ECONOMICS BULLETIN

Learning to deal with risk: what does reinforcement learning tell us about risk attitudes?

Albert Burgos
Universidad de Murcia

Abstract

People are generally reluctant to accept risk. In particular, people overvalue sure gains, relative to outcomes which are merely probable. At the same time, people are also more willing to accept bets when payoffs involve losses rather than gains. I consider how far adaptive learning can go in explaining these phenomena. I report simulations in which adaptive learners of the kind studied in Roth Erev (1995, 1998) and Borgers Sarin (1997, 2000) deal with a problem of repeated choice under risk where alternatives differ by a mean preserving spread. The simulations show that adaptive learning induces (on average) risk averse choices. This learning bias is stronger for gains than for losses. Also, risk averse choices are much more likely when one of the alternatives is a certain prospect. The implications of a learning interpretation of risk taking are explored.

Citation: Burgos, Albert, (2002) "Learning to deal with risk: what does reinforcement learning tell us about risk attitudes?." *Economics Bulletin,* Vol. 4, No. 10 pp. 1−13
Submitted: May 7, 2002. **Accepted:** May 28, 2002.
URL: http://www.economicsbulletin.com/2002/volume4/EB−02D80010A.pdf

1 Introduction

Although the idea of modeling people as stimulus-response mechanisms shaped by learning forces has a long history in psychology, it has only become popular in economics in the last ten years or so. Reinforcement learning models have been used recently either to explain experimental findings in strategic encounters (as do several references reviewed in Camerer and Ho (1999)), or to give learning theoretical foundations of population dynamics used in evolutionary game theory (Börgers and Sarin (1997, 2000)). These models consider the adaptive behavior of goal oriented agents which need not be subjective expected utility maximizers, or indeed maximizers of any sort. In fact, choices are a (stochastic) function of earlier payoffs, while these payoffs are again a function of earlier choices. As a result, one gets a sequence analysis of actions and payoffs in which risk attitudes does not appear explicitly, but only 'between the lines'.

This paper tries to make explicit the reactions to risk **induced** by reinforcement learning processes. Then, one is in a position to analyze how far such processes provide sufficient structure to tie down the set of possible reactions to risk.

The paper reports the result of computer simulations[1] in which reinforcement learners deal with pairwise choices between risky prospects with the same expected value. Any consistent pattern showing more propensity to choose one alternative rather than the other is interpreted as reflecting risk preference. The central issue here is the possible connection between the attitudes toward risk resulting from adaptive learning and some patterns of actual choices in experiments that have been extensively documented in experimental economics.

In a first block of simulations, learners choose between a certain prospect and an uncertain one with equal expected value. I test for risk aversion over positive and negative payoffs. Experimental and field data studies have shown greater risk aversion for gains than for losses. Reinforcement learners show behavior consistent with this **reflection effect**. The second block of simulations is designed to induce violations of the independence axiom of expected utility theory. These involve a test of the **certainty effect** or Allais ratio paradox (Allais (1953), Kahneman and Tversky (1979)). Simulated learners, too, violate the independence axiom in the Allais-type direction.

2 Background

Reinforcement models have been developed largely in response to observations by psychologist about human behavior and animal behavior. All of these models are built upon quan-

[1] Computations in the paper were performed by the author in GAUSS. Detailed programs can be obtained on request.

tifications of Thorndike's (1898) **law of effect**: choices that lead to good (bad) outcomes are more (less) likely to be repeated. Implicit in the law of effect is that choice behavior is probabilistic: A reinforcement learner is characterized by an initial probability vector over alternatives and, subsequently, after each learning trial by a revised probability vector. In this context, 'learning' means updating the probabilities of taking each alternative on the basis of the outcomes experienced. Namely, if $p_i(t)$ is a specific learner's probability of choosing alternative i on trial t, a learning model is defined as a rule specifying how $p_i(t)$ is transformed to $p_i(t+1)$.

Applications of reinforcement learning to economic behavior can be organized around two models: The first is developed by Börgers and Sarin (1997, 2000). It is a version of the classic Bush and Mosteller (1955) linear adjustment model in which reinforcements can be either positive or negative, depending on whether the realized payoff is greater or less than the agent's 'aspiration level', which in turn follows a linear adjustment process too. The second model is proposed by Roth and Erev (1995, 1996, 1998). It is a more highly parametrized version of a quantification of the law of effect based on average returns proposed by Herrnstein (1961, 1970), and adds to the Börgers and Sarin's model an interesting feature: learning curves get flatter over time, a fact known as the **power law of practice** (Blackburn (1936)). I now present theses models in detail for choices involving only two alternatives.

