Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria

By Ido Erev and Alvin E. Roth*

We examine learning in all experiments we could locate involving 100 periods or more of games with a unique equilibrium in mixed strategies, and in a new experiment. We study both the ex post ("best fit") descriptive power of learning models, and their ex ante predictive power, by simulating each experiment using parameters estimated from the other experiments. Even a one-parameter reinforcement learning model robustly outperforms the equilibrium predictions. Predictive power is improved by adding "forgetting" and "experimentation," or by allowing greater rationality as in probabilistic fictitious play. Implications for developing a low-rationality, cognitive game theory are discussed. (*JEL* C72, C92)

Game theory has traditionally been developed as a theory of strategic interaction among players who are perfectly rational, and who (consequently) exhibit equilibrium behavior. This approach has been complemented by evolutionary game theory, which, motivated by biological evolution, seeks to understand how equilibria could arise in the long term by selection among generations of players who need not be rational or even conscious decision makers. Somewhere in between are models of learning, which consider the adaptive behavior of goal-oriented players who may not be highly rational, both to provide foundations for theories of equilibrium and to model empirically observed behavior.

The present paper considers how well simple learning models, motivated by the psychology of learning, can model the interaction of players who must learn about the game and each other in the course of playing the game, over time spans that may not be long enough to lead to equilibrium. Our goal will be to model observed behavior, starting with behavior observed in experimental settings. (In the conclusion we will also consider the implications of this approach for applied economics in naturally occurring, nonexperimental settings.) We will show that a wide range of experimental data can be both well described *ex post* and robustly predicted *ex ante* by a very simple family of learning theories.

Economists have traditionally avoided explaining behavior as less than rational for fear of developing many fragmented theories of mistakes. Part of the attraction of highly rational models is the idea that there may be many ways to be less than rational, but only one way (or in light of the equilibrium refinement literature perhaps only a few ways) of being highly rational. In this view, the success in economics of the assumptions of utility maximization and equilibrium behavior is in large part due to the prospect that they may provide a useful approximation of great generality, even if they are not precisely correct models of human behavior (cf., Roth, 1996a).

* Erev: Faculty of Industrial Engineering and Management, Technion, Haifa, Israel 32000, and Department of Economics, University of Pittsburgh, Pittsburgh, PA 15260 (e-mail: erev@techunix.technion.ac.il); Roth: Department of Economics, Harvard University, Cambridge, MA 02138, and Harvard Business School, Boston, MA 02163 (e-mail: aroth@hbs.edu; http://www.economics.harvard.edu/faculty/roth/roth.html). The work of both authors is partially supported by grants from the National Science Foundation. We have benefitted from helpful conversations with Yoella Bereby-Meyer, Nick Feltovich, Daniel Gopher, Joachim Meyer, Ayala Cohen, Dan Hamermesh, and Shmuel Zamir. Yoella Bereby-Meyer also contributed to the design and programming of the new experiment. We are indebted to Barry O'Neill, Jack Ochs, and Amnon Rapoport for access to unpublished parts of their data. The present version reflects numerous comments by three anonymous referees on several earlier drafts. This work was completed while Roth was at the University of Pittsburgh.

Similarly, the development of learning theories of considerable generality will be most likely if it turns out that learning does not have to be modeled in a fundamentally different way in each game. One of the chief purposes of the present paper is to investigate whether this is likely to be the case. As we will see, the evidence supports the conjecture that a simple model of learning may have quite general application. We will also discuss some limitations of the models presented here, and our conjecture that these may primarily have to do with the sometimes complex strategy space in which relatively simple kinds of learning may be going on. (Just as commodities and states of the world need to be carefully modeled if utility theory is to be useful as a general tool, a general theory of learning will not free us from the need to model specific environments.)

Learning in strategic environments presents some phenomena not found in individual decision-making because the environment in which each individual gains experience includes the other players, whose behavior changes as they, too, gain experience. At least in the intermediate term, the effect of experience appears to depend on features of the strategic environment different from those which determine equilibrium: Experience leads to quick convergence to equilibrium in certain games, but has little effect in other games with similar equilibria.[1] Experience even appears to lead behavior *away* from equilibrium in certain matrix games with mixed strategy equilibria considered in the present paper.

In Roth and Erev (1995) we showed that a simple model of individual learning could capture this range of behavior. We considered three games with similar perfect equilibrium predictions, in two of which experimental subjects were observed to converge quickly to the perfect equilibrium prediction, while no sign of such convergence was observed in the third. The learning model we studied exhibited the same kind of behavior in each game as the experimental subjects did, and did so using the same parameter choices for all three games, in a way that helped explain why games with similar equilibria might elicit different behavior. (Games with similar equilibria may be quite different from one another away from equilibrium, and so players who start away from equilibrium may learn very different things.)

Recently there have been a number of other papers which compare the predictions of various learning models to the learning observed experimentally in games.[2] Collectively these papers powerfully begin to make the case that learning models have great potential for describing observed behavior. Some of these papers also use the observed behavior in a game to compare different learning models, most typically by fitting the parameters of each model to the data, and testing which provides the best fit for each game studied. They generally support the idea that models in which individuals perform probabilistically in ways that respond to their experience are likely to outperform simple deterministic models.

The present paper builds on this emerging consensus. In keeping with our goal of studying the robustness and predictive power of learning models, we will take a somewhat different approach. We explore the possibility that a simple reinforcement learning model can be used to predict, as well as explain, observed behavior on a broad range of games, without fitting parameters to each game. We start with the basic one-parameter model examined in Roth and Erev, and then ask which psychological assumptions have to be added to the basic model in order to more accurately account for the observed behavior.

[1] For experimental data see e.g., the market and ultimatum games studied in Roth et al. (1991), both of which have a unique subgame-perfect equilibrium which gives all the wealth to one side of the market, and both of which have other equilibria which support the full range of distributions between the two sides. Behavior in the market game robustly and quickly converged to the perfect equilibrium, while behavior in the ultimatum game, equally robustly, showed no signs of approaching the perfect equilibrium. For a comprehensive survey of experimental results, see the *Handbook of Experimental Economics* (John Kagel and Roth, 1995).

[2] See for example Yin-Wong Cheung and Daniel Friedman (1995, 1996), James Cox et al. (1995), David Cooper and Nick Feltovich (1996), Fang-Fang Tang (1996a, b, c), and Colin Camerer and Teck-Hua Ho (1998a, b).

We concentrate first on a class of games for which the necessary psychological assumptions may be simple and easy to quantify, namely repeated matrix games with unique, mixed strategy equilibrium in which repetition does not create opportunities for players to cooperate.[3] For this purpose, we have assembled and analyzed a data set consisting of all experiments we could locate involving play of 100 periods or more of games with a unique equilibrium in nontrivial mixed strategies.[4] The reason for looking for so many periods of play is to observe intermediate-term as well as short-term behavior. The data sets we have assembled report repeated play of 11 games, under a variety of experimental conditions, from the experiments of Patrick Suppes and Richard C. Atkinson (1960); David Malcolm and Bernhardt Lieberman (1965); Barry O'Neill (1987); Rapoport and Richard B. Boebel (1992); Jack Ochs (1995). For the experiments from the 1960's, we use data at a useful level of disaggregation contained in the published reports, and for the others we have obtained data at the individual level from the authors. We also consider a new, twelfth data set, from an experiment we conducted on one of the games studied by Suppes and Atkinson, to examine the robustness of some of the observed results to different experimental conditions.

Games with a unique, mixed strategy equilibrium present a difficult test case, both for theories of equilibrium and of learning, because at equilibrium no player has positive incentives to play the equilibrium probabilities. (But away from equilibrium some player has positive incentives to change his behavior.) Another reason for looking at games with a unique equilibrium is that finite repetition of such games does not increase the set of equilibria, so the repeated game has a unique equilibrium, which can be achieved in stage-game strategies. Thus in principle the stage-game strategies may be adequate to model the strategy sets of the players of the repeated game. And the experiments in this data set were designed to concentrate on stage-game strategies. Nine of the 12 games are constant sum (and so finite repetition does little to enlarge the scope for cooperation or retaliation compared to the stage game). And the three nonconstant sum games were played under conditions (to be described) which limited the use of repeated-game strategies.

Ideally we would like to be able to predict behavior at every level of aggregation or disaggregation, for every game, for any length of play. Since the models we consider are computational, we can use them to simulate each experiment and predict the probability of each action at each period. We will then compare the predictions of different learning models and of equilibrium by computing the mean-squared deviation (MSD) of the predicted and observed behavior, period by period, for each game, both for all subjects and for individual pairs (when individual-level data are available). For each model and each of the 12 experimental data sets we consider we will perform two tests of descriptive power and one test of predictive power, as follows. First, we will find the best parameters for minimizing the MSD over all games, and compute the MSD for each game using these parameters. Then we will find the best parameters for minimizing the MSD for each of the 12 games separately (i.e., by looking at a model which replaces each parameter of the original model with 12 distinct parameters, one for each game). Finally we will test the predictive power of each model on each of the 12 games, by estimating the model's parameters on the data from the other 11 games, using the model to predict behavior in the game of interest, and comparing the predicted path of behavior with the observed path.

[3] In Section VI we briefly consider games in which players can reciprocate. In follow-up studies we consider games and individual decision tasks with dominant strategies. These latter studies show that the reinforcement learning models considered here also do well in capturing individual learning phenomena observed in games with pure strategy equilibrium and in (individual choice) probability learning experiments (see a review of this literature in Wayne Lee, 1971). Yoella Bereby-Meyer and Erev (1997) consider this individual choice literature in a study built upon the current results. Other follow-up studies focus on individual learning in a complex task with delayed outcomes (Erev et al., 1997) and probabilistic signals (Erev, 1998).

[4] A "nontrivial" mixed strategy is one in which at least two strategies are played with positive probability. We also excluded a game with a unique, mixed strategy equilibrium in which all strategy choices were equally likely (Amnon Rapoport and David V. Budescu, 1992).

The main results of this paper will be that a one-parameter reinforcement learning model outperforms the equilibrium prediction for all values of its one parameter. The model's descriptive and predictive power is further improved by incorporating (into a three-parameter reinforcement model) psychological assumptions about experimentation and forgetting that facilitate responsiveness to a changing environment (i.e., an adaptive opponent). We also consider a four-parameter belief-based model which explicitly adds to the reinforcement model responsiveness to a changing environment in the manner of probabilistic fictitious play, and show that it, too, improves on the one-parameter reinforcement model, although not on the three-parameter reinforcement model.

The paper is organized as follows. Section I presents the 12 experimental data sets. We observe that: (1) in 5 of the 12 games equilibrium predicts badly: average choice probabilities, pooled over all rounds, are closer to random choices than to the equilibrium predictions;[5] (2) initial learning trends often move away from the equilibrium predictions; (3) in most cases of initial movement away, behavior moves towards the equilibrium after sufficiently long play; and (4) there is large between-pair variability that is not eliminated by experience.

Section II motivates the reinforcement learning approach, and evaluates a basic one-parameter model. This section demonstrates that the basic model robustly (over the entire range of its parameter) outperforms the equilibrium predictions and captures the initial learning trends, but it fails to account for the late direction change and the between-pair variability.

Section III examines the value of adding to the model the two additional parameters introduced in Roth and Erev to model "experimentation" and "forgetting." Both parameters contribute to the model's descriptive power. The three-parameter model captures the conditions under which the direction of the learning trend is changed, and accounts for the observed between-pair variability.

Section IV examines a modification of the basic model that makes it more like belief-based models of learning (e.g., probabilistic fictitious play in the manner of Drew Fudenberg and David K. Levine, 1997b). This includes an information parameter (that determines the extent to which subjects respond to information beside the payoffs they have actually received; see Camerer and Ho, 1998a), a maximization parameter (that determines subjects' tendency to optimize), and a habit parameter (that weighs previous actions). On our data, only the habit parameter contributes to the model's descriptive power. Consistent with this conclusion, within-subject analysis reveals that individual subjects are better described as reinforcement learners than expectation learners on this class of games. Comparisons of all the models' predictions are made in Section IV on the aggregate data, and in Section V on the individual data.

Section VI briefly considers the case of games for which the present models will have to be extended. The most challenging of these will be games in which repetition creates opportunities for cooperation, for which a more detailed investigation of the empirically observed repeated-game strategies will be needed. We consider how such an investigation will be related to research in cognitive psychology.

Section VII discusses how the kind of adaptive models we consider here might contribute to applied economics, and Section VIII concludes.

I. The Data

A few words are in order about why we concentrate on data gathered by other experimenters. One of the great benefits of experimental economics is that investigators can easily collect new data well designed to test particular hypotheses. However there is a danger that investigators will treat the models they propose like their toothbrushes, and each will use his own model only on his own data. More subtly, there is a danger that in making the

[5] Note again that our sample of games is not a random sample, rather it is a sample of games selected by a variety of experimenters, which elicited widely varying behavior, including specifically good and bad performance of the equilibrium predictions.

many decisions that go into an experimental design, an investigator will unconsciously be guided, by the same intuition which motivates the model he considers, to make design choices that promote behavior of the kind predicted by the model. (This danger is only partially attenuated when an investigator selects experiments done by others, if there is room for his intuition to guide which experiments are selected.) Thus, while in Roth and Erev we explored data we had generated ourselves, in the present paper we chose to "tie our hands" by exploring the entire set of available experiments concerning long runs of games with unique equilibria in mixed strategies. These were conducted under widely varying experimental conditions, by investigators with widely varying theoretical dispositions (and who reached quite different conclusions from one another on the basis of their observations). In particular, the data were collected under an unusually wide range of conditions involving the information of the players (from full information in some treatments to others in which participants did not even know they were playing a game), and the manner in which they were paid (from monetary to nonmonetary rewards, delivered deterministically or stochastically). These data thus provide a universe on which we can test claims of robustness regarding both the games and the conditions under which they are played.

A. *The Aggregate Learning Curves and the Equilibrium Predictions*

For each of the experiments to be described next (except the new one which is described separately), the left-hand column in Figures 1–3 presents the aggregate experimental results and the equilibrium predictions (the right-hand columns are simulation results to be discussed later). The payoff matrices are presented at the left of the figures. Each cell within the figure's frame is a graph that has the probability of a certain choice (ranging from 0 to 1) on the Y axis, and the rounds of the experiment (organized into blocks as in the data of that experiment) on the X axis. For the 2×2 games, the mean probability with which players 1 (row players) and players 2 chose their first strategy (A) is plotted over time in Figures 1 and 3. For the games with more strategies the choice probabilities of the asymmetrical strategies are presented in Figure 2. Player 1 choices are indicated by triangles, player 2 choices by squares. The equilibrium predictions for players 1 and 2, respectively, are given by the triangle and square at the far right of each cell in column 1.

Suppes and Atkinson (1960)—minimal information: The top four rows in Figure 1 present experiments conducted by Suppes and Atkinson to test their "Stimulus-Sampling" theory of learning. This theory can be interpreted as an even simpler reinforcement learning model than those studied here, in that it is limited to the case of two possible outcomes—"reinforcement," and "no reinforcement."

