Learning to Respond: The Use of Heuristics in Dynamic Games[1]

Mikhael Shor

Owen Graduate School of Management
401 21st Avenue South
Nashville, TN 37203
Mike.Shor@owen.vanderbilt.edu

September 2001

Abstract

While many learning models have been proposed in the game theoretic literature to track individuals' behavior, surprisingly little research has focused on how well these models describe human adaptation in changing dynamic environments. Analysis of human behavior demonstrates that people are often remarkably responsive to changes in their environment, on time scales ranging from millennia (evolution) to milliseconds (reflex). The goal of this paper is to evaluate several prominent learning models in light of a laboratory experiment on responsiveness in a low-information dynamic game subject to changes in its underlying structure. While history-dependent reinforcement learning models track convergence of play well in repeated games, it is shown that they are ill suited to these environments, in which sastisficing models accurately predict behavior. A further objective is to determine which heuristics, or "rules of thumb," when incorporated into learning models, are responsible for accurately capturing responsiveness. Reference points and a particular type of experimentation are found to be important in both describing and predicting play.

JEL Classification: D83, C91, C73

Keywords: learning, limited information, responsiveness

[1] I am indebted to the members of CREED, The Netherlands, for their generous hospitality while pursuing this research. This work was supported by the National Science Foundation under Grant No. 9730162.

1. Introduction

Human responsiveness to changing environments has been well-studied in many disciplines. Psychologists have long analyzed human and animal adaptation to environmental variations. Computer scientists have applied principles of responsiveness both to the design of "intelligent" software and to the design of software and networks responsive to human learning. Sociologists have analyzed adaptive group behavior, an inquiry applied to organizational theory and institutional decision making in business settings (Dooley, 1997; Carley and Lee, 1998; Levitt and March, 1988). The notion of the "learning organization" (Hayes, Wheelwright, and Clark, 1988) emphasizes the responsiveness of business units to changes in their market environment.

The historical economic approach to decision making, rooted in the rational actor paradigm and assuming unbounded cognitive ability is quite separated from the psychologists' perspective on human learning. A recent literature, concerned with building learning models rooted in classic psychological principles such as the law of effect (Roth and Erev, 1995), bounded rationality (Simon, 1957), and aspirations (Selten 1991, Karandikar, Mookherjee, Ray, and Vega-Redondo, 1998) has led toward a unification of psychological principles with the economic view of an agent.

While many authors have evaluated the ability of learning models to explain observed human behavior in repeated games,[2] surprisingly little research has focused on how well these models track individuals' adaptation in dynamic settings in which the underlying payoff matrix changes over time. Empirical research on adaptive behavior demonstrates that people are often remarkably responsive to changes in their environment (Payne, Bettman, and Johnson, 1993; Schunn and Reder, 1998). The goal of this paper is to analyze some common learning models in light of laboratory experiments on responsiveness in low-information dynamic games. A further objective is to determine which heuristics of these various models help capture responsiveness.

Heuristics are "rules of thumb" representing principles for deciding among competing alternatives. Since heuristics incorporate only general principles of behavior, they are tactics for approaching a problem, not fully represented strategies (see Pearl, 1984, for a review). Thus, while heuristic-based approaches are likely to select better actions among competing strategies, they do not do so necessarily in an optimal fashion. As any person knows from personal experience, people are subject to the same fault.

---

[2] See, for example Mookherjee and Sopher, 1994; Roth and Erev, 1995; Van Huyck, Battalio, and Rankin, 1996; and Erev and Roth, 1998.

The experiments considered here (from Friedman, Shor, Shenker, and Sopher, 2000) isolate responsiveness from other behaviors by instituting a very simple learning environment. Subjects participate in a real-time monopoly quantity-setting game. A change in the demand curve during the experiment is unobservable to subjects except through its payoff effects. In agreement with the psychological literature, subjects react quickly to the change in the payoff function. The learning models we consider, including reinforcement learning, evolving aspirations, satisficing, and responsive learning automata, however, differ in how well they capture this adaptability.

This paper proceeds as follows. In the following section, section 2, results of the experiment on learning in a low-information dynamic setting are discussed. Next, in section 3, the heuristics implied by these results motivate the selection of learning models to be analyzed. Section 4 contains a comparison of the performance of these learning models, both in closeness of fit to the observed data, and in prediction of subjects' behavior in a similar experiment. Section 5 contains concluding thoughts and implications for the design of learning models.

2. Experiments

The data come from an experiment by Friedman, *et. al.* (2000) on dynamic decision making in low-information environments. Design features of the experiment included very limited information, a dynamic, changing real-time setting, and a simple, noise-free environment. Subjects were given no information about the structure of the game, the underlying payoff function, the number of players, or the stability of the environment. Subjects were not informed of what a "reasonable" payoff was, nor did they know the bounds on the payoff function at any given time. Further, while they were aware that the payoff function may change during the experiment, subjects were not informed of the source or timing of these changes.

The experiments were computerized, run within web browsers, and in real time. Short periods, one second in length, and variations in the underlying payoffs provided a changing, dynamic setting. A subject's selected action would remain in effect until changed, which could be done at any time. Actions were chosen from a grid of 101 strategies, $\{0,1,\ldots,100\}$, by moving a slider provided on screen. Payoff information was presented every second, and a history of payoffs was also provided on the user interface. The length of the experiment was ten minutes, not including instructions. While a seemingly short time, the experiment permitted 600 periods, which

is substantially more than other individual decision-making experiments known to the author. Further, the short length avoids boredom, which may lead to excessive experimentation.[3]

The underlying game was a simple quantity-setting monopoly game with linear demand. A subject's strategy choice was mapped into payoffs according to the payoff function $\Pi_t = aq_t - bq_t^2$ where q_t is the action chosen at time t. Two different treatments were run. In each treatment, the game began with the same values of a and b and then the payoff function changed once, at seven minutes for the first treatment, and at five minutes for the second treatment (Table 1). The major difference between the two treatments is whether the change in the payoff function is noticeable to the subject, i.e. if it changes a subject's payoffs at equilibrium. In Treatment 1, a subject playing the optimal strategy of 40 before the change will instantly see her payoffs rise from 60 to 88 at seven minutes when the payoff function changes. On the other hand, in Treatment 2, a subject playing her equilibrium strategy will see no change in her payoffs when the payoff function changes. Only by sufficiently exploring the strategy space (playing a strategy above 45) can the change in the payoff function be recognized. Hence, the two treatments differentiate between how people recognize changes in their environment, solely through changes in current payoffs, or through experimentation.

A total of 56 subjects participated in the first treatment, and 22 subjects in Treatment 2. Subjects were very responsive in both treatments (Figure 1). In Treatment 1, the median player recognized the change in the payoff function, and quickly learned the new equilibrium (within 100 seconds). In Treatment 2, the path of play was similar, with a slightly longer delay after the payoff change, due to the fact that the change was not noticeable until a subject experimented.