2.1 The Börgers & Sarin (B&S) model

In this model, $p_i(t)$ changes as a result of the alternative chosen and the outcome observed at $t-1$. If the outcome exceeds the aspiration level, then the probability associated with the action increases. If the outcome falls below the aspiration level, then the probability associated with the alternative chosen decreases. The size of the change in $p_i(t)$ is proportional to the size of the difference between the outcome and the aspiration level. Formally, if $\rho(t)$ denotes the aspiration level (or reference point) at time t, probabilities of choosing each alternative evolve in the following way: if payoff at date t is $x \geq \rho(t)$, then the reward associated with x is $R(x,t) = x - \rho(t) > 0$, and for $i = 1, 2$

$$p_i(t+1) = \begin{cases} [1 - \lambda R(x,t)]p_i(t) + \lambda R(x,t) & \text{if } i \text{ was chosen at } t, \\ [1 - \lambda R(x,t)]p_i(t) & \text{otherwise.} \end{cases}$$

If, however, payoff at t is $x < \rho(t)$, the reward $R(x,t)$ is negative, and

$$p_i(t+1) = \begin{cases} [1 + \lambda R(x,t)]p_i(t) & \text{if } i \text{ was chosen at } t, \\ [1 + \lambda R(x,t)]p_i(t) - \lambda R(x,t) & \text{otherwise.} \end{cases}$$

The parameter λ controls the effect of rewards in $p_i(t+1)$, and can assume any value guaranteeing that the absolute value of $\lambda R(x,t)$ always lies (strictly) between zero and one. The greater the value of λ, the faster the adaptation.

In addition to the probability vector, the aspiration level is also linearly adjusted in the direction of the outcome experienced. Thus,

$$\rho(t+1) = (1-\beta)\,\rho(t) + \beta x,$$

where $0 \leq \beta < 1$. Initial values for $\rho(1)$ can be set to any value. Thus, the model is controlled by the initial probabilities and the three parameters: the aspiration level's parameters ($\rho(1)$ and β) and the learning parameter λ.

2.2 The Roth and Erev (R&E) model

In R&E's model, $p_i(t)$ changes on the basis of the history of returns that have been obtained from the two alternatives. This memory of the average return from each alternative is modified by the effect of reference points, forgetting (or recency), and experimentation, yielding what R&E call 'propensities'. Again, let $R(x,t) = x - \rho(t)$ be the reward associated with x. Thus, if at date t the decision maker receives a payoff of x, then the propensity to play each alternative is updated by setting

$$q_i(t+1) = \begin{cases} \max\{\nu, (1-\phi)\,q_i(t) + (1-\varepsilon)\,R(x,t)\} & \text{if } i \text{ was chosen at } t, \\ \max\{\nu, (1-\phi)\,q_j(t) + \varepsilon R(x,t)\} & \text{otherwise.} \end{cases}$$

Parameter ν represents a small "cutoff" value guaranteeing that propensities remain positive. Parameters ϕ and ε have behavioral meaning: ϕ (recency) slowly reduces the importance of past experience, and ε (experimentation) prevents the probability of choosing any alternative from going to zero.

The probability of choosing alternative i in period t is proportional to past average propensities, i.e. for both i,

$$p_i(t) = q_i(t)/\left[q_1(t) + q_2(t)\right].$$

Thus, the ratio of the probabilities of choosing each alternative equals the ratio of their propensities. The task of determining initial propensities is reduced by setting $S(1) = \left[q_1(1) + q_2(1)\right]/X$ where X is the absolute value of the mean return associated with the problem given uniformly distributed choice probabilities. According to this, initial propensities follow from the initial choice probabilities and the strength parameter $S(1)$, which controls for the weight of initial tendencies. Notice that when $S(1)$ is high (i.e. initial propensities are strong) learning will be lower than when $S(1)$ is low.

Finally, two additional parameters, ω^- and ω^+, control the adjustment of the reference point following negative and positive rewards (compared to the reference point), respectively:

$$\rho(t+1) = \begin{cases} (1-\omega^-)\,\rho(t) + (\omega^-)\,x & \text{if } x < \rho(t), \\ (1-\omega^+)\,\rho(t) + (\omega^+)\,x & \text{if } x \geq \rho(t). \end{cases}$$

R&E calibrate their model against experimental data on two-player matrix games with mixed-strategy equilibria played repeatedly but against different opponents each time (in order that the one shot character of each game be preserved). The calibration of the general seven-parameter model appears in Erev and Roth (1996). The exact values taken from Table 3 in their paper are $S(1) = 3$, $\phi = 0.001$, $\varepsilon = 0.2$, $\rho(1) = 0$, $\omega^{+} = 0.01$, $\omega^{-} = 0.02$, and $\nu = 0.0001$.

