ to both players. The other possible agreement (corresponding to outcome **F,** with both choosing left) is an *unequal division* yielding $(9, 3)$, 9 points to player 1 and 3 to player 2. If both choose differently the game results in a *conflict* and both players receive the conflict payoff 0 (outcomes **G** or **H**). Outcome **G** is also called *anticonflict* because it results if both players make the proposal that is more favorable to the opponent.[6]

The *left subgame* starts with a choice for player 1 to either end the game, and thus *takes his outside option* (also called OUT) with payoffs $(7, 4)$ or – by taking his right choice (called IN) – to play a subgame that is identical (up to the embedding) to the right bargaining game. The latter subgame will be referred to as the *left bargaining game* or *the bargaining game preceded by the outside option*. Our bargaining game with outside option is a variant of an example by van Damme (1989). van Damme discusses a symmetric battle of the sexes game with two symmetrically unequal divisions $(3, 1)$ and $(1, 3)$ where an outside option $(2, 0)$ is added for player 1. Our bargaining game instead has an unequal division $(9, 3)$, an equal division $(6, 6)$, which is a focal point in the bargaining game, and an outside option yielding $(7, 4)$. The payoffs in our game were chosen in such a way that the two solution concepts discussed in the next subsection lead to two different outcomes in the left subgame.

To compare how the presence of the outside option affects the play of the bargaining game we introduced a random move which decides whether the outside option is available to player 1 or not.

2.1 Normative solutions: Forward versus Backward induction

In this subsection we compare two normative solutions for the extensive game, backward induction based on the focal point and forward induction. We assume here that the payoffs at the terminal nodes are the true utilities of the players and that this is common knowledge. Behavioral models modifying these assumptions are considered in Section 5.

The solutions we discuss are subgame perfect Nash equilibria in pure strategies.[7] To

[5] This terminology was not used to explain the game to subjects in the experiment. We use it here to gain some flexibility when discussing the experiment and its results.

[6] The first author heard the term "anticonflict" first in a talk given by Harsanyi. He compared anticonflict to the situation of two gentleman who want to pass a narrow entrance and each proposes to the other "Please go first!".

[7] To ease the discussion, mixed strategy equilibria are ignored in the following.

find all such equilibria, we must first select solutions to the left and the right bargaining game. Each bargaining game has two pure strategy Nash equilibria, corresponding to the two agreements. The solution selected for the left bargaining game determines via backward induction whether a rational player 1 will choose his outside option or not. Thus we see that there are four subgame perfect Nash equilibria in the extensive game overall.

Of these four only two are *subgame consistent*, namely those where the same agreement is chosen for both bargaining games. Subgame consistency is based on the idea that a solution concept must always induce identical solutions in identical subgames, regardless of how they are imbedded, (see Harsanyi and Selten (1988) for a formalization). One could argue that a solution concept based on backward induction should be both subgame perfect and subgame consistent, independently of how the subgames are embedded (compare the discussion in Kohlberg and Mertens (1986), section 2.3). Backward induction requires to solve the subgames *in isolation*. Hence it seems natural so solve identical subgames in the same way and then iteratively extend the solution found to the whole game.

The equal division $(6, 6)$ seems to be a natural solution, a focal point for the bargaining game in isolation (see also Subsection 5.2.2). A *normative theory* selecting this equilibrium outcome is theory of risk dominance Harsanyi and Selten (1988). Since there is no conflict between risk dominance and payoff dominance in this game, Harsanyi and Selten's general theory of equilibrium selection would choose this equilibrium as well.

If we accept the equal division as the only solution in both bargaining games, then *backward induction based on the focal point* prescribes unique strategies for both players in the extensive game of Figure 2. Both players must take their right choices in the two bargaining subgames and player 1 must take his outside option in the left subgame, given that 7 points for OUT is more than the 6 points player expects to receive in the bargaining subgame. These strategies induce outcome **A** if nature selects the right subgame and otherwise in outcome **I.** Because this solution is subgame consistent it can be shown to be a solution according to Harsanyi and Selten (1988).[8]

Contradicting subgame consistency van Damme (1989) (and also Kohlberg and Mertens (1986)) argue that the embedding of a subgame may be important for how it has to be played. The presence of the outside option may affect how the bargaining game is played and can make one of its two pure strategy equilibria focal.

The forward induction argument applies to the left subgame of our extensive game as follows. It is never optimal for player 1 to move into the bargaining subgame and to propose the equal division. He would thereby get either 6 or 0 while the outside option

[8] Finding this solution requires only two essential steps of reasoning. Players have to accept the equal division as the solution to the bargaining game and player 1 must decide on his outside option as the consequence.

yields him 7. Therefore player 2, when he observes that the bargaining subgame is reached and he expects his opponent to act rationally, should conclude that player 1 intends to propose the unequal division. He should hence agree to the unequal division, since 3 is better than 0. A rational player 1 should be able to follow this line of reasoning and anticipate that his opponent would propose the unequal division. Consequently, he should play the bargaining subgame and propose the unequal division, getting 9.

Hence forward induction implies outcome **B** in the left subgame of Figure 2. It does not contradict the choice of the focal point outcome **I** or the selection of any other equilibrium in the right subgame.[9]

The forward induction outcome **B** arises most naturally from many normal form concepts. It is the unique strict (hence evolutionary stable) equilibrium of the reduced normal form of the left subgame, and also results from strategic stability and the iterated elimination of weakly dominated strategies.[10]

3 The Design of the Experiment

The experiment was conducted at the Bonn Laboratory of Experimental Economics in the years 1989 and 1990. In total 154 students, mostly of economics and law, participated in 13 independent sessions which lasted, including instructions, between three and four hours. With the exception of a one person, who participated in sessions 1 and 9, no participant took part in more than one session. The extensive game was played between subjects via a computer network. The subjects made their choices by selecting a move in the extensive game as it was presented graphically to them on the computer screen. The extensive game was presented in a style similar to the graph in Figure 2. A move was highlighted by simultaneously highlighting all edges in the graph belonging to it. Computer programs were developed for most of the instructions except for a last summary part, which was always done verbally in a classroom. While the instructions made subjects familiar with the extensive form, they did not see the actual payoffs of the game before the experiment started. Similarly, we refrained from any interpretation of the game as a "bargaining

[9] The contradiction between forward induction and subgame consistency is inherent. It is not due to the selection of the focal point equilibrium in the bargaining game but to the fact that *some selection* among the Nash equilibria was made for the bargaining game. Suppose, we would, for whatever reason, consider $(9, 3)$ as the only solution to the bargaining game. If we then replace the outside option for player 1 by an outside option for player 2 with payoffs $(2, 5)$ forward induction would select outcome $(6, 6)$ and not $(9, 3)$.

[10] It is strictly dominated for player 1 to choose to play the bargaining game and to propose the equal division in this game. If this strategy is eliminated, it becomes weakly dominated for player 2 to propose the equal division. If this strategy is also excluded taking the outside option becomes weakly dominated for player 1. Thus three levels of reasoning lead to the forward induction outcome **B**.

game with outside option". We wanted the subjects to approach our extensive game like a parlour game which often has highly abstract rules and can only be understood by gaining experience through play. Thus our data reveal genuine learning and how subjects familiarize themselves with the a new strategic conflict.[11]

Each session had twelve participants, except for session 10 which had nine, and sessions 11 and 13 which had each eleven participants. If the number of participants was even, players were equally divided into players 1 and 2, maintaining these roles throughout the session. In each period subjects in role 1 were randomly matched with subjects in role 2 and typically the basic game would be played simultaneously by six different pairs of subjects. If the number of participants was odd, one randomly selected subject in role one had to sit out.

Sessions differed by two main design features: 1. the information on the outcomes of plays given after each period, as described further below, and 2. whether player 1 (in sessions 9, 10, and 11) or player 2 (in all other sessions) moved first in both 2x2 bargaining subgames. We could not find any significant differences between the treatments and hence pool the data.[12]

	Player 2	Player 1
Statistics	1 - 8	9
no Statistics	12, 13	10, 11

FIGURE 3: Variations of the experimenal design in Sessions 1 - 13. In most, but not all, sessions a summary statistics was given and player 2 moved first in the bargaining subgames.

