The Speed of Learning in Noisy Games: Partial Reinforcement and the Sustainability of Cooperation

The Speed of Learning in Noisy Games:
Partial Reinforcement and the Sustainability of Cooperation

By Yoella Bereby Meyer and Alvin E. Roth

Ben Gurion University of the Negev, and Harvard University

January 26, 2006

Abstract: In an experiment, players' ability to learn to cooperate in the repeated prisoner's dilemma was substantially diminished when the payoffs were noisy, even though players could monitor one another's past actions perfectly. In contrast, in one-time play against a succession of opponents, noisy payoffs increased cooperation, by slowing the rate at which cooperation decays. These observations are consistent with the robust observation from the psychology literature that partial reinforcement (adding randomness to the link between an action and its consequences while holding expected payoffs constant) slows learning. This effect is magnified in the repeated game: When others are slow to learn to cooperate, the benefits of cooperation are reduced, which further hampers cooperation. These results show that a small change in the payoff environment, which changes the speed of individual learning, can have a large effect on collective behavior. And they show that there may be interesting comparative dynamics that can be derived from careful attention to the fact that at least some economic behavior is learned from experience.

Acknowledgements: This research has been partially supported by a grant from the National Science Foundation. This paper has benefited from the comments of , Guillaume Frechette, Ido Erev, Muriel Niederle, and anonymous referees on an earlier draft.

Just as the hypothesis that prices are at equilibrium can be used to generate fruitful comparative statics predictions, the observation that some kinds of behavior are learned from experience has the potential to generate predictions about comparative dynamics. It is therefore useful to ask in what kind of economic environments we should expect learning to occur quickly, or slowly, and how changes in the speed of learning might affect the behavior that emerges in complex strategic environments such as repeated games. We will focus here on learned cooperation (and learned non-cooperation).

Theories of reinforcement learning in strategic environments, which recently have gained attention in economics[1], suggest that players' ability to learn to cooperate will be hampered if the payoffs they earn are noisy, even when they can noiselessly monitor each other's actions. This is related to the observation, long known in the psychology literature, that partial reinforcement of actions slows learning, especially early learning, i.e. when subjects are inexperienced, *(*see e.g. Solomon Weinstock, 1958, and for a review see Donald Robbins, 1971).

In the psychology literature, an action is said to be "partially reinforced" if it is rewarded only some of the time, in contrast to actions that are rewarded every time they are taken, which are said to be fully (or "continuously") reinforced. An early general conclusion from this literature was that learning "proceeds somewhat more rapidly and reaches a higher final training level under continuous reinforcement than under partial reinforcement" (William Jenkins and Julian Stanley, 1950).[2] That is, variance in how often an action is rewarded slows learning [3]. So the distinction between partial and full reinforcement is a potentially important one for economics, since few economic decisions in natural environments yield deterministic outcomes.

We will compare games having deterministic payoffs with games having probabilistic payoffs with the same expected value. The "partial reinforcement learning" hypothesis is

[1] See e.g. Alvin Roth and Ido Erev, 1995; Erev and Roth, 1998; Drew Fudenberg and David Levine 1998; John Duffy and Nick Feltovich, 1999; Feltovich 2000.

[2] Although they also noted that, "With prolonged training partially reinforced groups may approach the same level of proficiency …" as fully reinforced groups of subjects**.** In Stanley's (1950) experiment, total reinforcement was kept constant when comparing various partial and full reinforcement conditions, by having correspondingly more periods when reinforcement was random. In the experiments reported here, we keep the expected payoff constant, by having larger prizes when there is a smaller probability of reward.

[3] Of course the change from full to partial reinforcement changes not only the variance in how *often* an action is rewarded, but also the variance in how *much* an action is rewarded. See Erev et al. (1999) for a reinforcement learning model that models variance in payoffs as a determinant of learning speed.

that *whatever learned behavior we see in the deterministic games will develop more slowly in the probabilistic games*.

We will see that a small change in the payoff environment, which induces a change in the speed of individual learning, can have a large impact on whether mutual cooperation is learned. This in turn will raise some questions about how to model the payoffs in economic environments in which learning is likely to be an important influence on behavior.

It has been customary to model the payoffs in games as expected payoffs of one sort or another, at least since the axiomatization of expected utility by John von Neumann and Oskar Morgenstern (1944).[4] The results of the present paper will suggest that, since variance in early experience can change early behavior, it may not always be innocuous to model payoffs as expectations, even when the income stream will consist of many small payoffs. Variance may change behavior in a way that cannot be explained by the more conventional approach of assuming that risk may simply change the relative *desirability* of different outcomes (risk aversion).[5]

The hypothesis that partial reinforcement slows learning in games has different consequences for different kinds of games. When players have dominant strategies, as in the prisoner's dilemma, it suggests that, as in individual choice tasks, the path of play will simply change more slowly as variance in the frequency of rewards is increased. That is, a player who has a dominant strategy has an opportunity to learn to play it regardless of what other players are (simultaneously) learning, so the player faces an environment similar to the individual choice tasks that make up the bulk of the psychology literature on learning.[6].

Specifically, players in deterministic prisoner's dilemma games in which players are rematched with different players each period are observed to cooperate less as they gain experience (see e.g. Russell Cooper, Douglas Dejong, Robert Forsythe and Thomas Ross

[4] Expected payoffs are also used in biological models of evolutionary games. In these models, payoffs are not in utility, but expectations reflect the assumption that variance in payoffs will be smoothed by large numbers of interactions. Note also that while reinforcement models of learning take as input realized payoffs, other kinds of learning models such as fictitious play can be implemented largely in terms of expected payoff (although in their dependence on realized payoffs to form expectations they have much similarity with reinforcement models) see e.g. Fudenberg and Levine, 1998; Feltovich, 2000; Colin Camerer and Tek Ho, 1999.

[5] Indeed, in our one-period game condition with probabilistic payoffs, the fact that the payoffs are binary lotteries (Roth and Michael Malouf, 1979) implies that expected utility maximizers whose utilities are linear in the small payoffs in the deterministic condition (cf. Matthew Rabin, 2000) would have exactly the same expected utility from corresponding outcomes in both the probabilistic and deterministic conditions.

[6] See e.g. Wayne Lee (1971) for an overview of individual learning in probability matching.

1996). So the prediction of the partial reinforcement hypothesis is that players will learn to play their dominant strategy more slowly in games with noisy payoffs than in games with deterministic payoffs, i.e. they will learn not to cooperate, but more slowly than in the deterministic game.

