Journal of Mathematical Psychology **42**, 356–384 (1998)
Article No. MP981225

Experimental Results on the Centipede Game in Normal Form: An Investigation on Learning

Rosemarie Nagel*

Department of Economics, Universitat Pompeu Fabra, Barcelona, Spain
E-mail: nagel@upf.es

and

Fang Fang Tang

Department of Economics and Statistics, National University of Singapore,
E-mail: ecstff@nus.edu.sg

We analyze behavior of an experiment on the centipede game played in the reduced normal form. In this game two players decide simultaneously when to split a pie which increases over time. The subjects repeat this game 100 times against randomly chosen opponents. We compare several static models and quantitative learning models, among them a quantal response, model reinforcement models and fictitious play. Furthermore, we structure behavior from period to period according to a simple cognitive process, called learning direction theory. We show that there is a significant difference in behavior from period to period whether a player has decided to split the pie before or after the opponent.

INTRODUCTION

In this study, we report experimental results on centipede games in reduced normal form, played repeatedly against changing opponents. The main purpose of this study is to compare different learning models. Originally, the centipede game

* Corresponding author: Rosemarie Nagel, Department of Economics, Universitat Pompeu Fabra, Ramon Trias Fargas 24-27, 08005 Barcelona, Spain, E-mail: nagel@upf.es. We thank Antonio Cabrales, Colin Camerer, Gary Charness, Ido Erev, Nick Feltovitch, Michael Mitzkewitz, Alvin Roth, Abdolkarim Sadrieh, Reinhard Selten, and two anonymous referees for helpful comments, Tibor Neugebauer for providing his computerprogram for this experiment and organizing the experiments. Special thanks also to the Bonn Laboratorium fuer experimentelle Wirtschaftsforschung for their support running the experiments. We thank Serguei Maliar and Nick Feltovitch for providing parts of the simulations and Richard McKelvey for the computations of the quantal response equilibria. Financial support from the German Science Foundation through SFB303 (University of Bonn) to Nagel and Tang and the postdoctoral fellowship of the University of Pittsburgh to Nagel are gratefully acknowledged.

0022-2496/98 $25.00

(Rosenthal, 1981) has been discussed mainly in extensive form in the theoretic and experimental literature. The *game in extensive form* is as follows (see also Fig. 1): two players, called player A and player B, decide sequentially whether to split a given pie in a predetermined way or whether to pass the splitting option to the other player. If a player passes the decision to the other player, the pie increases in size. Passing can be done only a finite number of times. Once a player splits the pie, the game is over with that player gaining the higher share of the pie and the other player obtaining the smaller share. All standard game-theoretic equilibrium concepts predict that the pie is split at the first decision node of player A, the first player.

In the *reduced normal form game*, which is strategically equivalent to the extensive form game, players decide simultaneously at which of his decision nodes to "take" the opportunity to split the pie. This form of decision making has an advantage for studying learning in centipede games. The experimenter is informed about the intended "Take-node" by both players. Thus, in the repeated normal form games we are able to study the change of behavior of a player from round to round in a more precise way than can be done in the extensive form game. Furthermore, this kind of structure allows us to repeat long centipede games much faster than repeated sequential move games. However, while there might be substantial differences in behavior in the extensive form game and in the normal form game, we *do not address* this question here and leave that to a later study. Here we only concentrate on learning in the normal form game. Since after each period the players only obtain information as in the extensive form game, we hope that the structure of behavior we find will be valuable for the understanding of behavior in the extensive form centipede games. The centipede game in extensive form has gained attention in game theory in the discussion on the limitations of common knowledge of rationality and backward induction (see, e.g., Aumann, 1992; Binmore, 1988). Cressman and Schlag (1996) and Ponti (1996) discuss stability criteria of an evolutionary model in the centipede game.

McKelvey and Palfrey (1992), Fey, McKelvey, and Palfrey (1994), Zauner (1996), and McKelvey and Palfrey (1995b) have analyzed actual behavior in the extensive-form centipede game. As in our experiments, behavior is quite different from the game theoretic solution. These authors explain the data in terms of equilibrium models with errors. Learning is discussed as reduction of errors or reduction of uncertainty over time. As a consequence of that kind of learning, behavior should converge to the Nash-equilibrium of the stage game. We apply the quantal response model (McKelvey & Palfrey, 1995a) for our data set and compare it with other models, some static models, a class of simple adaptive models, called reinforcement learning models, and fictitious play. Furthermore, we formulate hypotheses according to a qualitative learning direction model, in the spirit of Selten and Stoecker (1986). Testing these hypotheses we reveal features of period to period behavior that have not been discussed in other studies on the centipede game. There seems to be a clear-cut separation of behavior whether a player has chosen the lower number or has chosen the higher number of the match. Learning is not necessarily related to convergence to equilibrium. All models we consider have been applied by different authors to various experimental games in the economic literature.

The paper is organized as follows. Section 2 presents the game under consideration and Section 3 the experimental design. Section 4 gives a summary statistic of the data. Section 5 discusses quantitative static models and learning models and compares the performance of these models in the light of the data. Section 6 analyses the actual data and the basic reinforcement model with respect to hypotheses of a qualitative learning theory—learning direction theory. Section 7 concludes.

1. THE GAME UNDER CONSIDERATION

Consider the extensive form game displayed in Fig. 1. At each node (numbered from 1 to 12) either player A or player B has to decide whether to "take" or to "pass." Once a player chooses "take," the game is over and the payoffs are determined by the payoff vector at that "Take-node."

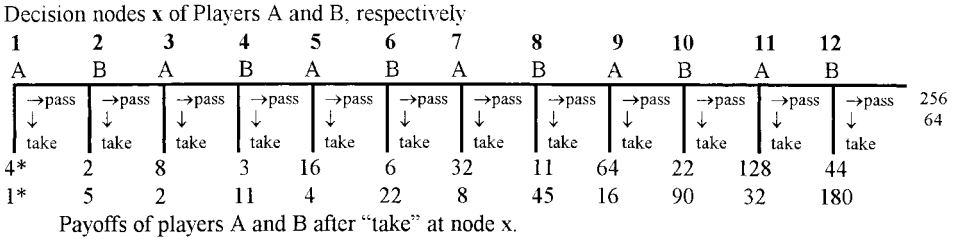

FIG. 1. The centipede game in extensive game form (* = Nash-equilibrium payoffs).

The reduced normal form we study experimentally in this paper is presented in Fig. 2. The main difference between the reduced normal form and the extensive form is that in the normal form game both players have to decide *simultaneously* at which node to take or whether to pass till the end. Therefore within a game, when making a decision, the players do not know whether or not the opponent has passed at early nodes. However, since the players repeat the game, they gain this

		Player B (even numbers)						always "pass"
		2	**4**	**6**	**8**	**10**	**12**	**14**
	1	4 * / 1	4 / 1	4 / 1	4 / 1	4 / 1	4 / 1	4 / 1
Player A	**3**	2 / 5	8 / 2	8 / 2	8 / 2	8 / 2	8 / 2	8 / 2
(odd	**5**	2 / 5	3 / 11	16 / 4	16 / 4	16 / 4	16 / 4	16 / 4
numbers)	**7**	2 / 5	3 / 11	6 / 22	32 / 8	32 / 8	32 / 8	32 / 8
	9	2 / 5	3 / 11	6 / 22	11 / 45	64 / 16	64 / 16	64 / 16
	11	2 / 5	3 / 11	6 / 22	11 / 45	22 / 90	128 / 32	128 / 32
always "pass"	**13**	2 / 5	3 / 11	6 / 22	11 / 45	22 / 90	44 / 180	256 / 64

FIG. 2. Reduced normal form payoff matrix of the centipede game (* = Nash-equilibrium payoffs).

experience from game to game instead within a game. The simultaneity removes any explicit sequential "reciprocity." If a player played repeatedly against the same opponent, there could be reciprocity from period to period.

The strategies α (for the row (A)-player are odd numbers between 1 and 13, $\alpha \in \{1, 3, ..., 13\}$; and the strategies β for the column (B)-player are even numbers between 2 and 14, $\beta \in \{2, 4, ..., 14\}$; numbers 1 to 12 indicate the take node of the extensive form. Strategies 13 and 14 correspond to the strategy "always pass" in the extensive form, resulting in the highest payoff for player A.[1] Since some cells contain the same payoff vectors, players cannot always infer what an opponent has chosen; in the extensive form game, a player who takes before his opponent, does not know opponent's move, except at node 12. We maintain this information of the extensive form game by not telling the subjects what the other player has done if he cannot infer it from the matrix.

Both games are strategically equivalent and have a unique outcome. In the extensive form game, the players have to take at their first opportunity following the backward induction logic. The equilibrium is subgame-perfect. There are also mixed equilibria with the same outcome, player A "takes" the first node and player B mixes between later nodes with a positive probability such that player A has no reason to deviate from his strategy. In the normal form game, the players choose strategy 1 and 2, respectively. There is only one form of elimination process to reach this equilibrium[2]: Strategy 14 is a weakly dominated strategy (see normal form payoffs for player B). Thus, player B will not play 14. If player A believes that, he can delete 13, assuming that player B is rational. Then, player B should eliminate 12 and so on. This process is called iterative elimination of weakly dominated strategies with strategies 1 and 2 being singled out. However, always "passing" by both players or player A always "passing" and player B taking strategy 12 (taking at the last node) are the Pareto optimal strategy combination. If the game is finitely repeated against the same opponent or repeated against changing opponents, the pure one shot equilibrium is maintained and there are no other pure equilibria.

2. EXPERIMENTAL DESIGN

We ran five sessions of the centipede game in reduced normal form with payoffs as shown in Fig. 2. Both the first endnode and the last endnode have the same payoff vector as in the 6-move game in McKelvey and Palfrey (1992). Each player has seven choices; player A chooses an odd number from 1 to 13 and player B an even number from 2 to 14.[3] Each session involved 12 subjects (six for type A and

[1] In the nonreduced normal form game, a player has to decide for each of his node whether to take or to pass. This is according to the definition of a *strategy in the extensive form game*, allowing for an enormously large strategy space. However, all strategies with the same first intended Take node produce the same payoff. Passes at later nodes have no influence on the outcome. McKelvey and Palfrey run pilot studies on the nonreduced normal form. More than 95% of the chosen strategies were monotonic in Take behavior, i.e., after the first intended Take node all other nodes were followed by a Take.

