Bonn Econ Discussion Papers

Discussion Paper 18/2008

Learning in experimental 2 x 2 games

by

Thorsten Chmura, Sebastian J. Goerg,
Reinhard Selten

Dezember 2008

Bonn Graduate School of Economics
Department of Economics
University of Bonn
Kaiserstrasse 1
D-53113 Bonn

Financial support by the
Deutsche Forschungsgemeinschaft (DFG)
through the
Bonn Graduate School of Economics (BGSE)
is gratefully acknowledged.

Deutsche Post World Net is a sponsor of the BGSE.

Learning in experimental 2×2 games[‡‡]

Thorsten Chmura , Sebastian J. Goerg , Reinhard Selten

December 16, 2008

Abstract

In this paper we introduce four new learning models: impulse balance learning, impulse matching learning, action-sampling learning, and payoff-sampling learning. With this models and together with the models of self-tuning EWA learning and reinforcement learning, we conduct simulations over 12 different 2×2 games and compare the results with experimental data obtained by Selten & Chmura (2008). Our results are two-fold: While the simulations, especially those with action-sampling learning and impulse matching learning successfully replicate the experimental data on the aggregate, they fail in describing the individual behavior. A simple inertia rule beats the learning models in describing individuals behavior. (97 words)

Keywords: Learning, Action-sampling, Payoff-sampling, Impulse balance, Impulse matching, Reinforcement, self-tuning EWA, 2×2 games, Experimental data

JEL Classification: C72, C91, C92

[‡‡] Financial support by the Deutsche Forschungsgemeinschaft is gratefully acknowledged.

I Introduction

It is known that rational learning, in the sense of Bayesian updating, leads to the stationary points of the Nash equilibrium (e.g. Kalai and Lehrer, 1993). But it also known that actual human behavior not necessarily converges to the Nash equilibrium. In fact, a vast body of literature indicates situations in which standard theory performs not as a good predictor for subjects' behavior in experiments (e.g. Brown & Rosenthal, 1990, Erev & Roth, 1998).

A recent publication by Selten & Chmura (2008) documents the predominance of behavioral stationary concepts regarding the descriptive power . In the paper the concepts of impulse balance equilibrium (Selten & Chmura, 2008), payoff-sampling equilibrium (Osborne & Rubinstein, 1998) and action-sampling equilibrium (Selten & Chmura, 2008) outperform Nash equilibrium as well as quantal response equilibrium (McKelvey & Palfrey, 1995) in describing the decisions of a population in twelve completely mixed 2×2 games.

The three behavioral stationary concepts of action-sampling equilibrium, payoff-sampling equilibrium and impulse balance equilibrium contain precise description of stationary behavior and thus they are predestined to be used as the basis of learning models. It is obvious that if human behavior tends (in the short run) to other stationary points than Nash equilibrium, learning mechanisms leading to theses points are a promising approach.

The main purpose of this paper is to introduce four new learning models which are based on the behavioral reasoning of payoff-sampling equilibrium, action-sampling equilibrium and impulse balance equilibrium and test them in the environment of twelve repeated 2×2 games. Hereby, the learning rules have to meet two challenges: First, do they reproduce the aggregate behavior of a human population and second do they adequately describe the observed behavior of a single individual? For comparison we include the models of reinforcement learning (Erev & Roth, 1998) and self-tuning experience weighted attraction learning (EWA) (Ho, Cramerer & Chong , 2007) into our study.

We conduct simulations with the learning models and the twelve 2 × 2 games experimentally investigated in Selten & Chmura (2008). The simulations replicate the exact situation of the 2 × 2 experiments. In each simulation run, eight agents, four deciding as row players and four deciding as column players, are randomly matched each round over 200 rounds. In each simulation run one game is played and one learning model is applied. To judge the predictive power on the aggregate level we compare the distribution of choices in the simulation runs with the data from Selten's & Chmura's 2 × 2 experiments.

In addition we evaluate the explanatory power of the learning models for each participant of the 2 × 2 experiments, separately. For each of the 864 subjects we compared the actual decision in every round with the decision predicted by the learning model given the subject's history. To judge the power of the learning models we introduce a benchmark which all learning models should beat. This benchmark is the inertia rule, which predicts for each round the same choice as executed in the round before.

Our results are twofold, while our models are able to capture the distribution of decisions on the aggregate level, they fail to explain the individual data. On the aggregate level the learning models of impulse matching learning and action-sampling learning have the smallest distance to the experimental data, while the concepts of self-tuning EWA and reinforcement learning have biggest. On the individual level all learning models fail to beat the inertia rule.

The rest of the paper is organized as follows: In section II we will introduce the models impulse balance learning, impulse matching learning, action-sampling learning and payoff-sampling learning. In addition we will briefly deal with reinforcement learning and self-tuning EWA. Afterwards, in section III, we will recapitulate the experiment conducted in Selten & Chmura and introduce our measurements of predictive success for the aggregate data and for the individual data. Subsequently, section IV gives our results and section V summarizes and concludes the paper.

II The Learning Models

In this section we will introduce four new learning models, which are based on the behavioral stationary concepts discussed in Selten & Chmura (2008). The concepts to be introduced are: impulse balance learning, impulse matching learning, action-sampling learning and payoff-sampling learning. In addition to the new learning models, the more established concepts of reinforcement learning (c.p. Erev & Roth, 1998) and self-tuning EWA (Ho, Cramerer & Chong , 2007) are briefly explained.

Three of the discussed models, namely action-sampling learning, payoff-sampling learning and self-tuning EWA are parametric concepts. In case of action sample learning and payoff sample learning the parameter is the sample size. Self-tuning EWA is based on the multi-parametric concept of experience weighted attraction learning (Cramerer & Ho, 1999). Self-tuning EWA replaces two of the parameters with numerical values and two with experience functions. The remaining *"parameter λ measures sensitivity of players to attractions"* (p. 835 Cramerer & Ho, 1999). The version of reinforcement learning theory examined here does not have any parameter and the initial propensities are not estimated from the data.

For the sampling learning models we will not determine the optimal sample size, but apply the sample sizes which determined the best fit for the related stationary concepts to the data in Selten & Chmura (2008). In case of the action-sampling learning this is the action-sampling equilibrium and in case of the payoff-sampling learning this is the payoff-sampling equilibrium. The parameter of self-tuning EWA is determined in such a way that it leads to the best fit over all data and over all games.

In the literature parametric concepts are usually fitted for each game separately. We believe that this gives an unfair advantage to one-parameter theories over parameter free ones, especially in the case of 2×2 games where only two relative frequencies are predicted. Adjusting one parameter separately for each game so to speak does half the job. Therefore we base our analysis on one estimate for all games in case of the self-tuning EWA and in case of the sampling learning rules we take the parameter for the stationary concepts estimated in Selten & Chmura (2008) over all games.

A Impulse balance learning

Impulse balance learning relates to the concepts of impulse balance equilibrium (Selten, Abbink & Cox 2005 and Selten & Chmura 2008) and learning direction theory (Selten & Buchta, 1999). After a decision and after the realization of the payoffs the behavior is adjusted to experience. Selten and Buchta explain the concept by the example of a marksman aiming at a trunk: *"If he misses the trunk to the right, he will shift the position of the bow to the left and if he misses the trunk to the left he will shift the position of the bow to the right. The marksman looks at his experience from the last trial and adjusts his behavior [...]."* (p. 86 Selten & Buchta, 1999).