3 Testing for the "reflection effect"

A common result in experiments conducted to elicit certainty equivalents for lotteries is that revealed risk preferences show greater risk aversion when payoffs are positive than when payoffs are negative, a phenomenon labelled as the "reflection effect". The central issue here is whether adaptive learning generates risk biases in the direction of showing more risk aversion for positive payoffs than for negative payoffs, as experiments show. This gives rise to the following hypothesis.

Hypothesis 1: *Adaptive learners exhibit greater risk aversion for gains than for losses.*

To examine this question, I simulate the choice situation between two pairs of alternatives. Each pair consists of one lottery, denoted S for "safer," which pays qx with certainty, and a mean preserving spread of S, denoted R, for "riskier," which pays x with probability q and nothing otherwise. Measure for evaluating Hypothesis 1 consists of the comparison of the ex-post probability of choosing S over R when x is positive (gains) and when x is negative (losses) starting from a situation in which both choices are equiprobable. Figure 1 shows the average results involving 1000 simulated learners over the first 500 trials for the calibrated R&E model. The simulations assume that $q = 0.5$, and the figures compare (i) a case of negative outcomes in which $x = -1$ with (ii) a case of positive outcomes in which $x = 1$.

The simulations clearly support Hypothesis 1 for the case of the calibrated R&E model. Although the model ultimately yields equal probabilities of choosing the safer alternative for gains and for losses, it requires a substantial amount of experience—about 200 trials— to reach such a result. Risk aversion for gains is weak, and learning ultimately results in risk neutrality, but in the short run adaptive learning produces risk seeking when the returns lie in the negative domain.

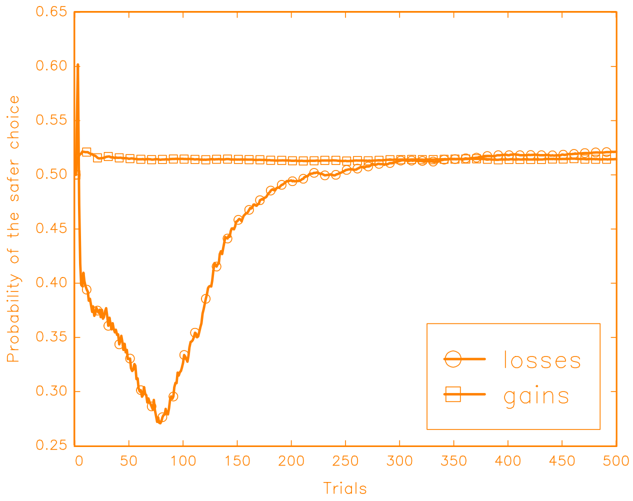

FIGURE 1. COMPARING RISK PREFERENCES FOR GAINS AND LOSSES, R&E MODEL

Figure 2 shows the results of a simulation of the B&S model when $\lambda = 0.3$, $\rho(1) = 0$, and $\beta = 0.05$. These values have been chosen to approach a similar distribution for gains and for losses in about 200 choices, the same amount of experience needed in the calibrated R&E model.[2] Again, choosing the less risky alternative is more likely when possible outcomes lie in the positive domain than when they are negative. However, unlike the case of the R&E model, learning that conforms to the B&S model produces risk averse behavior both for losses and for gains, and this behavior persists by the end of the 500 trials.

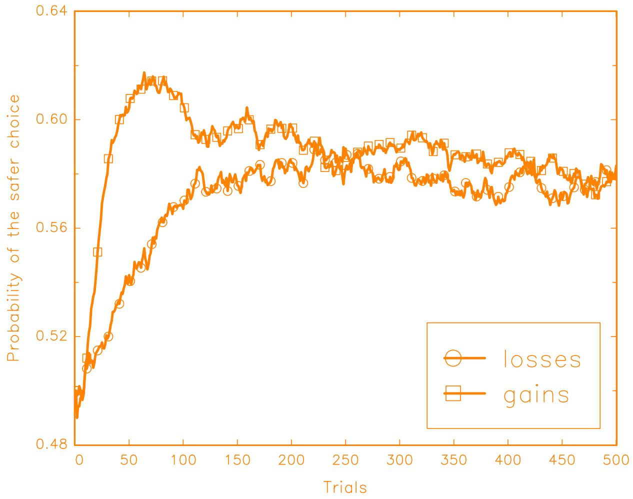

FIGURE 2. COMPARING RISK PREFERENCES FOR GAINS AND LOSSES, B&S MODEL

[2] Since the B&S model does not employ intermediate propensities and therefore the speed of learning quickly exceeds that of the R&E model, this implies that initial learning must be slower than in the calibrated R&E model.