[2] See also Glazer and Rubinstein (1996) who discuss the equivalence of solving normal form and extensive form games.

[3] We also ran five sessions of a game with four actions for each player, with the same payoff structure as in the 6-move game of McKelvey and Palfrey (1992). We will not discuss the results here.

six for type B) from various departments of the University of Bonn. Each subject participated only in one session maintaining the same role throughout the session. The interaction between subjects was via computer terminals. After the assistant had read the instructions aloud (see Appendix), subjects were randomly allocated to one of the two types (type A or type B). Subjects were informed that the same one-shot game was repeated for 100 periods and in each period, a player of one type was randomly matched with a player from the other type without identifying the players to each other. Instead of displaying the normal form, we presented and explained a table as shown in the instructions (see Appendix): The lower number (column 1) of a matched pair determines the payoff for each player, which is shown in column 3 for player A and column 4 for player B. Since 14 can never be the lower node, it is not shown in the table. Column 2 presents the possible choices of the opponent in case he chooses the higher number. For example, if the lower number is 9, player A must have chosen it, and player B has chosen either 10, 12, or 14. This way we reduce the normal form to the diagonal cells and one cell of each row below the diagonal. In the last column the total sum of the pie is stated. We additionally mentioned that the pie increases by about 40% with each increasing lower number. The person who chooses the lower number of the match receives about 80% of the pie.

During a session, columns 1, 3, and 4 of the table in the instructions were always displayed on the screen. Player A had seven colored buttons numbered 1, 3,..., 13 to choose from by clicking one button with the mouse or pressing the number on the keyboard. The choices from player B were also displayed as gray buttons, with no consequence if pressed by a mouse button. Player B screen was similar with numbers 2, 4,..., 14 as colored buttons.

After each round, each player is informed of the lower number of his match (this he can also infer from the payoffs), the corresponding payoffs to both players, and his own cumulative payoff. Thus, the player who chooses the lower number is *not* informed about the choice of the opponent, as in the centipede game in extensive form. During the session their own complete history was displayed if requested by a mouse click.[4]

At the end of a session the total points were converted into DM (2 points = 0.01 DM), and paid privately to each subject. Each session lasted about 1 to 1 1/2 h. Average payoffs were 17.70 DM (about $12.60). Note that the sessions by M/P (1992) who repeated each six-move game 10 periods took also 1 h.

3. POOLED RESULTS

Many patterns of the behavior found in the extensive form game study by McKelvey and Palfrey (M/P, 1992) and repeated in Zauner (1996) can also be recognized in our study. Table I shows the relative frequencies of choices of players A and B for each session.

[4] Subjects examined their history about 6% of the time, both in the first 50 periods and the last 50 periods.

TABLE I

Relative Frequencies of Players A and B Choices Pooled over All Periods

Player A		Session 1	Session 2	Session 3	Session 4	Session 5	All
Choices	1	0.005	0.002	0.008	0.000	0.008	0.005
	3	0.005	0.008	0.035	0.000	0.030	0.016
	5	0.053	0.022	0.107	0.003	0.083	0.054
	7	0.325	0.073	0.347	0.117	0.445	0.261
	9	0.258	0.333	0.342	0.427	0.297	0.331
	11	0.192	0.293	0.130	0.380	0.130	0.225
	13	0.162	0.268	0.032	0.073	0.007	0.108
Player B							
Choices	2	0.022	0.003	0.005	0.003	0.010	0.009
	4	0.027	0.008	0.012	0.002	0.035	0.017
	6	0.140	0.053	0.132	0.018	0.220	0.113
	8	0.313	0.118	0.467	0.317	0.438	0.331
	10	0.312	0.382	0.282	0.355	0.225	0.311
	12	0.143	0.217	0.095	0.215	0.047	0.143
	14	0.043	0.218	0.008	0.090	0.025	0.077

• All strategies, but 1 and 3 in session 4, have strictly positive relative frequencies.

• The weakly dominated choice 14 is selected with positive probability (7.7% across all sessions; most of it is accounted for by two players in session 2, who decided after round 12 and round 44, respectively, to always choose 14). The estimated frequency of altruists (player who always pass) in M/P (1992) is 5%.[5]

• Choice 1 is chosen 0.5% of time in our game, compared with 0.7% in M/P-6-move games and 7% in 4-move games. Thus, the longer a centipede game is, the fewer equilibrium strategies are chosen, although this is not significant between the 6-move and our 14-move game. There is a significant difference on a 2% level between our game and the 4-move game of M/P, using Mann–Whitney–U-test.

• Modal behavior is concentrated at the middle choice 7 and choice 9 for players A, or middle choice 8 and choice 10 for players B. This is similar to the coordination games where middle choices are most frequent (see van Huyck *et al.*, 1992). In the first round the highest frequency is at 9 and 10 for the two player types.

• There is no clear trend at the sessional level. Also in M/P (1992) convergence to equilibrium is questionable (see Fig. 3 which shows the average take node in

[5] McKelvey and Palfrey (1992) explain this low estimation of altruists according to their theoretical model; the reason is that part of passing at the last move is attributed to errors in action and not to altruism.

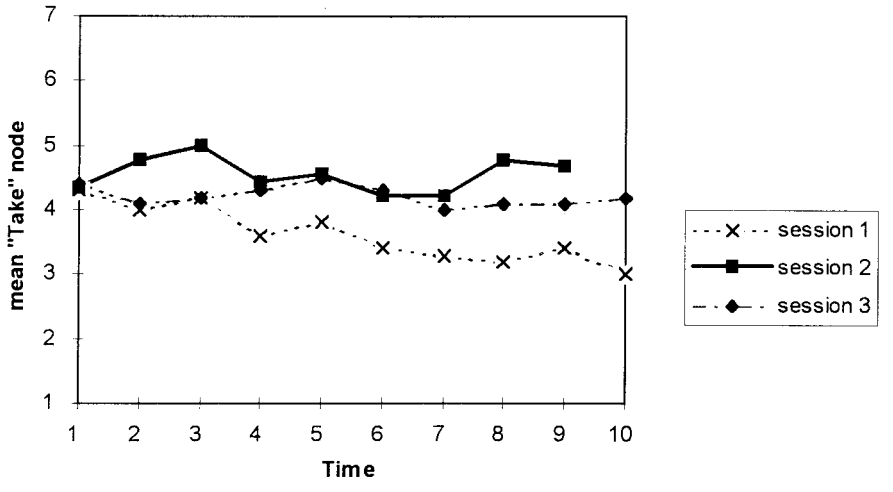

FIG. 3. Mean behavior over time in the extensive form centipede games (6-move games). In session 2 there were only 9 periods played with 18 subjects. Data source: McKelvey and Palfrey (1992).

each period, separately for each session of M/P). Note, that these authors interpret their results as an indication of convergence to equilibrium, which might be due to the fact that they aggregate the data.

Figure 4 shows for each session the average lower choice in 10 period blocks (average lower choice across all six pairs over 10 periods) over time. One can see that only in session 4 there is a slight downward trend. In neither session is there a clear movement toward the Nash equilibrium. One reason might be that there are only six players on each side, supergame effects, and thus cooperation cannot be completely excluded. However, in McKelvey and Palfrey (1992) and Nagel and

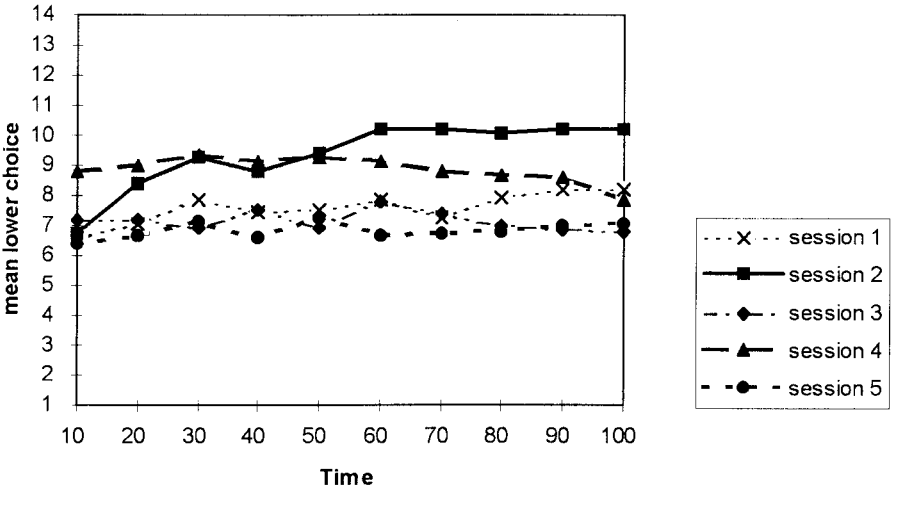

FIG. 4. Mean behavior across all pairs pooled over 10 periods, separately for each session of the normal form centipede experiment.

Sadrieh (1998) players of one type met an opponent only once during 9 or 10 periods, and there is no significant downward movement either, based on single session results. Thus, we see no reason to presume that our results are caused entirely by supergame effects. Note also that the question whether or not there is convergence to equilibrium is not the main purpose of this study. In the next sections we study whether we can detect learning patterns over time and compare different learning models.

4. LEARNING MODELS

Learning is usually defined as a systematic change of behavior according to past information. In this section we discuss several learning models that have been applied to various game experiments. We follow the quantitative methods used in Tang (1996), who compared 18 different quantitative learning models for 3×3 normalform games, and Chen and Tang (1998). One of the several kinds of models we consider is a basic reinforcement model which has been introduced into experimental economics by Roth and Erev (1995).[6] It has been refined for example by Erev and Roth (1998), Stahl (1996), and generalized by Camerer and Ho (1997). The later model contains fictitious play and reinforcement models as special cases. In Section 6 we try to give some reasons why the extensions of this model cannot be directly be applied to our dataset. We discuss three basic variants of reinforcement models which have performed "best" in Tang (1996) and compare them with the performance of some static models and the generalized fictitious play model. We cannot directly compare the learning direction theory to these models because we leave it in its qualitative form. Therefore, we discuss this theory in Section 6.