Each session consisted of two parts, using the strategy method Selten (1967):

In the first part the basic game was played strictly sequentially according to the game tree in 50 rounds.[13] In sessions 1 to 9 each player was informed after each round not only about the outcome (i.e. the terminal node reached) in his own play, and the corresponding payoff received, but also of the outcome in the other simultaneous plays by other pairs. A subject not matched would still receive the information about the outcomes in all plays of the period. We hoped to ease learning this way. Since all subjects could condition their behavior on this public information, strategic behavior as familiar from the folk theorem

[11] For a much more detailed description of the design see Balkenborg (1994).

[12] "Virtual observabily" does not seem to be an issue in our experiment. It seems that the order of moves only matters when there is nothing else to distinguish the players.

[13] Between round 25 and round 26 there was a break in which they were offered a coffee and the instructors made sure that there was no communication about the game

cannot be excluded. For this reason we conducted four further sessions, sessions 10 - 13, were only the information on the own play was given. The computer did not provide information on past plays. Instead subjects were given forms which allowed them to keep track of all this information over time. However this was not enforced and often not done.

In the second part players had to select strategies; i.e., they had to select a choice for each of their information sets in the game.[14] Once the strategy had been selected, the subject had to confirm it and send it off. When all strategies of all subjects had been submitted, the computer would evaluate each strategy of a subject in role 1 against all strategies submitted by subject in role 2 (and a randomly selected move by nature). A subject received the average payment from all the plays in which he participated. In session 1 to 9 a subject was not only told at the end of a period which terminal nodes were reached in his plays, but also how often each terminal node was reached in all (typically 36) plays in the period. Again, the latter information was not given in the last four sessions.

At the end of the session each participant received in private his total gain in German Marks with an exchange rate that varied between sessions subject to our budget constraint. In sessions 1 - 9 it varied between 0.09 and 0.11, yielding average payoffs about 35 DM ($23 at the time) for subjects in role 1 and 25 DM ($17) for subjects in role 2. In sessions 10 to 12 the exchange was 0.14 and 0.16 in session 13 with correspondingly higher average payoffs.

4 Results

In this section we discuss first the behavior in the main part of the experiment, in which the game was played sequentially, and then the final part, where the subjects submitted complete strategies. Unless stated otherwise, we use a sign test over all 13 independent sessions at the significance level of 5% to test a 1-hypotheses of the form that number n is higher than number m against the zero hypothesis that n is as often higher than m than that it is lower. We also give average behavior over all sessions. We will mention if the test does not yield significance for the first eight sessions alone where the treatments are identical. As stated above, we pool over all 13 sessions since differences in the treatments seem not to affect the results.

[14] Initially it was only vaguely announced that there would be a further brief second part. Instructions were then given after the first part had concluded.

4.1 Results for the main part

Table 11 and 12 show the frequencies with which each terminal node is reached in each session and overall, pooled over all periods, for the left and for the right subgame.

We will first discuss the results for the right subgame, then for the left subgame followed by a comparison of behavior between the two bargaining subgames. The subsection concludes with the behavior over time.

4.1.1 The right subgame

The right bargaining game was selected by the chance move in 1817 (49%) out of a total of 3700 plays of the basic game in all thirteen sessions.

Result 1 In the right subgame the focal outcome (**E**) with payoffs $(6, 6)$ is reached overall in 86%. In session 11 64% of plays resulted in this outcome. In all other sessions the percentage is at least 80% (see Figure 4). A sign test rejects the hypothesis that the outcomes **F, G** or **H** together are observed as often as the equal division outcome **I** in favor of the hypotheses that **I** is observed more often ($p \leq 0.000244$).

Note that the unequal division was only reached in 1% of all cases.

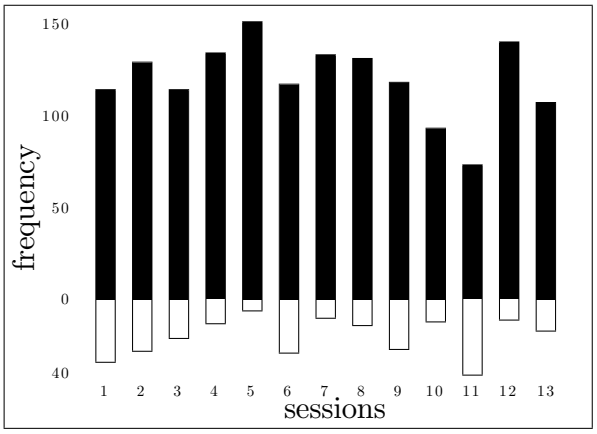

FIGURE 4: For the right subgame the figure shows for each session how often the equal division outcome (indicated by black bars, above) and the other outcomes (white bars, below) were reached in absolute numbers in all plays of the subgame in all 50 rounds of the main part by all pairs.

Result 2 Subject in role 1 chose "Left" more often than subjects in role 2 (11% versus 4%). As Figure 5 shows, this holds for all sessions but sessions 1 and 5, and is significant ($p \leq 0.0225$).[15]

[15] In session 1 players 2 propose "Left" 22 times. However, 16 were made by two subjects who obviously did not understand the rules of the game or the handling of the keyboard. In session 5 both player types choose "Left" with the same (low) frequency and therefore the focal outcome was reached in 96\%.

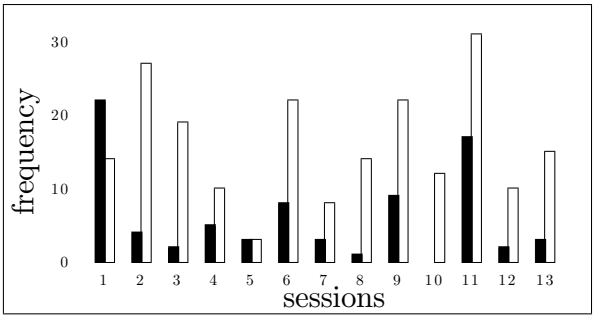

FIGURE 5: Unequal division proposals in the right bargaining subgame by subjects in role 1 (white bars, left) and in role 2 (black bars, right), details as in Figure 4.

Overall we observe very little coordination failure in the right bargaining subgame. Only very few subjects in role 1 try repeatedly to achieve the unequal division (see Balkenborg (1994)).

4.1.2 The left subgame

Result 3 The outside option is observed in each session more often than any of the other outcomes of this subgame together, on average with 88% ($p \leq 0.000244$). In each session this outcome was reached in at least 82% (see Figure 6).

The forward induction outcome is reached in less than 2% of all plays.

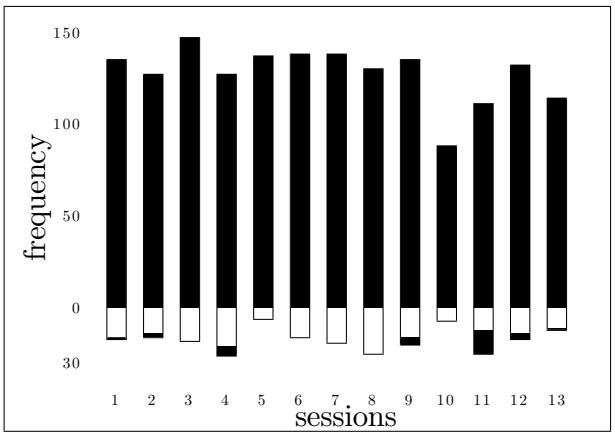

FIGURE 6: Number of outside options outcomes (black bars, above), forward induction outcomes (black bars, below, often zero) and other outcomes (white bars, below) for the right subgame, details as in Figure 4.

The left bargaining subgame In particular, the left bargaining subgame is not reached very often. Conditional on reaching it there is much less coordination than in the right bargaining subgame. The unequal split (**B**) is reached in 13% and the equal split (**E**) is reached in 38% of the cases. In nine session outcome **E** is reached more often than outcome **B**, in two sessions less often and in two sessions equally often ($p \leq 0.0654$).

Result 4 Subjects in role 2 chose "Right" (on average 82%) more often than "Left" ($p \leq 0.00342$). This holds in all but one session (see Figure 7).