The data suggest a number of characteristics of play. First, experimentation was quite common. Subjects spent a substantial proportion of time trying suboptimal strategies well after learning the equilibrium. Second, experimentation was not, in general, an occasional deviation from the optimal strategy. Instead, subjects would enter "experimentation phases" in which they would sample the entire strategy space. Friedman, *et. al.* termed a common pattern of experimentation "arrhythmic heartbeat patterns," as equilibrium play was occasionally disturbed by a period of experimentation in which subjects would sample the full strategy space below the equilibrium, and then above (or vice versa) resembling heartbeats when plotted against time.

[3] Instructions were provided on screen and took an average of eight minutes. While the experiment discussed here lasted ten minutes, subjects continued participating for a total of fifty minutes in a related experiment. See Friedman, *et. al.* (2000). This continued participation beyond the game reported here avoids endgame effects.

Hence, experimentation is autocorrelated, and not controlled by an independent random draw in each period, as most commonly modeled in the learning literature.

In informal post-experiment interviews, subjects indicated that learning consisted of discovering what constituted a good payoff and then attempting to maintain payoffs in that range, suggesting satisficing behavior (Karandikar, *et. al.*, 1998). Further, when the notion of a "good" payoff changed, subjects assumed that the environment was different and reinitiated the learning process. Interestingly, subjects suggested that they did not pay much attention to historical payoffs when it became clear that the payoff structure changed. Hence, selective use of history, aspirations or reference points, and the nature of experimentation all may be relevant traits for analyzing subjects' play, and are a primary motivation for selecting the models to consider in this analysis.

3. Learning Models

Given the low-information design of the experiment and the nature of this investigation, models rooted in rational optimization are not considered. The fact that subjects are not privy to the underlying structure of the game nor have any information about the payoff matrix implies that forward-looking learning models may not applicable, and learning should occur through some adaptive, or backward-looking mechanism. Further, in light of experimental support for such "myopic" learning (see note 2), the analysis of backward-looking learning models is interesting in its own right. Hence, we selected models of learning representing a variety of assumptions about the heuristics, or behavioral patterns, that subjects may exhibit. Table 2 surveys the models considered, and the heuristics they incorporate. These are discussed in more detail below.

3.1 Heuristics

A cornerstone of the psychological learning literature holds that if people are motivated by past events, then they should react positively to good outcomes and negatively to poor ones. Hence, the models considered here all incorporate Thorndike's classic *law of effect* (Thorndike, 1898, Broadbent, 1961). Generally termed reinforcement learning, Thorndike's principle may be summarized as follows: an action which performs well, or results in high payoffs, will be used more often, while an action which performs poorly will be used with less frequency in the future.

Reinforcement models of learning all capture the notion that people learn from the rewards, or payoffs, received and attributed to a particular action. While it is generally accepted that any learning model in low-information settings should incorporate Thorndike's law of effect

in some form, learning models may still be differentiated on at least three levels: the roles of history, reference points, and experimentation. We consider each of these in turn.

3.1.1 History or memory. The role of history in learning is rooted in Thorndike's second principle, the *law of exercise*. Closely related to frequency (Watson, 1914) and the power law of practice (Blackburn, 1936), the law of exercise holds that actions used more often will carry stronger reinforcement. This has an important implication for responsiveness as learning is initially quite fast, but eventually becomes more sluggish. As the "weight of history" becomes greater, it becomes harder to change a strategy that has been performing well historically.

For example, consider Roth and Erev's (1995) basic one-parameter model of reinforcement learning. Each strategy is assigned a "propensity" which is simply the sum of all payoffs received from that strategy over the course of the game, plus some initial value. The probability of using a strategy in any period is proportional to its propensity. Suppose that a player has only two strategies, *A* and *B*, and that payoffs range from 0 to 1. After a few periods, if the propensity of strategy *A* is 4 and of strategy *B* is 1, then the probability of playing strategy *A* in the next period is 0.8. However, a few periods of achieving high payoffs from *B* can easily shift these probabilities. If after some time, the propensities for *A* and *B* are 400 and 100, respectively, the probability of strategy *A* is still 0.8, but it will remain near 0.8 for many periods to come regardless of the relative performance of the two strategies.

Consider a player's probability distribution over her pure strategies in period *t*. History dependence of a learning model may be determined by considering whether the transition from the t^{th} period to the probability distribution in the $t+1^{th}$ period depends only on the last payoff received, or on time-dependent parameters. If the probability distribution over actions at time *t* depends only on the probability distribution, action, and outcome at time *t*-1, then learning is not dependent on history, or is *memoryless*. Hence, the formulation $p_t=f(p_{t-1},a_t,\pi_t;\theta)$, with p, a,and π representing the probability distribution, action taken, and payoff received, and θ a set of parameters, implies a Markovian property, with the transition from one probability distribution to another depending only on last period's play. If a model depends on time explicitly (for example, incorporating a learning term which diminishes over time), or implicitly, as in the previous example, by having probabilities reflect the whole of past experience, then such models will be termed history dependent. In this sense, all variants of the Roth-Erev reinforcement learning

model are history dependent, due to the construction of propensities.[4] The other learning models under consideration perturb the probability distribution over strategies directly by incorporating the payoff from the last period, and hence are not history dependent.

3.1.2 Reference Points. While all of the models considered generate higher probabilities for strategies with "good" outcomes and lower probabilities for strategies resulting in "bad" outcomes, the notion of a good or bad outcome is not absolute, and depends on one's point of reference. New employees of a company might consider a $1,000 bonus at year's end a positive reinforcement, leading to greater loyalty. The CEO receiving the same compensation will certainly view the bonus as a bad reinforcement.

Reference points were introduced to economics as a representation of bounded rationality in the form of satisficing (Simon, 1955, 1957). Referring specifically to environments in which agents may have little information about possible payoffs, Simon suggests that people may develop aspirations and evaluate strategies based upon whether they yield payoffs higher or lower than this satisficing level.[5] In general, the probability of repeating a certain action decreases if the resulting payoff is less than the aspiration. This notion was operationalized into a learning model by Karandikar, *et. al.* (1998).

Reference points are incorporated more broadly than satisficing models. In some formulations of the Roth-Erev model, the payoff used for updating propensities is the realized payoff minus some reference point. In this sense, reference points simply scale payoffs downward, implying that even if all payoffs in the game are positive, relatively low payoffs may be negative reinforcements. Roth and Erev also consider variable reference points, which evolve as the game progresses. Similarly Karandikar, *et. al.* (1999) allow aspiration levels to evolve in the direction of realized payoffs. This captures the idea first put forward by Tinkelepaugh (1928, in an experimental study of monkeys), that "individuals" learn not only about the payoff implications of various actions as the game progresses, but also learn what a "good" payoff is.