Suppose that the first of two actions has been chosen in a period and this action was not the best reply to the action played by the other player. Then the player receives an *impulse* towards the second action. This impulse is the difference between the payoff the player could have received for his best reply minus the payoff actually received given the decision by the other player in this period. The player does not receive an impulse if his action was a best reply against the other player's decision.

To incorporate loss aversion, the impulses are not calculated with the original payoffs but with transformed ones. In games with two pure strategies and a mixed Nash equilibrium each pure strategy has a minimal payoff and the maximum of the two minimal payoffs is called the pure strategy maximin. This pure strategy maximin is the maximal payoff a player can obtain for sure in every round and it forms a natural aspiration level. Amounts below this aspiration level are perceived as losses and amounts above this aspiration level are perceived as gains. In line with prospect theory (Kahneman & Tversky, 1979) losses are counted double in comparison to gains. Thus, gains (the part above the aspiration level) are cut to half for the computation of impulses. Figure 1 is taken from Selten & Chmura (2008) and illustrates the transformation of the payoffs by the example of game 3.

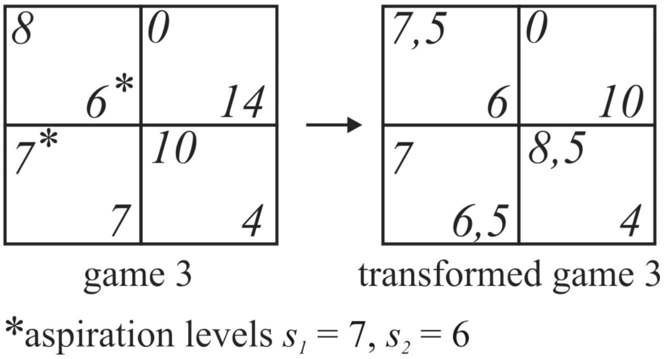

Figure 1: Example of matrix transformation as given in Selten & Chmura (2008)

Impulse balance learning can be described as a process in which a subject builds up impulse sums. The impulse sum $R_i(t)$ is the sum of all impulses from j towards i experienced up to period $t-1$. The probabilities for playing action 1 and 2 in period t are proportional to the impulse sums $R_1(t)$ and $R_2(t)$:

$$p_i(t) = \frac{R_i(t)}{R_1(t) + R_2(t)} \text{ , for } i = 1, 2 \tag{A.1}$$

The impulses from action j towards action i in period t is as follows:

$$r_i(t) = \begin{cases} max[0, \pi_i - \pi_j] & \text{, if the chosen action is } j \\ 0 & \text{else.} \end{cases} \tag{A.2}$$

for $i, j = 1, 2$ and $i \neq j$. Here, π_i is the payoff for action i given the matched agents decision and π_j the one for action j. Afterwards the impulse sums are updated with the new impulses:

$$R_i(t+1) = R_i(t) + r_i(t) \tag{A.3}$$

In the first round all impulse sums are zero $R_1(1) = R_2(1) = 0$ and until both impulse sums are higher than zero the probabilities are fixed to $p_1(t) = p_2(t) = 0.5$.

B Impulse matching learning

This learning model is very similar to impulse balance learning. In fact, in our 2×2 setting the resulting stationary point of impulse matching learning is the same as of impulse balance learning. But for other types of games both concepts do not necessarily lead to the same stationary points. Therefore we treat the impulse matching learning as a self contained model. As in the case for the impulse balance learning impulse matching learning is applied to the transformed matrix, described in section A.

The idea of an impulse is different in impulse matching. Here it is assumed that after a play a player always receives an impulse to his ex-post optimal strategy, the best reply to the pure strategy chosen by the other player. Thus an impulse from j towards i is defined as a payoff differences, regardless of the player's own action. This means that (A.2) has to be replaced by the equation (B.2).

$$r_i(t) = max[0, \pi_i - \pi_j] \tag{B.2}$$

The equation (B.1.) and (B.3.) are identically to (A.1) and (A.3) respectively. As before π_i is the payoff of action i and π_j is the payoff of action j given the matched player's decision.

The name impulse matching is due to the fact that this kind of learning leads to probability matching by player one if the probabilities p_1 and $(1 - p_1)$ on the other side are fixed and the payoffs for the player is one if both players play the strategy with the same number (one or two) and zero otherwise. Probability matching has been observed in early learning experiments, e.g. Estes (1954).

C Payoff-sampling learning

Payoff-sampling learning relates to the stationary concept of Osborne & Rubinstein (1998) which was first applied to experimental data in Selten & Chmura (2008). The behavioral explanation of the stationary concept is that a player chooses her action after sampling each alternative an equal number of times, picking the action that yields the highest payoff.

To implement this behavior payoff-sampling learning is based on samples from earlier periods. Therefore the agent draws two samples $(s_1(t), s_2(t))$ of earlier payoffs, one sample with payoffs from rounds in which she chose action 1 and one with payoffs from rounds in which she chose action 2. The samples are randomly drawn with replacement and a fixed sample sizes of $n = 6$.[1] In the following $S_1(t)$ and $S_2(t)$ denote the payoff sums in $s_1(t)$ and $s_2(t)$, respectively.

After the drawing of the samples, the cumulated payoffs $S_1(t)$ and $S_2(t)$ are calculated and the action with the higher cumulated payoff is played, if there is one. If the samples of both possible actions have the same cumulated payoff the agent randomizes with $p_1 = p_2 = 0.5$.

$$p_i(t) = \begin{cases} 1 & \text{if } S_i(t) > S_j(t) \\ 0.5 & \text{if } S_i(t) = S_j(t) \\ 0 & \text{else} \end{cases} \tag{C.1}$$

for $i, j = 1, 2$ and $i \neq j$.

As before $p_i(t)$ is the probability of playing action i in period t. At the beginning and until positive payoffs for each action have been obtained at least once, the agent chooses both actions with equal probabilities, i.e. $p_1 = p_2 = 0.5$.

[1] Recall that $n = 6$ leads to the optimal fit for the payoff-sampling equilibrium to the experimental data in Selten & Chmura (2008).

D Action-sampling learning

Action-sampling learning relates to the idea of the action-sampling equilibrium of Selten & Chmura (2008). According to action-sampling equilibrium a player takes in the stationary state a fixed size sample of the pure strategies played by the other players in the past and optimizes against this sample.

In the process of action-sampling learning the agent randomly takes a sample $A(t)$ of n earlier actions $a_1, ..., a_n$ of the other player. In the following we are keeping n fixed to 7.[2] Let $\pi_i(a_j)$ be the payoff of action i if the opponent plays action a_j. For $i = 1, 2$ let $P_i(t) = \sum_{j=1}^{7} \pi_i(a_j)$ be the sum of all payoffs of the player for using her action i against the actions in this sample.