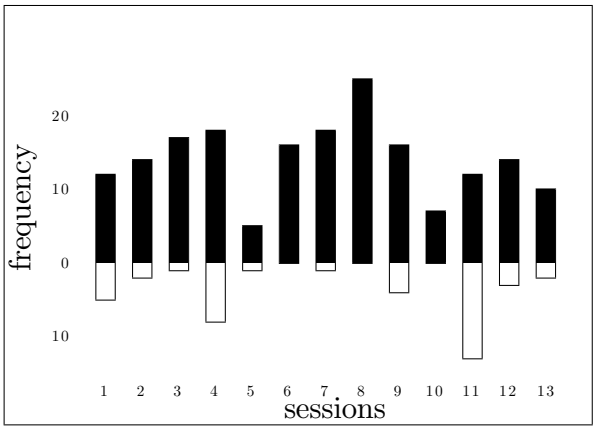

FIGURE 7: Equal divison proposals (black bars, above) and unequal division proposals (white bars, below) in the left bargaining subgame by subjects in role 2, details as in Figure 4.

In contrast, as shown in Figure 8, the behavior of subjects in role 1 in the left bargaining subgame varies from session to session. Only very few subjects chose repeatedly to play the left bargaining subgame (see Balkenborg (1994)). They predominantly proposed the unequal division in six session and the equal division in 4 sessions. In three sessions the numbers are equal. Counted over all session the unequal division was proposed in 57%. Comparing the behavior of the two players we have:

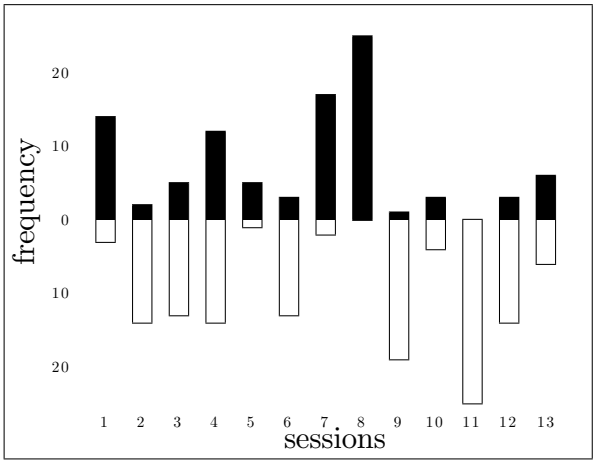

FIGURE 8: Equal divison proposals (black bars, above) and unequal division proposals (white bars, below) in the left bargaining subgame by subjects in role 1, details as in Figure 4.

Result 5 Subjects in role 1 chose "Left" more often than subjects in role 2 (128 times or 57% versus 40 times or 18%) ($p \leq 0.0386$). This result does not hold in two sessions and in one session no player chose left**.**

4.1.3 Comparison of play in the left and right subgame

Result 6 Subjects in role 1 try to reach the unequal division more often in the right subgame by playing left than they try to reach it in the left subgame by forgoing the outside option and then choosing left. Although the averages are close (11% versus 7%) the result holds for 11 sessions and is significant ($p \leq 0.0225$).

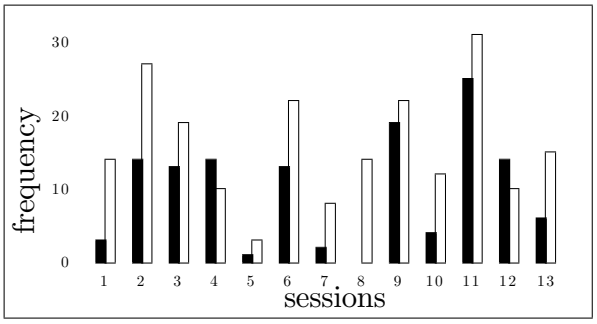

FIGURE 9: Number of times subjects in role 1 used the forward induction strategy in the left subgame (black bars, left) and number of times they proposed the unequal division in the right subgame (white bars, right), details as in Figure 4.

Comparison of play in the left and right *bargaining* subgame Our results so far indicate that the behavior of subjects is consistent with the backward induction solution and Harsanyi and Selten's solution equilibrium theory. The next results show, however, that behavior in the two bargaining subgames is markedly different and insofar contradicts subgame consistency. One must keep in mind, of course, that the absolute frequencies for the left bargaining subgame are very small.

Result 7 Subjects in role 1 tend to propose the unequal division more often (57%) in the left bargaining subgame than in the right bargaining subgame (11%). This holds in all but one session ($p \leq 0.00342$).

Result 8 Subjects in role 2 tend to propose the unequal division more often (18%) in the left bargaining subgame than in the right bargaining subgame (4%). This holds in 10 sessions while in one session no role 2 player chose left ($p \leq 0.0386$). However, we do not get significance for the first eight sessions alone. But we do not get significance for the first eight sessions alone. On the first 8 sessions a Wilcoxon test yields a p-value of 7%.

Result 9 There is more miscoordination in the left bargaining subgame (49%) than in the right bargaining subgame (14%). The fair outcome is reached less often in the left than in the right bargaining subgame (38% versus 86%). Both results holds in all but one session ($p \leq 0.00342$).

It is not significant that the unequal division is reached more often in the left than in the right bargaining subgame. The comparison of the averages (13% versus 1%) is misleading here since there are several sessions without forward induction play.

4.1.4 Behavior over time

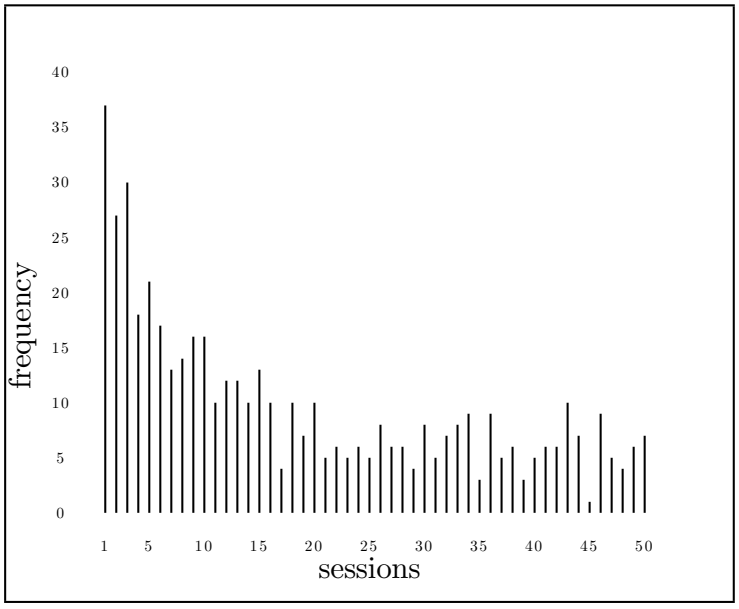

FIGURE 10: Number of plays that did not result in the outside option or the equal division in the right subgame per period. The data are aggregated over all 13 sessions.

Aggregated over all sessions 74 pairs played the basic game in each round (10 sessions with 6 pairs and 1 session with 4 pairs and 2 sessions with 5 pairs) per period. Figure 10 shows for each period how many of these plays resulted in outcomes different from

the (7,4) payoff in the left game and the (6,6) payoff in the right game. We see that in the first period exactly 1/2 of all plays did not result in one of these outcomes with a steep decline immediately after that period. After period 5 there are no more than 25% atypical outcomes and after period 15 no more than 13.5%. From then on the picture looks pretty stationary, centered around 5 atypical results per period. The picture is however misleading insofar, as there are severe differences with respect to atypical behavior for the different sessions in the later periods: A χ-test rejects at a 5% significance level the hypothesis that the data can be assumed to be in all sessions independently and identically distributed in the second half of the main part of the experiment.[16]

4.2 The Final Part

In the five rounds of the final part of the experiment subjects had to hand in complete strategies for the extensive game. Table 13 and 14 shows the frequencies of outcomes for both subgames (parallel to tables 11 and 12). Tables 15 and tables 16 show the strategies for each type of player, respectively.[17] With respect to outcomes the aggregate frequencies are surprisingly similar to the main part of the experiment, except that here in the right bargaining subgame the outcome (9,3) is reached only in session 1. All results we described in the previous subsections hold here with the exception that it is no longer significant that subjects in role 1 tended to use the forward-induction strategy for the left subgame less often than they chose to propose the unequal division in the right subgame.

This similarity with the main part of the experiment implies that most players 1 choose the outside option in the left game (89%) and both players choose right in the right game (91% players 1 and 96% players 2). The highest frequencies for a single strategy of player 1 are ORR with 66%, followed by OLR with 17% and for player 2 are rr with 77%, followed by lr with 19%. All other strategies are typically played with less than 10% in the single sessions and less than 5% across all sessions.