To detect the best performing model given the experimental data, we calculate the quadratic deviation measure (QDM) for each proposed model. This measure takes the quadratic difference between the *actual choice vector of a subject in a period* and the *choice vector predicted by the model*. This measure is a proper scoring rule; i.e., it does not give the forecaster any incentive to ignore the results of the measure or to influence the results (see Brier (1950) and Yates (1990); also see Selten (1997) for an axiomatic analysis).[7]

Let $\mathbf{c}_A(t) = (c_{A1}, c_{A3}, ..., c_{A13})$ be the choice vector of player A who chooses odd numbers from 1 to 13 and $\mathbf{c}_B(c_{B2}, c_{B4}, ..., c_{B14})$ the choice vector of player B who chooses from even numbers from 2 to 14 at round t with

[6] In many papers in which the reinforcement model is applied the authors do not compare different learning models. Instead they want to see whether this basic model predicts some features of mean behavior over time.

[7] The more common method for comparison of models in experimental economics is the maximum likelihood estimation (MLE). The relationship between QDM and MLE is as between nonparametric vs parametric statistic. For MLE one usually assumes a distribution underlying the data and thus the parameters are estimated and have a certain variance. Not so for QDM where the parameters are usually calculated by the minimum distance of the actual data and the prediction by a model. Hence such a parameter value does not have a variance. Unfortunatly, so far there is no discussion at hand about the comparison of the two methods.

$$c_{A\alpha}(t) = \begin{cases} 1, & \text{if strategy } \alpha \text{ is chosen in round } t \\ 0, & \text{otherwise} \end{cases};$$

$$c_{B\beta}(t) = \begin{cases} 1, & \text{if strategy } \beta \text{ is chosen in round } t \\ 0, & \text{otherwise.} \end{cases}$$

Let $\mathbf{p}_A(t) = (p_{A1}, p_{A3}, ..., p_{A13})$, similarly for B, denote the predicted choice probability vector of a particular model for player A at round t. Then the quadratic deviation for subject A in round t is

$$QDM_A(t) = \sum_{\alpha=1}^{13} [c_{A\alpha}(t) - p_{A\alpha}(t)]^2, \qquad \forall A \in \{1, 2, ..., 6\}, \quad \alpha \in \{1, 3, ..., 13\}, \qquad (1)$$

similarly for $B \in \{7, 8, ..., 12\}$, $\beta \in \{2, 4, ..., 14\}$. $QDM(t)$ of a player is between 0 and 2, since the choice of a player can coincide completely with the strategy predicted by a model or at worst if the model predicts that a strategy has to be chosen with probability 1 and the actual choice of a player does not coincide with it, the deviation reaches the maximum 2. The average quadratic deviation measure per player and period for an entire session is

$$\text{QDM} = \text{QDM}_A + \text{QDM}_B = \left[\sum_{A} \sum_{t=1}^{100} \text{QDM}_A(t) + \sum_{B} \sum_{t=1}^{100} \text{QDM}_B(t) \right] \Big/ 12*100. \qquad (2)$$

Clearly, the smaller the QDM score of a model, the better is its prediction.

4.1. Four Static Benchmark Models

The following four static models are strictly speaking not learning models, but may serve as benchmark models to judge the relative performance of the dynamic learning models. The static models are the equilibrium model, a quantal response model, the uniform random model, and the individual-observed frequency model.

To analyze behavior in economic experiments, we usually start with the question how well the game theoretic solution describes actual behavior. Here we use the stage game subgame-perfect equilibrium, given that the players repeatedly play the same game against changing opponents in finite times. To calculate the equilibrium the entire structure of the game has to be considered.

The calculation of the QDM for the *subgame-perfect equilibrium model* per session and period is straightforward:

$$p_{A\alpha}(t) = \begin{cases} 1, & \text{if } \alpha = 1 \\ 0, & \text{otherwise,} \end{cases} \qquad p_{B\beta}(t) = \begin{cases} 1, & \text{if } \beta = 2 \\ 0, & \text{otherwise;} \end{cases}$$

therefore, $\text{QDM} = (200 - 100*(f_1 + f_2))/100$, where f_1 and f_2 are the relative frequencies of equilibrium strategies 1 and 2 within a session. Note that the value of $\text{QDM}(t)$ for a player is either 0 or 2.

Another equilibrium model is the quantal response model (McKelvey and Palfrey (M/P, 1995a)). It assumes that players do not play the equilibrium strategies of the

original game, but make mistakes which can be interpreted as calculation errors of expected payoffs. Given the error rate which is common knowledge, an equilibrium is calculated. For any given $\lambda \geqslant 0$, the logit quantal response function is used, which is also known in the study of individual choice behavior (Luce, 1959):

$$p_{A\alpha}(t) = \frac{e^{\lambda \pi_{A\alpha}(t)}}{\sum_{\kappa=1}^{13} e^{\lambda \pi_{A\kappa}(t)}} \qquad \forall \text{A, } \alpha \text{, and similarly for B,} \tag{3}$$

where λ is a parameter inversely related to the error rate. For $\lambda = 0$ the uniform random distribution results and for $\lambda \to \infty$, the equilibrium of the original game is approached. $\pi_{A\alpha}$ is the expected payoff for strategy α given the probability distribution of strategies β. M/P (1995b) examine the dynamics of behavior with the quantal response model by calculating λ-values for period blocks 1–10, 11–20, etc., and check the evolution of the values. We also did this and did not find that the λ parameters increase over time.

The next useful benchmark model is the *uniform random model*, which assumes that a player chooses each of his seven strategies with the same probability in every round. Any selected or preferred model should certainly perform better than this model since it does not rely on the structure of the underlying game or situation. The vector of predicted choice probability is $\mathbf{p}_A(t) = (\frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7}, \frac{1}{7})$ $\forall$A $\forall t$ and similarly for B. Hence, the quadratic deviation measure for a session is

$$QDM = \left\{ \sum_{A} \sum_{t=1}^{100} \sum_{\alpha=1}^{13} [c_{A\alpha}(t) - \tfrac{1}{7}]^2 + \sum_{B} \sum_{t=1}^{100} \sum_{\beta=2}^{14} [c_{B\beta}(t) - \tfrac{1}{7}]^2 \right\} \Big/ 12*100$$
$$= [(1 - \tfrac{1}{7})^2 + 6(0 - \tfrac{1}{7})^2] = 0.86. \tag{4}$$

The fourth static model, the *individual-observed frequency* (*mean*) *model*, is based on the actual frequency distribution of each player. This means that we compare a player's behavior in a period with the "prediction" given by his relative frequency vector over all periods. The observed frequency distribution minimizes the QDM for any subject and therefore is the best static challenge for a learning model. Since it is constructed ex post and has six free parameters for each subject, we suppose that a *QDM* of another model with only one or two parameters that comes near this measure is better. The frequency distribution of an individual is given by $\mathbf{p}_A(t) = (f_{A1}, f_{A,3}, ..., f_{A13})$, $\forall A$, similarly for B, where $f_{A\alpha} = \sum_{t=1}^{100} c_{A\alpha}(t)/100$ is the actual relative frequency that subject A chooses strategy α.

4.2. *Reinforcement Models*

The next kind of models we study belong to a class of stochastic dynamic models which consider the payoffs received in a period. They are probably the most basic learning models, first developed in the psychological literature (see Hull, 1943; Bush and Mosteller, 1955).[8] Cross (1973 and 1983) introduced this so-called *reinforcement model* into the economic literature; however, the major influence on experimental

[8] Reinforcement learning has similarities with machine learning (see Barto *et al.*, 1983; Sutton, 1992) or as classifier system (Holland, 1975; Holland *et al.* 1986).

economics was introduced by the papers by Arthur (1991) and Roth and Erev (1995). The family of models they use has met with success in predicting behavior in some experiments, despite (or perhaps because of) the low level of rationality they attribute to individuals.[9] The basic idea is that over time, players play better strategies (strategies leading to higher realized payoffs) more often, and worse strategies less often. We will evaluate three basic reinforcement models without using the extensions made by various authors. In Section 7 we will give a justification for this.

The first dynamic model is the linear form *relative-payoff-sum model* (RPS). Define $M_{A\alpha}(t)$ and $M_{B\beta}(t)$ as the discounted payoff sum or propensity of player A and B, respectively,

$$M_{A\alpha}(t) = qM_{A\alpha}(t-1) + c_{A\alpha}(t)\,\pi_{A\alpha}(t), \qquad \alpha \in \{1, 3, ..., 13\} \text{ and similarly for } B, \quad (5)$$

where $q \in [0, 1]$ is the time/memory discount factor or forgetting parameter. Note, when $q = 0$, the model degenerates to the *repeat last choice model*. $\pi_{A\alpha}$ is the payoff for strategy α, given the strategy of the opponent in that period. In the most rudimentary case $q = 1$, if strategy α is chosen in period t, the payoff sum increases, whereas the payoff sum stays put for strategies that are not chosen in period t.[10] The predicted probability for player A at round $t+1$ is

$$p_{A\alpha}(t+1) = \frac{M_{A\alpha}(t)}{\sum_{k=1}^{13} M_{Ak}(t)}; \qquad \text{similarly for player B.} \quad (6)$$

In addition to estimating the parameter q from the data, we also have to initialize the discounted payoff sum of the first round. One restriction is that all strategies have to be chosen with positive probability in the first round if they should be in the choice-set of a player in the future. We choose to initialize the first round probabilities to uniform distribution over all choices; the initial values of the propensities are $M_{A\alpha}(0) = 50$ and $M_{B\beta}(0) = 50$ which minimize the *QDM*. (We have tried various combinations of initial propensity values selected from 1 to 500; the differences are negligible. It seems that the effects of initial propensity values, if not selected to be too extreme, tend to vanish in 100 runs. This also holds for the two following models).