Therefore, in period t the player chooses her action 1 or 2 according to

$$p_i(t) = \begin{cases} 1 & \text{if } P_i(t) > P_j(t) \\ 0.5 & \text{if } P_i(t) = P_j(t) \\ 0 & \text{else} \end{cases} \tag{D.1}$$

for $i, j = 1, 2$ and $i \neq j$.

At the beginning the probabilities are set to $p_1 = p_2 = 0.5$ until both possible actions were played by the opponent agents.

E Reinforcement-Learning

The reinforcement learning is one of the oldest and well established learning models in the literature, refer to Harley (1981) for an early application in the field of theoretical biology.

In our reinforcement model a player builds up a payoff sum $B_i(t)$ for each of his actions 1 and 2 according to the following formula:

[2] As mentioned above $n = 7$ leads to the highest fit of the action-sampling equilibrium to the data in Selten & Chmura (2008).

$$B_i(t+1) = \begin{cases} B_i(t) + \pi_{(t)} & \text{if action } i \text{ was chosen in } t \\ B_i(t) & \text{else.} \end{cases} \tag{E.1}$$

Here $\pi(t)$ is the payoff obtained in period t. After an initial phase in which both possible actions are used with equal probabilities the probability of choosing action i in period t is given by:

$$p_i(t) = \frac{B_i(t)}{B_1(t) + B_2(t)} \tag{E.2}$$

This model presupposes that all payoffs in a player's payoff matrix are positive with the possible exception of one. All twelve games considered here have this property. In the first round the initial payoff sums $B_i(t)$ are zero. The initial phase ends as soon as each of both possible actions has been used at least once. The player chooses both possible actions with equal probabilities $p_1 = p_2 = .5$. Only from then on rule $E.2$ is applied.

For games with negative payoffs this approach is not adequate. For example in the model used by Erev & Roth (1998) the payoff $\pi(t)$ in $E.1$ was replaced by $\pi(t) - \pi_{\min}$, where $\pi_{\min}$ is the smallest possible payoff of the player. Moreover they estimated initial values $B_i(0)$ from the data. We did not do this since we are only interested in models with at most one parameter.

F Self-tuning EWA

Self-tuning EWA was introduced by Ho, Camerer & Chong. It is based on the experience weighted attraction model, but estimates the parameter of this model with several functions. Of all models discussed in the paper at hand, self-tuning EWA is the most complex one.

The decisions are made according to *attractions* $A_i(t)$ for each strategy. The attractions depend on an experience weight, a change-detector function and an attention function. For more details on the attraction updating function refer to the appendix.

The probability of playing action i in period t depending on the attractions is calculated as a logit response function:

$$p_i(t) = \frac{e^{\lambda A_i(t-1)}}{\sum_{j=1}^{2} e^{\lambda A_j(t-1)}}$$

Here, λ is the response sensitivity and this parameter must be specified to fit to the empirical data. We searched for one λ to yield the best fit over all 12 games. Our measurement of the predictive success is the quadratic distance Q, which will be explained in more detail in the next chapter. Figure 2 gives the quadratic distance for the different values of lambda.

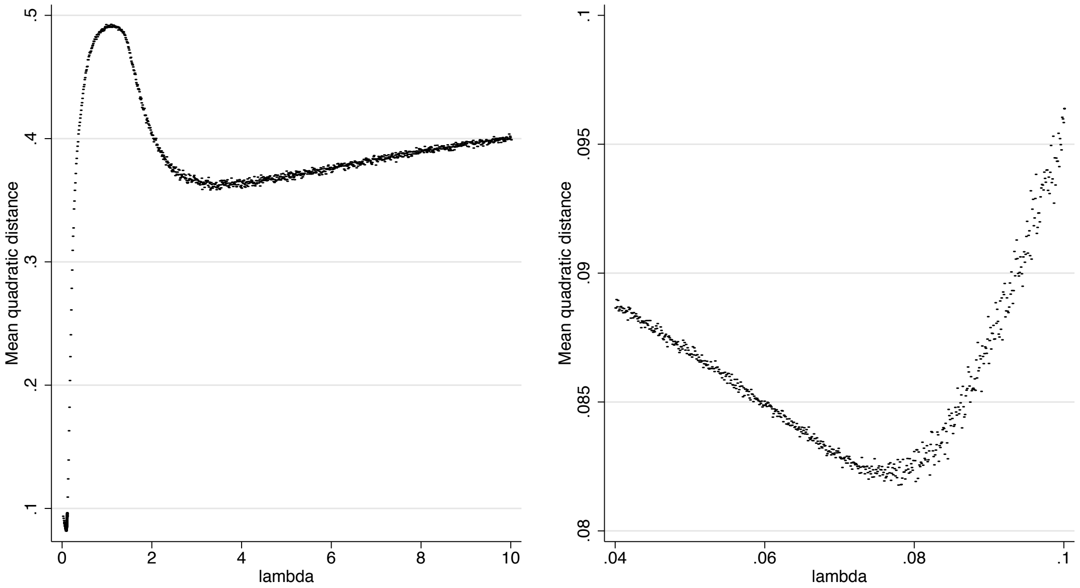

Figure 2: Quadratic distances of self-tuning EWA for different lambdas, each point represents the mean quadratic distance over 500 simulations. Left figure for $0 \leq \lambda \leq 1$ and right figure for $.04 < \lambda < .1$

The left part gives the quadratic distance for $.04 < \lambda < .1$ and the right one for all tested lambdas between 0 and 10. Each point in both graphs represents the mean quadratic distance over all twelve games with 500 simulations runs per game with one specific lambda value. The value leading to the smallest quadratic distance is $\lambda = 0.0778$.

To be consistent with the other models we have choosen not to estimate any additional values. Therefore the initial attractions were set to $A_1 = A_2 = 0$.

III Design

A Games and Experiments

The experimental data, which are compared with the simulations, are those on which the paper by Selten & Chmura (2008) is based. In their study twelve 2×2 games were experimentally investigated. To cover a broad field of games, six constant and six non-constant sum games were played. Figure 9 shows the twelve games used in the experiment. The constant sum games are shown on the left side of the figure and the non-constant sum games on the right side.

Note that the first six games have the same best response structure as the second six games and that the concepts of action-sampling equilibrium and Nash equilibrium only depend on this best response structure. Thus the predictions of Nash equilibrium are the same for the first and the second six games. The same holds true for the action-sampling equilibrium.

Each game was played by matching groups consisting out of eight subjects. The role of the subjects were fixed for the whole experiment, thus four subjects decided as column players and the other four as row players. At the beginning of each round row and column players were randomly matched. After each of the 200 rounds subjects received feedback about the other player's decision, their own payoff, the period number and their own cumulative payoff. The game played was known by all subjects.

For each constant sum game twelve independent matching groups were gathered, for each non-constant sum game six independent matching groups were gathered. Overall 864 subjects participated.