In those sessions where some subjects in role 1 used the forward induction strategy repeatedly in the main part of the experiment there is a visible effect on the behavior of subjects in role 2 in the final part: The more the forward induction strategy was used in the main part, the more often subjects in role 2 tended to propose the unequal division in the final part. For each session we can look at the number of times subjects in role 1

[16] For various time ranges between period 20 and 50 we calculated how many "typical" outcomes (**O** in the left subgame or **E** in the right subgame) and how many atypical outcomes where obtained in each period and calculated the χ-statistic from these values. The lowest χ-statistic we found was for the time range from period 25 to period 44, where the χ-statistic was 40.8.

[17] A strategy specifies a vector of three choices for player 1: 1. outside option In or Out, 2, L or R in the left bargaining subgame, L or R in the right bargaining subgame. For player 2 it is a vector of two choices for the left and right subgame, respectively.

chose the forward induction strategy in the main part relative to the number of times this subgame was played. For the final part of each session we can look at the number of times subjects in role 2 chose to propose the unequal division for the left bargaining game relative to the number of strategies handed in. The Kendall rank correlation coefficient for these two groups of numbers is $\tau = 0.63$. Thus the ranks are positively correlated and a rank correlation test yields a p-value of 0.13%. Consequently, the hypothesis that the ranks are uncorrelated can be rejected. Still, the number of times subjects in role 2 proposed the unequal division in the final part was never high enough to make it worthwhile for subjects in role 1 to use the forward induction strategy.

5 Some behavioral theories to explain the results

Our data seem to clearly reject the forward induction solution in favor of backward induction based on the focal point. Substantial empirical evidence in existing experimental literature often reject, however, predictions based on backward induction since it requires a too complicated reasoning or is in conflict with fairness consideration. Since fairness considerations seem to matter for our results it is important to know what the recently developed descriptive theories of fairness imply for our games. Unless otherwise mentioned, our discussion refers to the left subgame with the outside option.

5.1 Fairness theories, and levels of reasoning

In the following we explore the implications of the fairness theories by Fehr and Schmidt (1999), Bolton and Ockenfels (2000) and Charness and Rabin (2002) for our model.

All three fairness theories assume that players maximize utility (or "motivational") functions which depend not only on their own monetary gains but also on that of the opponent. A player is assumed to perceive some additional disutility if the opponent gets

more than he does or vice versa.[18, 19] In all theories different "types" of a player have different attitudes towards fairness and hence different utility functions. In the theory by Charness and Rabin a player may perceive additional disutility if he accommodates "bad" behavior of the opponent. The assumption that players can be of different types turns the games studied into genuine games of incomplete information. In all theories a mild equilibrium refinement concept (sequential equilibrium of perfect Bayesian equilibrium) is used to determine a solution. In the games we study here, perfect Bayesian equilibria, sequential equilibria, extensive form perfect Selten (1975) and normal form perfect equilibria all have the same outcomes. We will hence briefly speak of "fairness equilibrium" when we refer to any such equilibrium in one of these incomplete information games with fairness motivated players.

In order to calculate equilibria in the Bayesian games, priors over the types must be given. Fehr and Schmidt provide bold calibrations for the distributions of types while Bolton and Ockenfels largely abstain from such calibrations and assume only that all relevant types have positive mass. We will need here one assumption on the distribution of types which seems to be in agreement with many experimental data sets, in particular on ultimatum games. This assumption is implied by the Fehr-Schmidt calibration and consistent with the Bolton-Ockenfels model. Under this assumption the fairness models considered here are dominance solvable and have only two Nash equilibrium components. Below we will consider first dominance solvability, i.e., the iterated elimination of weakly dominated strategies. Recall that the forward induction solution is obtained in three steps by iteratively eliminating weakly dominated strategies.

Relevant for us is that all theories assume a certain fraction of subjects in the role of player 2 to reject $(9, 3)$ in favor of $(0, 0)$. In the basic Charness-Rabin model, which is an

[18] More precisely, they behave as if they would perceive some disutility.

[19] Let $x > 0$ be the payoff of the player at a terminal node and let $y > 0$ be that of the opponent. Then utility is in the Fehr-Schmid model utility

$$u = x - \alpha \max [0, y - x] - \beta \max [0, x - y]$$

where the parameter α describes how much one dislikes other having more ("negative inequality aversion") and β how much one dislikes others having less than oneself ("positive inequality aversion"). In the Bolten-Ockenfels utility takes the form

$$u\left(x, \frac{x}{x+y}\right)$$

where the utility function is increasing in the first argument and single peaked in the second argument with a maximum at 1/2.Charness and Rabin extend the Fehr-Schmid model by subtracting a further term

$$... - \gamma\sigma \min [0, y - x]$$

with $\gamma \geq 0$ where $\sigma = 1$ if the opponent behaved nasty and $\sigma = 0$ otherwise.

extension of the Fehr-Schmidt model as far as the utility functions are concerned, the set of types, for whom proposing the unequal division in the bargaining subgame is dominated, can only be larger than in the Fehr-Schmidt model: Suppose a player 2 observes that player 1 has chosen to play the bargaining subgame. Then there are two possibilities. Either player 1 is a "nice person" who aims for the equal division. In this case player 2 should propose the equal division too. Or it is a player with nasty intentions heading for the unequal division. In that case proposing the unequal division would give player 2 disutility because he would reward bad behavior. If this disutility, together with his inequality aversion, is high enough he will propose the equal division even if it brings him zero. Regardless of whether it is inequality aversion or the disutility from accommodating bad behavior that motivates a type to reject $(9, 3)$ in the ultimatum bargaining game, if he does so in the ultimatum bargaining game, he should also reject $(9, 3)$ in our game.

What fraction of the types of player 2 can be reasonably expected *not to* propose the unequal division or to reject $(9, 3)$ in an ultimatum game? Based on experimental results in ultimatum games Fehr and Schmidt estimate the fraction to be 40-70%.[20] Clearly, there is much variance in the data and results change with details in the procedure. However, any fraction above two ninth (or 22%) of these highly inequity averse types is enough for the following argument.

5.1.1 Dominance solvability in the left subgame

We now show that the iterated elimination of weakly dominated strategies leads in four steps to a unique fairness equilibrium which is a partially separating equilibrium. This equilibrium is chosen by most refinement concepts discussed in the literature on signalling games.

For those highly inequity averse types of player 2 who prefer $(0, 0)$ in favor $(9, 3)$ it is a weakly dominated strategy to propose the unequal division.[21] Once this strategy is eliminated (step 1) for all these types (at least 22%), it becomes strictly dominated for *every* type of player 1 to forgo the outside option and try to reach the *unequal division*. He would gain less than $\frac{7}{9} \times 9 = 7$ in expectation and therefore less than with the outside option.

If IN and left is eliminated for all types of player 1 (step 2), it becomes weakly dominated for *all* types of player 2 to propose the unequal division and thus this strategy can be eliminated (step 3). Finally, (step 4) only those players 1 will enter the bargaining

[20] Our example is a knife-edge case for their calibration. They assume 30% of types to have the inequity aversion parameter $\alpha = 0.5$ and hence to be indifferent between $(9, 3)$ and $(0, 0)$, see Fehr and Schmidt (1999), p. 844.

[21] For all such types it is a *strictly* dominated strategy to do so in the right subgame.

game who are sufficiently fairness minded to prefer $(6, 6)$ over $(7, 4)$ while all other types will take the outside option.

Proposition 1 *Assume that there are more than two ninth of the types of player 2 for whom proposing the unequal division is weakly dominated. Suppose that there is a positive mass of types of player 1 who prefer (6,6) over (7,4). Then the fairness model based on the bargaining game with outside option is dominance solvable in four steps. The solution yields an isolated Nash equilibrium. It is partially separating for the types of player 1. Sufficiently selfish types choose the outside option while sufficiently fairness minded types choose to bargain and to propose the equal division. All types of player 2 propose the equal division when the bargaining subgame is reached, which happens with positive probability.*

The Nash equilibrium just described is perfect and strategically stable. Also Harsanyi and Selten's theory would select this equilibrium in the fairness model.