It is reasonable to expect that if the difference in riskiness associated with the two alternatives were reduced (by increasing q), so would be the reflection effect. This is indeed the case for R&E learners. Figure 3 shows the difference between gains and losses in the probability of choosing the safer alternative after 75 trials[3] as a decreasing function of q. For the B&S model, however, this relation is non-monotonic (see figure 4).

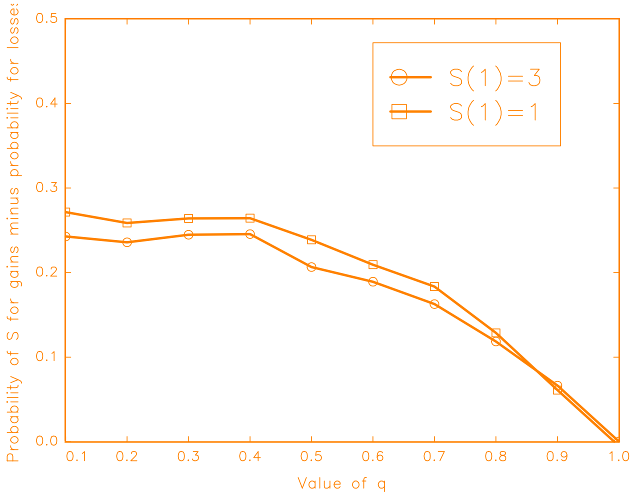

Figure 3. Difference between risk preferences for gains and losses in the R&E model after 75 trials

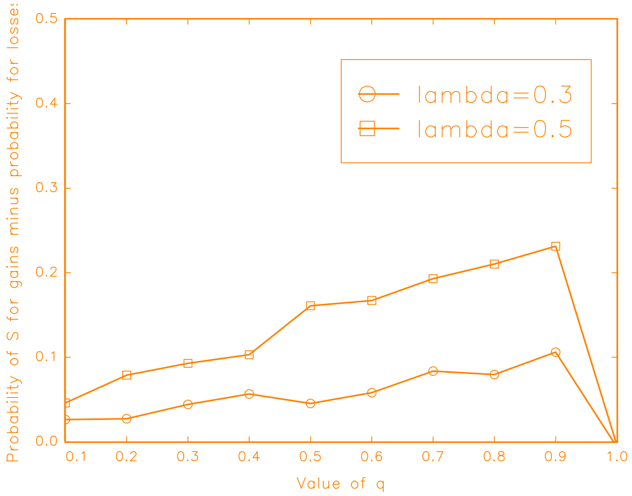

Figure 4. Difference between risk preferences for gains and losses for the B&S model after 100 trials

[3] This was the amount of experience at which the difference between $p(t)$ for gains and for losses in the calibrated model attained its maximum when $q = 0.5$. The B&S model with $\lambda = 0.3$, $\rho(1) = 0$, and $\beta = 0.05$ required more experience (about 100 trials) to reach its maximum.

Figures 3 and 4 above contain also information regarding the effect on risk taking of the learning rate (i.e. parameters $S(1)$ and λ) keeping the remaining parameters as in simulations 1 and 2. Low values of $S(1)$ and high values of λ are associated with fast learning. As might be expected, fast learning accentuates the tendency towards risk aversion in the positive domain, slow learning reduces it—provided the initial probabilities of choosing the various alternatives are equal.

4 Testing for the "certainty effect"

Among the experimental challenges to expected utility theory, perhaps the most systematically observed violation of the theory refers to the systematic 'over-valuation' of outcomes that are considered certain, relative to outcomes which are merely probable. This anomaly (from the expected utility theory point of view), appears in a class of systematic violations of expected utility theory known as the certainty effect or Allais ratio paradox (see for instance, Kahneman and Tversky (1979 p. 266)). In this section I test for this effect in adaptive learning.