Constant sum games

Game 1

	L	R
U	10 8	0 18
D	9 9	10 8

Game 2

	L	R
U	9 4	0 13
D	6 7	8 5

Game 3

	L	R
U	8 6	0 14
D	7 7	10 4

Game 4

	L	R
U	7 4	0 11
D	5 6	9 2

Game 5

	L	R
U	7 2	0 9
D	4 5	8 1

Game 6

	L	R
U	7 1	1 7
D	3 5	8 0

Non-constant sum games

Game 7

	L	R
U	10 12	4 22
D	9 9	14 8

Game 8

	L	R
U	9 7	3 16
D	6 7	11 5

Game 9

	L	R
U	8 9	3 17
D	7 7	13 4

Game 10

	L	R
U	7 6	2 13
D	5 6	11 2

Game 11

	L	R
U	7 4	2 11
D	4 5	10 1

Game 12

	L	R
U	7 3	3 9
D	3 5	10 0

The payoffs for the column-players are shown in the lower right corner,
the payoff for the row-palyers are shown in the upper left corner.
Abbreviations used: L Left, R Right, U Up, D Down

Figure 3: The twelve 2 × 2-games taken from Selten & Chmura (2008).

The main goal of the present paper is to find learning algorithms which can replicate the human behavior in this twelve games. To evaluate this problem we compare the simulations with the experiments on the aggregate level and on the individual basis.

B Measure of Predictive Success on the Aggregate Level

On the aggregate level everything is kept the same as in the experiment except that instead of real participants now computer agents interact. Each agent interacts according to her history and to one learning model over 200 rounds. In each round eight agents with fixed roles, four deciding as row players and four as column players are randomly matched.

After each round they receive feedback about the matched agent's decision and their payoff. Since none of the learning models makes use of the round number and since the calculation of the cumulated payoff can be done by the agents themselves this information is not provided to the agents. It is crucial that the agents do not receive more information than the subjects in the experiment did.

All learning models include stochastic elements. To avoid the influence of statistical outliers 500 simulation runs per game are conducted. In each simulation run all agents act in accordance with one learning model, thus our data set obtained by the simulations consists out of 500 simulations per game and learning model.

To measure the predictive success on the aggregate basis, we will compare the mean frequencies of U and L in the simulations with the mean frequencies obtained in the experiments by means of the quadratic distance.

The mean quadratic distance Q is the average quadratic distance over all 12 games and over all 500 simulations for each of these

$$Q = \frac{1}{12}\sum_{i=1}^{12}\left(\frac{1}{500}\sum_{n=1}^{500}(s_{in}^{L} - f_i^{L})^2 + (s_{in}^{U} - f_i^{U})^2\right),$$

whereas s_{in} is the frequency for L or U in game i and simulation run n and f_i the mean frequency for L or U observed in the experiments with game number i.

The predictive success of a learning model increases with a decrease of the mean quadratic distance, i.e. the smaller the mean quadratic distance is the better does the learning theory fit the experimental data on the aggregate level.

C Measure of Predictive Success on the Individual Level

To judge the performance on the individual level we compare the individual decisions in every round with the predicted decisions or predicted probability by the learning rule, given the history of the subject.

To measure the predictive success of the learning theories describing the behavior of a single individual we apply the quadratic scoring rule. It was first introduced by Brier (1950) in the context of weather forecasting. The rationale behind the quadratic scoring rule is that for each round a score is determined which evaluates the nearness of the predicted probability distribution to the observed outcome.

In Selten (1998) the quadratic scoring rule is axiomatically characterized. The characterizing properties of the quadratic scoring rule as described in Selten (1998) are: symmetry, elongational invariance, incentive compatibility and neutrality. Symmetry means that the score of a theory must not depend on the numbering on the names of the decision alternatives. Elongational invariance assures that the score of a theory is not influenced by adding or leaving an alternative which is predicted with a probability of zero. Incentive compatibility requires that predicting the actual probabilities yields the highest score. Finally, neutrality means that in the comparison of two theories among which one is right in the sense that it predicts the actual probabilities and the other is wrong the score for the right theory does not depend on which of the two theories is the right one. This means that the score does not prejudge one of the theories depending on the location of the theory in the space of probability distribution.

We apply the quadratic scoring rule to measure the predictive success of a theory for every period and subject separately and then add up over subjects, rounds and games. Accordingly a score depending on the predicted probabilities and the actually observed action is computed. In order to compute the score the observation is interpreted as a frequency distribution where for the chosen action the relative frequency is one and for the not chosen action zero. Thus the quadratic score $q(t)$ of a learning theory for subject choosing action i in period t is given as:

$$q(t) = 2p_i(t) - p_i(t)^2 - (1 - p_i(t))^2$$

Here $p_i(i)$ is the predicted probability of the learning theory. The predicted probability of the learning theory is calculated by applying the theory's learning algorithm on the whole playing history of this player. If no history is available we assume that the player randomizes with .5.

The concepts of action-sampling learning and payoff-sampling learning always return the probability of one for one of the possible actions. Which action is chosen depends on the randomly drawn sample. Therefore we calculate the probability of drawing a sample that commands playing action 1 or action 2 as the predictions of theses two concepts.

If a player decides completely in line with the prediction of the theory he receives a score of 1 if he decides in complete contrast to the prediction the theory he receives a score of -1.

The mean score $\bar{q}$ is given as the mean of $q(t)$ over all 200 rounds, 12 games and 108 subjects groups of 8 subjects each. Of course $\bar{q}$ must be in the closed interval between -1 and $+1$.

IV Results

In this section we will first take a look at the simulations and the experiments on the aggregate level. We will start with the relative frequencies for U and L observed in the simulations with the different learning models and compare them with the experimental data. Then we will take a closer look at the simulations and start by comparing the results obtained in the constants sum games with the results in the non-constant sum games. Afterwards we will investigate how the learning models perform in the original matrices and in the transformed matrices. Thereafter we will compare the overall mean quadratic distances to the experimental data. We will conclude our examination on the aggregate level by testing the robustness of the overall result over time and therefore compare the performance of the learning rules in the first and second 100 rounds.

The second part of this section deals with the individual behavior. There we will check for the subjects in the 2×2 experiments how well they conform in the average to each of the learning theories.

A Aggregate Behavior

Table 1 gives the observed mean frequencies for each game and simulation type and as well as the observed ones in the experiments. For the experimental games 1 to 6 the mean frequency observed in a game is based on the observed frequencies in twelve independent matching groups, for games 7 to 12 it is based on the observed frequencies in six independent matching groups. Each matching group consists out of eight subjects. For each learning type and game the mean is based on 500 simulation runs, which produced 500 independent matching groups per game. Each matching group consists of eight agents.

As mentioned before games 1 to 6 and games 7 to 12 have the same best response structure. The concept of action-sampling equilibrium depends only on this structure and therefore leads in Selten & Chmura to the same predictions in the constant and an non-constant sum games. Since action-sampling learning is based on best replies, it does not surprise, that the frequencies in the simulations with games 1-6 are very similar to those with games 7-12. For all other learning models different frequencies are observed in the constant and non-constant sum games.