Conditional on the bargaining subgame being reached outcome **E** will be reached in this equilibrium for sure. Based on experimental dictator game results (see the references in Fehr and Schmidt (1999)) one might expect the fraction of positively "fair" players to be around 20%, but we observe much less, around 5%. We do observe frequent fairness play by some of the subjects only in three out of 13 experiments (namely in sessions 1, 7 and 8). When a fairness player is around, the forward induction strategy is practically never used, an observation which fits well with the equilibria just described. One could argue that this equilibrium plays some role for these sessions. However, a sign test would reject in our data the hypotheses that outcome **E** is reached with at least 70% of all times the subgame is reached. Overall we would still conclude that the "hyperfair" equilibrium is not a good description of the data. This is consistent with evidence from other experimental literature (see Nagel (1995), Camerer (2003) which shows that three or four levels of reasoning, as required here for the equilibrium, are practically not observed).

5.1.2 Further Nash equilibria in the left subgame

For the left subgame we only eliminated iteratively *weakly* dominated strategies and hence other Nash equilibria then the one found can – and often do – exist. It is easy to see, by following the chain of dominance arguments from above, that the only Nash equilibrium of the fairness model where the bargaining subgame is reached with positive probability is the partially separating equilibrium we have already encountered. Consequently the only other type of equilibrium is a pooling equilibrium where all types of player 1 choose the outside option.

Proposition 2 *Under the assumption of Proposition 1, there can exist further fairness*

equilibria of the bargaining game with the outside option. In any of these equilibria the outside option is taken with certainty.

This solution fits better with the data from our experiment.

Proof. To show existence of the perfect equilibria as described, consider trembles of (the types of) player 1 where the subgame is reached with very small probability and then the unequal division is proposed with a conditionally very high probability. All sufficiently selfish types of player 2 will then propose the unequal division if the bargaining subgame is reached. Provided this fraction is so high that even the most fairness-motivated type of player 1 does not want to enter (here the details of the distribution and the type of utility functions matter, extremely fairness minded types may have to be ruled out), we have a perfect Nash equilibrium.

Concerning uniqueness, consider a Nash equilibrium where the bargaining subgame is reached with positive probability. By assumption, there is at least a probability of two ninth that player 2 will propose the equal division and so no type of player 1 is going to enter and propose the unequal division. Thus all types of player 2 would have to propose the equal division in a best reply. Correspondingly, a positive fraction of types of player 1 will enter. The equilibrium outcome obtained coincides with the dominance solution. The only way to get a different type of equilibrium is by not having the bargaining subgame reached, because then it can be rational for selfish types of player 2 to propose the unequal division. ■

5.1.3 The right subgame

For the right subgame to be dominance solvable, we need a much stronger assumption on the fairness attitude of player 2, which is, however, consistent with many experimental data sets on the ultimatum game. A possible parallel assumption for player 1, that is inconsistent with the data, leads to the same result.

Proposition 3 *Suppose that more than 60% of types of player 2 or more than one third of players 1 prefer* $(0,0)$ *over* $(9,3)$*. Then the fairness model based on the isolated bargaining game has a unique rationalizable outcome and hence a unique Nash equilibrium outcome which agrees with the equal division outcome and is reached with three steps of elimination of* strictly *dominated strategies.*

5.1.4 Fairness theories in other forward induction experiments

Cooper, DeJong, Forsythe, and Ross (1993), Schotter, Weigelt, and Wilson (1994), Brandts and Holt (1992), and Brandts and Holt (1995) report evidence on experiments with games

where a battle-of-the-sexes-game is preceded by an outside option. They find mild evidence in favor of forward induction, although the outside option is still taken with a high probability. However, conditional on the battle-of-the-sexes game being played subjects tend to play the equilibrium advantageous for the player with the outside option. The results strongly suggest that forward induction would fair better if we replaced the outcome $(6, 6)$ in our game with the outcome $(3, 9)$. If true, this would contradict the predictions of the fairness theories which do not depend on this change, since fairness oriented player 2 would favor $(3, 9)$over $(9, 3)$.

Thus the focal point outcome - and hence fairness - seems to matter for our results, although the fairness theories are rejected. Interestingly, a related phenomena arises in the mini ultimatum games Falk, Fehr, and Fischbacher (2003). When a proposer selects the division $(8, 2)$ favoring him out of the only two allowed divisions $(8, 2)$ and $(5, 5)$, he is often rejected. He is rejected less often when he chooses the division $(8, 2)$ out of the only two allowed divisions $(8, 2)$ and $(2, 8)$ (see). In the first, but not in the latter case his behavior is regarded unfair. Thus intentions matter. Overall we observe that we can use the fairness theories based on intentions by Charness and Rabin (2002) to explain both our results and that of Cooper, DeJong, Forsythe, and Ross (1993). The extension of fairness theories is not only relevant for ultimatum games but equally for forward induction experiments.

5.1.5 Fairness and Quantal Response

Even a fair player 1 will choose to bargain and propose the equal division in the left subgame only, if he expects player 2 to choose right with a sufficiently high probability. In quantal response equilibria (McKelvey and Palfrey (1995), McKelvey and Palfrey (1998)) players make naturally errors. It is easy to see that if the error probability is sufficiently high, even a fair player 1 will intend to stay out. Consequently there remain one fairness equilibria remains, where all types of player 1, fair or selfish, choose the outside option.

5.2 Equilibrium selection based on limited levels of reasoning and learning.

The following theories provide alternative explanations why forward induction might not work in our experiment. We consider first a simple argument assuming one level of reasoning which leads immediately to the outside option outcome for the left subgame. For the right subgame without the outside option we either need an argument assuming two levels of reasoning or the level 1 argument combined with a learning model. One may be tempted to estimate, as has been done for many other games, learning models for our

data. However, the task is not very insightful because most play is anyway in equilibrium and the few deviations observed are caused by very few players (see Balkenborg (1994)). Hence we abstain from it here.

5.2.1 Simple Priors

The following argument is in line with the arguments in Brandts and Holt (1993). We consider the left bargaining game with the outside option. Suppose that both players believe that if the bargaining subgame were reached the opponent would choose any of his two strategies with equal probability. Suppose every player plays a best reply against this belief, so there is one level of reasoning. Then player 2 would expect payoff 1.5 by going left and 3 by going right. He would hence choose right in the bargaining subgame. Player 1 would expect 3 from going into the subgame followed by playing right, 4.5 from going into the subgame followed by playing left and 7 from taking his outside option. He would hence take the outside option in a best reply. Thus we obtain immediately a subjective equilibrium (in the sense of Fudenberg and Levine (1998)) where the outside option is taken. As long as player 1 does not go into the subgame the beliefs of both players are not questioned. If there are both level 1 types (who play right) and level zero types (who mix 50-50) of player 2, player 1's expectation that it is not worthwhile to play the subgame would be confirmed if he would occasionally experiment and play it.

While the theory does explain our results, it would also lead to the outside option outcome in the experiment by Cooper, DeJong, Forsythe, and Ross (1993) and hence contradict their result. It seems that fairness matters in our experiment, although in a way which is not explained by the Fehr-Schmidt and Bolton-Ockenfels approach.

5.2.2 Risk dominance equilibrium

In the isolated bargaining game level-1-reasoning, in the sense that each player plays a best reply against a naive, uniform prior, would not result in equilibrium play. However, a level 2 argument based on a uniform prior would lead to the risk-dominant equilibrium. We present here a slightly simplified argument from Harsanyi and Selten (1988). The argument is easiest understood for a 2×2-bimatrix game as given below. Hereby the numbers a_i, b_i $(i = 1, 2)$ are assumed to be positive so that (TOP, LEFT) and (BOTTOM, RIGHT) are the (only) pure strategy equilibria of the game. The equilibrium with the higher Nash

product a_1a_2 or b_1b_2 is the risk dominant Nash equilibrium.

	LEFT	RIGHT
TOP	a_2 a_1	0 0
BOTTOM	0 0	b_2 b_1

Player 1 plays a best reply against the expected strategy of his opponent. He believes that his opponent plays a best reply against some belief and that all beliefs of his opponent are equally likely. Each possible belief – and hence each "type" – of his opponent is described by the probability $0 \leq p \leq 1$ with which the opponent beliefs that player 1 is playing BOTTOM. Thus player 1 beliefs that p is uniformly distributed on $[0, 1]$. A rational player 2 will choose LEFT if he assigns probability $p < \frac{a_2}{a_2+b_2}$ to his opponent playing BOTTOM and he will choose RIGHT if he assigns probability $p > \frac{a_2}{a_2+b_2}$ to his opponent playing BOTTOM. Given his uniform prior, player 1 therefore expects player 2 to play LEFT with probability $\frac{a_2}{a_2+b_2}$ and RIGHT with probability $\frac{b_2}{a_2+b_2}$. Thus player 1 expects payoff $\frac{a_1a_2}{a_2+b_2}$ by playing TOP and $\frac{b_1b_2}{a_2+b_2}$ by playing BOTTOM. He will play TOP when $a_1a_2 > b_1b_2$ and BOTTOM in the opposite case. A symmetric reasoning yields that player 2 will play LEFT if $a_1a_2 > b_1b_2$ and RIGHT in the opposite case. Thus both players choose in accordance with the risk dominant equilibrium.