We also report on the results of a variant of the basic RPS model, the *power-reinforcement model* which was independently developed by Tang (1995, 1996) and Chen, Friedman, and Thisse (1996). The difference to the former model is that it uses the power transformation of the basic RPS model with the predicted probability of choosing a strategy by player A,

$$p_{A\alpha}(t+1) = \frac{[M_{A\alpha}(t)]^r}{\sum_{k=1}^{13} [M_{Ak}(t)]^r} \qquad \forall A, \alpha; \quad \text{similarly for player B,} \quad (7)$$

[9] This model does particular well in mixed motive games (see, for example, Tang, 1996; and Erev and Roth, 1998). Yan and Tang (1998) show that it is better than fictitious play in public good experiments. Camerer and Ho (1997) show that reinforcement does worse than their more general model for data on different normal form games.

[10] See Chen and Tang (1998) and Erev and Roth (1998) for adjustment of the model if negative payoffs are involved in a game.

where r is a nonnegative constant. When $r=1$, the probabilities are as in the RPS model. When $r=0$, the model degenerates to the uniform random model. When $r>1$, the relative weight of the discounted payoff sum is scaled up and when $r<1$ it is scaled down.

The third reinforcement model we consider is the *exponentialized-RPS model* with the predicted probability

$$p_{A\alpha}(t+1)=\frac{e^{\lambda M_{A\alpha}(t)}}{\sum_{k=1}^{13} e^{\lambda M_{Ak}(t)}} \qquad \forall A, \alpha, \tag{8}$$

where $\lambda \geqslant 0$ can be interpreted in a similar way as the power parameter r; when $\lambda=0$, the uniform random model results.

Chen and Tang (1998) mention that an advantage of this functional form is that negative payoffs can be treated the same as positive payoffs. In our case (with non-negative payoffs) the model performs worse than the other two reinforcement models. This model has also been applied by Camerer and Ho (1997), Mookherjee and Sopher (1996), and Weisbuch, Kirman, and Herreiner (1996). It can be interpreted as a dynamic version of the quantal response model of McKelvey and Palfrey (1995a). However, it is not an equilibrium model, since the probability $p_{A\alpha}(t)$ does not depend on the probability distribution of the opponent and no fixed point is calculated.

For the RPS model, the optimal discount factor q minimizing QDM was searched at the grid size 0.05 which is fine enough; the power parameter r for the power-RPS model and λ were searched at a grid size of 0.001 for the static quantal response model and the exp.-RPS.

4.3. *Generalized Fictitious Play*

Fictitious play (Brown, 1951; Robinson, 1951) is a dynamic learning model or algorithm that predicts convergence in the centipede game towards the equilibrium within some finite number of rounds. The reason is straightforward; in each period a best response to the frequency distribution of the entire past is calculated. Therefore the highest choice can never be best response; 14 will not be selected, except maybe in the initial rounds where each choice has a positive probability. Hence, strategy 14 will sooner or later die out. Then 13 will not be best response anymore and so on. Slowly, but surely the choices will become lower and lower over time. We already know that the data does not show this characteristic. However, since it is a learning model with a long tradition in economic theory we do not want to ignore it. Another reason for its application is that it contains the Cournot model as a special case. This is also an important benchmark learning model which states that in each period a player gives best response to the opponent's choice of the previous period.

According to Brown (1951) and Robinson (1951), in each period, a player makes a best response to the frequency distribution of the opponent(s) he observed in the entire past. Due to the random matching scheme we used, we will not follow this approach but instead calculate player's best response to the frequency distribution

of the entire group of his opponents. Note that our players do not get this kind of information. This learning model is a population learning model.

The best response $x_A(t)$ for player A, given the probability vector of players B, $\mathbf{p}_B(t) = (p_{B1}(t), p_{B3}(t), ..., p_{B13}(t))$, is

$$x_A(t) = \left\{ m+1 \text{ for } \max \left\{ 4 \text{ if } m=0, \sum_{\beta=2}^{m} p_{B\beta}(t)\, \pi_{A\beta} + \left(1 - \sum_{\beta=2}^{m} p_{B\beta}(t)\right) \pi_{Am+1} \right\}, \right. \tag{9}$$

for $m = \{0, 2, ..., 14\}$, $\beta = \{2, 4, ..., 14\}$. $\pi_{\alpha x}$ is the payoff for player A, where $x \in \{\beta, m+1\}$ is the lower choice of a match. The dynamics of the decision process is determined by a retrospective learning rule. For some discount factor, δ, we assume that players A predict the probability of the players B strategies, $\mathbf{p}_B(t+1)$ according to

$$p_{B\beta}(t+1) = \frac{g_{B\beta}(t+1)}{\sum_{b=2}^{14} g_{Bb}(t+1)}$$

$$\text{with} \quad g_{B\beta}(t+1) = \begin{cases} \dfrac{\delta g_{B\beta}(t-1)}{1 + \sum_{u=1}^{t-1} \delta^u}, & \text{for} \quad \beta \neq x_B(t) \\ \dfrac{\delta g_{B\beta}(t-1) + 1}{1 + \sum_{u=1}^{t-1} \delta^u}, & \text{for} \quad \beta = x_B(t), \end{cases} \tag{10}$$

where $g_{B\beta}(t)$ is the frequency of strategy β for the entire past. This means that in period $t+1$ a player A updates the frequency vector of player B by adding 1 to that choice of the vector of the entire past that has been best response for B in period t. The other frequencies remain the same.

If there are several best responses, their frequencies are updated with equal weights. When $\delta = 0$, the Cournot model results, choosing best response strategy x_A, x_B, to the opponent's last choice. When $\delta = 1$, we obtain the general fictitious play model. With $0 < \delta < 1$, more recent observations influence the present choice with higher weights than the distant past.

As for the dynamic models mentioned above, the initial frequency distribution is the uniform random distribution. The discount factor, $\delta \in [0, 1]$, was searched at a grid size of 0.01. In the next section we describe the performance of the different learning models given our data set.

4.4. Performance of the Suggested Learning Models

Table II states the average *QDM* scores pooled over all sessions for each model, together with the initial values and the calculated parameter values; the first four rows contain the information about the static models, the next three rows, about the reinforcement models (individualistic models), and the last row contains the population learning model. Table III summarizes the results of *QDMs* for the single

TABLE II

Average Quadratic Distance Measures of the Learning Models, Parameter Values, and Initial Frequencies

Type	Model	Initial value	Parameters	Parameter values	QDM (avg.)
Static models	Equilibrium	Stage game equilibrium	—	—	1.99
	Quantal response	Individual frequency distribution	λ	0.1745	0.79
	Random	(1/7, ..., 1/7)	—	—	0.86
	Mean	Individual frequency distribution			0.54
Individualistic models	RPS	50*1/7	q	0.90	0.57
	Power-RPS	50*1/7	(r, q)	(0.58, 0.8)	0.56
	Exp.-RPS	50*1/7	(λ, q)	(0.018, 0.85)	0.60
Population model	Generalized fictitious play	(1/7, ..., 1/7)	δ	0.95	1.32

sessions. The *QDM*-score is between 0 and 2. The lower score is obtained if the prediction coincides perfectly with the observation. The upper score results if the model predicts a pure strategy which is not chosen at all.

Not surprisingly, the equilibrium model performs worst in explaining the behavior of the players. This is easily seen from the frequency distribution of choices given in Table I. Both equilibrium strategies, 1 or 2, are rarely selected. The random model does not explain the data well either. The static quantal response version is slightly better. However, the null-hypothesis that it is indistinguishable from the random model can

TABLE III

Quadratic Distance Measures of Each Learning Model in the Single Sessions

	Model	Session 1	Session 2	Session 3	Session 4	Session 5	QDM avg.
Static models	Equilibrium	1.97	2.00	1.99	2.00	1.99	1.99
	Quantal response	0.82	0.86	0.74	0.72	0.78	0.79
	Random	0.86	0.86	0.86	0.86	0.86	0.86
	Mean	0.54	0.47	0.58	0.50	0.58	0.54
Individualistic models	RPS	0.58	0.48	0.64	0.56	0.61	0.57
	Power-RPS	0.57	0.47	0.63	0.55	0.58	0.56
	Exp.-RPS	0.59	0.52	0.66	0.60	0.63	0.60
Population model	Generalized fictitious play	1.51	1.45	1.17	1.29	1.19	1.32

be rejected only on a 10% level with any conventional test. If we calculate a different parameter λ for each session then the QDMs of the quantal response model is always lower than the random model. The worst dynamic model is the fictitious play model; one reason might be that it is a deterministic model and thus the QDM(t) of a player is either 0 or 2.[11] However, note also that this model predicts that choices converge steadily towards the equilibrium over time, contrary to the evolution over time shown in Fig. 4. Since δ is far away from 0 it is clear that the Cournot model would predict the behavior more poorly.

The best performing models are the Individual-Observed Frequency (Mean) model and the power-RPS. The null hypothesis that there is no difference cannot be rejected on a 5% level, when applying the permutation test.[12] Both models are significantly better than the basic RPS and the exponential RPS-model on a 5% level under the permutation test. Note, however, that the Mean model is not a prediction model, since it takes the aggregated data over all periods as a "prediction." It serves just as an extreme benchmark.

In the following we will analyze the basic RPS-model in terms of a simple cognitive process. This might help to improve this simple model for explaining the centipede data or other similar games with a similar form of information given to the subjects.

5. A SIMPLE COGNITIVE PROCESS—LEARNING DIRECTION THEORY

The following analysis of actual behavior and the reinforcement model are guided by a simple qualitative theory, called learning direction theory, which has been applied for describing behavior for many games (see, for example, Selten, Stoecker, 1986; Nagel, 1995; Selten and Buchta, 1998; Cachon and Camerer, 1996). Stahl (1996) incorporated elements of this theory in his reinforcement model.

So far we have shown that a simple updating rule, the reinforcement model, explains the behavior better than the equilibrium model, the random model, and fictitious play. The updating in the reinforcement model relies entirely on the payoff streams realized. No other detail of the game is necessary. The following analysis of the reinforcement model will show that on a different dimension (separating take and pass by a player) the model does not perform as well as when take and pass situation are combined. This additional analysis might help to improve the performance of RPS models even further.