Suppes and Atkinson assumed that being "correct" is a reinforcing event, and did not use monetary rewards in the treatments we consider. In each trial of the experiments considered below, subjects were asked to choose between two keys, and then (within a few seconds) received a binary feedback (indicating whether they were correct or not). The feedback was probabilistically determined by the payoff matrix. For example, in the condition whose payoff matrix is presented in the top row of Figure 1, each payoff unit increases the probability of a "correct response" feedback by $1/6$.[6]

Suppes and Atkinson studied the effect of the payoff matrix and of subjects' information about it. Four of their experimental conditions involved matrix games with unique, mixed strategy equilibrium, and these conditions are considered here.

The top row of Figure 1 corresponds to the "mixed strategy" experimental condition in Chapter 3 of Suppes and Atkinson (first described in Atkinson and Suppes, 1958). The

[6] Because each payoff in the game matrix is a probability of being reinforced, this is an early example of a binary lottery payoff. Although it was not intended in this case to control for hypotheses involving expected-utility maximization, it has the effect of allowing us to interpret the predictions of such hypotheses (of which mixed strategy equilibrium is one) without having to worry about risk aversion. For a history of the use of binary lottery designs to allow the predictions of expected-utility hypotheses to be interpreted unambiguously, see Roth (1995), particularly pages 40–49 and 81–85.

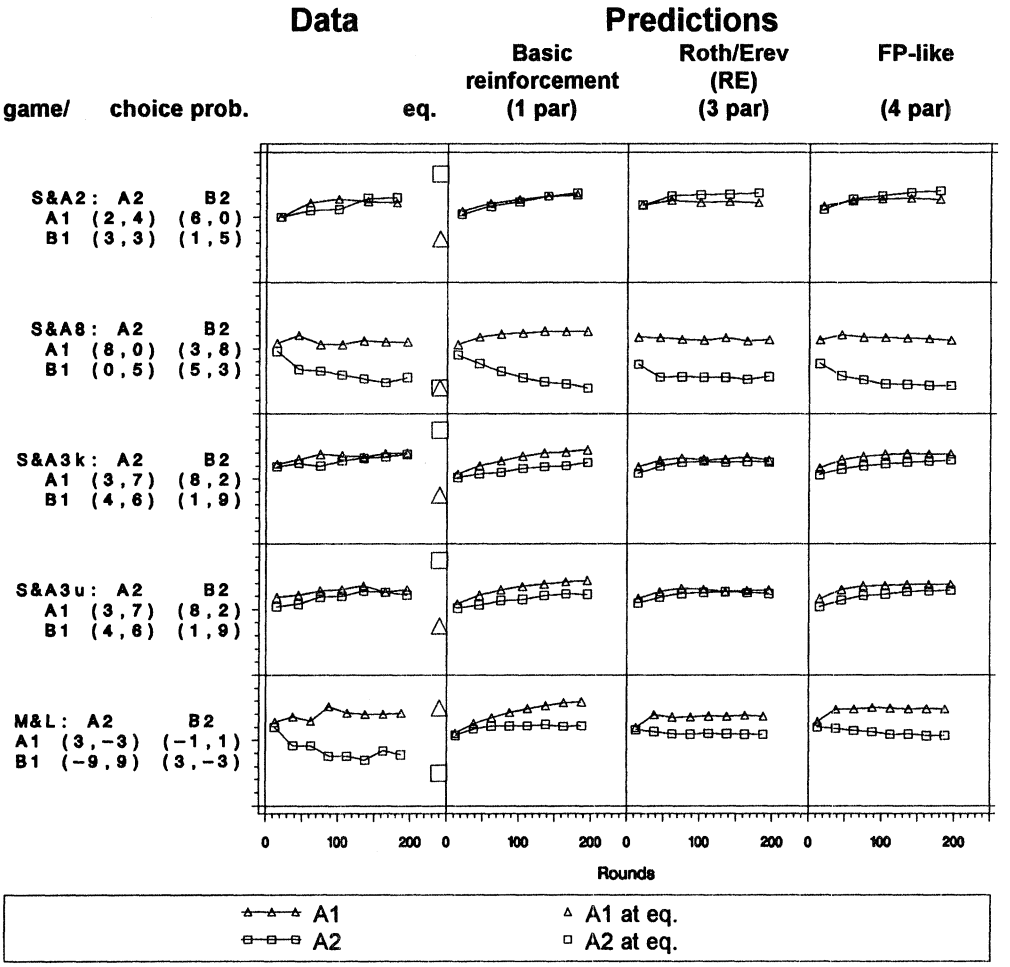

FIGURE 1. REPEATED 2 × 2 GAMES (SUPPES AND ATKINSON [S&A], 1960; MALCOLM AND LIEBERMAN [M&L], 1965)

Notes: In the top four games each payoff unit increases the probability of winning (by $^1/_6$ in S&A2, by $^1/_8$ in S&A8, and by $^1/_{10}$ in S&A3k and S&A3u). In M&L payoffs were directly converted to money. Each cell in the left-hand column presents the experimental results: The proportion of *A* choices over subjects in each role (grouped in 5 to 8 blocks) as a function of time (200–210) trials in all cases. The three right-hand columns present the models' predictions in the same format. The equilibrium predictions are presented at the right-hand side of the data cells.

game played in this condition, referred to here as game S&A2, has a unique, mixed strategy equilibrium in which player 1 chooses A1 with probability $^1/_3$ and player 2 chooses A2 with probability $^5/_6$. It was played by 20 pairs of subjects for 200 rounds. The subjects were not informed that they were playing a two-person game. They were told that their task, in each of the 200 trials, was to predict which of two lights will be turned on. Subjects were run in pairs and, as described above, the probability of a "correct" response was determined by the game payoff matrix. Thus, although the subjects did not know that they were playing a game, the game is a description of the reinforcement structure.

Suppes and Atkinson presented the choice proportions in blocks of 40 trials. The results (see the data in the top-left panel of Figure 1) show that player 2 appears to move toward the

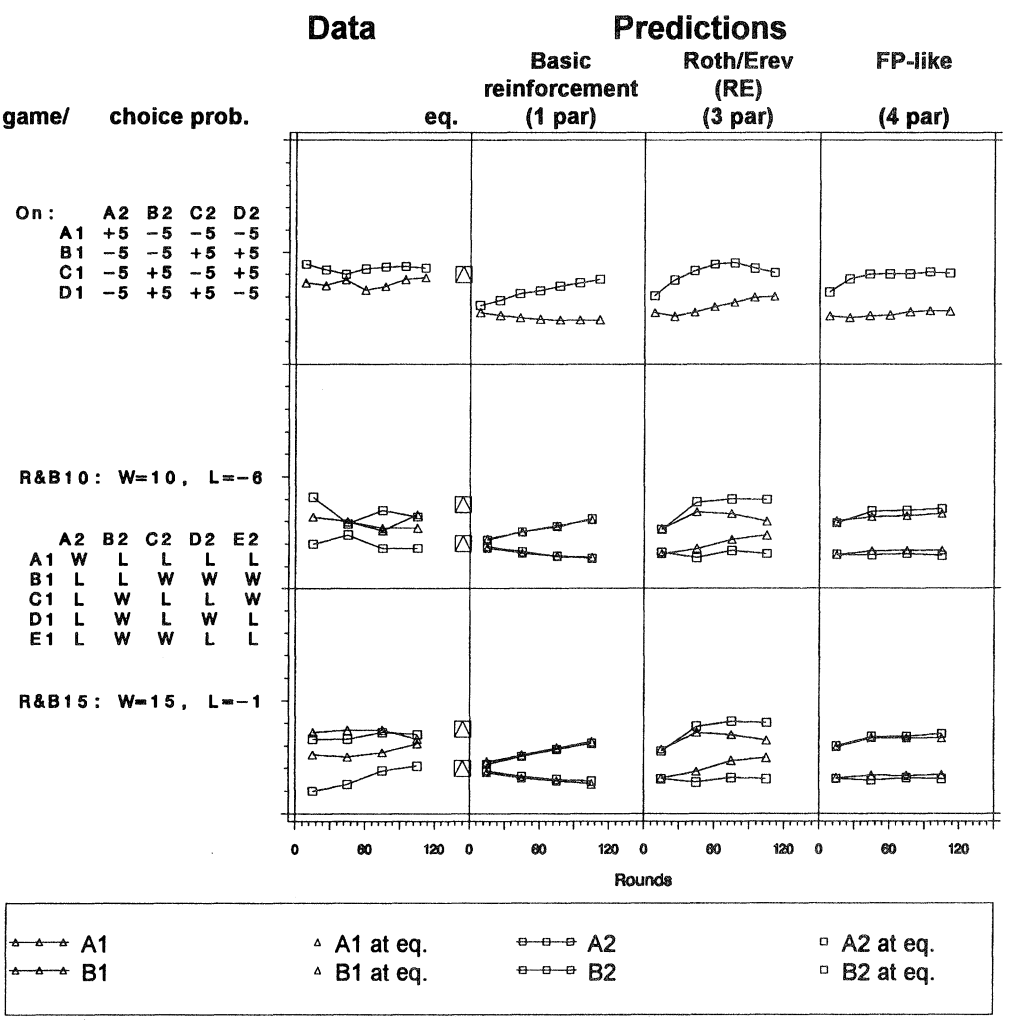

FIGURE 2. REPEATED 4 × 4 (O'NEILL [On], 1987) AND 5 × 5 (RAPOPORT AND BOEBEL [R&B], 1992) GAMES

Note: The curves show predicted and observed choice probabilities (7 blocks of 15 trials in game On, and 4 blocks of 30 trials in the R&B games).

equilibrium prediction (the proportion of A2 choices increases with time). Player 1 initially moves away from the equilibrium. Only in the last two blocks is the proportion of A1 choices reduced.

The data graph in the second row of Figure 1 summarizes the results of a condition reported in Chapter 4 of Suppes and Atkinson, in which players knew they were playing a game, but did not know the payoff matrix. At the equilibrium of this game (S&A8) both players choose *A* with probability 0.2. This game was played by 20 pairs of subjects for 210 rounds. Subjects were told that they were playing a two-person game in which they were to predict which of two lights would turn on. They were told that the correct answer depended on their response, on the other subject's response, and on a random event. As in game S&A2, the probability of a "correct" response was determined by the payoff matrix, which was not presented to the subjects.

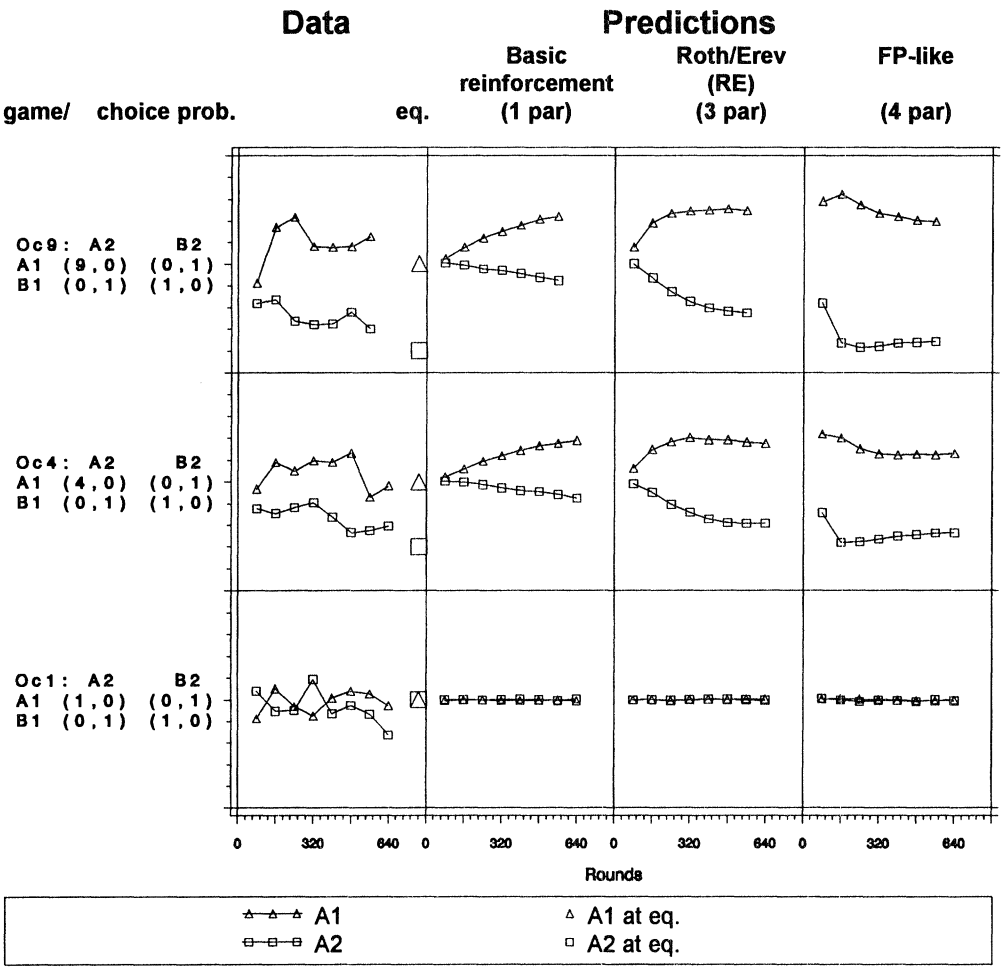

FIGURE 3. TWO-POPULATION 2 × 2 GAMES (OCHS [Oc], 1995)

Note: Each cell presents the probability of *A* choices in blocks of 80 trials (7 blocks in game Oc9, and 8 blocks in the other games).

The results (summarized by the proportions of *A* choices in blocks of 30 trials) are similar to the results obtained in game S&A2. Whereas one of the players (player 2) quickly learns to approach the equilibrium prediction, the other (player 1) initially moves away from the equilibrium.

Suppes and Atkinson (1960)—matrix effect: In order to evaluate the effect of explicit presentation of the payoff matrix, Suppes and Atkinson compared two experimental conditions. The control group played game S&A3 (see Figure 1) under the game/prediction condition in which game S&A8 was played. Since the game was unknown to the subjects in this condition, we refer to it as game S&A3u. The experimental condition, referred to as S&A3k, was identical with the exception that the payoff matrix was known to the subjects. Twenty pairs participated in game S&A3u, and 20 pairs participated in game S&A3k. Both games were run for 210 rounds (and the data in Figure 1 are presented in 7 blocks of 30 trials). At equilibrium player 1 chooses A1

with probability $^3/_8$, and player 2 chooses A2 with probability $^7/_8$.

Suppes and Atkinson observed a very weak effect due to the presentation of the payoff matrix on the overall choice probabilities. In both groups, one of the subjects (player 2) moves toward the minimax prediction, whereas the second subject (player 1) moves away.

A replication study—To evaluate the robustness of Suppes and Atkinson's results we ran a replication of condition S&A3u (Suppes and Atkinson, 1960) with two procedural variations: our subjects were paid for their performance and were run for more (500) rounds. This game was selected because it provides the sharpest contrast between the predictions of the equilibrium and experimental results.