A simple test of the certainty effect involves choices between a prospect S, with r chance of y, or $1 - r$ chance of 0, and a prospect R, with qr chance of x, or $1 - qr$ chance of 0. Probability q is kept fixed, and experiments measure the effect of r on choices. The null hypothesis is that choices are independent of r, as the independence axiom of expected utility theory predicts. In the classic experiments x and y are positive, and y is chosen to be equal to (or slightly smaller than) qx. Thus, any risk averse expected utility maximizer would chose S for all r (or else R for all r if the subject were a risk seeker). The certainty effect occurs when S is chosen in problems with $r = 1$ and R is chosen in problems with $r < 1$. Starting as early as in Allais (1953), researchers have found considerable evidence that this effect is a systematic property of risk attitudes, for a wide range of parameter values. This provides the basis for the following hypothesis.

Hypothesis 2: *When faced with choice problems which mimic experimental tests of the certainty effect, adaptive learners exhibit Allais-type behavior.*

I employed the same simulation procedure as in the previous section. Figure 5 illustrates the certainty effect using the calibrated R&E model. It shows the effect of r on probability of choice of the safer alternative, S, over 500 trials, where $x = 1$, $q = 0.5$, and $y = qx$. As the hypothesis states, certainty accentuates the tendency toward risk aversion, though the impact on choices probability is low (always below 1%).

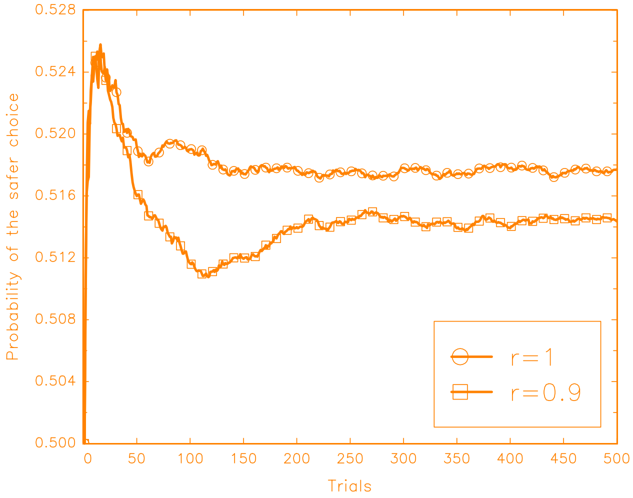

FIGURE 5. COMPARING RISK PREFERENCES FOR CERTAIN AND UNCERTAIN OUTCOMES, R&E MODEL

Figure 6 shows a much stronger certainty effect associated with the B&S model where again $\lambda = 0.3, \rho(1) = 0$ and $\beta = 0.05$. With these parameters, a decrease in the value of r from 1 to 0.9 leads to a difference in the average probability of choosing S which remains above 5% for more than 200 trials. The relation between the ratio r and the difference between the probabilities of choosing S and R for two different values of the learning parameter λ is shown in Figure 7.