Bounded rational agents might be more clever and use more information about the environment than supposed by that model. In particular they might try to *search*

[11] Chen and Tang (1998) correct for this bias with two different deviation measures of all models. However, they find that the performance ranking still holds.

[12] For permutation test (also know as the Fisher randomization test), is a nonparametric version of a difference of two means t-test (see Siegel and Castellan, 1988, pp. 151–155). It compares the difference between the means of two independent small samples (in our case with 5 for each of two learning model). The sum of the 5 means of the QDMs of one model is calculated. The probability that this model has a smaller mean than the other sample is just the frequency of all possible sums of any 5 out of the 10 means smaller or equal the actual sample sum divided by all possible permutations. This test uses all of the information in the sample, and thus has a power-efficiency of 100%. It is among the most powerful of all statistical tests.

for better strategies given the information in a period. This is the basic element of a simple theory which has been called "learning direction theory" developed in Selten and Stoecker (1986) in experiments on repeated prisoners' dilemma supergames. This kind of reasoning can be incorporated into quantitative models as done by Stahl (1996) for a beauty contest game and Camerer and Ho (1997) for other normal form games. Here, we want to look at the hypotheses in isolation in order to get a full picture of the reasoning process. Furthermore, we will explain why the model used by Camerer and Ho (1997) reduces to the basic reinforcement model when applied to our experiment.

The basic idea of the simple cognitive reasoning process is the following: Observing his payoff in the previous period and obtaining all additional information according to the rules of the game, a player considers in an ex-post reasoning process whether he could have improved his payoff by a different strategy, x; i.e., was there a better strategy $x < x_{t-1}$ or $x > x_{t-1}$, given the behavior of the opponent(s)? If a player intends to change his strategy he should change it in the "right" direction. Since other influences might also guide the behavior, one might only expect a weak conformance to directional learning; that is, in case of a change, more choices will go in the right direction than in the wrong direction (see also Selten & Buchta, 1998). For this kind of reasoning, the structure of the payoff function has to be known to the player. However, the player does not need to know where exactly the optimum is. It is supposed that the player only uses the information of the previous period.

In the centipede game each player is informed after a match whether or not he has chosen the lower number; in extensive form language he knows whether he took earlier ("take") or later ("pass") than the opponent.

After "take" he knows that a lower choice would have produced a lower payoff. Since he is not informed about the opponent's choice, he does not know whether he has already chosen the optimal number or whether higher payoffs would have been possible by a higher choice. He can only make a guess. Of course, after a choice 12 and 13, the players A and B, respectively, have no uncertainty, and 14 can never be the lower choice. After lower choice 1 or 2 a player cannot decrease his choice. We formulate the following hypothesis[13]:

Hypothesis (1). After "take" decreases of choices are less likely than increases (excluding observations after choices 1, 2, 12, and 13): $p(\text{increase} \mid \text{take}) > p(\text{decrease} \mid \text{take})$.

After "pass" he knows exactly what would have been best: his opponent's choice minus 1 (except after choice 2 for player B). Higher choices would have produced the same payoff.

Hypothesis (2). After "pass" increases of choices are less likely than decreases, excluding choices 1, 2, 13, and 14: $p(\text{decrease} \mid \text{pass}) > p(\text{increase} \mid \text{pass})$.

[13] Nagel (1994) formulated the following hypotheses for the analysis of the extensive form data by McKelvey and Palfrey (1992). They are similar as in Selten and Stoecker (1996), Mitzkewitz and Nagel (1993), Selten and Buchta (1998), Nagel and Vriend (1997), and others.

These two formulations can also be found for the analysis of behavior in repeated prisoner's dilemma supergames by Selten and Stoecker (1986, p. 54). Take is replaced by "player deviates sooner than his opponent after a string of cooperation" and pass is replaced by "player intended to deviate later than his opponent."

We also test the hypothesis whether decreases are more likely after "pass" than after "take" and increases are more likely after "take" than after "pass":

Hypothesis 3. $p(\text{increase} \mid \text{take}) > p(\text{increase} \mid \text{pass})$.

Hypothesis 4. $p(\text{decrease} \mid \text{pass}) > p(\text{decrease} \mid \text{take})$.

In the next section we analyze the data with respect to the proposed hypotheses and also analyze simulations of the reinforcement model in the light of the hypotheses.

5.1. Hypotheses Testing of the Actual Data and the Reinforcement Model

In table IV, we show two transition matrices resulting from the actual data pooled over all periods and five sessions. The first transition matrix contains the relative transition frequencies after a choice was the lower number in a match (previous period was "take") and the second transition matrix shows those frequencies after a choice was the higher number in a match (previous period was "pass"). For example, the main diagonal presents the relative frequencies of transition behavior when choices are the same in period t and $t+1$ (unchanged behavior). Cell a_{ij} gives the relative transition frequencies from choice i in period t to choice j in period $t+1$, and i and j either odd or even. Recall that player A chooses odd numbers and player B chooses even numbers and therefore there are never entries from odd to even numbers or the reverse. The last column states the number of observations of choices i in period t.

In order to check whether an individual player agrees with the hypotheses we calculate the frequencies of increases, unchanged, and decreases of choices, separately after a "take" and a "pass" in a match. This means that we aggregate the frequencies above the main diagonal, on the diagonal, and below the diagonal, respectively, for each player, separating transition frequencies after "take" and "pass."

Unchanged behavior in $t+1$ is most likely after a choice greater than 6 in period t. We come back to this point when we look at transition behavior in the beginning versus the end of a session. At low choices increases are more likely than unchanged behavior. Zauner (1996) predicts that after about 80 periods behavior will converge to equilibrium. However, if increases in $t+1$ are always very high after low choices in t, then his prediction is questionable. Selten and Stoecker (1986), who study the end-effect behavior (i.e., defecting after a string of cooperation) in the repeated prisoners dilemma supergame, wonder about the following: "Even if it is clear from the data that there is a tendency of the end-effect to shift to earlier periods, it is not clear whether in a much longer sequence of supergames this trend would continue until finally cooperation is completely eliminated [or whether] ... the intended deviation period ... finally converges to a stable limit." The end-effect here is just the choice of a player. If the end-effect shifts to the first period (choice 1), then in the

TABLE IV

Transition Matrices Separating "Take" and "Pass"

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total observ
Choice in period $t+1$ after "Take" (own choice in t was lower choice of match)																
c	1	*0.07*		0.29		0.21		0.07		0.21		0.07		0.07		14
h	2		*0.28*		0.08		0.32		0.08		0.12		0.04		0.08	25
o	3	0.04		*0.09*		0.44		0.13		0.18		0.09		0.02		45
i	4		0.11		*0.11*		0.40		0.15		0.15		0.06		0.02	47
c	5	0.01		0.06		*0.20*		0.47		0.15		0.08		0.03		156
e	6				0.05		*0.32*		0.41		0.14		0.06		0.01	296
	7			0.01		0.04		*0.60*		0.28		0.07				617
i	8				0.01		0.05		*0.56*		0.36		0.02		0.01	594
n	9					0.01		0.08		*0.62*		0.26		0.03		545
	10						0.01		0.12		*0.73*		0.14		0.01	353
t	11									0.17		*0.60*		0.23		173
	12						0.03		0.05		0.07		*0.83*		0.02	59
	13											0.09		*0.91*		46
	14							not possible								2970
Choice in period $t+1$ after "Pass" (own choice in t was higher choice of match)																
c	1							Not possible								
h	2															0
o	3	1.00														1
i	4				*0.50*						0.50					2
c	5			0.60		*0.20*				0.20						5
e	6		0.08		0.23		*0.15*		0.33		0.18		0.03			39
	7	0.01		0.06		0.25		*0.48*		0.10		0.06		0.04		156
i	8		0.01		0.04		0.29		*0.49*		0.15		0.04		0.01	388
n	9			0.01		0.04		0.33		*0.48*		0.11		0.02		446
	10		0.01		0.01		0.08		0.40		*0.40*		0.06		0.03	572
t	11	0.01				0.02		0.10		0.31		*0.43*		0.12		490
	12				0.01		0.03		0.10		0.21		*0.54*		0.11	364
	13					0.01		0.05		0.10		0.34		*0.50*		276
	14								0.06		0.10		0.19		*0.65*	231
																2970

Note. Each entry is a relative frequency of transition behavior using the actual data. The underlined entries are the *modal off diagonal frequencies.* Italic cells indicate the frequencies of unchanged behaviour.

next centipede game, the player most likely does not stay there. This might support the hypothesis that in prisoner's dilemma supergames complete elimination of cooperation is also unlikely. Instead it suggests that there is some stable limit of cooperation as shown in a mathematical analysis by Selten and Stoecker.

Note also that unchanged behavior is more likely after "take" than after "pass." This has also been observed in ultimatum games where we distinguish between acceptance and rejection of offers (see Mitzkewitz & Nagel, 1993). The reason might be that the information is different after the two conditions "take" or "pass"

in the centipede game or "acceptance" or "rejection." After a "pass" it is clear that a decrease of the choice would have been better. In ultimatum games after "rejection" it is clear that an increase of an offer would have increased the likelihood of acceptance. This is not true after a "take" or an "acceptance of an offer." In these cases one knows for sure that decreases of take nodes or increases of offers should not be done and the going in the opposite direction only *might improve* payoffs.