The subjects were 20 undergraduate students at the Technion. They were run in ten pairs. Both pair members were seated in front of the same computer. They were separated by a plastic divider so that each member could see only half of the screen. They received 10 Shekels (about $3) for showing up, and were told that they would play a game on the computer in which they could earn more money. In each of the game's 500 trials they had to select one of two keys.[7] They were told that their choice and that of the other player determined their probability of winning, and that on the average whenever one person wins the other person loses. The payoff matrix was identical to the probabilistic matrix of the game S&A3, with a win worth 0.01 Shekel and a loss worth 0. The payoff matrix was not shown to the subjects. Each subject could see on the screen his/her cumulative and last trial payoffs. The 500 rounds took about 45 minutes. The average final payoff was 35 Shekels ($12).

The results of the replication study, referred to as S&A3n, practically coincided with the original study. For the 210 periods that were run in the original study, the average distance between the replication and the original learning curve measured by mean-squared deviation was only 0.2. And in the remaining 290 rounds both curves continued with a slow, noisy upward movement. The proportion of *A* choices in the last 100 trials was 0.65 for player 1, and 0.67 for player 2. Learning curves of individual pairs in this study are presented in Figure 4 and discussed shortly. [Note added in proof: While checking the galleys we detected a minor bug in the computer program used to run the replication study, that occurred when a player pressed the two keys in a single trial. This occurred in about 3 percent of the trials and was more or less uniformly distributed over pairs and over time. We ran another replication study after correcting this bug. The results were practically identical.]

Malcolm and Lieberman (1965): The study conducted by Malcolm and Lieberman was designed to test the descriptive power of the minimax model. The payoff matrix was explained to the subjects, and the payoff units were chips that were converted to money at the conclusion of the experiment. Nine pairs of subjects participated in 200 replications of the game.

The fifth row in Figure 1 presents the payoff matrix (game M&L) and the results. At the equilibrium of this game player 1 chooses A1 with probability $^3/_4$, and player 2 chooses A2 with probability $^1/_4$. Malcolm and Lieberman presented the choice proportions in blocks of 25 trials. Experience led both players toward the equilibrium prediction, but player 1 appears to learn faster, reaching equilibrium by the forth block, whereas player 2 approaches equilibrium only slowly.

O'Neill (1987): O'Neill argued that the research conducted prior to his study cannot be used to reject the minimax prediction because it involves strong additional assumptions. For example, Suppes and Atkinson explicitly assumed that "being correct" has a utility, and Malcolm and Lieberman assumed that the utilities are linear in money. O'Neill designed a careful experiment that avoids these assumptions. Twenty pairs of subjects played a 4×4 zero-sum matrix game (see top of Figure 2) for 105 rounds. The game was described as a simple card game. In each round, each subject chose a card (that stood for one of the four strategies) and the payoff was determined by the payoff matrix presented at the top of Figure 2 (which was verbally explained to the subjects). Because each subject can receive one of only two possible payoffs, there is no op-

[7] ⟨*A*⟩ or ⟨*Z*⟩ for the subject on the left, and ⟨6⟩ or ⟨3⟩ on the numerical key pad for the player on the right.

portunity for choices by expected-utility maximizers to be influenced by nonlinearities (risk preferences) in their utility functions.[8] Note that for both players three of the four strategies (B, C and D) are symmetrical. At equilibrium both players are expected to choose A with probability 0.4, and to choose each of the other strategies with equal probability (0.2).

At the aggregate level, the results (summarized in Figure 2 by the proportion of A choices in blocks of 15 trials) appear to support the static equilibrium prediction.

Rapoport and Boebel (1992): Rapoport and Boebel utilized O'Neill's careful design to study behavior in two versions of a 5×5 constant sum matrix game (bottom panels in Figure 2). In the first experimental session ten pairs of subjects played the game for 120 rounds under each of two payoff conditions. (The subjects then exchanged roles and played another 120 rounds in a second session. The data obtained in the second session are not presented here.[9] At equilibrium both players choose strategy A with probability $3/8$, strategy B with probability $2/8$, and each of the remaining (symmetrical) strategies with probability $1/8$. Rapoport and Boebel compared two experimental conditions. In the condition referred to as R&B15, W (player 1's profit in case of a "win") was 15, and L (player 1's profit in case of a "loss") was -1. In the condition referred to as R&B10, $W = 10$ and $L = -6$. The results are summarized in Figure 2 by the proportion of A and B choices in blocks of 30 rounds.

Unlike O'Neill's results, Rapoport and Boebel's results do not conform so closely to the equilibrium prediction. Yet, some movement toward equilibrium is observed.

[8] That is, when players play mixed strategies, all of the induced lotteries are binary lotteries.

[9] Rapoport and Boebel found no significant difference in behavior between sessions 1 and 2. And the models we consider here (when reinitialized at the beginning of session 2) did equally well at describing the observed behavior in either session. However the learning in these games (in both sessions) is fairly flat, so we do not want to suggest that for other games the behavior of experienced players who switch roles can be captured by our model without at least some attention to the effect of their prior experience.

Ochs (1995): Ochs' subjects were asked to state the proportion of "A" choices that they wished to make in the next ten games. Subjects were run in cohorts of eight players in each position for 56–64 trials of ten simultaneous games per trial, and they accumulated lottery payoffs to be used at the end to determine cash payoffs via a binary lottery payoff mechanism. In each trial, players were matched to new opponents using a quasi-random mechanism.

Three games were compared (see Figure 3). The equilibrium prediction implies that player 1 should choose strategy A with probability $1/2$ in all three conditions. Player 2 is predicted to choose strategy A with probability $1/10$ in the top game (Oc9), with probability $2/10$ in the middle game (Oc4), and with probability $1/2$ in the symmetrical game (Oc1). One cohort was run under each condition (game).

The experimental results are summarized in Figure 3 by the proportion of A choices in blocks of eight trials (80 games). Although Ochs' experimental design is very different from that of Suppes and Atkinson, his results show the main trends observed in their studies. In games Oc9 and Oc4 one of the two players (1) starts to move away from the equilibrium, and later moves back slowly.

B. *Individual Learning Curves*

One of the most interesting features of O'Neill's data set is that although the aggregate choice probabilities are very close to equilibrium, individual players' choices are not (cf., James N. Brown and Robert W. Rosenthal, 1990). So analysis of individual learning curves can reveal information that is lost in the analysis of the aggregate curves.

The two cells in the top row of Figure 4 present five, randomly selected, individual pairs from the O'Neill experiment and from our replication of the Suppes and Atkinson experiment. The X axis in each cell is the frequency of A2 choices by player 2 and the Y axis is the frequency of A1 choices by player 1. The right-hand cell presents five of the curves in O'Neill's data set. Each data point (seen as a point at which the curve changes direction) presents the average frequency over 35 rounds. The first block is marked by a triangle, the last block is marked by a dot, while

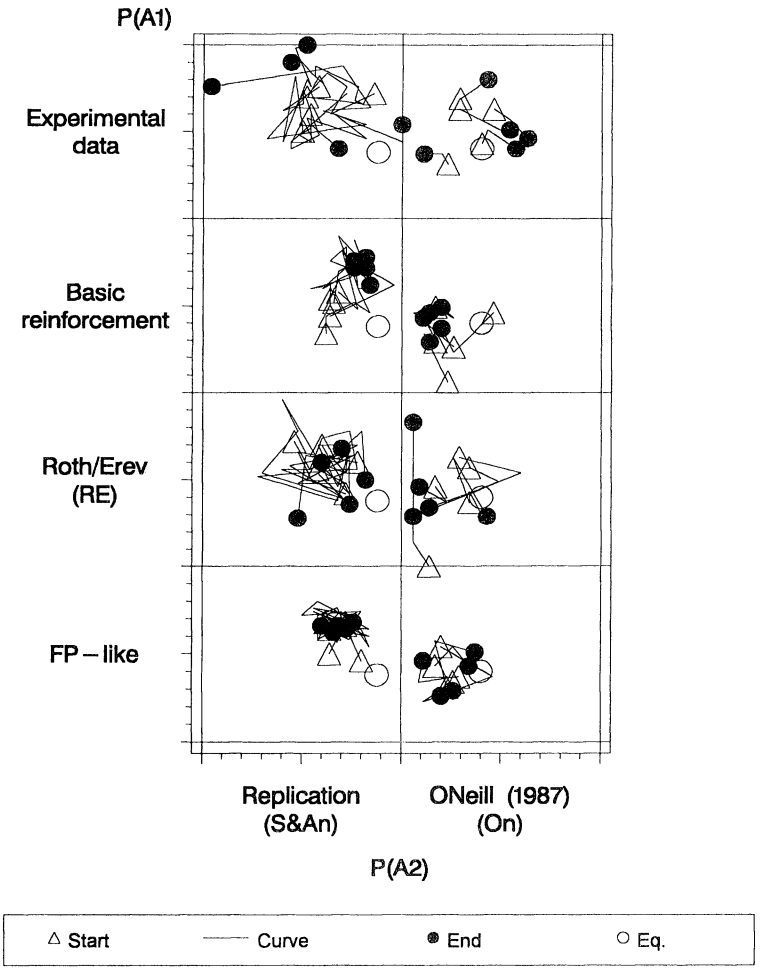

FIGURE 4. OBSERVED AND PREDICTED LEARNING CURVES OF RANDOMLY SELECTED INDIVIDUAL PAIRS IN THE REPLICATION STUDY (S&A3n) AND O'NEILL'S (On) GAMES

Note: Each curve shows the probability of A1 as a function of the probability of A2 for one of the pairs (in blocks of 50 trials for S&A3n and 35 trials for On).

the equilibrium is marked by a circle. The left-hand cell presents five of the individual pairs in the replication study (S&A3n) using the same format (in 10 blocks of 50 trials). Examination of these curves reveals high between-pair variation that is not diminished with time.

C. *Summary*

Four summary observations seem worthy of note. First, in some of the games the equilibrium prediction does very badly. This impression can be quantified by comparing the distance between the experimental curves, the equilibrium, and the random choice prediction (that all strategies will be chosen equally often). The distances were measured for each game by mean-squared deviation (multiplied by 100) scores (see Reinhard Selten [1998] for a discussion of this measure). The MSD scores were first calculated for each data block (as presented in Figures 1–3) and then averaged across blocks and player type in each ex-

periment. The results are presented in the top two rows of Table 1. Random choice beats equilibrium in 5 of the 12 games, and they tie in the sixth game. Over all games the equilibrium MSD score is 3.57 while the random choice score is 1.87, reflecting the fact that in the games where equilibrium does badly compared to random choice, it does very badly.[10]

Second, in most studies one of the two players initially moved away from the equilibrium predictions. But the equilibrium has some descriptive power. Whereas the learning away for the equilibrium is robust (i.e., continued 450 trials in game S&An), a late direction change towards the equilibrium prediction was observed in most cases. Finally, a robust between-pair variability that is not diminished with experience was observed even when the aggregate choice probabilities are at equilibrium.

II. Reinforcement Learning

Our main point of departure is the optimistic conjecture that the robust characteristics of human and animal learning behavior described in the psychological literature concerning individual decision makers may lead to a robustly descriptive model of human learning in strategic environments also. Thus, we start our attempt to account for the behavioral results described above with a simple model based on the most robust characteristics of individual behavior. We will proceed to augment the model with further psychological assumptions only after identifying aspects of the data that are not predicted by the simpler model.

Examination of the psychological learning literature led us (in Roth and Erev) to consider the following two basic principles as a starting point in the search for a model to approximate learning in games.

The Law of Effect (Edward L. Thorndike, 1898): Choices that have led to good outcomes in the past are more likely to be repeated in the future; and

The power law of practice (see e.g., J. M. Blackburn, 1936[11]): Learning curves tend to be steep initially, and then flatter.

Note that implicit in the law of effect is an additional psychological principle: *choice behavior is probabilistic.* This is one of the basic assumptions of most mathematical learning theories proposed in psychology (e.g., William K. Estes, 1950; Robert Bush and Frederick Mosteller, 1955; Duncan R. Luce, 1959; Suppes and Atkinson, 1960).

Where learning in games will differ from the individual learning literature is that we will have to concentrate on the behavior of *populations* of subjects, both when they are inexperienced and as they gain experience. In strategic environments the behavior of other subjects forms an important part of the environment faced by each subject.[12] Because different subjects may behave differently (especially when they are inexperienced) an important part of the environment may therefore be stochastic.

We begin with a basic, one-parameter model which will provide some initial benchmarks for comparisons with equilibrium and with slightly more elaborate models. The basic model also permits us to lay out the general framework (following Luce) within which all of the reinforcement models we consider are constructed.

A. *The General Framework, and a Basic, One-Parameter Reinforcement Model*

Initial propensities—At time $t = 1$ (before any experience has been acquired) each player n has an initial propensity to play his kth pure strategy, given by some nonnegative number

[10] The fact that random choice beats equilibrium in the games S&A3u, S&A3k, and S&A3n means that it does so for a fixed matrix played both with and without players' knowledge of the matrix, and with and without their being paid in cash for their outcomes, so the phenomenon is not isolated among those games with low information or among those without cash payoffs, but appears to be a property of the matrix of payoffs.

[11] As described in E. R. F. W. Crossman (1958).

[12] While the same can be said of evolutionary game theory, note that what makes learning potentially quite different from selection is the power law of practice, which has no parallel in modern theories of evolution (or in the replicator dynamics with which evolutionary game theory is most often studied). Individuals may learn more slowly as they gain experience, but there is no evidence that populations of organisms evolve more slowly.

TABLE 1—MSD SCORES (100 × MEAN-SQUARED DEVIATION—SMALLER IS BETTER) BETWEEN THE DIFFERENT PREDICTIONS (ROUND n OF THE DATA COMPARED TO ROUND n OF THE PREDICTION) AND THE EXPERIMENTAL RESULTS BY GAME AND AVERAGED OVER GAMES

Game: Model	S&A8	S&A2	S&A3u	S&A3k	S&A3n	M&L	On	R&B15	R&B10	Oc9	Oc4	Oc1	Mean over all games
Random	1.08	2.04	2.53	1.46	2.11	2.46	2.19	1.07	1.38	3.88	1.78	0.45	1.87
Equilibrium	6.92	7.18	7.27	7.56	6.14	2.11	0.14	0.45	1.03	2.22	1.37	0.45	3.57
Basic reinforcement:													
best fit (1 parameter)	0.16	0.30	0.31	0.11	0.57	2.27	1.81	0.98	0.73	2.71	1.54	0.48	1.00
by game (12 parameters)	0.07	0.24	0.14	0.10	0.41	1.89	0.33	0.50	0.16	2.34	1.54	0.41	0.68
prediction	0.16	0.30	0.31	0.11	0.57	2.27	1.81	0.98	0.86	2.71	1.64	0.48	1.02
RE:													
best fit (3 parameters)	0.38	0.18	0.12	0.07	0.31	1.24	0.72	0.65	0.33	1.54	1.09	0.48	0.59
by game (36 parameters)	0.05	0.10	0.04	0.05	0.25	0.21	0.32	0.35	0.11	1.34	0.99	0.37	0.35
prediction	0.67	0.26	0.19	0.09	0.39	1.24	0.87	0.83	0.48	1.54	1.17	0.51	0.69
FP-like:													
best fit (4 parameters)	0.34	0.20	0.16	0.09	0.37	1.26	1.05	0.71	0.44	2.04	1.48	0.42	0.71
by game (48 parameters)	0.05	0.09	0.04	0.03	0.29	0.45	0.04	0.14	0.17	1.70	1.19	0.28	0.37
prediction	0.77	0.44	0.24	0.09	0.37	1.54	1.24	0.71	0.44	2.10	1.65	0.45	0.84

Notes: Each of the first 12 columns of the table represents one of the games. The first two rows present the MSDs of the random choice and equilibrium predictions. Each of the other three panels summarizes the fit of one of the learning models. The first row in each panel displays the MSDs for the model in question using the parameters used in Figures 1–3. The second row shows the MSDs when the parameters are separately estimated for each game. The third row shows the accuracy of the prediction of the model when behavior in each of the 12 games is predicted based on the parameters that best fit the other 11 games. The final column gives the average MSD over all games, for each case, which is a quick summary statistic by which the models can be roughly compared.