But in more complex strategic environments, like the *repeated* prisoner's dilemma, defection is no longer a dominant strategy, and what a player learns from early experience of a game depends on what actions other players are choosing. The issues are clearest if the same pair of players will play for a fixed number of periods, known to both of them at the outset. We say a given pair of players plays the "n-period supergame" generated by a one-period game matrix if they each simultaneously make their choices in each of n periods, learning after each period the other player's choice the previous period, and each receives the sum of the n payments that are generated by the pairwise choices. When there are still future periods to play, each player can reward or punish the other for past actions, by cooperating or defecting in the future. Thus there are rewards for (conditional) mutual cooperation, but these diminish as the game nears the end, and in the last period the incentives are the same as in the one-period game [7].

A number of experiments have observed that players learn to reciprocate cooperative behavior as they gain experience with the repeated game. For example. Reinhard Selten and Rolf Stoecker (1986), James Andreoni and John Miller (1993), and Esther Hauk and Rosemarie Nagel (2001) observe that when people play n-period prisoner's dilemma supergames multiple times, against different players, they often learn to cooperate in the early periods of the supergame, but cooperation breaks down near the end of the supergame.[8] So if other players are slow to learn to cooperate, then the rewards of

[7] For this reason there is no cooperation at equilibrium in the simplest model of complete information repeated play with a fixed deadline, but see e.g. David Kreps et al (1982) for finite repeated games with incomplete information, and Roth and J. Keith Murnighan (1978) for repeated games with probabilistic termination, both of which are consistent with some cooperation at equilibrium.

[8] Because cooperation tends to break down when it is not reciprocated, cooperation is much more difficult to maintain if *actions are noisy*, i.e. if there is random error either in choosing actions or in monitoring others' actions; see e.g. Robert Axelrod and Douglas Dion, 1988; Jonathan Bendor, 1987; Bendor, 1993; Edward Green and Robert Porter, 1984; Per Molander, 1985; Miller, 1996; Barbara Sainty, 1999. That is, if actions are noisy, a player does not know whether another player's defection was an error or an intended choice, and strategies involving reciprocation (e.g. "tit for tat") can break down. The thrust of this literature on noisy actions is that the path to cooperation, which may not be too difficult in repeated games in which players can noiselessly monitor one another's behavior, is seriously complicated when actions are noisy. In this context, our results can be interpreted as showing that, since learning is slower when payoff variance increases,

cooperation will be less, which will further make cooperation difficult to learn. In particular, the learning hypothesis suggests that in two n-period prisoner's dilemma supergames with identical expected payoffs, cooperation in the early periods may be harder to achieve in the supergame with more variability in how often cooperation is rewarded, even when players can perfectly monitor one another's choices.

To test these hypotheses, we report an experiment with the prisoner's dilemma, that allowed subjects to gain experience either with the non-repeated game (in which players were rematched with new partners after every play of the game), or the repeated supergame (in which players played with the same partner for ten periods) in games that, holding the expected payoff fixed, either had deterministic payoffs (full reinforcement) or probabilistic payoffs (partial reinforcement).

Of course another difference between the deterministic and probabilistic conditions is that they have different information sets, and hence give the players different sets of strategies (since in the probabilistic condition players can condition their behavior on the outcome of the lotteries as well as upon past choices). So, as a control, we also consider games played under a third, "deterministic plus sunspots" condition, in which payoffs are deterministic, but participants also receive information about the outcome of the two binary lotteries ("sunspots"), which do not themselves influence the payoffs, but have the same distribution as the lotteries in the probabilistic condition . That is, the "deterministic plus sunspots" condition has the same payoffs as the deterministic condition, and the same information sets (and hence the same strategy sets) as the probabilistic condition. So it enables us to test if the difference between the deterministic and probabilistic conditions is due to slower learning due to increased variance (the learning hypothesis), or due to players' ability to condition on the outcome of the lotteries (the "strategic hypothesis").

To summarize, we examine a 2x3 experimental design, (one-time or repeated play) x (deterministic, deterministic with sunspots, or probabilistic payoffs). We expect from the results of prior experiments that in the deterministic non-repeated games, players will learn to cooperate less over time, and in the deterministic repeated game, players will learn to cooperate more over time in the early periods of each supergame. The hypothesis that

---

cooperation in repeated games is even more fragile to noise than has been thought, with noisy *payoffs* being an issue even when actions are unaffected by noise.

partial reinforcement slows learning therefore implies that, in games with probabilistic payoffs, having the same expected payoffs but more variance, players will learn more slowly. So the random-payoffs should elicit *more* cooperation in the non-repeated game, and *less* cooperation in the early periods of the repeated game than the comparable games with deterministic payoff.

If, instead, the difference between the deterministic and probabilistic conditions is primarily due to the difference in their information sets and strategy spaces, we would expect behavior in the deterministic plus sunspots condition to more closely resemble the probabilistic condition. However, if the difference is primarily due to slow learning associated with the random payoffs, then we would expect that behavior in the deterministic plus sunspots condition will more closely resemble the deterministic condition.

The results we report below confirm the predictions of the learning hypothesis, and further suggest that the effect may be more dramatic. We will see that in the repeated game, in which cooperation needs to be learned, increased variance may slow learning to the point that cooperation is hardly learned at all. Thus the play of two repeated games with the same expected payoffs may be affected quite profoundly by the variance of the random variable that counts how often the actions are rewarded.

The Experiments:

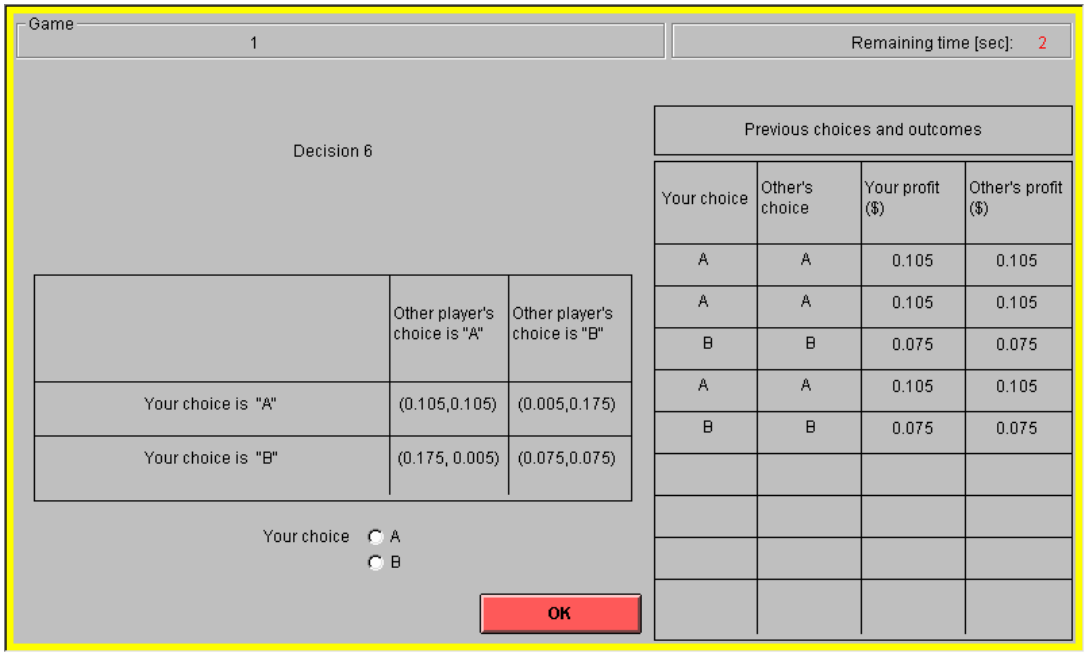
The experiments were programmed in z-Tree (Fischbacher, 1999). In each condition of both the non-repeated game and of the repeated game experiments, the players' payoffs in each period were represented to them by the matrix on the left side of Figure 1. (Figure 1 shows the screen that participants saw in the deterministic payoff condition of the repeated game experiment.) In the non-repeated game experiment, participants saw only the left side of this screen.