[4] In fact, the role of history can be made much stronger by incorporating the "extinction in finite time" principle of Roth and Erev's basic model (1995), which truncates low probabilities to zero, preventing the strategies from ever being used unless experimentation is explicitly incorporated.

[5] For a discussion of satisficing, see Gigerenzer and Todd, 1999.

3.1.3 Experimentation. At the heart of learning is a struggle between loss from intentionally playing sub-optimally in order to gain potentially valuable information, and using strategies currently believed to be optimal to reap maximum benefit in the short term. Responsiveness to an environmental variation is closely linked to experimentation. Faced with uncertain environments, people occasionally deviate from actions that they believe to be optimal in order to explore the strategy space. This notion of learning (the *exploration versus exploitation dilemma*), highlights the tradeoff between acquiring information about one's environment, and taking advantage of the information already acquired. Adaptation requires experimentation in any dynamic environment from control theory (Thrun, 1992a) to organizational learning (March, 1991). While all of the learning models considered incorporate experimentation in some form (since "trial and error" is a fundamental learning method), they differ in how it is modeled.

In computer science and artificial intelligence research, a distinction is drawn between directed and undirected experimentation. Undirected experimentation is achieved by admitting randomness into strategy selection by superimposing a probability distribution on the learning process. Such probabilities are often uniform, implying an equal chance of trying any strategy, or utility-driven, selecting strategies proportional to their expected utilities, or a combination.[6]

In contrast, directed exploration (see Thrun, 1992a, 1992b for a survey) implies using strategies that contribute most to the estimates of the underlying payoff function. Directed exploration incorporates experimentation in order to gain particular knowledge about the environment. Further, directed exploration implies that a learning model keeps track of the experimentation process as well as the learning process. Psychologists, in stark contrast to economists, almost exclusively imply the directed approach when referring to human experimentation. Two popular approaches to directed experimentation may be borrowed from the field of artificial intelligence: recency and full sampling.[7] Recency-based exploration[8] (Sutton, 1990) assumes that knowledge about the world decays, or decreases in informational value with

[6] A variant of the Roth-Erev reinforcement learning model, for example, adopts the combination approach, with strategies played in proportion to past payoffs, with an additional uniform experimentation probability.

[7] A third approach, error-based exploration, in which one experiments by playing strategies with the highest error or payoff variance, is not relevant in our environment with deterministic payoffs.

[8] A number of methods have been proposed for estimating the "exploration bonus" of a strategy based on recency. See Barto, Sutton and Watkins (1989) and Watkins (1989). Note that "recency" in this setting is quite opposite from the recency discussed in Roth and Erev (1995). While in their model, recency implies that more recently used strategies have a higher probability of being played, here we imply the opposite for experimentation. The more recently a strategy was used, the less informative value it has in exploration.

time. The longer the interval since an action has been tried, the more playing the action is expected to contribute to a decision maker's understanding of her environment.

Full sampling implies that decision makers initiate an experimentation phase in which enough of the action space is explored to gain good estimates of the payoff function. These estimates are then used to exploit the environment, or play a "best" strategy, until another period of experimentation is commenced. This is reflected in the analysis of Friedman, *et. al.* (2000), who find that subject experimentation is autocorrelated, not independent in each period. The authors present data characterized by "arrhythmic heartbeat patterns" which depict precisely the notion of full sampling. Many subjects embarked on periods of experimentation in which they sampled a broad portion of the strategy space.

Some of the learning models considered in this paper contain the "flavor" of directed experimentation through appropriate parameterization, though no model in economics appear to incorporate these exploration techniques directly. In the Roth-Erev reinforcement learning models with reference points, high initial propensities and high initial reference points can lead to recency exploration. Initially, as actions are tried, corresponding propensities decrease, making unexplored strategies relatively attractive. This method of directed exploration based simply on overestimating propensities was proposed by Kaelbling (1993) and is in the spirit of Gilboa and Schmeidler (1996) who show that overestimating aspirations can lead to optimization in the long run.

Gilboa and Schmeidler also suggest that long-run utility maximization may be achieved through occasional upward shocks to one's aspiration level. These "trembles" incorporated into the satisficing model of Karandikar, *et. al.* (1998) have a different effect on experimentation. An upward shock to one's reference point results in all strategies looking relatively bad to the decision maker until the reference point again settles down to a reasonable level. Hence, shocks in aspiration levels produce occasional periods of experimentation in the spirit of full sampling.

3.2 Models

3.2.1 Roth-Erev Reinforcement Learning. The initial formulation of Roth and Erev (1995) intended to incorporate the law of effect and power law of practice into a simple learning model. Proposed variants of the model (Roth and Erev, 1995; Erev and Roth, 1998) incorporated reference points, experimentation, and forgetfulness. Each strategy i in every period t has an associated propensity $\rho_t(i)$. Propensities are updated in the following manner. If, in period t, the player uses strategy i and receives payoff $\pi_t(i)$, then

$$\rho_t(i) = (1-\gamma)\rho_{t-1}(i) + (1-\varepsilon)(\pi_t(i)-\alpha_{t-1})$$

$$\rho_t(j) = (1-\gamma)\rho_{t-1}(i) + (\varepsilon/S)(\pi_t(i)-\alpha_{t-1}) \qquad j \neq i$$

where S is the number of pure strategies, γ is a forgetfulness or recency parameter, ε is the probability of experimentation, and α_t is a reference point.[9] The reference points, α_t, evolve according to the following rule:

$$\alpha_t = \lambda\alpha_{t-1} + (1-\lambda)\pi_t(i) \qquad \text{[Eq. 1]}$$

where λ is the persistence of reference points. Lastly, the probability of playing a strategy i in period t is given by:

$$p_t(i) = \rho_{t-1}(i) / \Sigma\rho_{t-1}(j)$$

While the model above consists of a total of five parameters, ε, γ, λ, and two initial conditions ρ_0, α_0, we consider six subsets of the parameter set.