Disregarding unchanged behavior, two observations are striking after "take": (1) In each row less than 20% of the transition frequencies are below the main diagonal. This means only a small percentage of players decreases their choices if they had chosen the lower choice of the match. Each player confirms Hypothesis 1 and, moreover, 21 players out of 60 decrease their choices in less than 2% of the cases after "take." Note that comparisons of transition frequencies after choices 1, 2, 12, and 13 are excluded from the first hypothesis. (2) The highest off-diagonal frequency is in the cells above and next to the main diagonal (underlined numbers) and its frequency is greater than the sum of the remaining frequencies to the right in the same row (not for choices 1, 2, 12, and 13). The hypothesis, that higher strategies are equally likely chosen, can be rejected in favor of the hypothesis, that the next highest strategy is more likely chosen than higher choices, after each choice 5 to 10 (binomial test, significance smaller 1%). Thus, players seem to be reluctant to increase their number by much from period to period. However, at low choices (below 5) there is a stronger tendency to increase more than just one step. This connects to the findings of Selten and Stoecker (1986). If the end effect occurred in the later part of the supergame and a player deviated to noncooperation before the opponent, then he most likely shifts upwards by one round. However, they have no observations for early end effect games. In the light of our game we suppose that defection in the PD-supergame already in the beginning rounds of a supergame is followed by defection more than one period later in the next supergame. The reluctance of shifting by "too much" can also be found in "beauty-contest" experiments (see Stahl, 1996).

After "pass" the situation is almost the reverse: (1) cells next to and below the main diagonal have the highest weight, indicated by the underlined numbers (except for choice 6, where there are few observations). The hypothesis that a decrease by one or more steps is equally likely can be rejected for behavior after choices 6 to 13 (binomial test, significance $<1\%$). No more than 20% of the observations are above the main diagonal (except for choices lower than 7, where there are few observations). All but five players confirm Hypothesis 2. Three of those exceptional players decreased their choices less than 10% for all cases, irrespectively after "take" or "pass."

Furthermore, increases are more likely after "take" than after "pass" (six players do not confirm this hypothesis) and decreases are more likely after "pass" than after "take" (all players confirm Hypothesis 4). In fact, the relative frequency of decreases after "pass" is about three times as high as after "take" after all choices; even after a "successful" choice 13, decreases are much less likely than after a "pass." The confirmation of the four hypotheses for most players suggest that the reactions after take and pass are very different.

Note that the patterns of learning direction theory do not necessarily imply learning to play closer to the equilibrium strategy (i.e., to decrease a choice). Furthermore, if players give the best reply to a belief of the existence of altruistic players and errors, as assumed in McKelvey and Palfrey (1992) the difference of behavior between Take and Pass should not be that sharp. Mixed strategies cannot explain these differences either. McKelvey and Palfrey (1992, p. 811) pointed out the behavior of two subjects who in fact behave as described by the learning direction theory. However, M/P did not notice that pattern. Instead, they called that an "interesting nonpattern in the data, ... inconsistent with the use of a single pure strategy .... Fairly common irregularities of this sort, ... would seem to require some degree of randomness to explain. While some sort of this behavior may indicate evidence of the use of mixed strategies, some such behavior is impossible to rationalize, even by resorting to the possibility of altruistic individuals or Bayesian updating across games."

In the following we analyze the basic reinforcement model with respect to the hypotheses of our simple cognitive process. In order to do so we run 200 simulations of the basic reinforcement model, with the parameter $q = 0.9$, initial uniform frequency distribution, and propensity weights 50 for each strategy as in Section 5. The updating after each period is similar as in the formula of the basic reinforcement model in Section 5. However, instead of updating the discounted payoff-sum with the payoff resulting from the actual choice of a player in a session, the updating follows from the choice resulting of the draw according to the probability distribution after each period,[14]

$$M_{Ai}(t) = qM_{\alpha}(t-1) + c'_{A\alpha}(t)\,\pi_{Ai\alpha}(t), \qquad \alpha \in \{1, 3, ..., 13\}; \quad \text{similarly for B,} \tag{11}$$

where $c'_{Ai\alpha}(t) = 1$ results from the draw of the probability distribution:

$$p_{A\alpha}(t) = \frac{M_{A\alpha}(t-1)}{\sum_{k=1}^{13} M_{Ak}(t-1)} \quad \forall A, \alpha; \quad \text{similarly for player B.} \tag{12}$$

Six players A are randomly matched with six players B in each period. We present the results of the simulations of the average lower choices over time, together with the pooled observation of the actual lower choices in each period (see Fig. 5). As one can see, the simulation captures the general trend of the data but converges too slowly and undershoots early and overshoots late. That it goes too slow is typical for the basic model as also seen in Roth and Erev (1995), where it takes about 100 periods for the simulations to reach period 10 data (however, see also footnote 14). Tables 5a, b show the transition matrices produced from the simulations of the reinforcement model.

[14] This is following the method used by Roth and Erev (1995) who are mainly interested whether the simulations of aggregated behavior with this simple model show similar patterns as the actual aggregated behavior over time. Also those authors who use the learning direction model have in mind to see how well a very basic model can predict important characterists in individual behavior within a large class of games. Clearly, the literature already shows that there are better models, yet these models contain as major parts elements of the basic reinforcement model and the learning direction model.

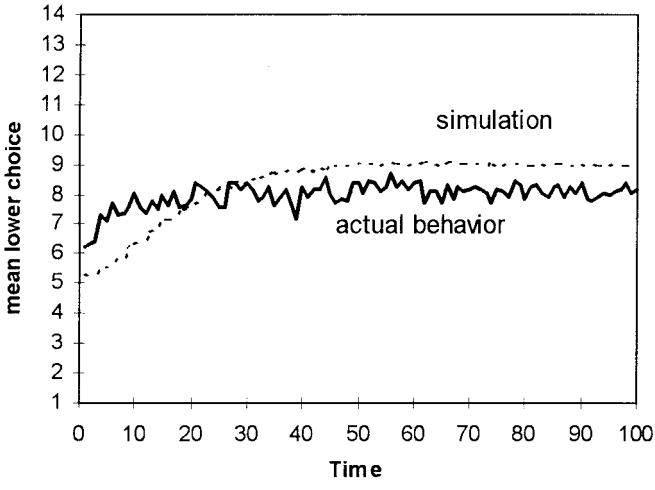

FIG. 5. Reinforcement simulation (mean lower choice), with forgetting parameter $q = 0.9$ and random initial frequencies, and actual mean lower choice per period pooled over all sessions.

As one can see, the asymmetry between "take" and "pass" is not as strong as in the actual data. Choices next to the main diagonal are only modal frequencies for choices higher than 6 (after take) or 8 (after pass); furthermore, the weights of these modal frequencies are not always greater than the remaining cells off the diagonal. Decreases after "pass" show higher frequencies than after "take," but they are only about twice as high as after "take," whereas in the actual data they are about three times as high. If we aggregated the frequencies after "take" and "pass" for the actual data and the simulations, the differences between actual behavior and simulation become much smaller. We know this already from the fairly low QDM measures calculated in the previous section which also does not distinguish for "take" and "pass." The chosen strategy is positively updated and the probabilities of the other strategies decrease by the same amount. Thus, the major improvement of this model would be to update strategies that are not chosen dependent on whether there was a "take" or a "pass". We mention in the following modifications of the reinforcement model, mentioned in the literature and the consequences for our data set.

Roth and Erev (1995) extend the basic reinforcement model by allowing for local experimentation. This means that the two adjacent strategies (here the adjacent strategies are ± 2 the choice i) of the realized action get also reinforced; some fraction of the round-t payoff of the realized action is subtracted from this propensity and added to the propensities of neighboring strategies. If so, decrease after "take" would be even more reinforced (also the increase after "pass"). Thus this kind of extension will clearly make it worse. Stahl (1996) and Camerer and Ho (1996) also use known information about payoffs of other strategies not chosen in a period for the updating. Chosen strategies are reinforced by their actual current payoffs. Non-chosen (hypothetical) strategies are also reinforced by the payoffs (weighted by a parameter) resulting from the current strategies of the opponents. This kind of updating has only been applied for normal form games, where the chosen strategies of the opponents are clear. Although players in our experiments play normal form

TABLE V

Transition Matrices Separating "Take" and "Pass"

		1	2	3	4	5	6	7	8	9	10	11	12	13	14	Total observ
Choice in period $t+1$ after "Take" (own choice in t was lower choice of match)																
c	1	*0.10*		0.11		0.12		0.17		0.18		0.18		0.15		60
h	2		*0.10*		0.11		0.12		0.18		0.21		0.17		0.12	64
o	3	0.09		*0.11*		0.11		0.17		0.19		0.18		0.16		64
i	4		0.07		*0.12*		0.13		0.19		0.22		0.16		0.11	78
c	5	0.05		0.06		*0.27*		0.17		0.19		0.14		0.12		102
e	6		0.04		0.05		*0.23*		0.22		0.21		0.16		0.09	162
	7	0.02		0.02		0.04		*0.64*		0.11		0.09		0.08		341
i	8		0.01		0.02		0.05		*0.49*		0.23		0.15		0.04	521
n	9	0.01		0.01		0.02		0.04		*0.81*		0.06		0.05		539
	10		0.01		0.01		0.03		0.13		*0.61*		0.17		0.04	531
t	11	0.01		0.01		0.01		0.03		0.05		*0.84*		0.05		292
	12		0.01		0.01		0.03		0.10		0.18		*0.62*		0.04	205
	13	0.02		0.03		0.03		0.04		0.06		0.07		*0.76*		43
	14							Not possible								3000
Choice in period $t+1$ after "Pass" (own choice in t was higher choice of match)																
c	1							Not possible								
h	2		*0.10*		0.14		0.15		0.15		0.20		0.14		0.12	5
o	3	0.10		*0.13*		0.11		0.20		0.19		0.11		0.16		6
i	4		0.11		*0.13*		0.13		0.20		0.17		0.13		0.14	14
c	5	0.09		0.11		*0.20*		0.15		0.16		0.16		0.13		15
e	6		0.08		0.10		*0.19*		0.21		0.18		0.13		0.12	29
	7	0.05		0.06		0.08		*0.53*		0.11		0.09		0.09		65
i	8		0.02		0.04		0.09		*0.49*		0.20		0.11		0.05	177
n	9	0.02		0.02		0.03		0.07		*0.77*		0.06		0.04		320
	10		0.02		0.02		0.05		0.19		*0.55*		0.13		0.04	481
t	11	0.01		0.01		0.02		0.05		0.06		*0.81*		0.04		585
	12		0.01		0.02		0.04		0.14		0.22		*0.53*		0.04	553
	13	0.01		0.02		0.02		0.05		0.07		0.07		*0.76*		569
	14		0.05		0.05		0.09		0.18		0.25		0.19		*0.18*	180
																3000

Note. Each entry is a relative frequency of transition behavior using the reinforcement simulations. The underlined entries are the *modal off diagonal frequencies*. Italic cells indicate the frequencies of unchanged behaviour.

games, they cannot always deduce from the payoff matrix what their opponent chose. This is the case for the player who chooses the lower choice in a match. However, one could apply Camerer and Ho's mechanism to restrict the updating to the observed information of the players: In a match in which a player has chosen the lower number, only those of his strategies equal to and below his current choice can be updated since the player knows these payoffs in that case. Of course, lower (hypothetical) choices produce lower payoffs. In case a player has chosen the higher number of his match, all strategies are updated, since he knows what the opponent

has chosen. The payoffs for (hypothetical) choices above his actual choice are the same as for his actual choice. The highest payoff is obtained for the choice just one below the opponents choice. This kind of updating would mean that after "take" decreases become more likely than increases, and after "pass" best response choices become most strongly updated. These choices might not be next and below the main diagonal. A similar point has been made by Vriend (1997) in connection with ultimatum games. A promising extension suggested by Camerer and Ho which was inspired by discussions about our data is to reinforce unchosen strategies after "take" by some elements of the set of forgone payoffs. This means, that after take, choices above the chosen take nodes are updated by for, e.g. the median payoff, or a combination of minimum or maximum payoff possibly gained with these higher choices. This will generate transition probabilities similar to those reported in Table 4 after take.