In our experimental conditions for one-time play, subjects played prisoner's dilemma games against different partners for 200 periods with either deterministic, probabilistic, or deterministic payoffs with sunspots.

In our experimental conditions for repeated play, subjects played 20 ten-period prisoner's dilemma supergames with different partners, with perfect (noiseless) monitoring

of one another's actions. Because players learn one another's choices after each period, there is no obstacle to reciprocation.

Figure 1: A feedback screen that participants saw before each decision, in the deterministic condition of the repeated game experiment



1. The non-repeated game experiment:

One hundred and seventy two participants, in 13 sessions of 10-20 participants, played 200 one period prisoner's dilemma games[9]. Pairings were anonymous (players sat at visually isolated computer terminals), and after each period players were re-matched with another partner. Players were assigned to one of the three experimental conditions (deterministic, probabilistic or deterministic payoffs with sunspots). We ran 5 sessions in the deterministic condition (10,10,10,18,18 participants), 5 sessions in the probabilistic

[9] The participants were not told the number of periods that they would play. Players earned a show-up fee of $15 plus their accumulated earnings in the games they played.

condition (10,10,10,20,20 participants) and 3 sessions in deterministic payoffs with sunspots (12,12,12, participants)[10,11].

In the probabilistic condition, the numbers in Figure 1 indicated the probability that each player would win a fixed amount of money ($1) in each period. For example, if in a period of the game the two players both cooperated, they would each immediately participate in a lottery that gave them a 10.5% chance of winning $1, and an 89.5% chance of winning 0 for that period. Their earnings in the experiment would be the sum of their earnings over the two hundreds periods in which they participated.

In the deterministic condition, in contrast, each player was credited each period with the number of cents in the matrix, i.e. with the expected value of the corresponding action in the probabilistic treatment. For example, in a period in which both players cooperated, each would be credited with $1 x .105 = $0.105, and their payoffs for the experiment would be the sum of their payoffs in each period.

Thus the expected payoffs of the players for a given pair of actions was identical in both treatments. Furthermore, because each player participated in 200 periods of play, the variance in expected payoffs between treatments would be small if players' actions were the same in both conditions of the experiment [12]. After each period the players were informed of their action, the action the other player had taken and their own payoff. In the probabilistic condition the payoffs would all be either 1.00 or 0.00, depending on the outcome of each player's (independent) lottery for the period. In the deterministic condition with sunspots, participants were paid as in the deterministic condition, and also saw the results of the two lotteries,.

Results:

Figure 2 shows the proportion of cooperation as a function of the period and the experimental condition (deterministic, probabilistic or deterministic with sunspots),

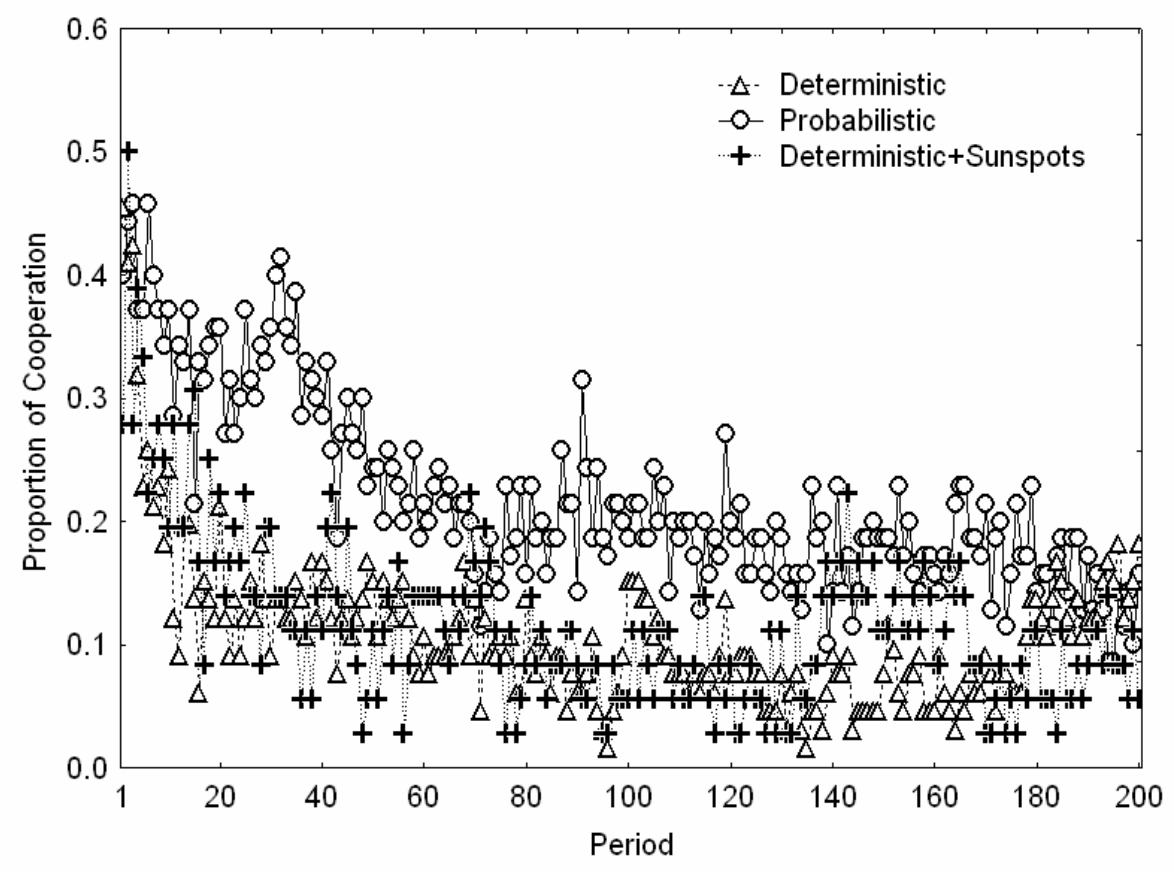
[10] In the deterministic payoffs with sunspots condition, subjects were told that their payoffs did not depend on the lotteries, which were related to another condition of the same experiment. The lottery distributions were explained as in the probabilistic condition.