Model	Parameters Included	
Basic	ρ_0	$\varepsilon = \gamma = \alpha_t = \lambda = 0$
Forgetfulness	ρ_0, γ	$\varepsilon = \alpha_t = \lambda = 0$
Experimentation	ρ_0, ε	$\gamma = \alpha_t = \lambda = 0$
Full model	ρ_0, γ, ε	$\alpha_t = \lambda = 0$
Fixed reference	ρ_0, α	$\varepsilon = \gamma = 0$, $\alpha_t = \alpha$
Evolving reference	ρ_0, α_0, λ	$\varepsilon = \gamma = 0$

3.2.2 Two-Stage World Resetting. Concocted after discussions with some of the subjects that participated in the experiment, this model is an extension of the basic Roth-Erev reinforcement learning procedure. Its construction is motivated by subjects informing the experimentalist that upon recognizing a shift in the world, they "reset" their learning, or do not consider historical payoffs. It is, in effect, the Roth-Erev basic model described above, incorporating a test of model fitness. A hypothesized subject maintains in memory not only propensities for each strategy but also estimates of the payoff function. When realized payoffs begin to differ substantially from the estimated, expected payoffs, the model resets the propensities to the initial value, ρ_0. While a

[9] Roth and Erev conjecture that payoffs might only need be generalized to the nearest strategies, such that $\rho_t(j) = (1-\gamma)\rho_{t-1}(i) + (\varepsilon/2)(\pi_{t-1}(i)-\alpha_t)$, $j=i\pm1$. In some preliminary simulations, this formulation made little difference in the models' predictions.

number of criteria exist for such resetting (e.g. Vulkan and Preist, 2000), the experimental setting does not provide much discrimination between these criteria. Since the payoffs in the experiment do not involve any noise or randomness in the payoffs, it is clear to most subjects when a change in the payoff function has occurred.

For the purpose of exposition, consider the following example construction. Payoff expectations for a strategy *i* are given by:

$$\pi_t^e(i) = \begin{cases} \varphi\pi_{t-1}^e(i) + (1-\varphi)\pi_{t-1}(i) & \text{if } i \text{ was used in period } t\text{-1} \\ \pi_{t-1}^e(i) & \text{otherwise} \end{cases}$$

where π_t is the payoff in period *t*, a superscript *e* denotes expectation, and φ measures the persistence of expectations, so that expected payoffs are simply a weighted average of all payoffs received from that strategy. A world is *understood* if

$$\left|\pi_t^e(s_t) - \pi_t(s_t)\right| < \varepsilon \quad \forall t \in \{\tau - k, ..., \tau\}, s_t \text{ used in period } t \qquad \text{[Eq. 2]}$$

i.e., if it predicts accurately for *k* consecutive periods. Once understood, a world is *changed* if

$$\left\{t \mid \left|\pi_t^e(s_t) - \pi_t(s_t)\right| > \varepsilon\right\} \geq \text{r}, s_t \text{ used in period } t$$

that is, if the prediction is substantially wrong in r periods. When the world changes, the model "resets." All parameters revert to the initial values as if the person begins learning a different task.

In our experimental setting, these parameters are not relevant. Without noise in the payoffs, either through the introduction of stochastic terms or interaction with other subjects, any value of φ would yield the same result since each strategy consistently results in the same payoffs, except during a singular change in the underlying payoff function. Similarly, any positive ε would have the same implications in this experimental design since the difference in [Eq. 2] would consistently equal 0 until the change in the payoff function occurs. The only parameter of import is *r* which determines how many periods after the change a subject will reset the learning model. For our purposes, any value of *r* between 1 and 30 yield very similar results in terms of model fitting and estimation. The results presented in the next section are for *r=5*.

While this model has a single parameter, ρ_0, for our purposes, it is not the intention to suggest that this model is "simple." In fact, for most purposeful applications, a criterion for change in the world would require a number of parameters, on top of the parameters inherent in the learning process itself. Even the simple procedure above requires four parameters just to judge the stability of the world. However, for the purposes of determining the role of history or memory in

learning in dynamic settings, this model provides a useful benchmark for the performance of reinforcement learning models.

3.2.3 Responsive Learning Automata. Learning automata were originally simple, one-period memory systems modeled after biological processes, and designed for solving control problems (Tsetlin, 1946; for a survey, see Narendra and Thatcher, 1989). One such simple learner for dynamic settings, the Responsive Learning Automata (Friedman and Shenker, 1996) preserves the low-memory, or no history-dependence property. Unlike the Roth-Erev model in which a dependence on history exists in the updating of propensities, the responsive learning automata's memory is encoded solely in the probability distribution over strategies. If, in period t, the player uses strategy i and receives payoff $\pi_t(i)$, then probability updating is governed by

$$p_{t+1}(i) = p_t(i) + \varepsilon\beta\pi_t(i)\sum_{j\neq i}\omega_t(j)p_t(j)$$

$$p_{t+1}(j) = p_t(i) - \varepsilon\beta\pi_t(i)\omega_t(i)p_t(i) \qquad j \neq i$$

where

$$\omega_t(j) = \min\left[1, \frac{p_t(j) - \varepsilon/S}{\varepsilon\beta\pi_t(i)p_t(j)}\right]$$

and β is a scaling parameter that captures the speed of learning. The probability of experimenting is denoted by ε, and again S is the number of pure strategies. The probability of playing the same strategy as in the previous period increases with the payoff received. All other strategies decrease in probability proportionally. However, no probability of any strategy is allowed to drop below some threshhold, ε/S, guaranteeing that experimentation is always possible.

3.2.4 Aspirations. Aspiration models incorporate the no-memory property of responsive learning automata with a reference-point based behavioral assumption. These models assume that in light of little information about the game and its attainable payoffs, people develop aspirations. A strategy is played more often if the resulting payoff exceeds this aspiration level, and less often otherwise. Further, aspirations may evolve in the direction of realized payoffs.

Aspiration-based learning models have received much attention since Selten (1991). Karandikar, *et. al.* (1998) propose a model in which a strategy is repeated as long as payoffs exceed aspirations. If payoffs fall short of aspirations, the strategy is repeated with some probability that is decreasing with the magnitude of the disappointment, or difference between the

aspiration and received payoff. Further, aspirations are subject to occasional trembles, which is the source of experimentation in the model.

With probability (1-ε) aspirations evolve according to Eq. 1, equivalent to the updating of reference points in Roth and Erev. However, in each period, with probability ε, the aspiration α_t "trembles," and is drawn from a uniform distribution over the feasible payoff space.[10] When a payoff does not exceed the aspiration level, probability updating is governed by the following rule (again, if strategy i is played at time t):

If $\alpha_t > \pi_t$:

$$p_{t+1}(i) = \left[1 + \beta(\alpha_t - \pi_t)\right]^{-1} p_t(i)$$
$$p_{t+1}(j) = \frac{p_t(j)}{1 - p_t(i)}\left(1 - \left[1 + \beta(\alpha_t - \pi_t)\right]^{-1} p_t(i)\right) \quad j \neq i \qquad \text{[Eq. 3]}$$

When the learner is satisfied, implied by the received payoff exceeding the aspiration level, two different models are considered. The first, due to Karandikar, *et. al.* (1998), revises probabilities only in the case of disappointment. If payoffs are above aspirations,

If $\alpha_t \leq \pi_t$ $\quad p_{t+1}(i)=1, \quad p_{t+1}(j)=0 \quad j \neq i$ [Eq. 4]