Another aspect which we feel worth mentioning is the transition behavior over time, that is, at the beginning versus at the end of the experiment. From the psychological literature, it is well known that learning takes place in the beginning of a

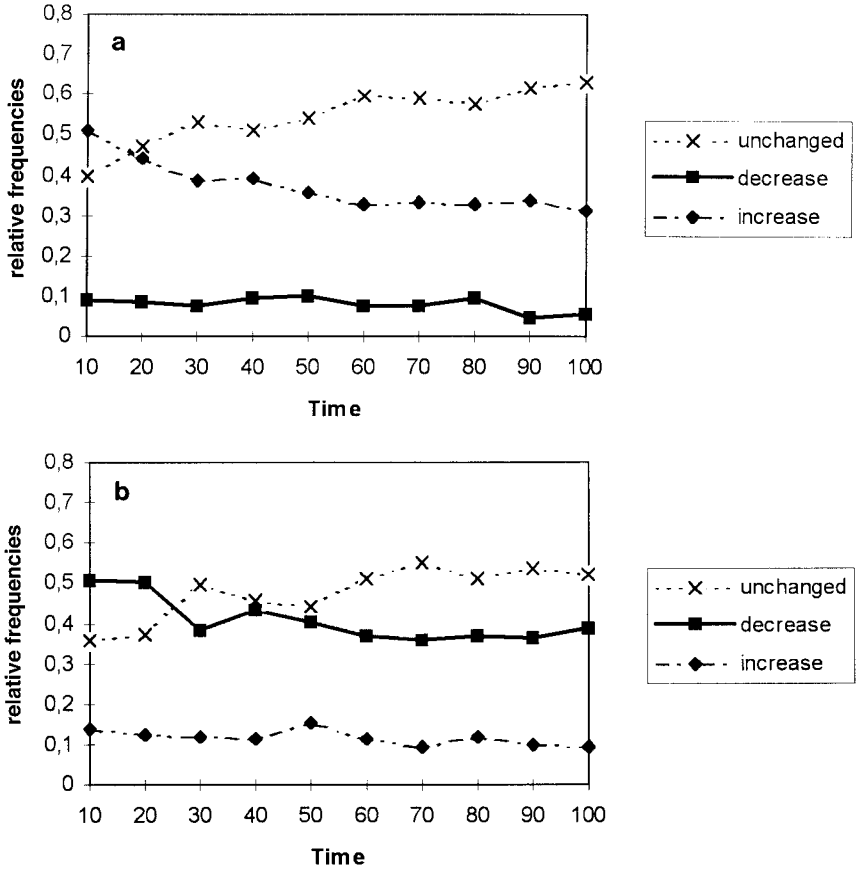

FIG. 6. Transition behavior of actual choices pooled over 10-periods, separately for "take" (a) and "pass" (b). E.g., the decrease line after "pass" for example means how many players in % chose a lower node after observing that their choice was higher choice of match.

new situation. In order to put this hypothesis in operation, we pool transition behavior across 10 periods and aggregate the relative frequencies of increased, decreased, or unchanged behavior across choices, respectively, separately for "previous period was take" and for "previous period was pass." In other words we aggregate the cells above, the cells below the main diagonal and the diagonal cells, respectively for each block of 10-period transition matrix. Figure 6 shows the development of the transition behavior (increased, decreased, unchanged) for each of the 10-period blocks pooled over all sessions, separately for "take" and "pass."

In the opening periods behavior according to learning direction theory has the highest frequency, which is the sum of relative frequencies of increases after take and those of decreases after pass. This holds for all five sessions, separately as well. A similar feature has been observed in Duffy and Nagel (1997). In the centipede game unchanged behavior receives increasing importance with highest frequency after about 50 periods in all sessions, except after "pass" in session 3. Thus, in the centipede game, most learning takes place in the beginning. This is an observation

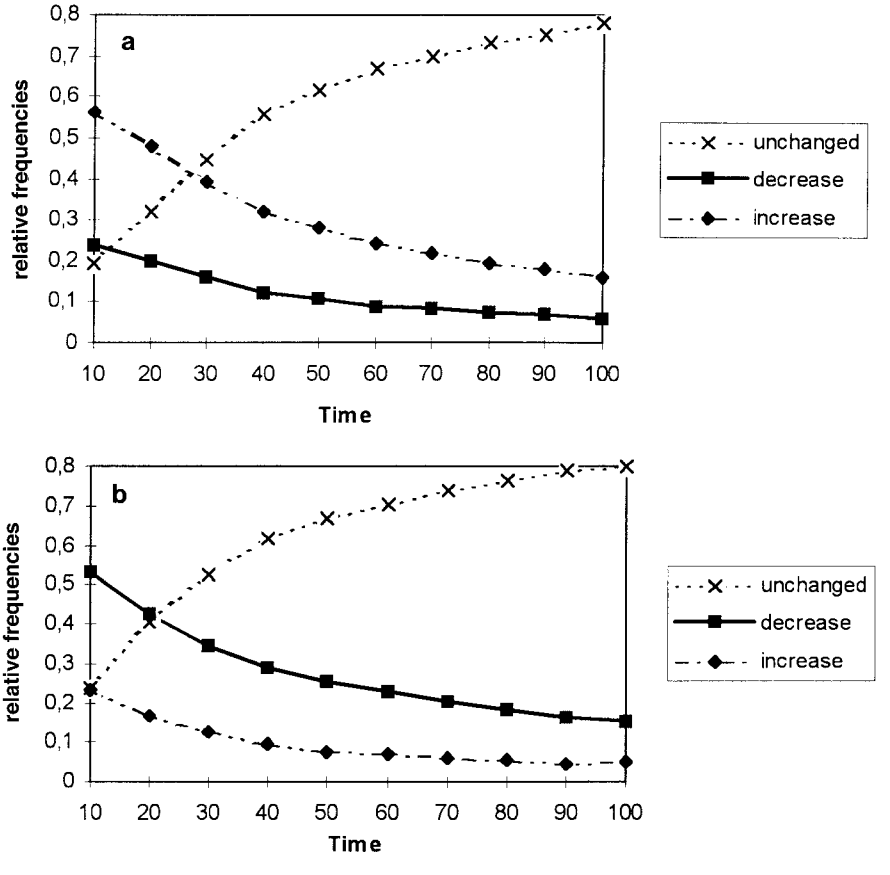

FIG. 7. Transition behavior of the reinforcement simulation pooled over 10-periods, separately for "take" (a) and "pass" (b). E.g., the decrease line after "pass" for example means how many players in % chose a lower node after observing that their choice was higher choice of match.

which is called the law of effect. Only after pass there is still substantial movement towards the equilibrium (see relative frequencies of decreases after pass). If unchanged behavior increased even more in a longer time horizon, it is questionable whether actual behavior would ever converge towards the one-stage equilibrium, as suggested by Zauner (1996). Clearly, the reinforcement model also predicts an increase of unchanged behavior. However, increases after Take and decreases after Pass are too much reduced and the difference between right or wrong direction is less sharp, especially after Take (see Fig. 7).

6. CONCLUSION

We have analyzed behavior on the centipede game played in the reduced normal form. The main advantage of the strategic form over the extensive form is that players have to reveal the intended take node, information that is interesting to study the adaptation of a player from period to period. We have compared different learning models which have been predominant in the experimental literature.

McKelvey and Palfrey (1992), Fey, McKelvey, and Palfrey (1994), and McKelvey and Palfrey (1996b) explain the behavior in centipede games by ex-ante rationality and by using equilibrium models with errors in beliefs or actions. We show that the two equilibrium models, the standard Nash equilibrium and the quantal response model (McKelvey & Palfrey (1996a) perform much worse than simple reinforcement models. This model is also better than fictitious play.

Another important aspect of this paper was to analyze behavior in terms of a simple ex-post reasoning process which prescribes in which direction the behavior should be changed from period to period. Because of the structure of the centipede game, we were able to discuss the qualitative learning direction theory in much greater depth than in any of the previous papers, where this theory has been applied. In particular, we were able to disaggregate transition behavior from period t to period $t+1$ after *each possible choice* in period t and also the transition behavior in earlier periods, in comparison to later periods. We found that most subjects conform on average to the qualitative learning theory. In the first periods this holds more often than in later periods, when unchanged behavior tends to dominate. The most robust finding is that decreases after Pass occur almost three times as much as after Take at each node; after Take increases are more likely than decreases and after Pass decreases are more likely than increases. This calls us to question the interpretation of the data by McKelvey and Palfrey (1992) that players adjust their behavior according to a model of incomplete information about altruists. Because unchanged behavior increases over time, Zauner's predictions of convergence towards the equilibrium is also questionable. At least we showed that they do not converge within 100 periods as he hypothesized. The extension to a simple reinforcement model introduced by Roth and Erev (1995) and Camerer and Ho (1997) cannot explain the differences of transition behavior after take and pass either. The later so far applies only to normalform games with exact information of opponents strategies. A modification of their model to extensive form games or games with extensive form information as in our game should be possible.