$q_{nk}(1)$. In our basic model, each player will be assumed to have equal initial propensities for each of his pure strategies, i.e., for each player n,

$$q_{nk}(1) = q_{nj}(1) \text{ for all pure strategies } k, j.$$

A reinforcement function—The reinforcement of receiving a payoff x is given by an increasing function $R(x)$. In the basic model, we take the reinforcement function to be

$$R(x) = x - x_{\min},$$

where $x_{\min}$ is the smallest possible payoff.

Updating of propensities—If player n plays his kth pure strategy at time t and receives a reinforcement of $R(x)$, then the propensity to play strategy j is updated by setting

$$(1) \quad q_{nj}(t+1) = \begin{cases} q_{nj}(t) + R(x) & \text{if } j = k \\ q_{nj}(t) & \text{otherwise.} \end{cases}$$

Probabilistic choice rule—The probability $p_{nk}(t)$ that player n plays his kth pure strategy at time t is

$$(2) \qquad p_{nk}(t) = q_{nk}(t)/\Sigma q_{nj}(t),$$

where the sum is over all of player n's pure strategies j.

Equation (2) is precisely Luce's linear probabilistic response rule. Note that the model satisfies the law of effect and the power law of practice. Pure strategies which have been played and have met with success tend over time to be played with greater frequency than those which have met with less success, and the learning curve will be steeper in early periods and flatter later [because nonnegative reinforcements imply $\Sigma q_{nj}(t)$ is an increasing function of t, so a reinforcement of $R(x)$ from playing pure strategy k at time t has a bigger effect on $p_{nk}(t)$ when t is small than when t is large].

This learning model has a certain resemblance to evolutionary dynamics (cf., John Maynard-Smith, 1982) even though they are not the "replicator" dynamics customarily associated with evolutionary models. (In fact this basic model was proposed as a quantification of the law of effect by Richard J. Herrnstein [1970], and considered as an approximation of evolutionary dynamics by Calvin B. Harley, 1981.[13]) The chief point of similarity with evolutionary dynamics is that the influence of other players' past behavior on any player n's behavior at time t is via the effect that their behavior has had on player n's past payoffs.

The single parameter of the basic model—It follows from the probabilistic choice rule [equation (2)] and our assumption that each player's initial propensities are all equal that at the initial period of the game each player chooses each of his strategies with equal probability. However, we have not made any assumption which fixes the sum of the initial propensities, which appears in the denominator of equation (2), and therefore influences the rate of change of choice probabilities, i.e., the speed of learning (which is also influenced by the size of the rewards). The basic model's sole parameter, $s(1)$, which we will call the *strength* of the initial propensities, is introduced to determine the ratio of these two determinants of the learning speed. Let X_n be the average absolute payoff for player n in the game. The initial strength parameter for player n is defined as $s_n(1) = \Sigma q_{nj}(1)/X_n$, and we assume that this is a constant for all players, i.e., $s_n(1) = s(1) > 0$ for all players n.

Note that this definition and the probabilistic choice rule yield the initial propensities $q_{nj}(1) = p_{nj}(1)s(1)X_n$, where $p_{nj}(1)$, the initial choice probability is given by $p_{nj}(1) = 1/M_n$, where M_n is the number of player n's pure strategies. Thus the initial propensities are determined by the observable features of the game and by the strength parameter $s(1)$.

Derivation of predictions—To derive the model's predictions for the experiments described above we conducted computer simulations designed to replicate the characteristics of each of the experimental settings. In each case the simulated players "participated" in the same number of rounds as the experimental subjects. Two hundred simulations were run for each game under different sets of parameters. At each round of each simulation the following steps were taken:

(i) Simulated players were matched (using the matching procedure of the experiment being simulated).
(ii) The simulated players' strategies were randomly determined via equation (2).
(iii) Payoffs were determined using the payoff rule employed in the experiment in question.
(iv) Propensities were updated according to equation (1).

Parameter estimation—A grid search with an MSD criterion was conducted to estimate the value of the free parameter, $s(1)$. That is, the simulations were run for a wide set of parameters, and the parameter that minimized the distance between the model and the data

[13] A family of closely related reinforcement models is studied by Tang (1996b), who finds they compare favorably to a number of other learning models in describing observed behavior in an experiment he considers. Another related model (in which probabilities are generated without propensities) was suggested by Bush and Mosteller (1955), and has been studied in economic contexts by John G. Cross (1983), W. Brian Arthur (1991, 1993), Michael W. Macy (1991), Tilman Borgers and Rajiv Sarin (1994, 1995), and Dilip Mookherjee and Barry Sopher (1997). This latter model may in fact be more closely related to replicator dynamics (see Jorgen Weibull, 1995), since it does not obey the power law of practice, i.e., since learning does not become slower over time.

(minimized the model's MSD score) was selected for each of the tests presented below.

B. *Aggregate Description and Prediction*

Best fit—The second column in Figures 1–3 presents the predictions of the basic learning model with the estimated parameter that best fit the data over all 12 games [$s(1) = 54$]. The distance (MSD score) between the best fit and the data by game and averaged over games is summarized in the first row of the basic model's statistics in Table 1. (The parameters were chosen to minimize the average score over all games: the MSDs reported in this row for each game all have the same parameter value.)

Note that the model's average score (1.0) is less than 30 percent of the equilibrium score. Since these scores represent error they imply that the equilibrium's error is more than three times the model's error. Yet, Table 1 also shows that only in 6 of the 11 games in which the predictions differ (all models considered here have the same predictions for game Oc1) does the basic model outperform the equilibrium prediction.

Sensitivity analysis—Sensitivity analysis reveals that the advantage of the model over the equilibrium is robust to the choice of the free parameter. The model's MSD score is below 1.5 (less than 50 percent the equilibrium score) as long as $s(1)$ is between 10 and 350. When $s(1)$ is very large the model predicts practically no learning, and equal choice probability among the strategies. The MSD distance from the data of this "flat" prediction is 1.87, i.e., it coincides with the random choice model. And the model's predictions are closer than the equilibrium prediction to the data for all positive values of $s(1)$.[14]

The value of game-specific parameters—The second row of the basic model's statistics in Table 1 presents the fit of a variant of the basic model that assumes that the strength parameter is affected by the game. That is, here we estimate the best fit separately for each game. Over games the MSD score of this 12-parameter model is 0.68. In the following sections we will see that this improvement is not large enough to justify the current 12-parameter model; models with fewer parameters have better scores.

Predictions—To evaluate the predictive power of the model, we predicted behavior in each of the 12 games without using that game's data. That is, the parameter [$s(1)$] was estimated based on the data of the other 11 games. The results (the "prediction" row in the statistics for the basic model in Table 1) show that for the basic model the predictive power is almost identical to the descriptive power. (This reflects the stability of the parameter estimates, which were not substantially changed by the removal of any one game from the sample.) Over the 12 games the average predictive MSD score is 1.02.

C. *Individual Learning Curves and Between-Subject Variability*

The second row in Figure 4 presents individual learning curves of random pairs of virtual subjects that were programmed to behave according to the current model [$s(1) = 54$]. Examination of these curves suggests that the virtual (basic) subjects are less variable and more homogeneous than the human subjects.

D. *Summary and Limitations*

The basic one-parameter model clearly outperforms the equilibrium prediction in accounting for average choice probabilities and initial learning trends. Yet (referring to Figures 1–3), the basic model fails to account for the late movement often seen towards equilibrium, and some characteristics of the individual curves (Figure 4). It seems that the basic model predicts a learning process that is less responsive to the opponent than the observed processes. Responsiveness to the opponent is expected to lead to a direction shift (when the opponent "moves" to the other side of the equilibrium), and to increased within-pair variability. The following sections introduce alternative extensions of the basic model that facilitate "responsiveness."

[14] For values of $s(1)$ between 0 and 10 the performance of the model is not monotonic.

III. Roth and Erev's (1995) Extension: The Three-Parameter RE Model

In Roth and Erev we introduced responsiveness to the model by adding two weaker psychological assumptions: experimentation, and a recency effect.[15] The first of these can be viewed as an extension of the law of effect (see e.g., B. F. Skinner, 1953; N. Guttman and H. Kalish, 1956; J. S. Brown et al., 1958).

Experimentation (or *Generalization*): Not only are choices which were successful in the past more likely to be employed in the future, but similar choices will be employed more often as well, and players will not (quickly) become locked in to one choice in exclusion of all others.

The second additional feature of individual learning modeled in Roth and Erev can be viewed as an interaction between the law of effect and the power law of practice.

Recency: Recent experience may play a larger role than past experience in determining behavior.

In Roth and Erev we called this "forgetting." Like generalization, recency is a robust effect considered and observed at least since John B. Watson (1930); see also Edwin R. Guthrie (1952).

These two assumptions were quantified in Roth and Erev by the following modification of equation (1), the updating function:

$$(1') \qquad q_{nj}(t+1) = (1-\phi)q_{nj}(t) + E_k(j, R(x)).$$

In 1′, ϕ is a forgetting (or recency) parameter which slowly reduces the importance of past experience, and E is a function which determines how the experience of playing strategy k and receiving the reward $R(x)$ is generalized to update each strategy j.

Experimental investigation of generalization suggests that strategies which subjects find "similar" to the selected strategy will be affected by the reinforcement. Brown et al. (1958) observed a normal generalization distribution. In games in which similarity of strategies can be linearly ordered (such as those studied in Roth and Erev) we chose a "three-step" function to approximate the generalization function, as follows:

$$E_k(j, R(x)) = \begin{cases} R(x)(1-\varepsilon) & \text{if } j = k \\ R(x)\varepsilon/2 & \text{if } j = k \pm 1 \\ 0 & \text{otherwise,} \end{cases}$$

where ε is an experimentation/generalization parameter. For games such as those in the present data set, when only two strategies are considered, or when the $M \geq 2$ strategies do not have an apparent linear order, a "two-step" function will be used:

$$E_k(j, R(x)) = \begin{cases} R(x)(1-\varepsilon) & \text{if } j = k \\ R(x)\varepsilon/(M-1) & \text{otherwise (where } M \text{ is the number of pure strategies).} \end{cases}$$

Another way to think of these two functions is that when the strategy sets allow similarity judgements to be made, players will generalize their most recent experience in a way that leads to experimentation among the most similar strategies.[16] When no similarity judgements can be made, players simply retain some propensity to experiment among all strategies.

Parameters—The model has three parameters: the strength parameter $s(1)$ (as in the basic model) and the experimentation and forgetting parameters ε and ϕ.

[15] The model proposed in Roth and Erev has subsequently been used to explore a number of data sets. See Gary Bornstein et al., 1994; John Dickhaut et al., 1995; Rosemarie Nagel, 1995; Ochs, 1995; Bornstein et al., 1996; Cooper and Feltovich, 1996; Feltovich, 1997; John Duffy and Feltovich, 1998; Robert Slonim and Roth, 1998.

[16] Thus this is an attempt to incorporate some structural information about the game into the learning model; in this respect, see also the "directional learning" approach of Selten and Joachim Buchta (1994).

A. *Aggregate Curves*

Best fit—A grid search revealed that the RE model best fit the data over the 12 games with the parameters $s(1) = 9$, $\varepsilon = 0.2$, and $\phi = 0.1$. The third column in Figures 1–3 graphs simulations of the RE model with these parameters. The summary statistics (for the RE model in Table 1) reveal that the addition of experimentation and recency parameters reduced the model's MSD distance to 0.59 over all games. Thus, this 3-parameter model outperforms the 12-parameter game-specific basic model. Table 1 also shows that the RE model outperforms the equilibrium prediction in 9 of the 11 relevant games. In addition to this quantitative improvement, the extended model captures the longer-term trends (movement toward the equilibrium) that are not captured by the basic model.

Sensitivity analysis—To evaluate the robustness of the model to the choice of parameters we asked how large is the subspace of the three-parameter space for which the model's fit is below 1.5. A grid search reveals that this criteria is satisfied for all the parameter sets (i.e., everywhere inside the cube) in which $0.02 < \varepsilon < 0.3$, $0 < \phi < 0.2$, and $0 < s(1) < 1000$ (there are also points outside this cube that satisfy the 1.5 criterion).

The value of game-specific parameters—The MSD score of the 36-parameter variant of the RE model with game-specific parameters is 0.35. Examination of Table 1 reveals that introducing game-specific parameters achieved the largest improvement in fitting the data in the games that involve negative payoffs (in games M&L and On in particular). This observation suggests that the apparent game effect may be a result of inaccurately modeling the effect of losses. We will return to this point in Section VII.

Predictions—When behavior in each of the 12 games is predicted based on the parameters that best fit the other 11 games, the MSD score is 0.69. This result suggests that the improvement of the RE model over the basic model is not a result of fitting more parameters. It seems that the forgetting and experimentation parameters capture robust properties of the data that facilitate prediction.

To evaluate the contribution of each of the added parameters we derived the predictions of the two reduced two-parameter models over all games. The overall predictive MSD distance of a "strength and experimentation" model (which fixes $\phi = 0$) is 0.75, and the predictive MSD score of a "strength and forgetting" model (fixing $\varepsilon = 0$) is 1.0. These results reveal that the addition of forgetting is useful only following the addition of experimentation. But then forgetting complements experimentation, as shown by the 0.69 overall predictive MSD for the three-parameter RE model.

B. *The Value of Estimating Initial Propensities*

In Roth and Erev we noted that in the ultimatum game the players' initial propensities can have a long-term effect on the learning process. Thus, in that game the prediction of the learning model can be improved by an assessment of these initial propensities. To evaluate the effect of the initial propensities in the current games we compared the fit of the model studied above (which assumes uniform initial probabilities) with a model that used the first block of data in each game as an estimate of the initial probabilities. To facilitate comparison the first block data was ignored in this analysis; without this block both variants of the RE model have only three parameters that are estimated from the data.

The effects of the estimated initial propensities are weak: They do not substantially affect the optimal parameters and the fit on these data. With estimated initial propensities the MSD is 0.55 compared to 0.59 with uniform initials. The estimated initial propensities improve the fit in 6 out of the 12 games. Given these results, and the cost of estimating initials (in extra parameters, and in difficulty of comparing models), we prefer to retain the assumption of uniform initial propensities in our analyses of these games.