The probability distribution over actions is only altered after a disappointment. Since the decision maker is not affected by the magnitude of the payoffs as long as payoffs exceed aspirations, this model captures satisficing behavior. Hence, the updating of probabilities according to [Eq. 3] and [Eq. 4], along with the trembling aspirations assumption, will be referred to as the *satisficing model* in the sequel. Borgers and Sarin (1995) consider a similar model, but an action in a given period is always a purely mixed strategy. The probability of playing a strategy not only decreases with the level of disappointment, but also increases with the level of surprise when payoffs exceed aspirations. Incorporating this notion into the above model, we replace [Eq. 4] with:

If $\alpha_t \leq \pi_t$:

$$p_{t+1}(i) = \left[p_t(i) + \beta(\pi_t - \alpha_t)\right]\left[1 + \beta(\pi_t - \alpha_t)\right]^{-1}$$
$$p_{t+1}(j) = p_t(j)\left[1 + \beta(\alpha_t - \pi_t)\right]^{-1} \quad j \neq i \qquad \text{[Eq. 5]}$$

This formulation, consisting of [Eq. 3] and [Eq. 5], as well as the aspirations updating with trembles, is termed the *evolving aspirations* model of learning in the sequel. It is similar in spirit to the Roth-Erev reinforcement learning model with reference points, since aspirations serve as a determinant of whether a strategy's probability should be revised upward or downward. In this sense, decision makers do not satisfice, since they react to ever greater payoffs. However, the

[10] In Karandikar, *et. al.*, discussion focuses on aspirations assumed to "tremble" locally. However, with deterministic payoffs, this implies that a strategy chosen in the first few periods would be repeated throughout the 600 periods of the experiment with probability close to 1, as local trembles would rarely lead to disappointment.

model is substantially distinct from the reinforcement models since it has no memory, and, with the inclusion of trembles in aspirations, it allows for full sampling experimentation.

4. Model Performance

The experimental design accommodated 101 possible actions in each period. Since the payoff function is continuous in strategies, subjects learned to generalize, or associate realized payoffs with strategies close to the strategy actually used. It is not apparent, however, how this generalization occurs.[11] Since the models considered in this paper do not generalize, but instead treat each strategy as entirely distinct, for the purpose of fitting and simulating the models, the game is reduced to ten strategies, {10,20,…,100}, and the experimental data is aggregated by mapping players' strategies into the next highest among the ten available.[12] To find the best parameters for each model, the mean squared deviation criterion (MSD) is used (Simon, 1956).[13] A total of 10,000 simulations over 1,000 sets of parameters were run for each model and for each of 56 subjects. For details of the simulation methodology, see the appendix.

Each model was fit to the data from Treatment 1 in two ways. Learning models were fit to the play of each individual subject, as well as to all of the data simultaneously. While a model that requires fitting parameters to each individual subject is perhaps of little relevance to economic forecasting, there is a pedagogical purpose for the exercise. Psychologists disagree about the source of variation in individual decision making. People may differ in the heuristics, or rules of thumb, that they employ. Alternatively, individuals might employ similar heuristics, but differ in particular learning or adaptation parameters.[14] Recognizing if learning models perform well on an individual level allows us to determine if these models represents heuristics common to most subjects, even if the exact value of the parameters representing that heuristic differ from person to

[11] For some of generalization methods, see Shepard (1987) and Staddon and Reid (1990). These studies do not provide functional forms for generalization, instead suggesting how like strategies are evaluated on a metric in "psychological space." (Shepard, 1987). Considering the myriad applications of generalization (e.g., auditory tasks, speech, visual problems) and numerous proposed mental processes (e.g., attributing realized payoffs to neighboring strategies, curve fitting, interpolation), any simple functional form of generalization is probably too specific to be globally useful.

[12] Early simulations using all 101 strategies demonstrate that none of the models track the data well over the 600 periods. Given a larger number of periods, however, the models act similarly to the results presented, but take substantially greater time to converge and react to changes in the payoff function.

[13] The probability distribution over actions is denoted by a vector p. If strategy i was used, $MSD = (1-p_i)^2 + \Sigma_{j \neq i} p_j^2$. For a discussion of some desirable properties of MSD, see Selten (1998)

[14] For example, Schunn and Reder (1998) study responsiveness in an experimental dynamic environment (air traffic control). They find that people employ similar strategies, but individual variations in inductive reasoning skill result in differences in values of parameters such as speed of adaptability.

person. Further, we may be able to distinguish between models of learning with good normative properties, and models that reflect the heuristics of actual decision-making.

The mean squared deviation scores for each model are presented in two forms (Table 3). Column A contains the lowest average MSD for each model when fit to each subject individually, while column B presents the MSD when each model is simultaneously fit to all of the data. Benchmark MSD scores are presented for the equilibrium prediction and for a random choice model, which assumes that each player selects each action with equal probability in every period.

4.1 Individual Fit

All of the models perform substantially better than either the equilibrium or random choice benchmark. However, comparing the models fit separately to each individual subject (Table 3, column A) the aspirations-based models appear to do quite well, followed by responsive learning automata and two variants of reinforcement learning, the Roth-Erev full model and the two-stage world resetting learner.

To compare the performance of the models, we inquire about the proportion of subjects for whom one model performs better than another (Table 4). Each entry in Table 4 is the proportion of subjects for whom the model in the row predicts better (in terms of lower MSD) than the model in the column. All learning models outperformed the random choice benchmark for every subject and most models substantially outperform the equilibrium prediction. Given that even one parameter models outperform the equilibrium for over 2/3 of the subjects, this may be more an affirmation of bounded rationality in low-information games than an endorsement of any particular model.

To gain insight into the heuristics being employed by subjects, the learning models in Table 4 are ordered by the number of competing models that they surpass in the accuracy of their fit for a majority of subjects (column “best”). For example, the first entry in the table, the satisficing model, obtains a lower MSD score than any other model for a majority of subjects. Interestingly, none of the top four models is history dependent, while the bottom six learning models incorporate propensities. Further, each of the top four models performs at least as well as each of the lower six. This suggests that history dependence is not an applicable heuristic for some real-time changing environments.

The two aspirations-based models perform substantially better than the rest. In fact, satisicing results in a lower MSD for at least three-fourths of all subjects, when compared to any non-aspirations-based model. This performance of the aspirations models is consistent with

"strategies" expressed by experimental subjects in informal interviews, many of whom suggested that they learn to be content with their current payoff, but every now and then wish to see if they can do better, reflecting both satisficing behavior and trembles in aspiration formation.