In our game, this two-level process of reasoning leads to the equal division outcome.

The same analysis can be made in the game with the outside option, provided player 2 optimizes against conditional beliefs and player 1 assumes a uniform prior over these conditional beliefs of his opponent. If player 1 would go into the 2×2 subgame he would expect payoff $\frac{3\times9}{3+6} = 3$ from going left and $\frac{6\times6}{3+6} = 4$ from going right. He would choose his outside option. Player 2, assuming that makes both choices in the subgame with equal probabilities, would choose right as before.

5.2.3 Learning

As we have discussed, playing best replies to a uniform level 1 prior leads for the left subgame immediately to the prediction that player 1 will choose the outside option. Simple learning models like fictitious play converge therefore in the first period. For the right subgame on the right this is not the case. We show next, as an alternative to the above level 2 argument, how fictitious play learning based on a naive prior leads to convergence on the equal division outcome. We consider here fictitious play with a Dirichlet prior (see Fudenberg and Levine (1998), Chapter 2) for the case where two players repeatedly play the right subgame. In period $t \geq 1$ player i then plays a best reply against the belief with

which his opponent is assumed to propose the unequal or equal division. It is given by the probabilities

$$\left(q^t_{U,-i}, q^t_{E,-i}\right) = \left(\frac{m_{U,-i} + n^t_{U,-i}}{m_{U,-i} + n^t_{U,-i} + m_{E,-i} + n^t_{E,-i}}, \frac{m_{E,-i} + n^t_{E,-i}}{m_{U,-i} + n^t_{U,-i} + m_{E,-i} + n^t_{E,-i}} \right)$$

Hereby $n^t_{U,-i}$ (respectively $n^t_{E,-i}$) is the number of times the opponent proposed the unequal (respectively equal) division in the past. $m_{U,-i}$ and $m_{E,-i}$ represent fictitious past experience and describe the initial prior. These numbers are parameters of the model. The larger they are, the less weight is put on actual experience. Once they are chosen, the path of fictitious play is deterministically determined. Let us now consider a naive prior, i.e. let

$$m = m_{U,1} = m_{E,1} = m_{U,2} = m_{E,2}$$

Then player 1 will propose the unequal division and player 2 the equal division in a best reply in the first period. If the opponent keeps proposing the equal division in the following periods, player 1 will switch his strategy after $m/2$ periods because then proposing the equal division becomes the best reply against the belief $\left(q^t_{U,-i}, q^t_{E,-i}\right) = \left(\frac{m}{2m+t}, \frac{m+t}{2m+t}\right)$ since $9m < 6\,(m+t)$ for $t > m/2$. If player 1 would keep proposing the unequal division, player 2 would switch to propose the unequal division later, after period m, because only then $3\,(m+t) > 6m$ holds. Thus, in all periods up to period $m/2$ player 1 would propose the unequal division and player 2 the equal division. Thereafter player 1 would switch to proposing the equal division, player 2 would no longer have to switch and play would have converged on the equal division equilibrium. How long it takes to converge depends on the parameter m in the model. For the left subgame fictitious play based on a naive prior would converge immediately on the outside option, as we remarked above.

It is not difficult, although also not very insightful, to extend the above analysis to the random-matching environment used in our experimental design.

The above explanations of our results are simple and appealing, but have some disadvantages. We already mentioned that the theory does not explain why there is convergence to the forward induction solution in the experiments by Cooper, DeJong, Forsythe, and Ross (1993) and Brandts and Holt (1992). Moreover, rote models of learning cannot easily explain why no convergence occurs or why convergence is much slower when the games are presented in normal form or similar (see Cooper, DeJong, Forsythe, and Ross (1993), Huck and Müller (2005), Caminati, Innocenti, and Ricciuti (2006)). Cognitive aspects, e.g. whether an outcome can be regarded as focal or as a suitable compromise, seem relevant for how quickly one observes convergence to an equilibrium, if at all.

In contrast to these simple deterministic models, learning models with randomness and experimentation will typically converge to the forward induction outcome in the very

long run because it is the unique strict equilibrium of the reduced normal form of the left subgame. Binmore and Samuelson (1999) argue, however, that both equilibrium components for the Dalek game are asymptotically stable for learning processes if an inward pointing drift is added. We believe that their considerations are important to explain the robustness of the outside option outcome in the long run but do not fully explain the speedy convergence we observe in our experiment in the short run.

6 Conclusions

We conducted an experiment where a forward induction hypothesis was clearly rejected in favor of backward induction based on a simple focal point, provided that the games are analyzed as one-shot games and that payoffs and monetary incentives are identified. Subjects had no difficulty to coordinate on the risk-dominant, equal division equilibrium in the 2×2 bargaining subgame. However, on a finer level we find differences in the play of the bargaining subgames with or without outside option, which are not in line with subgame consistency.

As Binmore, Proulx, Samuelson, and Swierzbinski (1995) summarize our finding:

> "The risks associated with a hard bargaining that is required to achieve an efficient outcome are avoided by inefficiently opting out."

Since there is some evidence for forward induction in other experiments for very similar games, we considered two behavioral approaches to explain the observations. One was learning based on naive priors which leads to the observed behavior in the long run. However, the less structured results when similar games are played in normal form (Cooper, DeJong, Forsythe, and Ross (1993), Muller and Sadanand (2003)) suggest that the result is not due to pure rote learning alone. We expect cognitive considerations, in particular whether subjects regard an outcome as fair or not, to matter as well. For this reason we studied what the recent fairness theories of Fehr and Schmidt (1999) and Bolton and Ockenfels (2000) (and also Charness and Rabin (2002)) tell us about our game.

These fairness theories rely conservatively on traditional game theoretic assumptions, in particular full sequential rationality, but use motivational or utility functions which are in better agreement with experimental observations. The use of incomplete information models allows to take account of subject heterogeneity. The approach used in the fairness theories has the advantage that one can draw conclusions from experimental results on dictator and ultimatum games for other types of games. For our outside option game the approach leads to the prediction that player 1 should predominantly take the outside option, as is observed in our experiment.

One disadvantage of the approach is that it does not, at least currently, model the cognitive limitations of subjects (see Binmore, McCarthy, Ponti, Samuelson, and Shaked (2002)). In our case the fairness model has one game theoretically very appealing equilibrium, a partially separating equilibrium, which is cognitively even more complex than the forward induction equilibrium of the original game. It requires four levels of reasoning, which is typically not observed in the experimental literature. To play this equilibrium subjects must understand that only very fairness minded types of player 1 will enter the bargaining subgame with the intention to coordinate on the equal division outcome. It should be a novelty that a "fair" equilibrium is not played because it is cognitively more complex than a simple backward induction equilibrium. Arguably, standard rationality is restored because fairness considerations are cognitively too complex.

Our application of the fairness models rests on the assumption that a sufficiently large proportion of players would reject a split $(9, 3)$ disadvantageous for them in favor of zero for both players, regardless of the context in which this decision is embedded. As one of the inventors of the fairness theories himself observed (see Falk, Fehr, and Fischbacher (2003)), this assumption is already violated in simple mini-ultimatum games. It is interesting to see that the same difficulty arises for forward induction experiments, when one compares our results with those of Cooper, DeJong, Forsythe, and Ross (1993).

In this paper we took an experiment designed to test equilibrium refinement or selection theories and discussed its results it the light of new behavioral theories. We feel that the discussion led to new insights and raised new questions on the interplay between coordination, fairness, learning and levels of reasoning which future research will need to address.