[11] We ran 3 sessions in the deterministic and the probabilistic condition in Israel and 2 sessions in the USA. The 3 sessions of the deterministic payoffs with sunspots condition were run in Israel. Since no difference was found as a function of place, we averaged across the different places.

averaging across the sessions. The proportion of cooperation was computed by averaging the choice of cooperation (C) responses for every period across individuals and across sessions.

[12] For example, if all players cooperated in every period, in the deterministic condition each would receive 200 x $.105 = $21, while in the probabilistic condition they would receive a mean payoff of $21 with a standard deviation of 4.3.

Figure 2: The proportion of cooperation as a function of the period and the experimental condition (Deterministic, Probabilistic or Deterministic with Sunspots).



The results of the deterministic and the probabilistic conditions are as expected, and the results of the deterministic with sunspots condition closely track those of the deterministic condition. When players received deterministic payoffs they learned faster to defect - to play the dominant strategy - than when they received a random payoff. As can be seen in Figure 2, in the first period the rate of cooperation in the conditions was similar. Throughout most of the experiment and almost until the end, the rate of cooperation in the probabilistic condition was higher than the rate of cooperation in the deterministic condition. Pairwise t-tests for independent samples on mean proportions of cooperation for blocks of 20 periods revealed a significant difference between the deterministic and the probabilistic condition in all blocks from the first to the ninth block ($p<0.05$). After 180 periods in both conditions the rate of cooperation was similar and was close to ten percent. Note that in the deterministic condition it took participants only 20 trials to reach this level

of cooperation, while in the probabilistic condition it took them 180 trials. Pairwise t-tests for independent samples on mean proportions of cooperation for blocks of 20 periods revealed no difference in all blocks between the deterministic and the deterministic with sunspots conditions.

The results for the nonrepeated prisoner's dilemma thus closely resemble the results reported in the psychology literature concerning learning in individual choice tasks. In the probabilistic payoff (partial reinforcement) condition, the learned behavior is much like the learned behavior in the deterministic payoff condition, but the learning is slower. And since the players are learning not to cooperate, there is more cooperation in the probabilistic condition; i.e. learning is slower, so the rate of cooperation diminishes more slowly.

The fact that the behavior in the deterministic with sunspots condition resembles the behavior in the deterministic condition supports the claim that the slower decline in the amount of cooperation is due to the noise in the payoffs and not the change it induces in the information sets.

We turn next to the more complicated strategic situation of the repeated game.

2. **The repeated game**

One hundred and ninety eight participants, in 11 sessions of 14-22 participants, played 20 consecutive prisoner's dilemma supergames with 10 periods in each supergame[13]. Pairings were anonymous (players sat at visually isolated computer terminals), and after each game players were re-matched with another partner. Players were assigned to one of the three experimental conditions (deterministic, probabilistic or deterministic payoffs with sunspots). We ran 4 sessions in the deterministic condition, 3 sessions in the probabilistic condition and 4 sessions in the deterministic payoffs with sunspots[14,15]. The payoff matrices were the same as the ones that were given to the participants in the non-repeated game. In

[13] The participants were not told the number of supergames that they were going to play, but they knew that each supergame would last for ten periods. Participants earned a showup fee of $15 plus their accumulated earnings in the 20 repeated games.

[14] The assignment of the participants to the deterministic or the probabilistic condition was random. The deterministic with sunspots condition was run later as a control condition.

[15] The extra, initial session for the deterministic condition was to verify that we had chosen a game that would reproduce the cooperative behavior observed by earlier investigators in the deterministic condition. With this in mind we chose a payoff matrix comparable to those earlier investigators: using the indices of cooperation proposed by Anatol Rapoport and Albert Chammah (1965), for our matrix they are r1=(R-P)/(T-S) =0.17 and

each condition, the players' payoffs in each period were represented to them by the matrix on the left side of Figure 1.

After each period the players were informed of both their own and the other player's payoff and actions. In the deterministic condition with sunspots, participants also saw the results of the two binary lotteries, each of which could yield either 1 or 0 with probability equal to the payoff in the appropriate cell of the payoff matrix.[16].

Figure 1 shows a feedback screen from the deterministic condition. Each player always saw what action the other player had taken in the previous period, as well as the payoffs received. In the probabilistic condition the display was the same, with exactly the same payoff matrix. On the right hand side of the screen the payoffs in the probabilistic condition would all be either 1.00 or 0.00, depending on the outcome of each player's (independent) lottery for the period.

The results of the probabilistic condition are dramatically different from the deterministic conditions. When players received deterministic payoffs, either with or without sunspots, our results reproduce those of Selten and Stoecker (1986), Andreoni and Miller (1993) and Hauk and Nagel (2001). Players learn in the first few supergames to cooperate early in the game, and to defect in the periods near the end. But in the random payment condition, even though players receive the same expected payoffs, and even though they can observe each others' actions, they do not achieve substantial levels of cooperation even after gaining experience with the repeated game.

Figure 3 shows the proportion of cooperation as a function of the period in the supergame for each session of the three experimental conditions, averaging the proportion of cooperation across 5 consecutive supergames. That is, one of the lines in the graph represents supergames 1-5, another supergames 6-10, a third for supergames 11-15 and a fourth for supergames 16-20. The proportion of cooperation was computed by averaging the choice of cooperation (C) responses for every period across individuals and 5 repeated games. There are two more lines: one represents the rate of cooperation in the very first supergame and the other represents the rate of cooperation in the last supergame.