Since we gave the extensive form information to the subjects, we hope that our analysis will inspire improvements of learning models for experiments of games played in extensive form or with extensive form information as in ours. The main difficulty of improving learning models for extensive form games is that a player might not receive information of behavior of opponents in subgames not reached in a period. This means that the updating of unchosen choices might have to be based on unobservable information.

APPENDIX

Instructions

— Each participant has to make a decision in each of 100 rounds.

— There are two different types: six participants are of type A and six are of type B.

— At the beginning of the experiment you will know your type, which is the same for the entire experiment.

— A type A always meets a type B and a type B always meets a type A.

— In each round, it is randomly determined which type A meets which type B.

— A and B simultaneously choose a number out of the following numbers:

— A chooses a number from $\{1, 3, 5, 7, 9, 11, 13\}$,

— B chooses a number from $\{2, 4, 6, 8, 10, 12, 14\}$.

— The smaller number of the chosen numbers (smaller choice) determines the payoff (according to Table VI).

— Table VI shows that the higher the "smaller choice," the higher the sum of the payoffs; the sum increases by about 40% if the "smaller choice" increases by 1.

— The sum is divided into a small and a high payoff:

— About 80% of the sum is provided to the person with the smaller choice,

— About 20% of the sum is provided to the person with the higher choice.

TABLE VI

Smaller choice	Possible higher choices of opponents	Payoff of A	Payoff of B	Sum
1	{2, 4, 6, 8, 10, 12, 14}	4	1	5
2	{3, 5, 7, 9, 11, 13}	2	5	7
3	{4, 6, 8, 10, 12, 14}	8	2	10
4	{5, 7, 9, 11, 13}	3	11	14
5	{6, 8, 10, 12, 14}	16	4	20
6	{7, 9, 11, 13}	6	22	28
7	{8, 10, 12, 14}	32	8	40
8	{9, 11, 13}	11	45	56
9	{10, 12, 14}	64	16	80
10	{11, 13}	22	90	112
11	{12, 14}	128	32	160
12	{13}	44	180	224
13	{14}	256	64	320

Information

— At the end of each round, each participant is informed about his result of the round:

— the lower choice

— the payoffs to A and B.

— You will not know with whom you were matched. You will know only about the choice of the other, if his number was the lower choice.

Payoffs

— The sum of the payoffs of all rounds of a participant is his total gain.

— The exchange rate is 0.01 DM for 2 points, thus 1000 points is 5 DM.

REFERENCES

Arthur, B. (1991). Designing economic agents that act like human agents: A behavioral approach to bounded rationality. *American Economic Review*, Papers and Proceedings, **81**, 353–409.

Aumann, B. (1992). Irrationality in Game Theory. In P. Dasgupta, D. Gale, O. Hart, & E. Maskin (Eds.), *Economic analysis of markets and games*, pp. 214–227. Cambridge, MA: MIT Press.

Binmore, K. (1988). Modeling rational players, I, II. *Economics and Philosophy*, **3**, 179–214; **4**, 9–55.

Barto, A. G., Sutton, R. S., & Anderson, C. W. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, **13**(5), 834–846.

Brier, G. W. (1950). Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, **78**, 1–3.

Brown, G. (1951). Iterative solution of games by fictitious play. In T. Koopsmans (Ed.), *Activity analysis of production and allocation*, pp. 374–376. New York: Wiley.

Bush, R. R., & Mosteller, F. (1955). *Stochastic models of learning*. New York: Wiley.

Cason, & Camerer, C. (1996). The sunk cost fallacy, forward induction and behavior in coordination games. *Quarterly Journal of Economics*, **111**, 165–194.

Camerer, C., & Ho, T. (1997). Experience-weighted attraction learning in games: A unifying approach. Caltech working paper 1003.

Chen, Y., & Tang, F. F. (1998). Learning and incentive compatible mechanism for public goods provision: An experimental study. *Journal of Political Economics*.

Chen, H., Friedman, J., & Thisse, J.-F. (1996). Boundedly rational Nash-equilibrium: A probabilistic approach. *Games and Economic Behavior*.

Cressmann, R., & Schlag, K. H. (1996). *The dynamic (in)stability of backwards induction*. University of Bonn, mimeo.

Cross, J. G. (1973). A stochastic learning model of economic behavior, *Quarterly Journal of Economics*, **87**, 239–266.

Cross, J. G. (1983). *A theory of adaptive economic behavior*. Cambridge: Cambridge University Press.

Duffy, J., & Nagel, R. (1997). On the robustness of behaviour in experimental "beauty-contest games", *Economic Journal*, **107**, 1684–1700.

Erev, I., & Roth, A. (1998), On the need for low rationality, cognitive game theory: Reinforcement learning in experimental games with unique mixed strategy equilibria. *American Economic Review*.

Fey, M., Mckelvey, R. D., & Palfrey, T. R. (1994). Experiments on the constant-sum centipede game. Caltech working paper.

Glazer, J., & Rubinstein, A. (1996). An extensive game as a guide for solving a normal game. *Games and Economic Behavior*, **70**, 32–42.

Gueth, W., Ockenfels, P., & Wendel, M. (1993). Efficiency by trust in fairness? Multiperiod ultimatum bargaining experiments with an increasing cake. *International Journal of Game Theory*, **22**, 51–73.

Van Huyck, J. B., Battalio, R. C., & Beil, R. O. (1990). Tacit coordination games, strategic uncertainty, and coordination failure. *American Economic Review*, **80**, 234–248.

Holland, J. H. (1975). *Adaptation in natural and artificial systems*. Ann Arbor: University of Michigan Press.

Holland, J. H., Holyoak, K. J., Nisbett, R. E., & Thagard, P. R. (1986). *Induction*: *Processes of inference*, *learning*, *and discovery*, Cambridge, MA: MIT Press.

Hull, C. L. (1943). *Principles of behavior*. New York: Appleton–Century–Crofts.

Kreps, D., Milgrom, P., Roberts, T., & Wilson, R. (1982). Rational cooperation in the finitely repeated prisoners' dilemma. *Journal of Economic Theory*, **27**, 245–252.

Luce, R. D. (1959). *Individual choice behavior*. New York: Wiley.

Mckelvey, R. D., & Palfrey, T. R. (1992). An experimental study of the centipede game. *Econometrica*, **60**, 803–836.

Mckelvey, R. D., & Palfrey, T. R. (1995a). Quantal response equilibria for normal form games. *Games and Economic Behavior*, **10**, 6–38.

Mckelvey, R. D., & Palfrey, T. R. (1995b). Quantal response equilibria for extensive form games. Caltech, mimeo.

Mookherjee, D., & Sopher, B. (1996). Learning and decision costs in experimental constant sum games. Mimeo.

Mitzkewitz, M., & Nagel, R. (1993). Experimental results on ultimatum games with incomplete information. *International Journal of Games Theory*, **22**, 171–198.

Nagel, R. (1994). Reasoning and learning in guessing games and ultimatum games with incomplete information: An experimental investigation. University Bonn disseration.

Nagel, R. (1995). Unraveling in guessing games: An experimental study. *American Economic Review* **85**(5), 1313–1326.

Nagel, R., & Vriend, N. (1997). *A study of adaptive behavior in oligopolistic market games*, Working Paper 230, Universitat Pompeu Fabra.

Nagel, R., & Sadrieh (1998). A comparison of behavior in extensive form & normal form centipede games. In preparation.

Neugebauer, T. (1994). *Experimente mit dem "Centipede"-spiel in normalform*, Diplomarbeit, University of Bonn.

Ponti, G. (1996). Cycles of learning in the centipede game. University College London. mimeo.

Robinson, J. (1951). An iterative method of solving a game. *Annals of Mathematics*, **54**, 296–301.

Rosenthal, R. (1981). Games of perfect information, predatory pricing, and the chain store paradox. *Journal of Economic Theory*, **25**, 92–100.

Roth, A. E., & Erev, I. (1995). Learning in extensive-form games: Experimental data and simple dynamic models in the intermediate term. *Games and Economic Behavior*, **8**, 164–212.

Selten, R. (1997). *Axiomatic characterization of the quadratic scoring rule*, Discussion Paper No. B-390, University of Bonn.

Selten, R., & Stoecker, R. (1986). End behavior in sequences of finite prisoner's dilemma supergames, a learning theory approach. *Journal of Economic Behavior*, 47–70.

Selten, R., & Buchta, J. (1998). Experimental sealed bid first price auction with directly observed bid functions. In D. Budescu, I. Erev, & R. Zwick (Eds.), *Games and human behavior*, *essays in honor of Amnon Rapoport*. Hillsdale, NJ: Erlbaum.

Siegel, S., & Castellan, N. J. (1988). *Nonparametric statistics for the behavioral sciences*. New York: McGraw–Hill.

Stahl, D. O. (1996). Bounded rational rule-learning in a guessing game. *Games and Economic Behavior*, **16**(2), 303–330.

Sutton, R. S. (1992). Introduction: The challenge of reinforcement learning. *Machine Learning*, **8**, 3/4, 225–227.

Tang, F. F. (1996). *Anticipatory learning in two-person games: an experimental study*. *Part II. Learning*. Sfb-Discussion Paper B-363, University of Bonn.

Vriend, N. (1997). Will reasoning improve learning? *Economics Letters*, *1997*, **55**(1), pp. 9–18.

Weisbuch, G., Kirman, A., & Herreiner, D. (1996). Market organization, mimeo.

Yates, F. J. (1990). *Judgment and decision making*. Englewood Cliffs, NJ: Prentice–Hall.

Zauner, K. G. (1996). *A payoff uncertainty explanation of results in experimental centipede games*. Working paper 96-030, University of New South Wales.

Received March 16, 1998