A Appendix

The following pages summarize the essential data from the experiment. For the main part in each session they show in Tables 11 and 12 how often the left and the right subgame (Γ_L and Γ_R) and each of their terminal nodes, as labelled in Figure 2, were reached in all plays of the game by any pair in the fifty rounds. A session with 12 subjects would thus have 300 plays. The numbers in brackets give the frequency relative to the number of times the relevant subgame was reached, i.e. it is the fraction calculated by dividing number before the bracket by the number in the last column. In the last column we aggregate the absolute frequencies and calculate relative frequencies again by dividing by the last column. We did not average average frequencies.

Tables 13 and 14 are constructed similarly for the final part, by evaluating all plays for all strategies of subjects in role 1 against all strategies of subjects in role 2 for each of the

five final periods. In a session with 12 subjects there would be $5 \times 36 = 180$ such plays. The strategies themselves chosen by the subjects in role 1 and 2 are given in the last two tables. O and I indicate "out" and "in". Choices for the left subgame are given first. The strategy ORL would, for instance, be the strategy of player 1 where he would choose the outside option in the left subgame, propose the equal division in the left bargaining subgame and the unequal division in the right bargaining subgame. In IRL he chooses to play the bargaining subgame.

In those session where the order of play in the bargaining subgames was interchanged the data are adjusted in the natural way.

Ses	A (outside option)	B (forward induction)	C (anti conflict)	D (conflict)	E (equal division)	Γ_L
1	135 (89%)	1 (1%)	4 (3%)	2 (1%)	10 (7%)	152
2	127 (89%)	2 (1%)	0 (0%)	12 (8%)	2 (1%)	143
3	147 (89%)	0 (0%)	1 (1%)	13 (8%)	4 (2%)	165
4	127 (83%)	5 (3%)	3 (2%)	9 (6%)	9 (6%)	153
5	137 (96%)	0 (0%)	1 (1%)	1 (1%)	4 (3%)	143
6	138 (90%)	0 (0%)	0 (0%)	13 (8%)	3 (2%)	154
7	138 (88%)	0 (0%)	1 (1%)	2 (1%)	16 (10%)	157
8	130 (84%)	0 (0%)	0 (0%)	0 (0%)	25 (16%)	155
9	135 (87%)	4 (3%)	0 (0%)	15 (10%)	1 (1%)	155
10	88 (93%)	0 (0%)	0 (0%)	4 (4%)	3 (3%)	95
11	111 (82%)	13 (10%)	0 (0%)	12 (9%)	0 (0%)	136
12	132 (89%)	3 (2%)	0 (0%)	11 (7%)	3 (2%)	149
13	114 (90%)	1 (1%)	1 (1%)	5 (4%)	5 (4%)	126
Σ	1659 (88%)	29 (2%)	11 (1%)	99 (5%)	85 (5%)	1883

TABLE 11: Main Part: Outcomes in the Left Subgame

Ses	F (unequal division)	G (anti conflict)	H (conflict)	I (equal division)	Γ_R
1	2 (1%)	20 (14%)	12 (8%)	114 (77%)	148
2	3 (2%)	1 (1%)	24 (15%)	129 (82%)	157
3	0 (0%)	2 (1%)	19 (14%)	114 (84%)	135
4	2 (1%)	3 (2%)	8 (5%)	134 (91%)	147
5	0 (0%)	3 (2%)	3 (2%)	151 (96%)	157
6	1 (1%)	7 (5%)	21 (14%)	117 (80%)	146
7	1 (1%)	2 (1%)	7 (5%)	133 (93%)	143
8	1 (1%)	0 (0%)	13 (9%)	131 (90%)	145
9	4 (3%)	5 (3%)	18 (12%)	118 (81%)	145
10	0 (0%)	0 (0%)	12 (11%)	93 (89%)	105
11	7 (6%)	10 (9%)	24 (21%)	73 (64%)	114
12	1 (1%)	1 (1%)	9 (6%)	140 (93%)	151
13	1 (1%)	2 (2%)	14 (11%)	107 (86%)	124
Σ	23 (1%)	56 (3%)	184 (10%)	1554 (86%)	1817

TABLE 12: Main Part: Outcomes in the Right Subgame

Ses	A (outside option)	B (forward induction)	C (anti conflict)	D (conflict)	E (equal division)	Γ_L
1	81 (98%)	0 (0%)	0 (0%)	2 (2%)	0 (0%)	83
2	57 (70%)	6 (7%)	1 (1%)	13 (16%)	4 (5%)	81
3	81 (93%)	1 (1%)	1 (1%)	3 (3%)	2 (1%)	87
4	86 (96%)	1 (1%)	0 (0%)	3 (3%)	0 (0%)	90
5	88 (99%)	1 (1%)	0 (0%)	0 (0%)	0 (0%)	89
6	99 (94%)	1 (1%)	0 (0%)	5 (5%)	0 (0%)	105
7	73 (88%)	0 (0%)	3 (4%)	0 (0%)	7 (8%)	83
8	79 (83%)	0 (0%)	0 (0%)	2 (2%)	14 (15%)	95
9	75 (74%)	5 (5%)	1 (1%)	18 (18%)	2 (2%)	101
10	49 (92%)	0 (0%)	0 (0%)	0 (0%)	4 (8%)	53
11	71 (83%)	4 (5%)	2 (2%)	4 (5%)	5 (6%)	86
12	86 (93%)	1 (1%)	0 (0%)	3 (3%)	2 (2%)	92
13	73 (82%)	0 (0%)	2 (2%)	7 (8%)	7 (8%)	89
Σ	998 (88%)	20 (2%)	10 (1%)	60 (5%)	46 (4%)	1134

TABLE 13: Final Part: Outcomes in the Left Subgame

Ses	F (unequal division)	G (anti conflict)	H (conflict)	I (equal division)	Γ_R
1	3 (3%)	24 (25%)	12 (12%)	58 (60%)	97
2	0 (0%)	0 (0%)	18 (18%)	81 (82%)	99
3	0 (0%)	0 (0%)	18 (19%)	75 (81%)	93
4	0 (0%)	4 (4%)	2 (2%)	84 (93%)	90
5	0 (0%)	0 (0%)	6 (7%)	85 (93%)	91
6	0 (0%)	0 (0%)	1 (1%)	74 (99%)	75
7	0 (0%)	6 (6%)	1 (1%)	90 (93%)	97
8	0 (0%)	0 (0%)	0 (0%)	85 (100%)	85
9	0 (0%)	0 (0%)	4 (5%)	75 (95%)	79
10	0 (0%)	0 (0%)	7 (15%)	40 (85%)	47
11	0 (0%)	0 (0%)	0 (0%)	64 (100%)	64
12	0 (0%)	0 (0%)	3 (3%)	85 (97%)	88
13	0 (0%)	5 (8%)	6 (10%)	50 (82%)	61
Σ	3 (0%)	39 (4%)	78 (7%)	946 (89%)	1066

TABLE 14: Final Part: Outcomes in the Right Subgame

Ses	OLL	OLR	ORL	ORR	ILL	ILR	IRL	IRR	Sum
1	5 (17%)	2 (7%)	1 (3%)	21 (70%)	0 (0%)	1 (3%)	0 (0%)	0 (0%)	30
2	1 (3%)	4 (13%)	0 (0%)	17 (57%)	5 (17%)	1 (3%)	0 (0%)	2 (7%)	30
3	2 (7%)	6 (20%)	2 (7%)	18 (60%)	1 (3%)	0 (0%)	1 (3%)	0 (0%)	30
4	0 (0%)	5 (17%)	0 (0%)	24 (80%)	1 (3%)	0 (0%)	0 (0%)	0 (0%)	30
5	1 (3%)	1 (3%)	0 (0%)	27 (90%)	1 (3%)	0 (0%)	0 (0%)	0 (0%)	30
6	0 (0%)	8 (27%)	1 (3%)	19 (63%)	0 (0%)	2 (7%)	0 (0%)	0 (0%)	30
7	1 (3%)	4 (13%)	0 (0%)	22 (73%)	0 (0%)	0 (0%)	0 (0%)	3 (10%)	30
8	0 (0%)	0 (0%)	0 (0%)	25 (83%)	0 (0%)	1 (3%)	0 (0%)	4 (13%)	30
9	1 (3%)	7 (23%)	0 (0%)	13 (43%)	1 (3%)	7 (23%)	0 (0%)	1 (3%)	30
10	3 (12%)	1 (4%)	0 (0%)	19 (76%)	0 (0%)	0 (0%)	0 (0%)	2 (8%)	25
11	0 (0%)	11 (37%)	0 (0%)	15 (50%)	0 (0%)	2 (7%)	0 (0%)	2 (7%)	30
12	1 (3%)	8 (27%)	1 (3%)	18 (60%)	0 (0%)	1 (3%)	0 (0%)	1 (3%)	30
13	0 (0%)	7 (23%)	1 (3%)	17 (57%)	1 (3%)	1 (3%)	1 (3%)	2 (7%)	30
Σ	15 (4%)	64 (17%)	6 (2%)	255 (66%)	10 (3%)	16 (4%)	2 (1%)	17 (4%)	385