4.2 Aggregate Fit

Each model was fit to data from Treatment 1, using the play of all subjects simultaneously. For each model, the set of parameters that minimized the average MSD over all subjects was selected. Unlike the individual fits, which allow a different set of parameters for each subject, aggregate model fitting does not capture individual idiosyncrasies in learning, but provide a single set of parameters that may encapsulate general behavior. These parameters then may be used to predict play in similar games. This ability of a model with a particular set of parameters to explain subsequent play in a variety of experiments was demonstrated by Roth and Erev (1995).

When fit to aggregate data, one would expect a given model to perform substantially worse than when fit to individuals. Given the 56 subjects in Treatment 1, a two-parameter model, for example, in effect uses 112 parameters when fit to individual data. While each model when fit to aggregate data still outperforms both benchmarks, the MSD scores of the models are substantially worse than the scores from individual fits (Table 3). The average model's best fitted MSD score increases by 25.6%. However, the ordering of the models remains largely unchanged.[15] Reinforcement learning models are the worst performing. Inclusion of reference points improves the fit of the models. Again, the best performing model is satisficing, which not only incorporates reference points, but is also not history dependent, and allows for experimentation in stages by an occasional tremble in aspiration levels. Analysis of both individually fit and aggregate data suggests that history-dependent learning models do not explain the data well.

Evaluation of models with varying formulations is important for assessing the relative value of different parameters and incorporated assumptions. We can compare the fit of the different Roth-Erev models, and the contribution of each parameter to the model's explanatory power (Figure 2). Beginning with the basic 1-parameter model, the arrows in Figure 2 show the improvement in mean squared deviation from the inclusion of additional parameters. Dotted lines lead to comparable memoryless models.

An observation that may be drawn is that γ, representing "forgetfulness" in the context of the Roth-Erev reinforcement model, contributes little explanatory power. The addition of γ to the

[15] This may suggest that people are rather idiosyncratic, but employ similar heuristics at different rates.

basic model decreases MSD by about 0.3%. Similarly, the addition of γ to a model already incorporating experimentation decreases MSD by about 0.2%. Experimentation, on the other hand, contributes greater explanatory power to both the basic model and a model with forgetfulness. However, the parameter contributing the most explanatory power to the basic model is a fixed reference point, lowering MSD by twice as much as experimentation. Reference points appear to play an important role in both the decision making process of subjects and the normative value of a model. After incorporating reference points into the model, additional explanatory power results from allowing those reference points to evolve. Finally, incorporating disappointment-based satisficing behavior causes the largest change in MSD, from 0.794 to 0.757.

We can compare the history-dependent models with models without memory of equal complexity, measured by the number of parameters. The dotted lines (Figure 2) reflect similar models that consist of the same number of parameters, do not incorporate history dependence, yet explain the data better. The two-parameter reinforcement learning model with experimentation may be compared to the responsive learning automata. Both incorporate experimentation as a minimum probability bound on each strategy, but responsive learning automata replaces a history-based parameter with a parameter representing the speed of learning. Despite an equal number of parameters, the responsive learning automata model generates a superior fit.

Numerical comparisons of models using the MSD criterion do not provide an adequate picture of how well the models describe subject behavior qualitatively. The path of play as predicted by each learning model was simulated using the parameters that minimize MSD (Figure 3). Numerically, both Roth-Erev models with reference points achieve lower MSD scores for aggregate fits. Yet, among all of the formulations of the reinforcement learning model (Figs. 3b-3g), only the full (three-parameter) model demonstrates responsive to the payoff change comparable to the data. This seeming paradox between quantitative and qualitative model comparisons can be explained by the inherent tradeoff between experimentation and exploitation. A model that does not incorporate a high degree of experimentation will be unable to respond well to environmental variations. However, since little time is spent exploring the strategy space, convergence is swifter and more robust. For this reason, the models with reference points quickly converge in both mean and median to the equilibrium strategy of four, while the full Roth-Erev model is affected by constant experimentation, pulling average play towards the central strategy. Since seven of the ten minutes of the experiment occur before any change to the payoff function,

the MSD score favors models which accurately track convergence over those that track the responsiveness in the last three minutes.

The remaining learning models all display both convergence to the equilibrium and responsiveness to the change in the payoff function (Figs. 3h-3k). The tradeoff inherent in any model of learning between strong convergence to the currently optimal strategy and experimentation significant enough to recognize changes in the environment may be examined by decomposing the MSD score into two time intervals (Table 5). A model with a relatively low MSD before the change in payoffs accurately tracks subjects' convergence, while a low MSD score after the change reflects a good fit to subjects' adaptation to the change. The basic, one parameter Roth-Erev model is comparable to the full model in performance prior to the change. However, the distinction between the basic (one parameter) and full (three parameter) models is quite stark if considering the post-change MSD score, where the one parameter model fails to predict better than even the random choice benchmark. Similarly, the addition of reference points to the basic Roth-Erev model, whether fixed or evolving, leads to very strong convergence initially, but the poorest responsiveness of any of the models considered.

Interestingly, the performance of the basic Roth-Erev model is benefited most not by the inclusion of experimentation parameters but through a notion of world-resetting. If the basic model "throws out" built-up propensities when the underlying payoffs change, its descriptive power rises considerably post-change, and even is a better descriptor of the data before the change in payoffs. The latter may sound counterintuitive; why should the two-stage world-resetting model perform better in the first part of the experiment when it is, up to the change in payoffs, equivalent to the basic model? Experimentation in the basic model is driven largely by the value of initial propensities. If a high amount of experimentation must persist in order to recognize environmental variations even after a good amount of time has elapsed, then the best-fitting estimated initial propensities must be large (ρ_0=300). The tradeoff is that large initial propensities will lead to slower convergence. Since the two-stage world resetting model is not hindered in this fashion, lower initial propensities provide a better fit (ρ_0=125).

Again, the intention is not to suggest that the two-stage world resetting model is a fair competitor, since its application in less simple settings would require a theory of how people perceive change, and a number of additional parameters to incorporate that notion. Further, the model was developed after the author was privy to the data and subjects' sentiments. Ironically, while its fabrication was driven by the results, it does not perform well when compared with

aspiration-based learning models (Figures 3i-3k). Both aspiration-based models track the data well throughout, but the satisficing model is the only model to accurately describe the dispersion of play. While indicating what the average player would do is an important task for any learning model, describing the dispersion of play is as important if not more so for many applications, especially if the efficiency or total payoffs in a game fall off substantially as players move away from equilibrium. The inner quartile of play, both in the experiment and that simulated by the satisficing model, is quite broad initially but quickly converges on the equilibrium strategy. Shortly after the change in payoffs, simulated and actual play once more becomes more volatile but again soon converge on the new equilibrium.