---

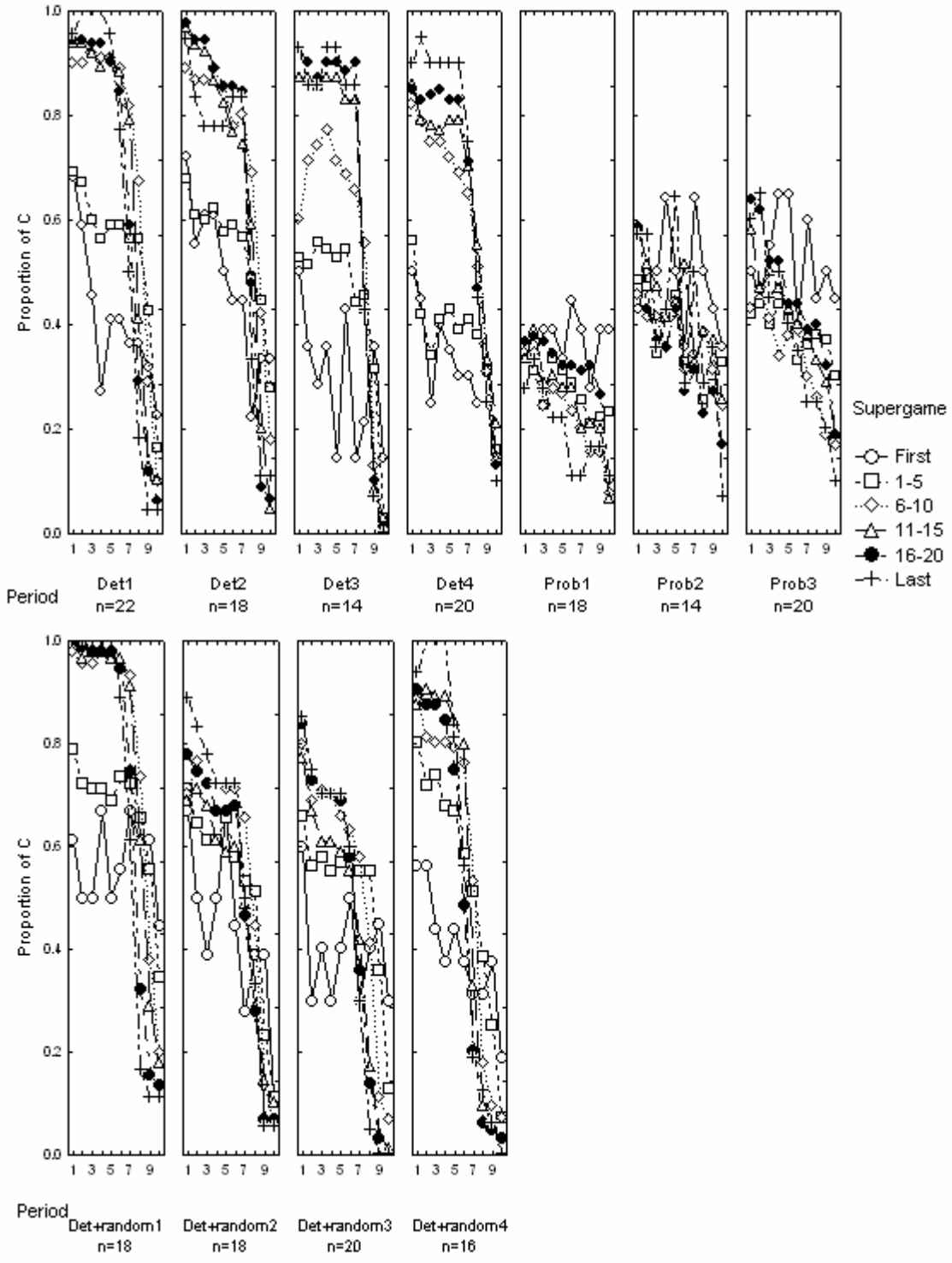
r2=(R-S)/(T-S)=0.58. These indices were close to the ones in Selten and Stoecker (1986) where r1=0.26 and r2=0.56 and in Andreoni and Miller (1993) where r1=0.25 and r2=0.58.

[16] The instructions regarding the sunspots events were as in the non-repeated game condition.

To understand the figures, look first at the line representing the first five supergames in the deterministic session at the far left of Figure 3, with n=22 subjects being matched to one another. In that session, the proportion of cooperation in the first period of the 10 period supergame climbed from around 0.7 for the first five supergames, to well over 0.9 for the last five supergames (represented by the red triangles). The same pattern of results can be seen in the deterministic condition with sunsposts.

But, while the players were learning to cooperate in period 1, they were also learning to stop cooperating by period 10: the proportion of cooperative choices declined in that session from about .15 in period 10 of the first five supergames, to about 0.05 in the last period of the final five supergames

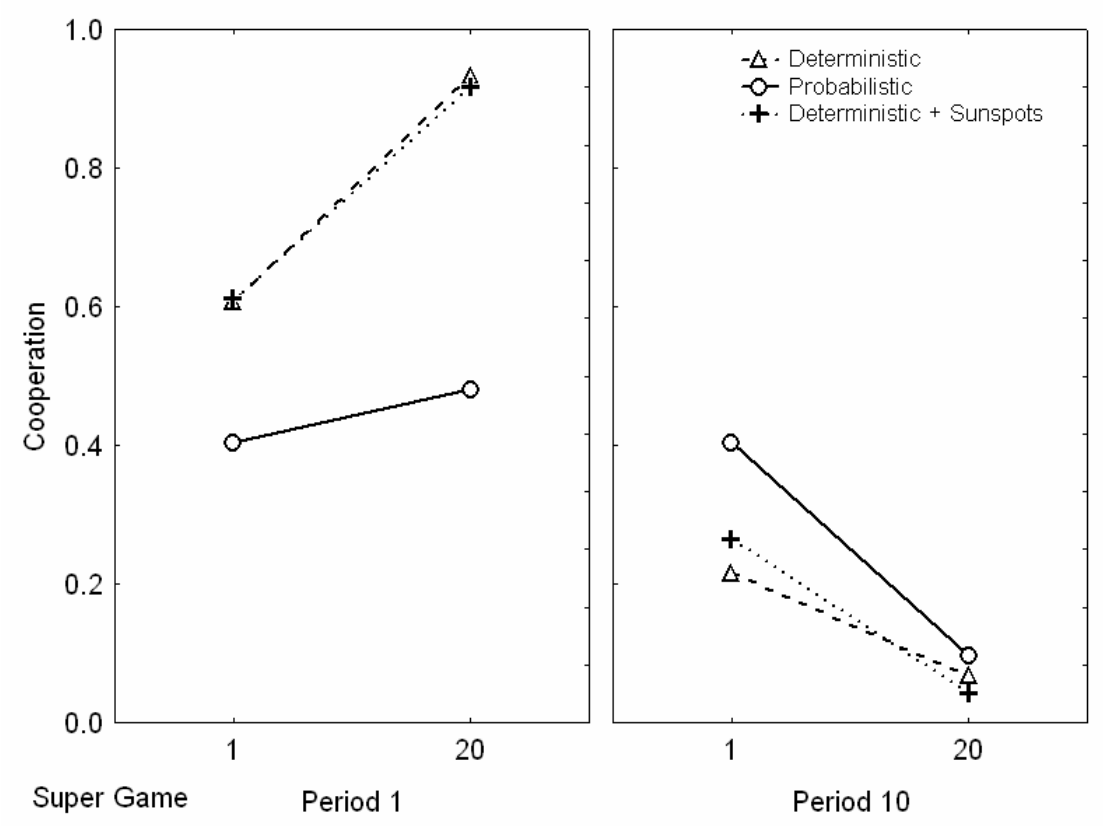
Figure 3: The proportion of cooperation in each session as a function of the period in a 10 period Prisoner's Dilemma supergame. Each line represents a group of 5 consecutive supergames, for the different experimental conditions, Deterministic payoffs (Det1-4), Probabilistic payoffs (Prob1-3) and Deterministic with Sunspots (Det1-4+). (Experimental sessions are presented separately to show that the phenomena are observed robustly, in every session, not just in aggregate.)