TABLE 15: Final Part: Strategy Choices of Subjects in Role 1

Ses	ll	lr	rl	rr	Sum
1	6 (20%)	0 (0%)	2 (7%)	22 (73%)	30
2	0 (0%)	11 (37%)	0 (0%)	19 (63%)	30
3	0 (0%)	12 (40%)	0 (0%)	18 (60%)	30
4	1 (3%)	10 (33%)	1 (3%)	18 (60%)	30
5	0 (0%)	3 (10%)	0 (0%)	27 (90%)	30
6	0 (0%)	4 (13%)	0 (0%)	26 (87%)	30
7	1 (3%)	2 (7%)	1 (3%)	26 (87%)	30
8	0 (0%)	0 (0%)	0 (0%)	30 (100%)	30
9	0 (0%)	10 (33%)	0 (0%)	20 (67%)	30
10	0 (0%)	0 (0%)	0 (0%)	20 (100%)	20
11	0 (0%)	10 (40%)	0 (0%)	15 (60%)	25
12	0 (0%)	8 (27%)	0 (0%)	22 (73%)	30
13	2 (8%)	1 (4%)	0 (0%)	22 (88%)	25
Σ	10 (3%)	71 (19%)	4 (1%)	285 (77%)	370

TABLE 16: Final Part: Strategy Choices of Subjects in Role 2

References

Balkenborg, D. (1994): "An Experiment on Forward- Versus Backward Induction.," SFB Disc. Paper B-268, University of Bonn.

Balkenborg, D., and K. H. Schlag (2006): "On the Evolutionary Selection of Nash Equilibrium Sets in Games," *J. Econ. Theory*, to appear, 000 – 000.

Binmore, K. (1987): "Modelling Rational Players, Part I," *Economics and Philosophy*, 3, 179–214.

——— (1988): "Modelling Rational Players, Part II," *Economics and Philosophy*, 4, 9–55.

Binmore, K., J. McCarthy, G. Ponti, L. Samuelson, and A. Shaked (2002): "A backward induction experiment," *Journal of Economic Theory*, 104(1), 48–88.

Binmore, K., C. Proulx, L. Samuelson, and J. Swierzbinski (1995): "Hard Bargains and Lost Opportunities.," SFB Discussion Paper B-319, University of Bonn.

Binmore, K., and L. Samuelson (1999): "Evolutionary Drift and Equilibrium Selection," *Review of Economic Studies*, 66, 363 – 393.

Bolton, G., and A. Ockenfels (2000): "ERC: A Theory of Equity, Reciprocety and Competition," *American Economic Review*, 90, 166 – 193.

Brandts, J., A. Cabrales, and G. Charness (2003): "Forward Induction and the Excess Capacity Puzzle: An Experimental Investigation," mimeo.

Brandts, J., and C. A. Holt (1992): "Forward Induction: Experimental Evidence from Two-Stage Games with Complete Information.," *Research in Experimental Economics*, 5, 119 – 136.

——— (1993): "Adjustment Patterns and Equilibrium Selection in Experimental Signalling Games.," *International Journal of Game Theory*, 22, 279 – 302.

Brandts, J., and C. A. Holt (1995): "Limitations of Dominance and Forward Induction: Experimental Evidence," *Economic Letters*, 49, 391 – 395.

Cachon, G. P., and C. F. Camerer (1999): "Loss-Avoidance and Forward Induction in Experimental Coordination Games," *The Quarterly Journal of Economics*, 111, 165 – 194.

Camerer, C. F. (2003): *Behavioral Game Theory: Experiments in Strategic Interaction*. Princeton University Press, Princeton, NJ.

Camerer, C. F., and E. Fehr (2006): "When Does "Economic Man" Dominate Social Behavior?," *Scince*, 47 – 52, 311.

Caminati, M., A. Innocenti, and R. Ricciuti (2006): "Drift Effect under Timing Without Observability: Experimental Evidence," *Journal of Economic Behavior and Organization*, forthcoming.

Charness, G., and M. Rabin (2002): "Understanding Social Preferences with Simple Tests," *The Quartely Journal of Economics*, 117, 817 – 869.

Cooper, R., D. V. DeJong, R. Forsythe, and T. W. Ross (1992): "Forward Induction in Coordination Games," *Economic Letters*, 40, 167 – 172.

——— (1993): "Forward Induction in the Battle-of-the-Sexes Game," *American Economic Review*, 83, 1303 – 1316.

Costa-Gomes, M. A., and V. P. Crawford (2006): "Cognition and Behavior in Two-Person Guessing Games: An Experimental Study," *American Economic Review*, 96, 1737 – 1768.

Crawford, V., U. Gneezy, and Y. Rottenstreich (2008): "The Power of Focal Points is Limited: Even Minute Payoff Asymmetry May Yield Large Coordination Failures," *American Economic Review*, 98, in press.

Falk, A., E. Fehr, and U. Fischbacher (2003): "On the Nature of Fair Behavior," *Economic Inquiry*, 41, 20–26.

Fehr, E., and K. Schmidt (1999): "A Theory of Fairness, Competition and Cooperation," *The Quarterly Journal of Economics*, 114, 817 – 868.

Fudenberg, D., and D. K. Levine (1998): *The Theory of Learning in Games*. MIT Press, Cambridge, Massachussets.

Harsanyi, J. C., and R. Selten (1988): *A general theory of equilibrium selection in games*. M.I.T. Press, Cambridge Mass.

Huck, S., and W. Müller (2005): "Burning Money and (Pseudo) First-Mover Advantages: An Experimental Study on Forward Induction," *Games and Economic Behavior*, 51, 109 – 127.

Johnson, E. J., C. F. Camerer, S. Sen, and T. Rymon (2002): "Detecting Failures by Backward Induction: Monitoring Information Search in Sequential Bargaining Experiments," *Journal of Economic Theory*, 104, 16 – 47.

Kagel, J. H., and A. E. Roth (1995): *Handbook of Experimental Economics*. Princeton University Press, Princeton, New Jersey.

Kohlberg, E., and J.-F. Mertens (1986): "On the strategic stability of equilibria.," *Econometrica*, 54, 1003 – 1037.

McKelvey, R. D., and T. R. Palfrey (1995): "Quantal Response Equilibria For Normal Form Games," *Games and Economic Behavior*, 10, 6–38.

——— (1998): "Quantal Response Equilibria for Extensive Form Games," *Experimental Economics*, 1, 9 – 41.

Muller, R. A., and A. Sadanand (2003): "Order of Play, Forward Induction, and Presentation Effects in Two-Person Games," *Experimental Economics*, 6, 5 – 25.

Nagel, R. (1995): "Unraveling in Guessing Games: An Experimental Study," *The American Economic Review*, 85(5), 1313–1326.

Ochs, J. (1995): "Coordination Problems," in *The Handbook of Experimental Economics*, ed. by J. H. Kagel, and A. E. Roth, pp. 195 – 251. Princeton University Press, Princeton, New Jersey.

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

Selten, R. (1967): "Die Strategiemethode zur Erforschung des eingeschränkt rationalen Verhaltens im Rahmen eines Oligopolexperimentes," in *Beiträge zur experimentellen Wirtschaftsforschung*, ed. by H. Sauermann, pp. 136 – 168. Mohr Siebeck, Tübingen.

——— (1975): "Reexamination of the Perfectness Concept for Equilibrium Points in Extensive Games," *International Journal of Game Theory*, 4, 25 – 55.

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

van Damme, E. (1989): "Stable Equilibria and Forward Induction," *Journal of Economic Theory*, 48, 476 – 496.

van Huyck, J. B., R. C. Battalio, and R. O. Beil (1993): "Asset Markets as an Equilibrium Selection Mechanism: Coordination Failure, Game Form Auctions, and Forward Induction.," *Games and Economic Behavior*, 5, 485–504.