C. *Does the Model Capture Nonlinear Trends in the Data?*

In order to examine whether the model captures the nonlinear trends in the data, an analysis of covariance was conducted. This analysis fits a linear curve to the probability

of A choices (thus, the B curves in the R&B games were not utilized in this analysis) of each player in each game (24 curves, each with two free parameters), and tests if the learning model's predictions can add significantly given the 48-parameter linear model. The addition of the RE model is highly significant ($F[1, 99] = 35.7$, $p < 0.0001$).

D. *Individual Curves*

The third row in Figure 4 presents individual learning curves of random RE players. These pairs appear to be closer to the experimental pattern (high variability that is not diminished with time) than the "basic" pairs.

IV. A Four-Parameter Generalization of Reinforcement Learning and Probabilistic Fictitious Play

The RE generalization of the basic reinforcement model makes it more responsive to changes in the opponent's behavior by adding experimentation and forgetting. In this section we explore a model in which responsiveness is added more explicitly, in the form of expected value calculations that allow a player to try to choose an action based on beliefs about opponents' behavior.

This model also facilitates comparison between the experimentally motivated reinforcement learning models considered above and belief-based models studied in the game theory literature. Like Camerer and Ho's experience weighted attraction (EWA) model, the current model is a generalization of reinforcement learning and the fictitious play (FP) model (George W. Brown, 1951). FP models a player as observing the past actions of the other players, and in each period choosing the action which maximizes his expected payoff under the assumption that other players will choose among their actions with the frequency observed up to that period. This is a deterministic model of behavior, originally proposed as an algorithm for computing equilibria (Julia Robinson, 1951), and we will see that it does not do as well as probabilistic models in tracking observed behavior.

In an effort to construct more descriptive belief-based models, a number of authors have considered probabilistic versions of fictitious play in which players have a higher probability of choosing an action the higher is its expected payoff according to beliefs formed as in fictitious play. Fudenberg and Levine (1997b Ch. 4) note that such a model coincides with a "stimulus-response" [reinforcement] model if the expected value of each action is taken to be its average return over past plays, and they suggest that this may be the natural way to adapt fictitious play to games in which players may not be able to observe other players' actions. We consider this case first.

The limited feedback case—We first consider the relationship between the basic reinforcement model and the FP model when the information available to the players is limited to the realized payoffs, by considering a model which generalizes them both, i.e., a model which for different values of its parameters coincides with one or the other. Take $REV_{nk}(t)$ to be the average return player n has received from those periods up to $t-1$ in which he has chosen action k. Then equation (1) of our basic model implies that his propensity $q_{nk}(t) = q_{nk}(1) + REV_{nk}(t)C_{nk}(t)$, where $C_{nk}(t)$ is the number of times player n has chosen action k up to time $t-1$. (Note that one of the key differences between a reinforcement model and an optimization model is that in a reinforcement model not only the average return on an action matters, but also the number of times it has been chosen.)

We can also replace the initial propensities $q_{nk}(1)$ with two initial "expectation" parameters and write $q_{nk}(t) = EV_{nk}(1)N_n(1) + REV_{nk}(t)C_{nk}(t)$, where $EV_{nk}(1)$ is player n's initial belief concerning the expected value of strategy k, and $N_n(1)$ is the "strength" of that belief.

Finally, it is convenient to define a "subjective reinforcement EV" as the sum of initial expectations and accumulated experience, namely

$$SREV_{nk}(t) = [EV_{nk}(1)N(1) + REV_{nk}(t)C_{nk}(t)]/N_{nk}(t),$$

where $N_{nk}(t) = N_n(1) + C_{nk}(t)$. Note that $q_{nk}(t) = SREV_{nk}(t)N_n(t)$. This implies that the probabilistic choice rule given by equation (2) can now be written as

$$(2') \quad p_{nk}(t) = [\,SREV_{nk}(t)N_{nk}(t)\,] \div \Sigma(\,SREV_{nj}(t)N_{nj}(t)\,).$$

Equation (2′) allows us to see that in the current limited information case the basic reinforcement model is distinguished from probabilistic fictitious play by the presence of the numbers $N_{nj}(t)$, which allow the number of times a strategy has been played in the past to influence the probability that it will be played in the future, instead of having this probability determined only by the expected values. To better see this, consider the two-parameter family of models given by equation (2″) below.

$$(2'') \quad p_{nk}(t) = [\,SREV_{nk}(t)^m(N_{nk})^h\,] \div \Sigma[\,SREV_{nj}(t)^m(N_{nj})^h\,],$$

where the parameter $m \geq 0$ can be interpreted as an indication of the degree to which player n maximizes based on his expectations, and the parameter $h \geq 0$ can be interpreted as measuring the force of habit, i.e., the force of past experience.

In a reinforcement model, h is positive, because past behavior influences current behavior through more than the expectations, while in a belief-based model h equals 0. For any sufficiently large value of m, when $h = 0$ (no force of habit) equation (2″) approximates (arbitrarily closely) the traditional FP model which makes the deterministic choice of the strategy with the highest expected value. When $m = 1$ and $h = 0$, equation (2″) describes a simple model of probabilistic fictitious play, in which actions are chosen proportionately to their expected payoff, with no regard to how often they have been played in the past. When $m = 1$ and $h = 1$, equation (2″) coincides with (2′), and with the basic reinforcement model given by (2).[17]

The complete information case—An additional difference between reinforcement learning and FP arises when the players receive complete information concerning their opponent's decisions, in which case reinforcement learning still models a player as being influenced only by the strategies actually played. Camerer and Ho point out that in this case the FP model implies that the players calculate the relevant Expected Values by the average returns that they *could have* received from choosing each of the strategies in the first $t - 1$ trials (under the fictitious play assumption that the other players' behavior is fixed). To calculate these fictitious *EV*s (*FEV*) set the initial values $f_{nk}(1) = q_{nk}(1)$ and,

$$(1'') \quad f_{nj}(t+1) = f_{nj}(t) + SR(x_j),$$

where $SR(x_j)$ is the reinforcement that player n would have received for choosing j in trial t [x_j is the payoff, and $SR(x_j) = x_j - x_{\min}$], and $f_{nj}(t+1)$ is the accumulated fictitious propensity up to trial $t + 1$ for player n to play strategy j.

This definition implies that $f_{nk}(t) = EV_{nk}(1)N_n(1) + FEV_{nk}(t)(t-1)$, where $FEV_{nk}(t)$ is the average return player n would have received from choosing action k in all periods until $t - 1$. And the "subjective fictitious *EV*" can be written as

$$SFEV_{nk}(t) = f_{nk}(t) / [\,N_n(1) + t - 1\,] = [\,EV_{nk}(1)N_n(1) + FEV_{nk}(t) \times (t-1)\,] / [\,N_n(1) + t - 1\,],$$

[17] In comparing different learning models via equation (2″) we are following a path explored by Cheung and Friedman (1996) and Camerer and Ho (1998a). Each of those papers looks at a parameterized class of models different from equation (2″) but similar in spirit. Cheung and Friedman (1996) consider a family of models which connect the dynamics of best reply (to the previous period's actions) to those of fictitious play. Our model is more similar to that of Camerer and Ho, although their model is more complex, with separate parameters for discounting past reinforcements and past beliefs. We certainly do not insist that our model is more correct; rather our intention is to use it as a basic model, comparable to the basic reinforcement model, to clarify the role that maximization and past experience (habit) can play.

where $f_{nk} = SFEV_{nk}[N_n(1) + t - 1]$ is the accumulated fictitious propensity.

Note that under the assumption that the other players' behavior is fixed (and in general in one-person games), the two subjective expected values estimates, $SFEV_{nk}(t)$ (which reinforces all strategies at each period) and $SREV_{nk}(t)$ (which reinforces only those strategies which are actually played), are unbiased. They are expected to lead to the same estimates when t is large enough. Yet, $SFEV_{nk}(t)$ uses more information.

To summarize all the differences between the basic reinforcement learning model and FP we will focus on a model that incorporates equations (1), (1″), and the following generalization of (2) and (2″):

$$(2''') \qquad p_{nj}(t) = \frac{[(d)SFEV_{nk}(t) + (1-d) \times SREV_{nk}(t)]^m (N_{nk})^h]}{\sum [(d)SFEV_{nj}(t) + (1-d) \times SREV_{nj}(t)]^m (N_{nk})^h]}$$

where $0 \leq d \leq 1$ is a weight parameter that determines the relative weight of the nonreinforcement information. With $d = 1$ (and $h = 0$ and large m) the model coincides with FP. With $d = 0$, $h = 1$, and $m = 1$ it coincides with the basic reinforcement learning model.

Parameters—In its general forms, the current model has $2M + 1$ initial propensity parameters for each player: Two parameters for each strategy, $q_{nk}(1)$ and $EV_{nk}(1)$, and one strength parameter $N_n(1)$. Yet, we can use the constraints $\Sigma q_{nk}(1) = s_n(1)X_n = \Sigma EV_{nk}(1)N_n(1)$, and the uniform/symmetrical initials assumptions utilized above to reduce the number of free initial-propensities parameters to one.

We start with the estimation of $q_{nk}(1) = s(1)X_n/M$ as stated above, and set $EV_{nk}(1)$ as the average reinforcement (for player n) in the game (average payoff minus $x_{\min}$). Following this simplification, the current model has four parameters: d, h, m, and $s(1)$.

In summary, three psychological assumptions distinguish the basic reinforcement model we consider from fictitious play. The first has to do with the information assumed to affect the implicit *EV*s. The parameter d (when it is positive) allows information about what strategies would have earned to enter the calculation, instead of only allowing information about what strategy choices did earn when actually chosen. The other two have to do with the absence of maximization ($m = 1$ rather than infinity), and the force of habit (h positive rather than 0).

A. *Aggregate Curves and the Value of the Added Sophistication*

The predictions of the FP-like model were derived under the assumption that when players' information is limited to their own payoffs (in the Suppes and Atkinson games and in our replication), choice probabilities are calculated with $d = 0$. For the remaining seven games the estimated value of d affects the model's predictions.

Best fit and the contribution of the different parameters—The "optimal" parameter set was found to be $d = 0.9$, $m = 1.5$, $h = 0.1$, and $s(1) = 27$. The MSD of this four-parameter model is 0.72—better than the basic model, but not as good as the three-parameter RE. The predicted curves are presented in the fourth column of Figures 1–3.

In evaluating the parameters' values it is important to note that the optimal fit function has an extremely flat optimum along the information parameter d. An almost equally good fit (MSD of 0.75) was obtained under the constraint $d = 0$, that forces the simulated players to ignore any information other than that contained in the reinforcement model.

These results are consistent with the observation that in the current setting (many repetitions of games having small number of strategies) the two sources of information (personal payoff versus all payoffs) are highly correlated.

The MSD score of a deterministic FP model (large m, $d = 1$, $h = 0$) with optimal $s(1)$ value (of 5) is 1.9. The addition of a probabilistic response rule ($m = 1$) improves the MSD score to 0.73.

In summary, only one of the three parameters that distinguish the basic model from FP, the habit parameter h, has a clear contribution to the model's descriptive power.

Sensitivity analysis—As noted above the predictions of the current model are relatively insensitive to the value of d. The subspace

of the four-parameter space for which the model's fit is below 1.5 include all values of d ($0 \leq d \leq 1$) and: $1 \leq m < 2$, $0 \leq h < 0.5$, and $0 < s(1) < 1000$.

The value of game-specific parameters—Over the 12 games the MSD score of the 48-parameter variant of the FP-like model with game-specific parameters is 0.37. As in the case of the RE model, the largest improvement due to game-specific parameters was achieved in the games that involve negative payoffs (cf., Table 1). [The "good" fit for game On1 (0.28) is, of course, a chance result. With high m values the model's predictions are extreme (0 or 1) and even the average of 100 simulations is noisy. With 1000 simulations the average prediction is 0.5 and the fit converges to 0.45.]

Predictions—The average prediction MSD score of the FP-like model is 0.84.

To evaluate (and quantify) the assertion that the advantage of the current model over the basic model is largely due to the effect of the habit parameter, we derived the prediction MSD score of a two-parameter "strength and habit" model (the FP-like model with the constraints $d = 0$, and $m = 1$). The MSD prediction score of this model is 0.85. Thus 0.15 (93 percent) of the 0.16 advantage in predictive power of the FP-like model over the basic model is obtained by the habit parameter. The two "rationality" parameters (maximization and FP-expectations) contribute together 0.01 (7 percent of the 0.16).

B. *The Effect of the Assumed Initial Beliefs*

To evaluate the robustness of these conclusions to the assumed initial beliefs, we derived the prediction of the FP-like model with the assumption of uniform initial beliefs. Under this assumption player n knows the payoff matrix and believes that his/her opponent will choose randomly among the possible strategies. This assumption did not improve the model's fit. The best MSD score was 0.75.

C. *Individual Learning Curves*

Samples of individual learning curves under the FP-like model are presented in the fourth row of Figure 4. This model appears to imply a reduction in between-pair variance that is not observed in the current data.

D. *Summary of the Model Comparisons on Aggregate Data*

Before moving on to individual data, a glance at Table 1 summarizes what we have learned. The one-parameter basic model outperforms the equilibrium predictions. The basic model's descriptive and predictive power are further improved by incorporating experimentation and forgetting into the three-parameter reinforcement model. The fact that the three-parameter model fit simultaneously to all games has a lower mean deviation (0.59) than does the one-parameter ($\times 12$) model fitted to each game separately supports the notion that it may be possible to find learning models which can be usefully applied to a variety of games, rather than having to construct or estimate models separately for each game.

Note once again that the set of games is not a random sample from the space of games, but rather a selection of games from experiments with very different conclusions about the performance of the equilibrium predictions. We can informally compare each of the models on the games in which equilibrium does badly and in which it does well by considering the performance on the five games in which equilibrium predicts less well than random choice, and on the seven games in which it predicts better. All of the models beat the equilibrium predictions on the games in which it does worse than random choice, and the multiparameter models outperform equilibrium even on the games in which it does better than random choice. We turn now to a consideration of the individual-choice data, including that from the game of O'Neill, in which the equilibrium predictions had the greatest success.

V. A Parameter-Free Comparison of Models on Individual-Subject Data.

The analyses summarized above compare the learning models' predictions of the behavior over an entire experiment to the aggregate data from each game. Looking at aggregate data has the advantage of smoothing some of

the variance found in individual subjects, both simulated and real. However the models we consider are individual-choice models, and we now turn to comparing their predictions to the individual-level data.

Looking at the individual data also has the advantage of letting us compare models with different unobservable parameters. As already noted, in the basic reinforcement model the unobservable parameter is the strength of the initial propensities. In the belief-based models, the unobservable parameters concern the beliefs with which subjects enter the first period of play. But these initial, unobservable parameters quickly become of only small importance if we can observe a subject's initial experiences, since these (observable) experiences soon become more important than initial propensities or beliefs.[18] In the analyses which follow, we set the strength of initial propensities in the basic reinforcement model to be the sum of payoffs received in the initial periods, and take a player's initial beliefs to be the frequency of other players' actions observed in those periods (or observed average payoffs). We then compare the predictions each model makes for each subject's choices, using for each period t the data of the subject's first t choices to predict the subject's next choice under the alternative models. Since we used the data collected in the experiment to derive the model's predictions (rather than simulation results) this technique is less sensitive to the choice of parameters.[19]

Five learning models and the equilibrium predictions (EQ) were compared in this analysis. The five models include: the basic reinforcement learning model, the traditional deterministic fictitious play model (FP), the probabilistic FP model (PFP) (the FP-like model with $m = 1$, $h = 0$, and $d = 1$), a FP model with the exponential probabilistic response rule (EFP), and a simple best reply (to the previous period's play) model (BR). Three studies, O'Neill, Ochs, and the new replication study (game S&A3n), for which we were able to obtain the choices made by individual subjects, were considered.