4.3 Model prediction

To assess the normative value of the learning models, we evaluate the predictive power of the parameters fit to Treatment 1. Two approaches for testing the predictive power of models have been adopted. The first, termed *in sample* (sometimes *post hoc* or *cross-validation*; see Mosier, 1951), uses parameter values estimated from a population of subjects in a given experiment to predict the behavior of a different population in the same experiment. *In sample* is useful for evaluating a model's stability - if a model calibrated to one population predicts well the behavior of another population in the same task then we might conclude that people are learning in similar ways when faced with this task. Busemeyer and Wang, (2000) suggest a methodological drawback to the *in sample* approach. Given that the data distributions across two populations faced with the same task are expected to be similar, the same models should perform well in both populations, hence not providing an adequate challenge to the models' predictive abilities.[16]

If a model is to have normative value, it should be able to provide some insight into how different players would perform in a task different from the one used to fit the parameters. The approach adopted in this paper, o*ut of sample* prediction, compares simulated data using the models appropriately fit to Treatment 1 to actual data from Treatment 2, which incorporates a different experimental design. The primary distinction, in terms of responsiveness, is that a player selecting the equilibrium action in Treatment 1 will notice a change in payoffs from that strategy when the environment changes, while a player in Treatment 2 will not. Another distinction is that in Treatment 2, the change in payoffs occurs halfway through the experiment, hence balancing the

[16] This is further complicated by the fact that models with more free parameters will generally fit better, and nested models will necessarily favor more parameters. Hence, the *in sample* approach will favor more complicated models.

relative weighing of initial convergence and responsiveness in the mean squared deviation scores whereas in Treatment 1, seventy percent of play occurred before the change.

The measure of how well the models predicted play in Treatment 2 was decomposed into predictive power before and after the change in payoffs (Table 6). The prediction of all of the models in Treatment 2 is systematically worse than the fit to Treatment 1. This certainly is not surprising given that for Treatment 2, a parameter-free out of sample comparison is used. Prior to the change in the payoff function, models involving aspirations or reference points all do well. The best prediction for initial play in Treatment 2 is derived from the fixed and evolving reference variants of reinforcement learning. However, these models again fail to capture responsiveness. With the exception of the full reinforcement learning model, the history-dependent learning models perform similar to or worse than the random choice benchmark after the change in the payoff function. Of the non-history dependent models, the worst performing is the responsive learning automata, the only memoryless model incorporating undirected, random experimentation.

In general, we can compare how well the models predict play both before and after the change in the payoff function in Treatment 2, given that the length of the experiment was the same on either side of the environmental change. All of the history dependent models have much higher MSD scores for the post-change part of the experiment than pre change (an average of 13% higher). For the responsive learning automata, the increase is five percent. In contrast, the two-stage world resetting, evolving aspirations, and satisficing models increase less than one percent in mean squared deviation, suggesting that they predict the initial convergence of subjects' play about as well as the responsiveness of subjects to the environmental change.

Models with higher MSDs than random choice after the change in parameters are the same models that showed little or no responsiveness in Treatment 1. In a graphical comparison restricted to models that surpass the random choice benchmark after the change in parameters (Figure 4), all of the models displayed initially converge to the equilibrium in median strategy and then eventually respond to the change in payoffs. However, while subjects' play again shows convergent behavior, in the sense that the inner quartile closes in on the equilibrium, most models do not track this behavior well, instead overestimating the variance of play. Only the satisficing model (Fig. 4f) displays comparable convergent behavior.

From visual inspection, the responsive learning automata model (Fig. 4d) appears to require the longest time to react to the change in the payoff function. The value of the fitted parameter, β, representing the speed of learning, is 0.0005, indicating slower adjustment times.

Even after initial reaction, the responsive learning automata model appears to stabilize on strategy 5 and only towards the end of the 600 periods begins to move towards the post-change equilibrium strategy of 6. Similar stepwise movement towards the new equilibrium is apparent in the evolving aspirations model (Fig. 4e) and, to a lesser degree, in the full Roth-Erev model (Fig 4b). Before the change in the payoff function, all of the models are placing greater probability on strategies near the equilibrium since these strategies result in higher payoffs. Hence, after the change in the payoff function, experimentation is more likely to occur with strategies near the former equilibrium, leading to an initial bias in favor of these strategies.

The two-stage world resetting model was introduced to capture the notion that subjects do not carry over history from one environment to the next, and that experimentation is a periodic phenomenon and not a random event in each period. Motivating this was the discovery that each of the four "learning opportunities" (two payoff functions in each treatment) appears to follow the same path of play. The average strategies used by experimental subjects for each of the four "learning opportunities," the first three minutes of each treatment, and the first three minutes after the parameter change in each treatment are presented (Figure 5). The closeness of the bottom two lines, representing the first three minutes of each treatment, is not surprising. Both treatments started identically. However, the comparison between the treatments after the change in the payoff function suggests that the learning process was similar, despite differences in the timing of the change, how noticeable the change was, and the payoff at the post change equilibrium. Further, the lower two lines appear as mirror images of the upper lines, suggesting that the process of learning did not change during the experiment. Subjects learned the first equilibrium following essentially the same path of play as the second. Strategic persistence is not observed, nor is the role of memory or history evident. The main distinction is that in Treatment 1, players reacted faster (began to move upwards earlier) to the change in payoffs. This, however, is due to the inability of players to notice the change in the environment until experimentation, in some form, occurred.

It is evident that aspirations-based models perform the best in both describing and predicting play in this dynamic experiment. Specifically, a variant of the model introduced by Karandikar, *et. al.* (1998) obtains the lowest MSD scores in descriptive and predictive roles, as well as qualitatively describes and predicts the aggregate path of play and its dispersion. An interesting question raised about this class of models is whether learning is benefited by occasional trembles in the aspiration level. Gilboa and Schmeidler (1996), in the context of a model with long-term memory, suggest that learning is improved through occasional upward

shocks as such shocks induce experimentation. Borgers and Sarin (1995), however, propose that in a model without history, trembles may lead to volatile, non-convergent play. To test the robustness of the satisficing model, one can investigate if subjects with higher tremble probabilities do in fact exhibit more volatile play or if the added experimentation leads to faster learning.

Given the parameter estimates for each individual subject for the satisficing model, the empirical distribution of aspiration tremble probabilities may be calculated (Figure 6). Interestingly, the distribution of tremble probabilities in the population of subjects is not significantly different from a uniform distribution on the interval [0,0.055] (Kolmogorv-Smirnov test yields a critical value of 80%). Subjects are separated into two groups, *low experimentation* and *high experimentation*, based on whether their tremble probabilities are above or below the empirical median value of 0.025. A graphical representation of median play for each subject group (Figure 7) demonstrates that more robust convergence is obtained for the low experimentation group, although both are comparable in responsiveness to change. This lends support to the argument that high tremble probabilities may decrease the robustness of convergence.