The left side of Figure 4 shows the change in cooperation in period 1, from the first to the last supergame, for the three experimental conditions. We found that in the deterministic condition the rate of cooperation in period 1 of the first supergame was .61 and it increased to .93 in the last supergame ($t(146)=4.62$, $p<0.001$)[17]. A significant increase (from .61 to .92) was also found for the deterministic condition with sunspots ($t(142)=4.38$, $p<0.001$). The two deterministic lines are almost identical, while the line representing the probabilistic condition is both lower and flatter.[18]

The right side represents the change in cooperation in period 10. In the deterministic condition cooperation dropped from .22 to .07 ($t(146)=2.59$, $p<0.01$). A significant drop in cooperation from .26 to .04 was also found for the deterministic payoff condition with sunspots ($t(142)=3.69$, $p<0.002$) .

Figure 4: Changes in cooperation in the first and last trial as a function of experience and condition



[17] We analyzed it using Statsoft, STATISTICA 7 ®

[18] In the very first period of the first supergame the probabilities of cooperation for the different sessions in the deterministic condition were: 0.68, 0.72, 0.5, 0.5 and for the different sessions in the probabilistic condition were: 0.27, 0.43, 0.5. Although the mean cooperation in the very first period was higher (0.61) for the deterministic than for the probabilistic (0.4) condition, if we look at deterministic sessions with the same level of cooperation as the probabilistic session (0.5), the difference between the conditions remains, and no learning is observed in the probabilistic condition.

That is, in the deterministic conditions, as players gained experience with the supergame they learned both to cooperate early and to defect late. This pattern was evident already in the early supergames, but it became much more pronounced with experience, with over 90% cooperation at the start of the last supergames, continuing at a high rate through the first few periods, and dropping sharply by periods 7 or 8, to end with almost no cooperation at all by period 10 of the last supergames.

The situation was dramatically different in the condition of the experiment with probabilistic payoffs. As Figures 3 and 4 show, in the probabilistic condition there is much less learning to cooperate in the early periods from one supergame to the next (and slower learning *not* to cooperate in the late periods already in the first supergame, although the level of late cooperation ends up about the same in the last supergame[19]). There was no significant change in cooperation in the first period, from the first to the last supergame (.4 to .48). A significant drop in cooperation from .4 in the first to .1 in the last supergame was found in period 10 ($t(102)=3.53$, $p<0.001$. Within a supergame there is more cooperation at period 1 than at period 10. And the total level of cooperation never rises above 60%, i.e. there is no more cooperation in early periods of the last supergames with random payoffs than there is in the first supergames with deterministic payoffs.

To get a clearer look at how the probabilistic payoff condition differed from the deterministic condition, we analyzed the extent to which participants conditioned their behavior on the outcome of their lottery, as well as on their partner's behavior. Table 1 shows the results

[19] Figures 3 and 4 show that even in the first probabilistic supergame the decline in cooperation from period 1 to period 10 is less than in the deterministic conditions, and so the cooperation in period 10 in the probabilistic condition starts at a higher level than in the deterministic condition.

Table 1: Probability of cooperation or defection at time t+1 (c_{t+1} = cooperation, d_{t+1} = defection) given that the participant cooperated or defected at time t (c_t = participant cooperated, d_t = participant defected) and won the lottery (+), compared to the case in which he or she lost the lottery (-) for a given choice of the partner (oc = partner cooperates, od = partner defects).

Condition	Probabilistic	Deterministic with sunspots
$P\{c_{t+1}\|(c_t \& oc_t \& +)\}$	**0.83**	**0.87**
$P\{c_{t+1}\|(c_t \& oc_t \& -)\}$	**0.68**	**0.85**
$P\{c_{t+1}\|(c_t \& od_t \& +)\}$	**0.57**	**0.125**
$P\{c_{t+1}\|(c_t \& od_t \& -)\}$	0.34	0.21
$P\{d_{t+1}\|(d_t \& oc_t \& +)\}$	0.71	0.79
$P\{d_{t+1}\|(d_t \& oc_t \& -)\}$	0.68	0.74
$P\{d_{t+1}\|(d_t \& od_t \& +)\}$	0.91	0.94
$P\{d_{t+1}\|(d_t \& od_t \& -)\}$	0.84	0.93

In the probabilistic condition, conditional on both players having cooperated in period t, a negative outcome of a player's lottery reduces his chance of cooperation to .68 (from .83, $t(46)=3.81$ $p<0.001$). And, even if the other player failed to cooperate at period t, a player who cooperated himself and (nevertheless) won his lottery has a .57 chance of cooperating, compared to only .34 had he lost his lottery[20]. So the effect of the noise in the payoffs--the outcome of the lottery-- on a player's choice of action for the next period is smaller than the effect of whether the other player cooperated or defected at the last period. But the

[20] We are aggregating here across supergames and, more importantly, periods within a supergame. As Figure 3 makes clear, probabilities of cooperation are also dependent on the period in the supergame.

outcome of the lottery nevertheless affects players' decisions, and as we have seen, this effect is sufficient to substantially reduce the joint learning to cooperate that goes on in the deterministic condition.

A different pattern of results emerged for the deterministic with sunspots condition. In this condition participants did not condition their response on the outcome of the random events. Conditional on both players having cooperated in period t, a negative outcome of a player's random event did not reduce the probability of cooperation. As expected, participants do condition their behavior on the other player's behavior. Conditional on having cooperated in the previous trial and winning the lottery, if the other player cooperated in the previous trial compared to if he defected, the probability of cooperation increased from 0.125 to 0.87 ($t(7)=6.67$, $p<0.001$).

This analysis indicates that while the information sets in the probabilistic and the deterministic with sunspots conditions are the same, participants in the sunspot treatment did not condition their behavior on the results of the random events.