It is surprising that self-tuning EWA yields relative frequencies very near to .5 for

		Impulse	Impulse	Action-	Reinforce-	Payoff-	self-tuning	Experiment
Game		balance	matching	sampling	ment	sampling	EWA	Selten & Chmura
		learning	learning	learning	learning	learning	learning	(2008)
1	L	.417	.574	.658	.342	.741	.501	.690
	U	.164	.063	.067	.126	.052	.501	.079
2	L	.417	.495	.589	.332	.514	.492	.527
	U	.283	.168	.231	.159	.069	.508	.217
3	L	.594	.770	.744	.498	.893	.519	.793
	U	.227	.157	.173	.135	.156	.483	.198
4	L	.581	.712	.656	.589	.854	.513	.736
	U	.309	.258	.343	.188	.315	.484	.286
5	L	.535	.631	.656	.554	.799	.507	.664
	U	.350	.297	.342	.233	.370	.492	.327
6	L	.539	.600	.529	.660	.765	.505	.596
	U	.420	.401	.407	.271	.464	.491	.445
7	L	.474	.638	.659	.392	.778	.539	.564
	U	.198	.099	.066	.164	.087	.468	.141
8	L	.485	.563	.589	.389	.752	.515	.586
	U	.337	.257	.230	.212	.254	.483	.25
9	L	.602	.770	.744	.530	.862	.538	.827
	U	.248	.185	.174	.164	.183	.449	.254
10	L	.602	.727	.656	.636	.850	.530	.699
	U	.335	.301	.342	.207	.308	.463	.366
11	L	.560	.647	.656	.606	.792	.521	.652
	U	.383	.354	.342	.287	.373	.474	.331
12	L	.556	.604	.528	.590	.631	.520	.604
	U	.458	.465	.406	.344	.603	.473	.439

Table 1: Relative frequencies observed in simulations and experiments for U and L

each of the twelve games. This is probably connected to the fact, that we estimate the free parameter of this model jointly for all games. However as we have already pointed out estimating parameters for each game separately would not be adequate.

Of all learning models only impulse matching learning and and action-sampling learning are quite close to their stationary counterparts after 200 periods. The quadratic distance between impulse matching learning and impulse balance equilibrium is smaller than 0.001 and the quadratic distance between action-sampling learning and action-sampling equilibrium is 0.004. The other distances between a learning rule and the related equilibrium are much greater, impulse balance learning (0.018), payoff-sampling learning (0.046), and reinforcement learning (0.159). Self-tuning EWA has a much higher distances towards all stationary concepts.

A.1 Constant Sum and Non-Constant Sum Games

Table 1 shows that the behavior of the subjects in the experiments differ in the constant (games 1 - 6) and non-constant sum games (games 7 - 12). Therefore, we will start comparing the predictive success of the learning models in constant and non-constant sum games. Figure 4 gives the mean quadratic distance in constant and non-constant games for each learning theory.

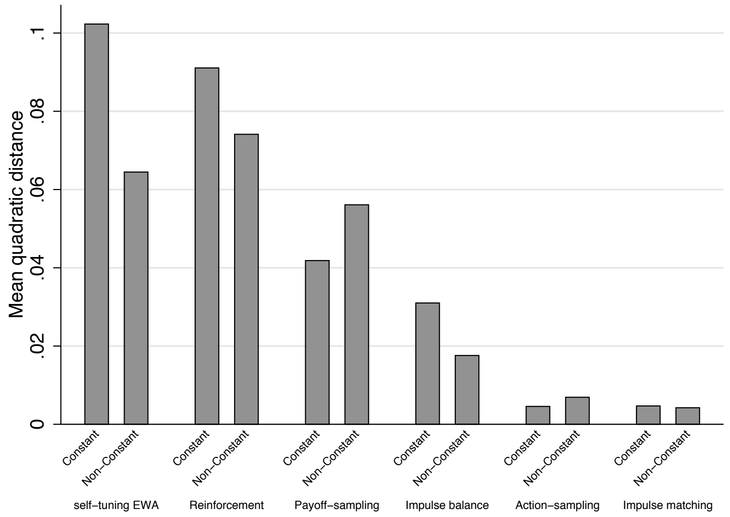

Figure 4: Mean quadratic distance in constant and non-constant sum games

The models of self-tuning EWA, reinforcement learning and impulse balance learning perform much better in the non-constant sum games. The concept of impulse matching learning performs slightly better in the non-constant sum games. In contrast, the two learning rules relying on samples, namely action-sampling learning and payoff-sampling learning, perform better in the constant sum games.

A.2 Original Versus Transformed Games

The concepts of impulse balance learning and impulse matching learning are applied to the transformed game rather than the original one. But the ideas behind these concepts could also be applied directly to the original games as well as the other concepts could be applied to the transformed games. Figure 5 shows the overall mean quadratic distances for self-tuning EWA learning, reinforcement learning, payoff-sampling learning, impulse balance learning, action-sampling learning and impulse matching learning applied to the original games and to the transformed games.

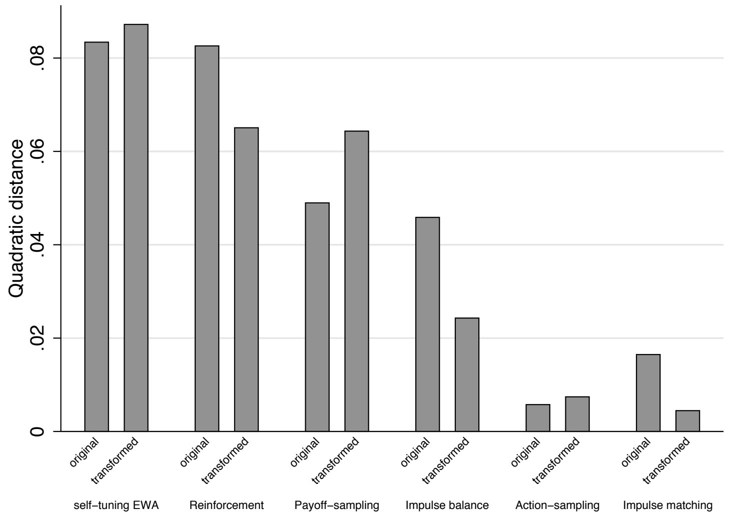

Figure 5: Mean quadratic distance in original and transformed games

For the original as well as for the transformed matrixes the results are based on 500 simulation runs per game and learning model.

It can be seen that impulse balance learning, impulse matching learning and reinforcement learning perform better when applied to the transformed games whereas self-tuning EWA learning, payoff-sampling learning and action-sampling learning do less well. While the improvement of impulse balance learning and impulse matching learning in transformed games is expected, the benefit of applying reinforcement learning to transformed games is unexpected. This improvement is substantial, in the original game the quadratic distance is 22% higher than in the transformed ones.

The theory of Roth and Erev (1998) already applies a transformation of the original game by replacing the payoff of a player by it's difference to the minimal value in her matrix. The transformation used here is different since it involves double weights for losses with respect to the pure strategy maximin. Nash equilibrium is the stationary concept corresponding to the reinforcement learning theory. However, in Selten & Chmura (2008) we did not observe an improvement of the predictive power of the Nash equilibrium applied to the transformed game rather the original one. It is interesting that the picture looks different for the simulations over 200 rounds.

A.3 Overall Comparison

Figure 6 gives the mean of the quadratic distance between the experiment and simulations over all games and rounds for self-tuning EWA learning, reinforcement learning, payoff-sampling learning, impulse balance learning, action-sample learning and impulse matching learning.