Twenty-five pairs participated in O'Neill's study. Each pair played the game 105 times. The present analysis focuses on the decisions made in rounds 6–105 (and uses the first five rounds to assess initials). To allow evaluation of the effect of the subjects' experience on the models' fit, these rounds were divided into two blocks (6–55 and 56–105).

At the first step of the analysis a vector of probabilistic predictions was obtained for each of the decisions, given each of the five models. Vectors had to be considered because the game involved four strategies. The EQ predictions were (0.4, 0.2, 0.2, 0.2) for all 100 rounds. The reinforcement learning predictions were calculated based on the cumulative reinforcements using equation (3). In a similar way the PFP and EFP predictions were calculated based on the fictitious play expected values $EV_{nk}(t) = FEV_{nk}(t)$. Like the equilibrium predictions, these models provide a vector of four probabilities. The FP predictions are deterministic (three 0's and one 1), a FP model predicts that player i will always choose the strategy that maximizes expected profit given player j's accumulated choice probabilities. According to the BR model, player i is

[18] This is because frequencies of play and accumulated payoffs are observable independently of other parameters. So in the analyses which follow we will look at "basic" variants of both reinforcement and belief-based models, rather than multiparameter models in which the estimated value of some parameters would depend on the estimated values of other parameters.

[19] Note that, to use the data this way, we are switching from the long-range predictions of the previous sections (which simulated the entire play of the game without using any of its data) to short-range predictions of period t choices given the data through period $t - 1$. In general we want long-range predictions (e.g., if someone asks for a prediction about a year from now, it is considered unprofessional in the soothsaying business to ask him to come back for the answer in 364 days). But concentrating on short-term predictions seems unavoidable for probabilistic models at the level of individual subject pairs, since a long-range simulation of a single pair might begin the first period with a different pair of choices than the subject pair in question, and from that different initial experience continue to diverge. In contrast, in the previous sections, when we did consider individual pairs, we generated multiple predicted pairs to gain information about the predicted variability. (Considering models at different levels of aggregation, of time as well as of subjects or of games, allows us to assess the usefulness of their predictions in different ways.)

expected to choose the strategy that maximizes profit given player j's last choice. When there was more than one, all maximizing strategies were assigned equal probabilities.

Two goodness-of-fit measures were then calculated for each prediction: a mean-squared deviation (MSD) score, and a proportion of inaccuracy (POI) score. The MSD is the mean-squared distance between the predicted and the observed vector. For example, if the observed vector is (1, 0, 0, 0) (that is, the subjects chose A) and the prediction is (0.5, 0.3, 0.1, 0.1), MSD = $[(1 - 0.5)^2 + (0 - 0.3)^2 + (0 - 0.1)^2 + (0 - 0.1)^2]/4 = 0.09$. The POI score returns the value 0 if the subject made the most likely choice under the model, the value 1 if the subject chose a strategy that differs from the most likely prediction, and $1 - 1/b$ if the model predicts that b strategies are equally likely and the subject chose one of them. (Thus the POI score judges all the models on the basis of their "deterministic" predictions, which should facilitate comparison of the deterministic models—fictitious play and best reply—and the stochastic, reinforcement learning models.)[20] At the final step of the analysis average MSD and POI scores were computed for each pair and for each game, and the various models were compared.

A similar three-part analysis was conducted for the 48 subjects (in three cohorts) that participated in Ochs' study and for the 20 subjects (in 10 pairs) that participated in the replication study. In Ochs' study we used the first trial (in which the game was played ten times) to assess initial propensities. Because in the replication study subjects were not informed of the payoff matrix, the expected values were computed as $EV_{nk}(t) = REV_{nk}(t)$ for the FP models, and since they did not observe the other players' action we take the best-reply rule to be the "win stay, lose change" rule (see e.g., David M. Messick and Wim B. G. Liebrand, 1995).

The mean goodness-of-fit scores and the comparison statistics are presented in Table 2. Statistical significance was computed in a paired t-test. The units of analysis were pairs for the O'Neill and replication data, and individuals for the Ochs' data. Note that larger values reflect worse fit. The data reveal that the basic reinforcement learning model outperformed the versions of fictitious play and best reply we consider in the last (second) block of all studies. In the O'Neill and Ochs' studies this second block advantage is significant. The observation that the relative fit of probabilistic FP models declines from the first to the second block suggests that the effect of the positive habit parameter increases as subjects gain experience. The insignificance of the difference between reinforcement learning and FP in the replication study is potentially interesting (recall that in this study subjects did not know the payoff matrix and their opponent's choices) but may also be a result of the smaller number of subjects.

The reinforcement model significantly outperforms the equilibrium predictions in Ochs and the replication data, but not in O'Neill's data. That the reinforcement models did not outperform the equilibrium predictions in O'Neill's game does not imply that subjects were insensitive to reinforcements in this game. Rather, this finding may be a result of the proximity of the initial propensities to the equilibrium. Support for this conjuncture was obtained in an analysis that compared each individual pair to simulations initialized with the pair's data. We initialized 100 simulations with the observed first 35 choices in each of O'Neill's 25 pairs. Regression analysis reveals that the model's predictions can be used to predict the observed dynamics. The dependent variable in this analysis is the change in the proportion of A choices between the first and the second block of 35 trials. Two predictors were compared for each of the 50 subjects: a prediction that the subjects will move toward the equilibrium, and the learning model prediction. The results indicate that whereas the dependent variable cannot be predicted based on the convergence to equilibrium predictions, it is significantly related to the simulation predictions ($t[49] = 2.95$, $p < 0.005$). [This is a different observation than Brown and Rosenthal's observation that although O'Neill's subjects were close (on the

[20] We include the POI score in order to answer the objection that models that make extreme predictions (e.g., deterministic models like fictitious play or best reply which predict probabilities of 0 or 1) cannot be adequately compared to models that make stochastic predictions.

TABLE 2—WITHIN-SUBJECT MODEL COMPARISON OF MSD AND POI (PROPORTION OF INACCURACY)

Data set:		O'Neill (1987)		Ochs (1995)		Replication (S&A3n)	
Model:	Block	MSD	POI	MSD	POI	MSD	POI
Basic (reinforcement learning)	1st	0.20	0.66	***0.13***	***0.37***	0.24	***0.40***
	2nd	0.18	0.61	***0.12***	***0.35***	***0.21***	***0.33***
FP	1st	0.26	0.71	0.21 ns	0.39 ns	0.39	0.41 ns
	2nd	0.24	0.69	0.20	0.37	0.31	0.34 ns
Best reply	1st	0.27	0.76	0.29	0.42	0.41	0.44
	2nd	0.26	0.74	0.33	0.45	0.38	0.40
PFP	1st	0.19	0.71	0.14 ns	0.39 ns	0.24 ns	0.41 ns
	2nd	0.19	0.69	0.15	0.37	0.22 ns	0.34 ns
EFP	1st	0.19 b	0.71	0.14 ns	0.39 ns	***0.24*** ns	0.41 ns
	2nd	0.19	0.69	0.15	0.37	0.22	0.34 ns
Equilibrium	1st	***0.18*** b	***0.61*** b	0.15	0.45	0.31	0.49
	2nd	***0.18*** ns	***0.60*** ns	0.14	0.42	0.30	0.51

Notes: Best fits are indicated by a bold italic font. (The basic reinforcement model provides the best fit for the Ochs and replication data, while the equilibrium gives the best fit for the O'Neill data.) In most cases the fit of the basic reinforcement model was significantly better (smaller score) than the fit of the alternative models. Exceptions are indicated by "b" when an alternative model significantly outperforms the basic model, and by "ns" when the difference is insignificant (at the 0.05 level in a one-tail paired t-test).

average) to the equilibrium, they did not exhibit the independence from one period to the next implied by the minimax strategy.[21]]

VI. Why Extensions of the Models Will Be Required on Larger Classes of Games

We have concentrated so far on a very simple class of games. The good results we have obtained make it plausible that reinforcement learning can serve as an engine to study behavior on a wider class of games. It is worth spending some time, even at this early juncture, to reflect on some modifications in the simple models we have so far considered which will be necessary to accommodate larger domains of games.

A. *Adjustable Reference Points*

Even a thought experiment suggests that the simple form we have assumed for the reinforcement function, $R(x) = x - x_{\min}$, is too simple to be very general. For example, if $x_{\min}$ were very much smaller than the average payoffs experienced in the game, then the simple constant reference point approach would lose the ability to distinguish between the most commonly experienced payoffs.[22] This is why a more general approach, with adjustable reference points $[R(x) = x - \rho(t)]$, seems necessary, despite the cost of added parameters.

Experiments on individual choice show that reference points can be important (e.g., Daniel

[21] The serial correlation in O'Neill's data and similar results (the negative recency effect reviewed by Lee, 1971), and the overalteration tendency (Rapoport and Budescu, 1992) are inconsistent with the current model when only stage-game strategies are considered. Yet, as shown by Rapoport et al. (1997) these phenomena can be accounted for by the assumptions that subjects consider "two-stage" strategies. The addition of such strategies does not effect the model's aggregate predictions, and does predict serial correlations.

[22] In a game with a dominated strategy that gave player 1 a payoff of −1,000,000, and with all other payoffs in the range of 1 to 10, our simple fixed-reference-point reinforcement function would mostly give reinforcements from 1,000,001 to 1,000,010.

Kahneman and Amos Tversky, 1979), and can be affected by previous outcomes (L. H. Tinkelepaugh, 1928). Tinkelepaugh's results suggest that the reference point moves towards the average reward. (He found that although lettuce is a positive reinforcer for inexperienced monkeys, monkeys who got used to a banana reinforcement behaved as if lettuce is a negative reinforcement.) In an earlier version of the present paper (Erev and Roth, 1996) we showed that a simple model with an adjustable reference point could increase the descriptive and predictive power of the reinforcement learning model.[23] (Much of the improvement was obtained for the games having negative payoffs.) But precisely how reference points should be modeled (so as to accommodate games which elicit different initial reference points in high-information environments, or in which reference points adjust at different speeds in low-information environments) remains a subject for future research. The question of reference points will arise both when we consider games with more variation in payoffs than those considered here, and when we consider with more strategic complexity. But for these games, discussed next, we also need to develop ways to better model strategies.

B. *Repeated Games in Which Stage-Game Strategies Are Not Enough*

The data we have just considered, from repeated play of games with a unique equilibrium in mixed strategies, has allowed us to see that a learning model has very substantial ability to describe and predict the data. This is all the more surprising because the strategy set we considered for each player consisted only of the stage-game actions. Thus these games allowed us to investigate learning, without having to investigate in detail the strategy sets employed by the experimental subjects.[24]

Before going on to draw general conclusions, we therefore want to emphasize that it will of course not generally be the case that learning behavior can be analyzed in terms of stage-game actions alone. An easy way to see this is to consider games whose repeated play leads to experimentally observed behavior that clearly depends on repeated-game strategies. For example, an experiment involving repeated play of the following version of the game of "Chicken" is reported by Anatol Rapoport et al. (1976).

	S_2	T_2
S_1	1, 1	−1, 10
T_1	10, −1	−10, −10

They report that a majority of the ten pairs of subjects they observed play the game for 300 rounds (without changing partners) quickly settled in to an alternating strategy, with the outcome changing back and forth between (S_1, T_2) and (T_1, S_2), and with the outcomes on the diagonal successfully avoided.[25]

It is apparent that this pattern of behavior cannot be achieved with stage-game strategies alone, since, for example, a player limited to stage-game strategies cannot remember whether it is his turn to play S or T (and so some diagonal outcomes would inevitably result from independently mixed stage-game strategies).

To put it another way, a learning model such as ours would certainly fail if it were restricted to the set of stage-game strategies alone. But such a restriction would clearly be artificial and undesirable in a general model of learning to play a repeated game.[26]

[23] This variant of the current reinforcement model was studied in Bereby-Meyer and Erev (1998), Erev (1998), Erev and Rapoport (1998), and Rapoport et al. (1998).

[24] That is not to say that subjects may not sometimes exhibit more complex repeated-game behavior even in such games; see e.g., the games studied by Mookherjee and Sopher (1994).

[25] In fact, they report an index $K = (S_1, T_2) + (T_1, S_2) - |(S_1, T_2) - (T_1, S_2)|$, where (S_i, T_j) is the percentage of observations which have that outcome. This index has a value of 100 only if (S_1T_2) and (T_1, S_2) are the only outcomes observed, and occur with equal frequency. They report (p. 157) that the values of K rise from 78 for rounds 1–50, to 99 for rounds 251–300, indicating that by the end of the game virtually all players were successfully alternating. (Rapoport et al. [1976] also report other variants of the game in which less alternation was observed.)

[26] Indeed, some of the success of our model on the non-constant sum games in this data set may be due at least in part to the fact that these games were run under conditions

On the other hand, simulations with our model show that it tracks this data quite well if the strategy sets of each player are taken to be the stage-game strategies plus the strategy of alternation.[27] But this is also an unsatisfactory approach, since selecting one repeated-game strategy from the multitude of possible strategies is like parameter fitting in a model with an enormous number of parameters. Consequently, while it will be both natural and necessary to model repeated-game strategies for repeated-game situations, a great deal of thought will be needed to do so in a systematic way which retains the predictive power of the model. We think this is where future work will have the greatest contact with cognitive psychology.

That being the case, a few remarks are in order about the relationship between strategies and learning, and between the approach explored here and current trends in the psychology literature.

The traditional game-theory approach to repeated games is that at the outset of the game each player chooses a strategy which determines his actions (perhaps probabilistically, in which case it is called a behavioral strategy) throughout the game, however long. While such strategies can in principle describe any behavior, it is not a helpful approach if we wish to study learning, since a player who chooses his strategy at the beginning, and never deviates from it, can hardly be said to learn. So, although the learning rules we consider can be thought of as behavioral strategies, our approach has been to concentrate on short-term strategies, and study how the learning rule selects among them differently over time. This is an approach which is also followed in the psychology literature.

The Law of Effect, which is the basis of the learning models supported in the present research, was initially proposed to describe the behavior of cats (Thorndike) and pigeons (Herrnstein), but we do not claim that choice behavior can be understood without considering "deeper" thought processes. Nor do we argue that people are no smarter than pigeons (cats are obviously smarter). Rather, we contend that it is useful to distinguish between the adaptive learning process and other relevant thought processes such as the strategies which players explore and learn about. This is in line with John R. Anderson's (1982, 1993) influential theory of cognition (the ACT* theory), which distinguishes between "production rules" and learning. The game-theoretic definition of strategy can be thought of as a description of a class of production rules (all strategies are production rules, but there are production rules in the ACT* theory which are not strategies).