To summarize the observations of the repeated games:

- In both the deterministic and probabilistic treatments, and starting from the first supergame, there is more cooperation in period 1 of the 10 period supergame than in period 10
- In the deterministic condition and in the deterministic with sunspots condition, players learn with experience with the 10 period supergame to cooperate more in the early periods and to cooperate less in the late periods, reproducing the results of Selten and Stoecker (1986), Andreoni and Miller (1993), and Hauk and Nagel (2001).
- In the otherwise identical probabilistic condition there is much less learning from one repeated game to the next; the "payoff noise" in that condition interferes with learning to cooperate; players condition their actions not only on the action of the other player, but also on the outcome of their lottery.

The apparent lack of learning (or very slow learning) to cooperate in the probabilistic condition of the repeated game, particularly when taken together with the opposite result of

more cooperation in the non-repeated games, adds strong support to the hypothesis that partial reinforcement slows learning.[21]

Conclusions:

As economists start to consider how to employ robust results from psychology in models of economic behavior, we may need to occasionally reexamine some of the most basic elements of economic models. In this paper we consider some of the limitations of representing payoffs as expected values or utilities, or related formulations. This is a subject that has been studied extensively in recent years from the point of view of static choice among lotteries: see e.g. Daniel Kahneman and Amos Tversky, 1979, or Rabin, 2000. Here we study it from the point of view of dynamic decisions in strategic environments.

That is, economists have been accustomed to thinking of variance as changing the *desirability* of lotteries, through risk aversion, but here we think of variance as changing the speed at which players learn about the strategic environment. The difference between the two approaches is clear when we compare our results for repeated versus non-repeated play. If increased variance (partial reinforcement) were making cooperation less desirable, we would see less cooperation when payoffs are probabilistic, for both repeated and non-repeated play. But instead, the prediction that variance in the frequency of reinforcement slows learning correctly predicts that, when payoffs are probabilistic, we will see *more* cooperation in the non repeated game (in which players in the deterministic condition learn not to cooperate), and *less* cooperation in the repeated game (in which players in the deterministic condition learn to cooperate).

Like utility maximization or alternative models of static preferences, learning models are simple approximations of complex behavior: no simple model will capture all the

[21] One additional alternative hypothesis that we considered is **Preferences for Fairness (inequality aversion):** Suppose that the loss of cooperation in the repeated probabilistic game (although not in the non-repeated game) is due not to learning, but because, unlike in the deterministic payoff condition, mutual cooperation does not always yield equal payoffs, since a player may lose the lottery while the other player wins. To test for this possibility we analyzed the probability of cooperation at time t+1, given that the player cooperated at time t, lost the lottery and the player's partner cooperated at time t and won the lottery, compared to the case in which both of them won the lottery. The fairness hypothesis suggests a much higher probability of cooperation after both players win. The probability of cooperation after only the other player won was .79, while the probability of cooperation after both won was .81. So payoff difference between the players, given the same actions, had no effect on the probability of cooperation.

behavior we observe, but simple models can serve as useful approximations for important aspects of learning, particularly those missed by static models of behavior.

Learning models are, loosely speaking, divided into two kinds, those that focus on learning facts (and updating beliefs about those facts), and those that focus on learning actions (or strategies and procedures). We have focused here on reinforcement learning models, which have quite old roots in the psychology literature, and which model the learning of actions.

The results of the repeated deterministic conditions suggest that participants learn contingent strategies, involving reciprocation, and not simple actions.. In the more modern psychology literature, reinforcement learning is sometimes modeled as the algorithm through which people learn to employ more complex cognitive strategies; see e.g. John Anderson (1990) and Erev and Barron (2005)[22]. These are very boundedly rational models of learning, and our results shed some light on bounded rationality. Some models of boundedly rational learning, such as fictitious play, model agents as making more limited calculations of the same sort as perfectly rational players. Since players have exactly the same information in both our probabilistic and deterministic (with sunspots) conditions (and thus can perform exactly the same calculations in both conditions), our results strongly support the contention of reinforcement learning models that the rewards a player *experiences* are critical to the observed differences between the probabilistic and deterministic conditions.

Since many strategic interactions do not have deterministic payoffs, our results suggest it may not always be innocuous to model payoffs in strategic environments as expected payoffs, even when there will be many interactions. Because early experience may influence subsequent behavior, the effect of the noise need not be averaged away by repetition.

Note again that the learning that goes on in strategic, economic environments is not a simple extension of the study of individual learning. In games, what is learned by a player early in the game depends on how others are behaving, and hence on what they are learning, and so the game provides a feedback loop between what players learn and what

---

[22] In the economics literature, rule learning is an example of modeling more complex strategies (e.g., Dale Stahl, 1999, 2000)

there *is* to learn. Therefore it is not surprising that the speed of learning can be so important, and that factors that have a small effect on the speed of individual learning can have a large effect on the whole path of play in strategic games.

One reason that learning has often played a secondary role in economic theory is that economists often consider markets and games after they have been running for a long time, so that early learning can be thought of as having given way to stable, equilibrium play. But as we've seen here, the nature of that long term stable behavior may depend on the initial learning environment. And, when we study markets that may have all new entrants, such as those that economists are increasingly being asked to play a role in designing, or markets that may have a steady stream of new entrants, it seems likely that the study of learning will become increasingly important.[23] Newly designed markets won't have any experienced players, and so their performance may depend critically on what they make it easy and fast and safe for participants to learn.

Bibliography

Abdulkadiroğlu, Atila, Parag A. Pathak, and Alvin E. Roth, "The New York City High School Match," American Economic Review, Papers and Proceedings, 95,2, May, 2005, 364-367.

Abdulkadiroğlu, Atila, Parag A. Pathak, Alvin E. Roth, and Tayfun Sönmez, "The Boston Public School Match," American Economic Review, Papers and Proceedings, 95,2, May, 2005, 368-371.

Anderson, John R. (1990). *The Adaptive Character of Thought.* Hillsdale, NJ: Erlbaum.

Ariely, Dan, Axel Ockenfels, and Alvin E. Roth, "An Experimental Analysis of Ending Rules in Internet Auctions," *Rand Journal of Economics*, forthcoming.

Andreoni, J. and Miller, J. H. (1993). Rational cooperation in the finitely repeated Prisoner's Dilemma: experimental evidence. *The Economic Journal*, **103**, 570-585.

Axelrod, Robert. and Douglas Dion, (1988). The further evolution of cooperation. *Science*, **9**, 1385-1389.