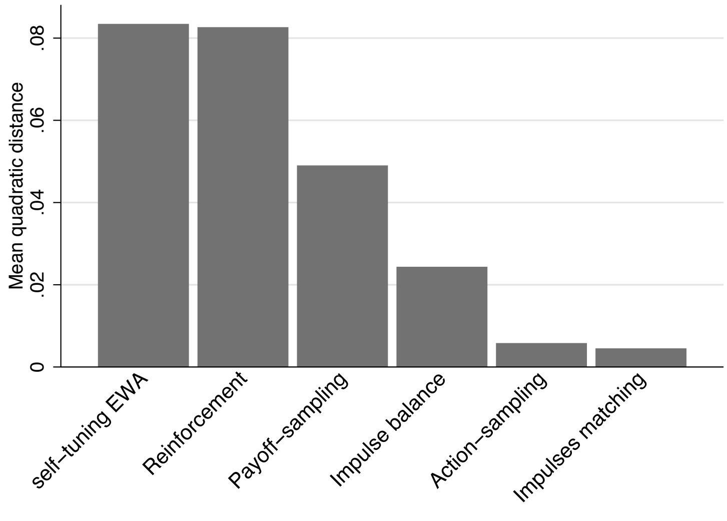

Figure 6: Overall mean quadratic distance over all games

The figure reveals an order of explanatory power. The order from worse to best (highest quadratic distance to lowest quadratic distance) is as follows: self-tuning EWA learning, reinforcement learning, payoff-sampling learning, impulse balance learning, action-sampling learning and impulse matching learning.

The difference between self-tuning EWA and reinforcement is very small and irrelevant. However the small difference between the two quadratic deviations does not mean that both theories make similar predictions. This can be seen in table 1. Recall figure 4, which demonstrates that self-tuning EWA performs better than reinforcement learning in the non-constant sum games, while reinforcement learning performs better in the constant-sum games.

The figure demonstrates that the concepts of self-tuning EWA and reinforcement fail to describe the aggregate behavior in the 2 × 2 experiments in contrast to the other concepts. Out of these new concepts especially the processes of action-sampling learning and impulse matching learning lead to results which are very close to subjects' behavior. Already the concept of payoff-sampling learning has a nearly 40% lower quadratic distance than self-tuning EWA and the quadratic distance of impulse matching is over 18 times smaller.

The order given by figure 6 is statistically robust. Because of the high number of observations, 6000 per learning type, all differences (even the slight ones between self-tuning EWA and reinforcement) are statistically significant on a high level (for all $p < 0.001$ two-sided Man-Whitney u-test).

A.4 Changes over Time

Learning processes are always dependent on time and history and therefore it is of interest to check whether our above results remain stable over time. To check stability of the order of explanatory power over time we compare the first hundred periods with the second hundred periods. Figure 7 gives the mean quadratic distances for periods 1-100 (left) and 101-200 (right) for the six learning models. Basis of the comparison is always the observed mean frequencies for the corresponding rounds (either round 1-100 or 101-200) in the experiments.

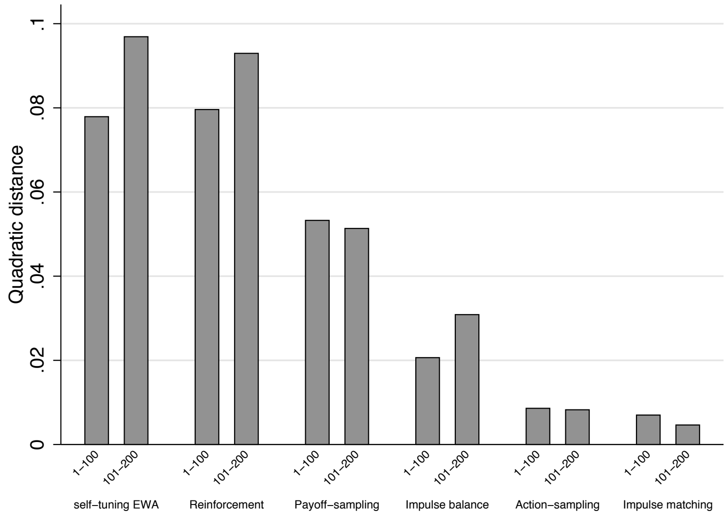

Figure 7: Mean quadratic distance over time

It is easy to recognize that in the second half of the simulation runs the explanatory power of self-tuning EWA, reinforcement learning and impulse balance learning decreases. For payoff-sampling learning and impulse matching learning the performance increases in the second half. The concept of action-sampling learning is rather stable over time and no relevant differences are observed over time.

For all theories the quadratic distance in the first and second half of the experiment differs significantly (two-sided Wilcoxon signed-rank test $p < 0.0000$). For action-sampling this is mainly due to the high number of observations.

The comparison over time confirms that the order of explanatory power obtained in the overall comparison. This order is stable as far as the better performing concepts of payoff-sampling learning, impulse balance learning, action-sampling learning and impulse matching learning are concerned. Only the direct comparison of reinforcement learning and self-tuning EWA changes over time. While self-tuning EWA performs better in the first half (rounds 1-100) reinforcement learning performs better in the second half (101-200).

B Individual Behavior

In this section we will take a closer look at subjects' decisions and whether they are in accordance with one of the learning theories. Therefore we will use the quadratic scoring rule, as introduced in section III.C. Recall, that in contrast to the quadratic distance the higher the value of the quadratic score the better the fit is.

In addition to the investigated learning rules we introduce one heuristic which we call the inertia rule. This rule commands to *"do exactly the same as in the preceding round"*. Of course this does not apply to the first period in which both possible actions are chosen with equal probabilities. The player is required to repeat the decision of the preceding period even if he deviated from this rule in the past. Obviously, the inertia rule is not a serious decision rule, but it serves as a benchmark that every learning rule should beat.

Figure 8 shows box plots of the mean quadratic scores in the 108 independent observations for each learning model and the inertia benchmark. The plot gives the median (the horizontal line in the box), the interquartile range (the box around the median), with the .75 percentile as the upper limit of the box and the .25 percentile as the lower limit. The whiskers describe the observations in the sample which are outside the inter quantile range and finally the dots describe the outliers, which are defined as values smaller or greater the 1.5 times lower or upper inter quantile range.