There are three relevant learning processes in the ACT* theory: (1) encoding past events; (2) converting knowledge into a production rule form; and (3) strengthening the production rules to affect their choices in the future. The third process is similar to the learning models we consider. It incorporates the law of effect and implies the power law of practice; however, it ignores cardinal payoffs.

The main criticism of the ACT* and similar approaches is that they involve too many unobservable production rules and learning processes. Our results can be viewed as a demonstration that in the context of matrix games with unique, mixed strategy equilibrium in which reciprocation is impossible, a very simple version of an ACT* system provides a good description of behavior. This system has well-defined, specific production rules (stage-game strategies), and one robust learning process.

The relative simplicity of the current model supports the conjecture that, when a well-defined economic environment is considered, an accurate model of the cognitive game (strategies and payoffs) is possible. Some support for this optimistic conjecture also comes from results accumulated in experimental decision-making research. It appears that many of the observed regularities can be described by few common cognitive strategies (see for example, Tversky and Kahneman, 1974; Jerome R. Busemeyer and

that interfered with the use of repeated-game strategies. Ochs' subjects were not playing repeatedly against the same opponent, and the nonconstant sum games in the Suppes and Atkinson data were run without full information about the game.

[27] For this purpose, the alternation strategy was formulated as "if the outcome at time t was off the diagonal, then at time $t + 1$ choose the action not chosen at time t; otherwise choose each action with probability 0.5".

In Jae Myung, 1992; John W. Payne et al., 1993).[28]

Thus, the approach taken here does not represent an attempt to revive behaviorism; i.e., we do not hope to explain strategic behavior without considering players' cognition. On the contrary, our approach can be thought of as an attempt to utilize knowledge that has been accumulated in game-theoretic and psychological research, toward the development of a low-rationality, cognitive game theory.

VII. Adaptive Game Theory and Applied Economics

We have concentrated here on modeling behavior observed in the laboratory. But the adaptive game-theory approach we consider would be of very limited interest to economists if it could not address the same range of issues in natural markets and economic environments which have made equilibrium game theory such an important tool of modern economic theory. So we conclude with some thoughts on how a well-developed, cognitively informed adaptive game theory will complement conventional game theory, both as a theoretical tool and as a tool of applied economics.

Consider the analysis of a market as some of the underlying "rules of the game" change.[29] The equilibrium approach is to calculate the equilibrium under the old and new rules, and suggests that we should anticipate a change from one equilibrium to the other. The adaptive approach tells us to also consider whether this transition might be very slow (the environment may have changed again, and all the current market participants may have died before equilibrium is reached) and whether, as we have seen in some of the experiments discussed here, the initial adjustment might even move in a direction opposite that predicted by the equilibrium comparative statics (in which case we might feel differently about the empirical evidence drawn from the period immediately following the transition).

Both traditional and adaptive game theory can be used to make specific predictions about complex natural environments. Both approaches require a model of the game, but if the game has enormous strategy spaces (as close modeling of complex environments would be likely to yield) then it will not in general be practical either to solve for equilibria or to simulate learning. So in practice, simple models must be constructed which approximate the actual strategic environment. The traditional game-theory approach, which in principle considers all strategies, gives modelers little guidance about how to do this, except that the model of the game itself must be simple enough for all its strategies to be considered. The cognitive approach to adaptive game theory suggests that modeling observed strategies (and not necessarily including all the logically possible strategies that use the same information) may be a fruitful alternative approach.

For some examples, consider the variety of annual markets (mostly entry-level professional labor markets) described and analyzed in Roth (1984, 1990, 1991) and Roth and Xiaolin Xing (1994). In each of these markets, there was a period in which the time at which transactions were completed moved slowly earlier from one year to the next, as agents reacted to their experience in the previous year's market, which gave the greatest rewards to certain kinds of employers if they made transactions just a little bit earlier than their competitors. The equilibrium of such a process yields all transactions at some very early date (see e.g., Nagel [1995] for a related experiment), but the observed behavior is slow movement towards the equilibrium—sometimes over a period of half a century (which is only 50 iterations of an annual market), rather than a rapid transition directly to equilibrium. This is just the kind of behavior

[28] And Erev et al. (1995), Sharon Gilat et al. (1997), Erev (1998), and Rapoport et al. (1998) show that learning among cutoff strategies can be accounted for by the present approach.

[29] If Schumpeter's characterization of "creative destruction" as the fundamental feature of capitalist economies is correct, we could argue that transition is the most common condition of markets. But even without making such an argument, there is no shortage of fundamental transitions, e.g., in labor markets when minimum-wage laws are changed, in markets for medical services when third-party payment systems are changed, etc.

we should expect from reinforcement learners.[30]

VIII. Concluding Remarks

Even the one-parameter reinforcement learning model we consider robustly describes and predicts the data from these games with mixed strategy equilibria better than the static equilibrium predictions, whether we are looking at predictions for the aggregate results of an entire experiment, for the paths of play of particular pairs of players, or for the individual decisions of each player. Adding some "responsiveness" to the model which allows it to adapt to the changing behavior of other players improves the predictive power. Adding this in a "higher rationality" way via belief-based models does not appear to have an advantage over "lower rationality" reinforcement models on this data set. It may therefore be useful at this point to take a step back from the particular games and data and models which are the subject of the present paper, and consider once again the differences and similarities between the adaptive learning approach and how traditional game theory might address the kinds of data we consider.

It is an empirical question whether a theory of very high rationality behavior may provide the basis for a predictive theory of observed behavior. Even the observation that existing notions of equilibrium may leave much to be desired from the point of view of prediction does not preclude the possibility that further developments of "high" game theory will provide more accurate predictions of observed behavior. It is in something of this spirit that Fudenberg and Levine (1997a) have reexamined some of the data reported in Roth and Erev (1995) with respect to a generalized notion of strategic equilibrium which they propose, and Richard D. McKelvey and Thomas R. Palfrey (1995) analyze some of the data considered here, with respect to a generalized notion of both the equilibrium and the game being played.

Rather than expanding the kind of equilibrium considered, one might consider how the preferences of the players may be systematically influenced by the nature of the game. That is, in some games we might consider whether players have particular preferences for fairness, cooperation, or reciprocity, and then proceed to remodel the payoffs of the game to reflect these preferences, and use conventional notions of equilibrium to predict the outcome. This is basically the approach explored in Ochs and Roth (1989); Camerer (1990); Gary Bolton (1991); Matthew Rabin (1993); Bolton and Axel Ockenfels (1997); Ernst Fehr and Klaus Schmidt (1997).

Like these other alternatives, the learning approach we have taken here retains the basic idea of noncooperative game theory, namely the strategic model of the environment. The present approach makes "lower" rationality assumptions than the traditional approach in two respects: (1) it does not assume that the players consider all the possible strategies, and (2) it does not assume that players are subjective expected-utility maximizers, or indeed maximizers of any sort. Preferences do not play any explicit role in our model, although the model is agnostic about where the initial propensities come from (e.g., there is room for preferences in the explanation of how players make their initial choices, although perhaps these can also be explained without preferences—see also Werner Guth [1995] in this respect).

The essential elements of our approach are a learning rule and a model of the game. For the learning rule, we have tried not only to avoid behavioral assumptions we know people do not conform to (e.g., universal hyperrationality), but also to incorporate some of the robustly observed properties of individual behavior from the psychology literature. Nevertheless, the simple rule used here and in Roth and Erev is meant to stand in for the large class of actual learning rules which subjects may employ. The surprisingly good results that can be obtained even with our very simple learning rule, and the roughly similar results obtained for related learning rules like probabilistic fictitious play when they are modified somewhat

[30] See also the analysis of the Marseille fish market by Gerard Weisbuch et al. (1996), who independently explore a reinforcement learning model related to the ones considered here.

to resemble reinforcement learning, argue that in many environments the results are sensitive only to the basic properties of the learning rule we consider—namely, that it is a probabilistic rule which obeys both the Law of Effect and the power law of practice.

Notice that the model of players as adaptive learners interacts with the model of the game (the strategy sets of the players, and the payoff structure) through the players' initial propensities to play each of their strategies. Unlike in equilibrium models, the strategic environment faced by any player, and what kind of feedback he gets from his choices, depends on what the other players are doing, particularly during the critical early periods when learning is fast. This only enhances the importance of usefully approximating the strategies used by the players, which we have argued (in connection with the game of "Chicken") will be the area of future research in which low-rationality adaptive game theory will need to interact most closely with cognitive psychology.[31]

In our analyses and comparisons we have been both estimating parameters and evaluating predictions. For distinguishing among models, we regard predictive power as of primary importance. This has to do both with the nature of the models we are considering, and what we want to use them for. All of the models we consider are approximations, and so are false at some level of detail. Conventional methods of hypothesis testing do little to illuminate whether a model is (despite not being a true description of the world) a useful approximation. And whether an approximation is useful depends on what it is to be used for. One reason we look at prediction is it seems to us that this is the weakness of current game theory that is most in need of being addressed as game theory is increasingly used to design new market mechanisms.[32]

Note that, like some of the experiments we considered, in many—if not most—naturally occurring markets and games, players will not know the full details of the game. They are unlikely to know in detail what all other players are doing, and even less likely to be able to observe all other players' payoffs. The reason that traditional game theory focuses so much attention on the special case when players have complete information about these things is that equilibrium predictions are easier to motivate and derive in the complete information case, and often have little empirical content in the incomplete information case (in which most outcomes may be consistent with some equilibrium). But the reinforcement learning models we consider are well suited to modeling learning in quite general informational environments. So adaptive game theory may well have implications for an even wider range of economic phenomena than the traditional approach.

Having concentrated on a particular family of learning models, we would be remiss if we did not remark on the recent and fruitful interest by both theorists and experimenters in many different aspects and models of learning. This makes us optimistic about the prospects for increasingly fruitful *interaction* between game theorists and experimenters. In this connection there have already been a host of interesting experiments, from many different points of view, which begin to demonstrate the promise of learning models for understanding observed behavior in strategic environments.[33]

In closing, in this paper we have taken some steps in the direction of a cognitive game theory. We have shown that a simple model of learning can organize a wide range of data, but have also noted that it will be necessary on many classes of games to pay more attention to players' thought processes as exhibited by

[31] See also the "strategy method" of Selten (1967) (and the related paper by Selten et al., 1988). For game-theoretic work in which a good deal of attention is paid to both learning and the modeling of strategy, see Nagel (1995) and Dale O. Stahl (1996a, b).

[32] See for example the Fall 1997 issue of the *Journal of Economics and Management Strategy,* which is devoted to papers related to the recent design of the Federal Communications Commission's auctions of radio spectrum, or see Roth (1996a) and Roth and Elliot Peranson (1997), which describe the recent redesign of the entry-level labor market for new American physicians.

[33] In addition to the papers already mentioned, a sample of notable recent work of this sort might include Vincent P. Crawford (1991, 1992); Jordi Brandts and Charles A. Holt (1992, 1993); Cheung and Friedman (1994); Cooper et al. (1994); John B. Van Huyck et al. (1994); Kenneth B. Binmore et al. (1995); Nagel (1995); Terry E. Daniel et al. (1996); Stahl (1996a, b).

the strategies they are able to consider. The robustness of our results suggests that it may be possible to study learning in games using simple general models, appropriately adapted to particular circumstances, rather than having to build or estimate special models for each game of interest. Finally, we have argued that the general approach of considering how particular games and economic environments influence the dynamics of learning is likely to contribute to making game theory as useful a part of applied economics as it already is a part of economic theory.

REFERENCES

Anderson, John R. "Acquisition of Cognitive Skill." *Psychological Review*, July 1982, *89*(4), pp. 369–403.

_____. *The architecture of cognition*. Cambridge, MA: Harvard University Press, 1993.

Arthur, W. Brian. "Designing Economic Agents That Act Like Human Agents: A Behavioral Approach to Bounded Rationality." *American Economic Review*, May 1991 (*Papers and Proceedings*), *81*(2), pp. 353–59.

_____. "On Designing Economic Agents That Behave Like Human Agents." *Journal of Evolutionary Economics*, February 1993, *3*(1), pp. 1–22.

Atkinson, Richard C. and Suppes, Patrick. "An Analysis of Two-Person Game Situations in Terms of Statistical Learning Theory." *Journal of Experimental Psychology*, April 1958, *55*(4), pp. 369–78.

Bereby-Meyer, Yoella and Erev, Ido. "On Learning to Become a Successful Loser: A Comparison of Alternative Abstractions of Learning in the Loss Domain." *Journal of Mathematical Psychology*, 1998 (forthcoming).

Binmore, Kenneth G.; Gale, John and Samuelson, Larry. "Learning to Be Imperfect: The Ultimatum Game." *Games and Economic Behavior*, January 1995, *8*(1), pp. 56–90.

Blackburn, J. M. "Acquisition of Skill: An Analysis of Learning Curves." IHRB Report No. 73, 1936.

Bolton, Gary. "A Comparative Model of Bargaining: Theory and Evidence." *American Economic Review*, December 1991, *81*(5), pp. 1096–136.

Bolton, Gary and Ockenfels, Axel. "ERC: A Theory of Equity, Reciprocity and Competition." Mimeo, Penn State University, 1997.

Borgers, Tilman and Sarin, Rajiv. "Learning Through Reinforcement and Replicator Dynamics." Mimeo, University College London, 1994.

_____. "Naive Reinforcement Learning with Endogenous Aspirations." Mimeo, University College London, 1995.

Bornstein, Gary; Erev, Ido and Goren, Harel. "The Effect of Repeated Play in the IPG and IPD Team Games." *Journal of Conflict Resolution*, December 1994, *38*(4), pp. 690–707.

Bornstein, Gary; Winter, Eyal and Goren, Harel. "Experimental Study of Repeated Team Games." *European Journal of Political Economy*, December 1996, *12*(4), pp. 629–39.

Brandts, Jordi and Holt, Charles A. "An Experimental Test of Equilibrium Dominance in Signaling Games." *American Economic Review*, December 1992, *82*(5), pp. 1350–65.

_____. "Adjustment Patterns and Equilibrium Selection in Experimental Signaling Games." *International Journal of Game Theory*, 1993, *22*(3), pp. 279–302.

Brown, George W. "Iterative Solutions of Games by Fictitious Play," in Tjalling C. Koopmans, ed., *Activity analysis of production and allocation*. New York: Wiley, 1951, pp. 374–76.

Brown, J. S.; Clark, F. R. and Stein, L. "A New Technique for Studying Spatial Generalization with Voluntary Responses." *Journal of Experimental Psychology*, April 1958, *55*(4), pp. 359–62.

Brown, James N. and Rosenthal, Robert W. "Testing the Minimax Hypothesis: A Reexamination of O'Neill's Game Experiment." *Econometrica*, September 1990, *58*(5), pp. 1065–81.

Busemeyer, Jerome R. and Myung, In Jae. "An Adaptive Approach to Human Decision Making: Learning Theory, Decision Theory, and Human Performance." *Journal of Experimental Psychology: General*, June 1992, *121*(2), pp. 177–94.