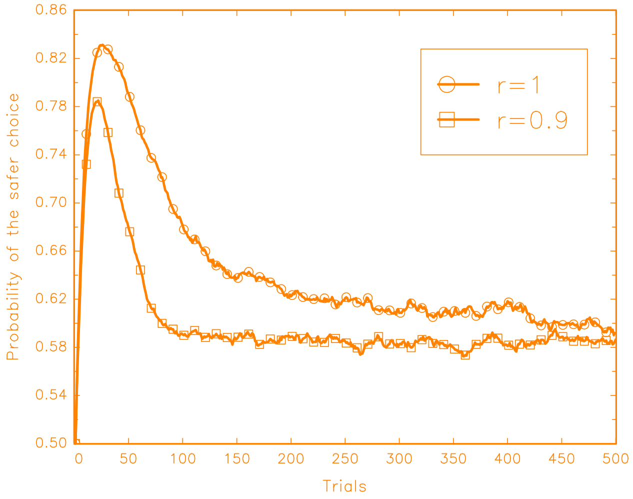

FIGURE 6. COMPARING RISK PREFERENCES FOR CERTAIN AND UNCERTAIN OUTCOMES, B&S MODEL

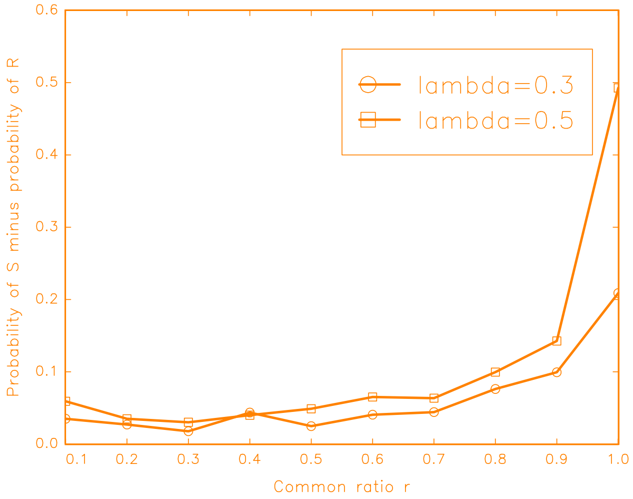

FIGURE 7. EFFECT OF r IN RISK PREFERENCES AFTER 100 TRIALS, B&S MODEL

5 Relevance to modelers?

As long as risk taking is interpreted within a framework of preference maximization, our understanding of risk attitudes is expressed in terms of features of utility functions. According to this interpretation, a considerable amount of time and effort has gone into the quest for models of preferences that try to accommodate at least the most widely observed behavioral regularities. Nowadays, however, the idea that there is a simple and coherent model of preferences that accommodates all experimental anomalies (from an expected utility point of view) is appearing more and more unattainable. See, for instance, Loomes (1998) and Selten, Sadrieh, and Abbink (1999).

A learning perspective provides a different view of the sources of risk aversion and risk seeking. Risk preference can be interpreted as a learned trait, rather than a consequence of prior utilities. Be that as it may, there is a spectrum of experience and expertise, with novices at one extreme and solid expected utility maximizers at the other, and little, if anything, was known about how to place a given choice problem along this spectrum. In this paper we have seen that this question depends, among other things, on whether the choices are about gains or losses, and payoffs are certain or uncertain. One lesson we draw is that Hypothesis 1 and 2 in the paper are likely to be found even among experienced subjects. Indeed, fast learning (i.e. in which behavior is modified rapidly in response to feedback) is especially likely to produce these biases. This fact may give pause to the traditional argument favoring the use of expected utility theory (and its special case, risk neutrality) on grounds of an appeal—often implicit—to learning forces.

A second lesson concerns which aspects of choice theory under risk deserve future research. An obvious follow-up of this paper is to conduct experimental study on the effect of learning on risk taking. Also, it should be checked if the same learning models used here can predict other important phenomena, such as branch independence, cumulative independence or violations of stochastic dominance.

References

Allais, M. (1953): "Le Comportement de l'Homme Rationnel devant le Risque, Critique des Postulats et Axiomes de l'École Americaine," *Econometrica,* **21**, 503–546.

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

Börgers, T., and R. Sarin (1997): "Dynamic Consistency and Non-Expected Utility Models of Choice Under Uncertainty," *Journal of Economic Theory,* **77**, 1-14.

——————(2000): "Naive reinforcement learning with Endogenous Aspirations," *International Economic Review,* **41**, 921-950.

Bush, R., and F. Mosteller (1955): *Stochastic Models for Learning,* New York, NY: John Wiley & Sons.

Camerer, C., and T. Ho (1999): "Experence-weighted Attraction learning in Games," *Econometrica,* **67**, 827–874.

Erev, I., and A.E. Roth (1996): "On the Need for Low Rationality, Cognitive Game Theory: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," mimeo, University of Pittsburgh. (Earlier version of Erev and Roth (1998).)

——————(1998): "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria," *American Economic Review,* **2**, 225-243.

Herrnstein, R.J. (1961): "Relative and Absolute Strength of Response as a Function of Frequency of Reinforcement," *Journal of the Experimental Analysis of Behavior,* **4**, 267-272.

——————(1970): "On the Law of Effect," *Journal of the Experimental Analysis of Behavior,* **13**, 244-266.

Kahneman, D., and A. Tversky (1979): "Prospect Theory: An Analysis of Decision under Risk," *Econometrica,* **47**, 263–291.

Loomes, G. (1998): "Probabilities vs. Money: A Test of Some Fundamental Assumptions about Rational decision Making," *Economic Journal,* **108**, 477–489.

Roth, A.E., and I. Erev (1995): "Learning in Extensive-Form games: Experimental Data and Simple Dynamic Models in the Intermediate Term." *Games and economic Behavior,* Special Issue: Nobel Symposium, **8**, 164-212.

Selten, R., A. Sadrieh, and K. Abbink (1999): "Money does not induce risk neutral behavior, but binary lotteries do even worse," *Theory and Decision,* **46**, 211–249.

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