The boxes in figure 8 are ordered from the highest median to the lowest and exactly the same order occurs if the models were ranked by the means. The plot reveals a clear order of predictive success, from best to worst: inertia rule, reinforcement learning, self-tuning EWA, impulse matching learning, action-sampling learning, payoff-sampling learning and impulse balance learning. The plot shows that the performance of the inertia benchmark is not only driven by the mean, but also by the median, the inter quantile range and by

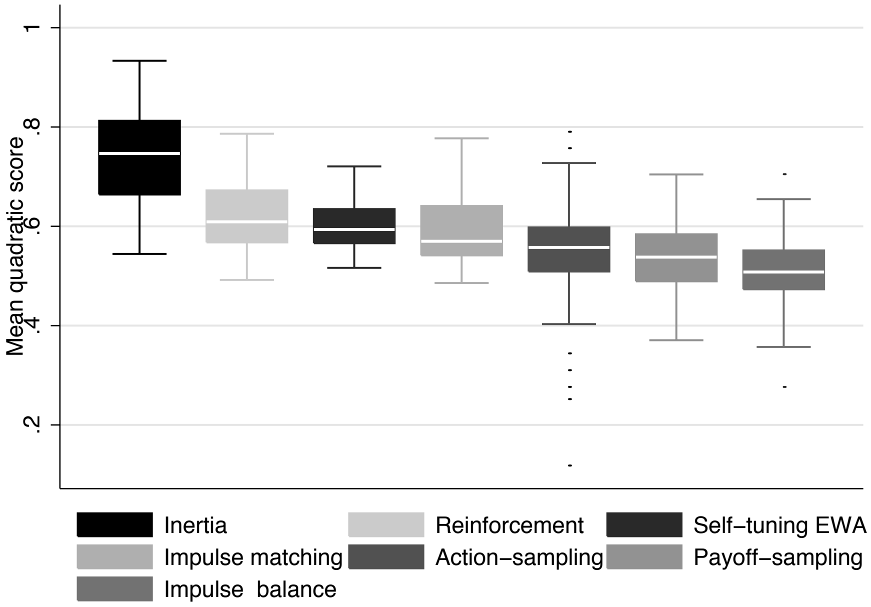

Figure 8: Box plots over the mean quadratic scores in the 108 observations with different learning models

the highest single mean score in one observation.

In 90 observations out of 108 independent observation groups the inertia benchmark has the highest score and in 9 cases it has the highest score together with the reinforcement learning. Self-tuning EWA, impulse matching learning and action-sampling learning obtain the highest mean score in three observations each. The mean score of the inertia benchmark is nearly 20% higher than the score of reinforcement learning and more than 45% higher than the score of impulse balance learning. Applying a two-sided Wilcoxon signed-rank test for the pairwise comparison of the mean scores over the independent observations reveals that the order given by the plot is statistically robust. All pairwise comparisons between two models are at least significant on the 1% level.

It is very remarkable that the inertia rule performs significantly better than all learning theories. Obviously all learning theories fail to meet the benchmark of the inertia rule. At least at first glance this is a devastating result. We must conclude that the learning theories do not really describe individual behavior.

Checking our data we find two main reasons for the failure of the learning rules. First, players tend to repeat a chosen action for quite a time before they switch to the other action. All the theories fail to incorporate this inertia. Of all the learning theories reinforcement learning performs the best probably because it has the the highest probability of not changing the decision (on the individual basis reinforcement lead to an overall probability of not changing of $p = .6271$). The second reason is that all investigated theories, although they might correctly model the systematic reasons for a change of choice, fail to forecast when exactly it will occur.

V Discussion

In this paper the models of impulse matching learning, impulse balance learning, action-sampling learning and payoff-sampling learning have been introduced and together with reinforcement learning and self-tuning EWA applied and tested in the environment of repeated 2 × 2 games.

The newly introduced learning models are based on the behavioral reasoning of payoff-sampling equilibrium, action-sampling equilibrium and impulse balance equilibrium, which had been successfully tested in experimental 2×2 games by Selten & Chmura (2008). Therefore the experimental dataset obtained by Selten & Chmura (2008) were used as a testbed for the learning models. The experimental data comprises aggregate and individual behavior in 12 completely mixed 2×2 games, 6 constant sum games with 12 independent subject groups each, and 6 nonconstant sum games with 6 independent subject groups each. Each subject group consists of eight participants being randomly matched over 200 periods.

The learning models had to prove whether they can replicate the aggregate behavior of the experimental population and whether they can explain the individual behavior of single subjects. For the comparison with the aggregate behavior 500 simulation runs per game and learning model were conducted. As in the experiment, 200 rounds with random matching and four agents deciding as row players and four agents as column players were simulated. Our measure of predictive power for the aggregate is the quadratic distance between observed relative frequencies in simulation runs and the mean frequencies observed in the experiments. For the comparison with the individuals' behavior the models were applied to the history of each participant. Then the actual decisions of every round were compared with the predictions of the learning models given the subject's history. For each subject and round a quadratic score, a measurement for the accuracy of a prediction, was calculated and averaged over rounds, subjects and games.

For our comparisons with the aggregate and the individual behavior we can conclude two main results:

Main Result 1: *The models of learning are able to replicate the aggregate behavior in our* 2×2. *In our study the models of impulse matching learning and action-sampling learning prove to be especially successful.*

The comparison of the six models yields the following order of predictive success from best to worst: Impulse matching learning, Action-sampling learning, Impulse balance learning, Payoff-sampling learning, Reinforcement learning, Self-tuning EWA learning. Due to the high number of simulation runs, this order is statistically robust, all pairwise comparisons with the two-sided Man-Whitney u-test are at least significant on the 0.1% level.

The predominance of the new models, impulse matching learning, action-sampling learning, impulse balance learning and payoff-sampling learning, over the established models of reinforcement learning and self-tuning EWA is stable over time and across the different game types (constant sum and non-constant sum games). One possible reason for the predominance of the new models, especially over self-tuning EWA is that we insisted on adjusting parameters as less as possible. A further interesting result is that for reinforcement learning the quadratic distance to the data is round about 22% lower if applied to the transformed matrixes instead to the original ones.

Main Result 2: *The models of learning are not able to adequately replicate the individual behavior in our* 2×2 *games. A simple inertia rule outperforms the sophisticated learning models.*

Overall we must conclude that all investigated learning models fail to describe the individual behavior. Although all models performed better than simple randomization with .5, they failed to beat our benchmark heuristic, which commands to *"do exactly the same as in the preceding round"*.

It may be the case that a learning theory is correct as far as the systematic reasons for a change of strategy are concerned, but nevertheless the exact timing of changing a strategy are very different from individual to individual. Moreover the timing may be influenced by the attention of the subject for the task which probably varies over time. This means that it depends on personality feature and uncontrollable influences from recent experiences outside the laboratory. However, the failure of the learning theories on the individual level does not mean that they are useless for the description of group behavior.

Actual learning algorithms are obviously not capable to describe individual human behavior, while they are able to describe cumulated human behavior in an appropriate way. Observing individual behavior is similar to observing an ant trail: Though one can describe the direction of the trail, it is hard to forcast the behavior of a single individual.

We are confident that our results are stable for a broad set of 2 × 2 games, yet our concepts still have to prove their power in other settings with different games.

Literatur

Brier, Glenn W. 1950. "Verification of Forecasts Expressed in Terms of Probability", *Monthly Weather Review*, 78(1): 1-3.

Brown James N. and Robert W. Rosenthal. 1990. "Testing the minimax hypothesis: A re-examination of O'Neill's game experiment", *Econometrica*, 58: 1065-1081.

Camerer, Colin F. and Teck H. Ho. 1999. "Experience-weighted attraction Learning in Normal Form Games", *Econometrica*, 67: 827-874.

Erev, Ido, and Alvin E. Roth. 1998. "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique Mixed Strategy Equilibria." *American Economic Review*, 88(4): 848-81.