[23] For a discussion of newly designed markets, see e.g. Roth (2002), Robert Wilson(2002), Paul Milgrom (2004), Atila Abdulkadiroglu et al. (2005a,b), Muriel Niederle and Roth (2005), Niederle, Deborah Proctor, and Roth (2006), Roth, Tayfun Sonmez and Utku Unver (2004, 2005a,b).. For markets with a stream of new entrants, consider e.g. auction markets such as eBay (and compare the behavior of experienced and inexperienced behavior reported in Roth and Axel Ockenfels, 2002 or Dan Ariely, Ockenfels, and Roth, forthcoming)..

Bendor, Jonathan. (1987). In good times and bad: Reciprocity in an uncertain world. *American Journal of Political Science*, **31**, 531-538.

Bendor, Jonathan. (1993). Uncertainty and the evolution of cooperation. *Journal of Conflict Resolution*, **37**, 709-734.

Camerer, Colin And Teck H. Ho, (1999). Experience-Weighted Attraction Learning in Normal Form Games. *Econometrica*, **67**, 827-74.

Cooper, Russell, Douglas V. DeJong., Robert Forsythe, andThomas W Ross, (1996). Cooperation without reputation: Experimental evidence from the Prisoner's Dilemma games. *Games and Economic Behavior* **12**, 187-218.

Duffy, John and Nick Feltovich, (1999). Does observation of others affect learning in strategic environments? an experimental study. *International Journal of Game Theory,* **28**, 131-152.

Erev, Ido, and Greg Barron. (2005) "On Adaptation, Maximization, and Reinforcement Learning Among Cognitive Strategies." *Psychological Review* October.).

Erev, Ido, Yoella Bereby-Meyer, and Alvin E. Roth, (1999). The effect of adding a constant to all payoffs: Experimental investigation, and implications for reinforcement learning models. *Journal of Economic Behavior and Organization*, **39**, 111-128.

Erev, Ido and Alvin E. Roth (1998). Predicting how people play games: Reinforcement learning in experimental games with unique, mixed strategy equilibria," *American Economic Review*, **88**, 848-881.

Feltovich, Nick (2000), "Reinforcement-Based vs. Beliefs-Based Learning Models in Experimental Asymmetric-Information Games," *Econometrica*, **68**, 605-642.

Fischbacher, U. (1999), z-Tree - Zurich Toolbox for Readymade Economic Experiments - Experimenter's Manual, mimeo

Fudenberg, Drew, and David Levine,.(1998) *The Theory of Learning in Games*, MIT Press.

Green, Edward J. and Robert H. Porter, (1984). Noncooperative collusion under imperfect price Information, *Econometrica*, **52**, 87-100.

Hauk, Esther and Rosemarie Nagel, (2001). Choice of partners in multiple two-person prisoner's dilemma games. An experimental study. *Journal of Conflict Resolution*, **45**, 770-793.

Jenkins, William. and Stanley, Julian (1950). Partial reinforcement. A review and critique, *Psychological Bulletin*, **47,** 197-234.

Kahneman, Daniel and Amos Tversky, (1979). Prospect theory: An analysis of decision under risk, *Econometrica*, **47**, 263-291.

Kreps, David, Paul Milgrom, , John Roberts, and Robert Wilson, (1982). Rational cooperation in the finitely repeated Prisoner's Dilemma. *Journal of Economic Theory,* **27**, 245-252.

Lee, Wayne. (1971). *Decision Theory and Human Behavior*. New York: Wiley.

Milgrom, Paul (2004), *Putting Auction Theory to Work*, Cambridge University Press.

Miller. John H. (1996). The coevolution of automata in the repeated prisoner's dilemma. *Journal of Economic Behavior and Organization*, **29**, 87-112.

Molander, Per. (1985). The optimal level of generosity in a selfish, uncertain environment. *Journal of Conflict Resolution*, **29**, 611-0618.

Niederle, Muriel and Alvin E. Roth, "The Gastroenterology Fellowship Market: Should there be a Match?," *American Economic Review,* Papers and Proceedings, 95,2, May, 2005, 372-375.

Niederle, Muriel, Deborah D. Proctor, and Alvin E. Roth, "What will be needed for the new GI fellowship match to succeed?" *Gastroenterology*, 130, January, 2006, 218-224.

Rabin, Matthew (2000), "Risk Aversion and Expected Utility Theory: A Calibration Theorem," Econometrica, 68(5), September, 1281-1292.

Rapoport, Anatol. and Albert M. Chammah, (1965). *Prisoner's dilemma*. Ann Arbor: Univ. of Michigan Press.

Robbins, Donald. (1971). Partial reinforcement: A selective review of the alleyway literature since 1960. *Psychological Bulletin*, **6**, 415-431.

Roth, Alvin E. "The Economist as Engineer: Game Theory, Experimental Economics and Computation as Tools of Design Economics," *Econometrica*, 70, 4, July 2002, 1341-1378.

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

Roth, Alvin E. and Michael W.K. Malouf, "Game-Theoretic Models and the Role of Information in Bargaining", *Psychological Review*, Vol. 86, 1979, 574-594.

Roth, Alvin E. and J. Keith Murnighan (1978). Equilibrium behavior and repeated play of the prisoner's dilemma. *Journal of Mathematical Psychology*, **17**, 189-198.

Roth, Alvin E. and Axel Ockenfels "Last-Minute Bidding and the Rules for Ending Second-Price Auctions: Evidence from eBay and Amazon Auctions on the Internet," *American Economic Review*, 92 (4), September 2002, 1093-1103.

Sainty, Barbara. (1999). Achieving greater cooperation in a noisy prisoner's dilemma: an experimental investigation. *Journal of Economic Behavior and Organization,* **39**, 421-435.

Selten, Reinhard and Rolf Stoecker, (1986). End Behavior in sequence of finite prisoner's dilemma supergames.: a learning theory approach. *Journal of Economic Behavior and Organization*, **7**, 47-70.

Stahl. Dale (1999). Evidence based rules and learning in symmetric normal-form games. International Journal of Game Theory, 28, 111-130.

Stahl. Dale (2000). Rule learning in symmetric normal-form games: Theory and evidence. Games and Economic Behavior, 32, 105-138.

Stanley, Julian. (1950). The differential effects of partial and continuous reward upon the acquisition and elimination of a running response in a two-choice situation. Unpublished Ed.D thesis, Harvard univ.

Von Neumann, John and Oskar Morgenstern, O. (1944). *Theory of Games and Economic Behavior*, Princeton University Press, Princeton.

Weinstock, Solomon. (1958). Acquisition and extinction of a partially reinforced running response at a 24-hour intertrial interval. *Journal of Experimental Psychology*, **56**, 151-158.

Wilson, Robert B (2002), "Architecture of Power Markets," Econometrica, July 2002, 70, 4, 1299-1340