Bush, Robert and Mosteller, Frederick. *Stochastic models for learning*. New York: Wiley, 1955.

Camerer, Colin. "Behavioral Game Theory," in Robin Hogarth, ed., *Insights in decision making: A tribute to Hillel J. Einhorn*. Chicago: University of Chicago Press, 1990, pp. 311–36.

Camerer, Colin and Ho, Teck-Hua. "Experience Weighted Attraction Learning in Normal-Form Games." *Econometrica,* 1998a (forthcoming).

_____. "Experience-Weighted Attraction Learning in Games: Estimates from Weak-Link Games," in David V. Budescu, Ido Erev, and Rami Zwick, eds., *Games and human behavior: Essays in honor of Amnon Rapoport's 60th birthday*. Hillsdale, NJ: Erlbaum, 1998b (forthcoming).

Cheung, Yin-Wong and Friedman, Daniel. "Learning in Evolutionary Games: Some Laboratory Results." Mimeo, University of California, Santa Cruz, 1994.

_____. "Individual Learning in Normal Form Games: Some Laboratory Results." Mimeo, University of California, Santa Cruz, December 1995.

_____. "A Comparison of Learning and Replicator Dynamics Using Experimental Data." Mimeo, University of California, Santa Cruz, October 1996.

Cooper, David and Feltovich, Nick. "Reinforcement-Based Learning vs. Bayesian Learning: A Comparison." Mimeo, University of Pittsburgh, 1996.

Cooper, David; Garvin, Susan and Kagel, John. "Signalling and Adaptive Learning in An Entry Limit Pricing Game." Mimeo, University of Pittsburgh, 1994.

Cox, James; Shacht, Jason and Walker, Mark. "An Experiment to Evaluate Bayesian Learning of Nash Equilibrium." Mimeo, University of Arizona, 1995.

Crawford, Vincent P. "An 'Evolutionary' Interpretation of Van Huyck, Battalio, and Beil's Experimental Results on Coordination." *Games and Economic Behavior*, February 1991, *3*(1), pp. 25–59.

_____. "Adaptive Dynamics in Coordination Games." Mimeo, University of California, San Diego, 1992.

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

Crossman, E. R. F. W. "A Theory of Acquisition of Speed-Skill." *Ergonomics*, November 1958, *2*(1), pp. 153–66.

Daniel, Terry E.; Seale, Darryl A. and Rapoport, Amnon. "Strategic Play and Adaptive Learning in the Sealed Bid Bargaining Mechanism." Mimeo, University of Arizona, April 1996.

Dickhaut, John; Mukherji, Arijit; Rajan, Vijay and Sevcik, Galen. "An Experimental Study of Learning, Equilibrium Refinements and Local Interaction in Games." Mimeo, University of Minnesota, 1995.

Duffy, John and Feltovich, Nick. "The Effect of Information on Learning in Strategic Environments: An Experimental Study." *International Journal of Game Theory,* 1998 (forthcoming).

Erev, Ido. "Signal Detection by Human Observers: A Cutoff Reinforcement Learning Model of Categorization Decisions Under Uncertainty." *Psychological Review*, 1998, *105*(2), pp. 280–98.

Erev, Ido; Gopher, Daniel; Itkin, Revital and Greenshpan, Yaacov. "Toward a Generalization of Signal Detection Theory to *N*-Person Games: The Example of Two-Person Safety Problem." *Journal of Mathematical Psychology*, December 1995, *39*(4), pp. 360–75.

Erev, Ido; Maital, Shlomo and Or-Hof, Ori. "Melioration, Adaptive Learning and the Effect of Constant Re-evaluations of Strategies," in G. Antoniedes, F. van Raaij, and S. Maital, eds., *Advances in economic psychology*. Sussex, U.K.: Wiley, 1997, pp. 237–53.

Erev, Ido and Rapoport, Amnon. "Magic, Reinforcement Learning, and Coordination in a Market Entry Game." *Games and Economic Behavior,* 1998, *23,* pp. 146–75.

Erev, Ido and Roth, Alvin E. "On the Need for Low Rationality, Cognitive Game Theory: Reinforcement Learning in Experimental Games with Unique Mixed Strategy Equilibria." Mimeo, University of Pittsburgh, 1996.

Estes, William K. "Toward a Statistical Theory of Learning." *Psychological Review*, March 1950, *57*(2), pp. 94–107.

Fehr, Ernst and Schmidt, Klaus. "How to Account for Fair and Unfair Outcomes—A Model of Biased Inequality Aversion." Paper presented at Gerzensee [Switzerland] Symposium on Economic Theory, July 1997.

Feltovich, Nick. "Learning and Equilibrium in an Asymmetric Information Game: An Experimental Study." Ph.D. dissertation, University of Pittsburgh, 1997.

Fudenberg, Drew and Levine, David K. "Measuring Players' Losses in Experimental Games." *Quarterly Journal of Economics*, May 1997a, *112*(2), pp. 507–36.

_____. *Theory of learning in games*. Draft, http://levine.sscnet.ucla.edu/papers/contents.htm (accessed March 10, 1997), 1997b.

Gilat, Sharon; Meyer, Joachim; Erev, Ido and Gopher, Daniel. "Beyond Bayes' Theorem: The Effect of Base Rate Information in Consensus Games." *Journal of Experimental Psychology: Applied*, June 1997, *3*(2), pp. 83–104.

Guth, Werner. "On Ultimatum Bargaining Experiments—A Personal Review." *Journal of Economic Behavior and Organization*, August 1995, *27*(3), pp. 329–44.

Guthrie, Edwin R. *The psychology of learning*. New York: Harper, 1952.

Guttman, N. and Kalish, H. "Discriminability and Stimulus Generalization." *Journal of Experimental Psychology*, January 1956, *51*(1), pp. 79–88.

Harley, Calvin B. "Learning the Evolutionarily Stable Strategy." *Journal of Theoretical Biology*, April 1981, *89*(4), pp. 611–33.

Herrnstein, Richard J. "On the Law of Effect." *Journal of the Experimental Analysis of Behavior*, March 1970, *13*(2), pp. 243–66.

Kagel, John H. and Roth, Alvin, E. *Handbook of experimental economics*. Princeton, NJ: Princeton University Press, 1995.

Kahneman, Daniel and Tversky, Amos. "Prospect Theory: An Analysis of Decision Under Risk." *Econometrica*, March 1979, *47*(2), pp. 263–91.

Lee, Wayne. *Decision theory and human behavior*. New York: Wiley, 1971.

Luce, Duncan R. *Individual choice behavior*. New York: Wiley, 1959.

Macy, Michael W. "Learning to Cooperate: Stochastic and Tacit Collusion in Social Exchange." *American Journal of Sociology*, November 1991, *97*(3), pp. 808–43.

Malcolm, David and Lieberman, Bernhardt. "The Behavior of Responsive Individuals Playing a Two-Person, Zero-Sum Game Requiring the Use of Mixed Strategies." *Psychonomic Science*, June 1965, (12), pp. 373–74.

Maynard-Smith, John. *Evolution and the theory of games*. Cambridge: Cambridge University Press, 1982.

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

Messick, David M. and Liebrand, Wim B. G. "Individual Heuristics and the Dynamics of Cooperation in Large Groups." *Psychological Review*, January 1995, *102*(1), pp. 131–45.

Mookherjee, Dilip and Sopher, Barry. "Learning Behavior in an Experimental Matching Pennies Game." *Games and Economic Behavior*, July 1994, *7*(1), pp. 62–91.

_____. "Learning and Decision Costs in Experimental Constant Sum Games." *Games and Economic Behavior*, April 1997, *19*(1), pp. 97–132.

Nagel, Rosemarie. "Unraveling in Guessing Games: An Experimental Study." *American Economic Review*, December 1995, *85*(5), pp. 1313–26.

Ochs, Jack. "Simple Games with Unique Mixed Strategy Equilibrium: An Experimental Study." *Games and Economic Behavior*, July 1995, *10*(1), pp. 202–17.

Ochs, Jack and Roth, Alvin E. "An Experimental Study of Sequential Bargaining." *American Economic Review*, June 1989, *79*(3), pp. 355–84.

O'Neill, Barry. "Nonmetric Test of the Minimax Theory of Two-Person Zerosum Games." *Proceedings of the National Academy of Sciences, USA*, April 1987, *84*(7), pp. 2106–9.

Payne, John W.; Bettman, James R. and Johnson, Eric J. *The adaptive decision maker*. Cambridge: Cambridge University Press, 1993.

Rabin, Matthew. "Incorporating Fairness into Game Theory and Econometrics." *American Economic Review*, December 1993, *83*(5), pp. 1281–303.

Rapoport, Amnon and Boebel, Richard B. "Mixed Strategies in Strictly Competitive Games: A Further Test of the Minmax Hypothesis." *Games and Economic Behavior*, April 1992, *4*(2), pp. 261–83.

Rapoport, Amnon and Budescu, David V. "Generation of Random Series in Two-Person Strictly Competitive Games." *Journal of Experimental Psychology: General*, September 1992, *121*(3), pp. 352–63.

Rapoport, Amnon; Erev, Ido; Abraham, Elizabeth V. and Olson, David E. "Randomization and Adaptive Learning in a Simplified Poker Game." *Organizational Behavior and Human Decision Processes*, January 1997, *69*(1), pp. 31–49.

Rapoport, Amnon; Seale, Darryl A.; Erev, Ido and Sundali, James A. "Coordination Success in Market Entry Games: Tests of Equilibrium and Adaptive Learning Models." *Management Science*, 1998, *44*, pp. 119–41.

Rapoport, Anatol; Guyer, Melvin J. and Gordon, David G. *The 2 × 2 game*. Ann Arbor, MI: University of Michigan Press, 1976.

Robinson, Julia. "An Iterative Method of Solving a Game." *Annals of Mathematics*, September 1951, *54*(2), pp. 296–301.

Roth, Alvin E. "The Evolution of the Labor Market for Medical Interns and Residents: A Case Study in Game Theory." *Journal of Political Economy*, December 1984, *92*(6), pp. 991–1016.

_____. "New Physicians: A Natural Experiment in Market Organization." *Science*, December 14, 1990, *250*(4987), pp. 1524–28.

_____. "A Natural Experiment in the Organization of Entry-Level Labor Markets: Regional Markets for New Physicians and Surgeons in the United Kingdom." *American Economic Review*, June 1991, *81*(3), pp. 415–40.

_____. "Introduction to Experimental Economics," in John Kagel and Alvin E. Roth, eds., *Handbook of experimental economics*. Princeton, NJ: Princeton University Press, 1995, pp. 3–109.

_____. "Individual Rationality as a Useful Approximation: Comments on Tversky's 'Rational Theory and Constructive Choice,'" in K. Arrow, E. Colombatto, M. Perlman, and C. Schmidt, eds., *The rational foundations of economic behavior*. London: Macmillan, 1996a, pp. 198–202; http://www.economics.harvard.edu/faculty/roth/rational.html.

_____. "Report on the Design and Testing of an Applicant Proposing Matching Algorithm, and Comparison with the Existing NRMP Algorithm." http://www.pitt.edu/economics.harvard.edu/faculty/roth/phase1.html, 1996b.

Roth, Alvin E. and Erev, Ido. "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term." *Games and Economic Behavior*, Special Issue: Nobel Symposium, January 1995, *8*(1), pp. 164–212.

Roth, Alvin E. and Peranson, Elliott. "The Effects of a Change in the NRMP Matching Algorithm." *Journal of the American Medical Association*, September 3, 1997, *278*(9), pp. 729–32.

Roth, Alvin E.; Prasnikar, Vesna; Okuno-Fujiwara, Masahiro and Zamir, Shmuel. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *American Economic Review*, December 1991, *81*(5), pp. 1068–95.

Roth, Alvin E. and Xing, Xiaolin. "Jumping the Gun: Imperfections and Institutions Related to the Timing of Market Transactions." *American Economic Review*, September 1994, *84*(4), pp. 992–1044.

Selten, Reinhard. "Die Strategiemethode zur Erforschung des Eingeschrankt Rationalen Verhaltens im Rahmen eines Oligopolexperiments," in Heinz Sauermann, ed., *Beitrage zur experimentellen Wirtschaftsforschung*. Tubingen, Germany: Mohr, 1967, pp. 136–68.

_____. "Axiomatic Characterization of the Quadratic Scoring Rule." *Experimental Economics*, 1998, *1*(1), pp. 43–62.

Selten, Reinhard and Buchta, Joachim. "Experimental Sealed Bid First Price Auction with Directly Observed Bid Functions." Discussion Paper No. B270, University of Bonn, Germany, 1994.

Selten, Reinhard; Mitzkewitz, Michael and Uhlich, Gerald R. "Duopoly Strategies Programmed by Experienced Players." Mimeo, University of Bonn, Germany, 1988.

Skinner, B. F. *Science and human behavior*. New York: Macmillan, 1953.

Slonim, Robert and Roth, Alvin E. "Learning in High-Stakes Ultimatum Games: An Experiment in the Slovak Republic." *Econometrica,* May 1998, *66*(3), pp. 569–96.

Stahl, Dale O. "Boundedly Rational Rule Learning in a Guessing Game." *Games and Economic Behavior*, October 1996a, *16*(2), pp. 303–30.

_____. "Evidence Based Rule Learning in Symmetric Normal-Form Games." Mimeo, University of Texas, April 1996b.

Suppes, Patrick and Atkinson, Richard C. *Markov learning models for multiperson interactions*. Stanford, CA: Stanford University Press, 1960.

Tang, Fang-Fang. "Anticipatory Learning in Two-Person Games: An Experimental Study, Part I, Equilibrium and Stability." Discussion Paper No. B-362, University of Bonn, Germany, March 1996a.

_____. "Anticipatory Learning in Two-Person Games: An Experimental Study, Part II, Learning." Discussion Paper No. B-363, University of Bonn, Germany, March 1996b.

_____. "Anticipatory Learning in Two-Person Games: An Experimental Study, Part III, Individual Analysis." Discussion Paper No. B-364, University of Bonn, Germany, March 1996c.

Thorndike, Edward L. "Animal Intelligence: An Experimental Study of the Associative Processes in Animals." *Psychological Monographs*, 1898, *2*(8).

Tinkelepaugh, L. H. "An Experimental Study of Representative Factors in Monkeys." *Journal of Comparative Psychology*, June 1928, *8*(3), pp. 197–236.

Tversky, Amos and Kahneman, Daniel. "Judgment Under Uncertainty: Huristics and Biases." *Science*, September 27, 1974, *185*(4157), pp. 1124–31.

Van Huyck, John B.; Cook, Joseph P. and Battalio, Raymond C. "Selection Dynamics, Asymptotic Stability, and Adaptive Behavior." *Journal of Political Economy*, October 1994, *102*(5), pp. 975–1005.

Watson, John B. *Behaviorism*, 2nd. Ed. Chicago: University of Chicago Press, 1930.

Weibull, Jorgen W. *Evolution and the theory of games*. Cambridge, MA: MIT Press, 1995.

Weisbuch, Gerard; Kirman, Alan and Herreiner, Dorothea. "Market Organization." http://www.lps.ens.fr/~weisbuch/mark.html (accessed May 21, 1996), 1996.