Estes, William K. 1954. "Individual Behavior in Uncertain Situations: An Interpretation in Terms of Statistical Association Theory." In *Decision processes*, ed. Robert M. Thrall, C. H. Coombs, and R. L. Davis, 127-137, New York: Wiley.

Harley, Calvin B. 1981. "Learning the Evolutionarily Stable Strategy", *Journal of theoretical Biology*, 89(4): 611-633.

Kahneman, Daniel and Amos Tversky. 1979. "Prospect Theory: An Analysis of Decision under Risk", *Econometrica*, 47(2): 263-291.

Kalain, Ehud and Ehud Lehrer. 1993. "Rational Learning leads to Nash Equilibrium", *Econometrica*, 61(5): 1019-1045.

McKelvey Richard D. and Thomas R.Palfrey. 1995. "Quantal Response Equilibria for Normal Form Games", *Games and Economic Behavior*, 10(1): 6-38.

Osborne Martin J. and Ariel Rubinstein. 1998. "Games with Procedurally Rational Players." *American Economic Review*, 88(4): 834-47.

Roth, Alvin E. and Ido Erev. 1995. "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term", *Games and Economic Behavior*, 8: 164-212.

Selten, Reinhard. 1998. "Axiomatic Characterization of the Quadratic Scoring Rule", *Experimental Economics*, 1: 43-62.

Selten, Reinhard, and Joachim Buchta. 1999. "Experimental Sealed Bid First Price Auctions with Directly Observed Bid Functions." In *Games and human Behavior: Essays in the honor of Amnon Rapoport*, ed. David Budescu, Ido Erev, and Rami Zwick, 79-104. Mahwah NJ: Lawrenz Associates.

Selten, Reinhard, and Thorsten Chmura. 2008. "Stationary Concepts for Experimental 2×2-Games", *American Economic Review*, 98(3): 938-966.

Teck H. Ho, Colin Camerer and Juin-Kuan Chong. 2007. "Self-tuning experience weighted attraction learning in games", *Journal of Economic Theory*, 133: 177-198.

Appendix - Not For Publication

Impulse matching and impulse balance

To show that the concepts of impulse balance equilibrium and impulse matching equilibrium lead to the same stationary points in case of the 2×2 games, we take a look at the structure of the investigated experimental 2×2 games, as introduced by Selten & Chmura (2008).

	L	**R**
U	$a_L + c_L$ b_U	a_R $b_U + d_U$
D	a_L $d_D + d_D$	$a_R + c_R$ b_D

Figure 9: The Structure of the Experimental 2x2-Games

The figure shows the transformed payoffs, the payoffs for the column-players are shown in the lower right corner and the payoff for the row-palyers are shown in the upper left corner. The following equations must be fulfilled: $a_L, a_R, b_U, b_D \geq 0$ and $c_L, c_R, d_U, d_D > 0$. In the following p_U and p_D are the probabilities of the row player for U and D and q_L and q_R are the probabilities for L and R by the column player. In the following we will only look at the row player, the behavior in equilibrium of the column player is calculated analogously.

In case of impulse balance equilibrium the expected impulses for each of the both strategy must be the same. Hereby, the row player receives only an impulse towards U for the proportion of plays in which he would choose down (given by p_D) and the other player at the same time would have chosen L (given by q_L). Therefore the expected impulse for U is given by $p_D q_L c_L$. Applying the same reasoning leads to $p_U q_R c_R$ as the expected impulse for D of the row player. Thus the impulse balance equation, which must be fulfilled in equilibrium is given as:

$$p_D q_L c_L = p_U q_R c_R$$

In case of impulse matching equilibrium, the row player receives always an impulse of c_L towards U if the column player plays L. The column player does so with a probability of q_L. In addition the row player always receives an impulse of c_R towards D if the column player choses R. The column player plays R with a probability of q_R. Impulse matching equilibrium is reached if the ratio of the two probabilities of U and D is the same as the ratio of expected impulses for U and D.

$$\frac{p_U}{p_D} = \frac{q_L c_L}{q_R c_R}$$

By transforming we obtain the impulse balance equation of impulse balance equilibrium:

$$p_D q_L c_L = p_U q_R c_R$$

Therefore, impulse matching equilibrium and impulse balance equilibrium have the same mixed stationary points in case of the described 2 × 2 games.

Self-tuning EWA

The decisions by the the players are done according to attractions A for each strategy. The probability for the k-th strategy of player n is calculated with a logit response function:

$$p_{nk}(t+1) = \frac{e^{\lambda A_{nk}(t)}}{\sum_{j=1}^{2} e^{\lambda A_{nj}(t)}}$$

where λ is the response sensitivity. λ is the parameter of this learning theory and must be specified to fit to the empirical data. In our case $\lambda = 0,079$ minimized the distance to the experimental data. The attractions are updated as described by the *EWA attraction updating function*:

$$A_{nk}(t) = \frac{\phi N(t-1)A_{nk}(t-1) + [\delta + (1-\delta)I(s_{nk}, s_n t)]\pi_n(s_{nk}, s_m(t))}{N(t)}$$

$I(x,y)$ is an indicator function, which is one if $x = y$ and zero if $x \neq y$. $N(t)$ is the *experience weight* and updated according to $N(t) = N(t-1)\phi(1-)\kappa + 1$. ϕ is a decay rate and detects changes in the learning environment. The *change-detector function* $\phi_n(t)$ is:

$$\phi_n(t) = 1 - 0.5(\sum_{k=1}^{2}[\frac{\sum_{\tau=t-W+1}^{t} I(s_{mk}, s_m(\tau))}{W} - \frac{\sum_{\tau=1}^{t} I(s_{mk}, s_m(\tau))}{t}]^2).$$

The W is the number of strategies played with positive probability in the Nash equilibria, in our case $W = 2$. The first term in the brackets counts how often strategy k was played by the others in periods $t - W + 1$ to t and divides by W. The second term is the relative frequency of the k-th strategy played over all periods. The forgone payoffs are weighted with δ which is calculated as $\delta_n(t) = \phi(t)/W$. The growth rate of the attractions is controlled by the *exploitation parameter* κ. Ho, Camerer and Chong calculate κ as a Gini coefficient of the probability inequity. In our 2×2 games this leads to the simple function of $\kappa_n(t) = 1 - 2\ min(p_{n1}, p_{n2})$.

The initial values of the functions are: $\phi(0) = \kappa(0) = 0.5$ and $\delta(0) = \phi(0)/W$ Over time these initial values are weighted with $\frac{1}{t}$ and the current function value with $\frac{t-1}{t}$ thus the influence of the initial values decreases over time. Thus the actual functions are given as: $\phi\hat{(t)} = \phi(0)\frac{1}{t} + \phi(t)\frac{t-1}{t}$; $\kappa\hat{(t)} = \kappa(0)\frac{1}{t} + \kappa(t)\frac{t-1}{t}$; and $\delta\hat{(t)} = \delta(0)\frac{1}{t} + \delta(t)\frac{t-1}{t}$

The initial attraction-level were set to $A_1 = 0$ and $A_2 = 0$.