5. Conclusion

History-based models of learning have performed quite well, explaining data from a variety of repeated games experiments (Roth and Erev, 1995). However, in real-time dynamic games, they appear ill suited. Subjects may be able to recognize a change in their environment, which leads to discarding much of what has been learned, as it may be inappropriate for the new setting. Nevertheless, evidence of persistence in strategy selection exists outside of this experiment. For example, corporate leaders often maintain a strategy that was successful in the past despite the strategy being suboptimal following an environmental shift (Audia, Locke, and Smith, 2000). Strategy persistence, therefore, may rely crucially on the beliefs of the decision makers. While in the experimental environment, subjects had little basis to believe that they could influence their environment, businesses may display greater persistence due to beliefs held by managers about their interrelationship with their environment.[17]

The incorporation of reference points provided the greatest contribution to explanatory and predictive power of the models considered. However, fixed reference points hinder responsiveness in dynamic environments. A critical limitation of fixed reference points is that they do not permit

[17] Such biases have been well documented. For example, *self-serving* biases (Heider, 1958) lead managers to attribute higher profits to their own ability, rather than environmental conditions, and *fundamental attribution* biases (Rotter, 1966) cause people to overestimate their ability to control the environment.

experimentation unless an environmental shift decreases the payoff at the former equilibrium. If the environment changes such that payoffs at the former equilibrium rise, then the model evaluates the strategy as "even better than before," further reinforcing that strategy and hindering adaptation.

In the field of artificial intelligence, practitioners have long realized that random, occasional deviations from the optimal strategy lead to slower learning than more directed experimentation techniques. Not surprisingly, nature may have learned this lesson well before the theoretician. However, modeling of human learners in economics generally maintains this assumption. That people's experimentation is not independent from one period to the next, nor as rare as often assumed in the literature was shown by Friedman, *et. al.* (2000). Instead, experimentation is often performed in stages, and models incorporating such experimentation outperform those that do not. What brings about such patterns may be referred to as "optimism in the face of uncertainty." High initial propensities, for example, suggest that untried, and hence uncertain strategies are played with high probability in early periods. Alternately, occasional upward shocks to one's aspiration or reference point beyond what currently is obtainable reflect an optimism, or a hope, that some other strategies may outperform what currently seems best.

A well-known maxim provides that a model is only as good as the assumptions that guide it, its extrapolation to novel situations, and the data that populate it. This paper has addressed the first two elements, differentiating the heuristic assumptions of a number of learning models, and evaluating both their *ex post* and *ex ante* descriptive power. However, while the experiment isolated responsiveness in a low-information, real-time dynamic framework, in most settings decision makers must discern between more subtle environmental variations, as well as distinguish between noisy payoff functions. Satisficing, directed experimentation, and lack of history dependence are necessary components of learning models which hope to accurately predict play in dynamic settings. However, the construction of such models flexible enough to be of normative use in economic theory requires a continuing fusion of psychology, control theory, and economics.

Appendix – Simulation Methodology

A1. Estimating Parameter Values

Simulations were run using the updating rules prescribed by each model, the actual payoff functions from the experiment, and the mean squared deviation (MSD) criterion. Parameters were chosen using an iterative grid procedure. For each model, a broad grid was chosen to permit 250 sets of values of the parameters. Specifically, for each parameter which is logically constrained (experimentation probability cannot be greater than one, for example), the initial grid covered the entire range of possible values. For other parameters with no logical ceiling (such as initial propensities) a maximum value was chosen large enough so that the highest two values of the grid would not minimize MSD for any subject. After this initial run, the inner quartile of the actual values which minimized subjects' MSD was used for the new bounds on the grid in the next iteration. A total of four iterations were used for each model, for a total of 1000 sets of parameter values. At each value, 10 simulations were run for each subject, and the MSD was averaged among them. Hence, a total of 10,000 simulations were used per subject, yielding 560,000 total simulations per learning model, to find the best parameters for each model.

A2. Generating Figures

Once the parameter values representing the best fit were found for each model, 1000 players were simulated for each treatment using those values. This data was used for the figures.

A3. Comparisons to Treatment 2

The parameters obtained from the above procedure were used to simulate players to compare to the behavior of subjects in the second treatment. A total of 1000 simulations were run for each model, each was compared to the actual data, and the mean absolute deviations presented are the average over the 1000 simulations.

References

Audia, Pino G., Edwin A. Locke, and Ken G. Smith (2000). "The Paradox of Success: An Archival and Laboratory Study of Strategic Persitence Following a Radical Environmental Change." *Academy of Management Journal*. Forthcoming.

Barto, Andrew G., Richard S. Sutton, and Chris J.C.H. Watkins (1989). "Learning and Sequential Decision Making." *CMPSCI Technical Report* 89-095. University of Massachusetts Department of Computer Science.

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

Borgers, Tilman and Rajiv Sarin (1995). "Naive Reinforcement Learning with Endogenous Aspirations." *Mimeo*. University College, London.

Broadbent, Donald E. Behavior. London: Eyre and Spottiswoode, 1961.

Busemeyer , Jerome R. and Yi-Ming Wang (2000). "Model Comparisons and Model Selection Based on Generalization Criterion Methodology." *Journal of Mathematical Psychology* 44: 171-189.

Carley, Kathleen M. and Ju-Sung Lee (1998). "Dynamic Organiations: Organizational Adaptation in a Changing Environment." In Joel Baum, Ed. Advances in Strategic Management Vol 15: Disciplinary Roots of Strategic Management Research. Stamford, CT: Jai Press, 269-297.

Dooley, Kevin J. (1997). "A Complex Adaptive Systems Model of Organizational Change." *Nonlinear Dynamics, Psychology, and Life Sciences* 1: 69-97.

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.

Friedman, Eric, Mikhael Shor, Scott Shenker, and Barry Sopher (2000). "Asynchronous Learning with Limited Information: An Experimental Analysis." *Mimeo.* Rutgers University Department of Economics.

Friedman, Eric, and Scott Shenker (1996). "Synchronous and Asynchronous Learning by Responsive Learning Automata." *Mimeo.* Rutgers University Department of Economics.

Gigerenzer, Gerd and Peter M. Todd (1999). Simple Heuristics That Make Us Smart. New York: Oxford University Press.

Gilboa, Itzhak and David Schmeidler (1996). "Case-based Optimization" *Games and Economic Behavior* 15: 1-26.

Hayes, Robert H., Steven C. Wheelwright, and Kim B. Clark (1988). Dynamic Manufacturing: Creating the Learning Organization. New York: The Free Press.

Heider, Fritz. The Psychology of Interpersonal Relations. New York: Wiley, 1958.

Kaelbling, Leslie P. (1993). *Learning in Embedded Systems*. PhD thesis. Cambridge, MA: MIT Press.

Karandikar, Rajeeva, Dilip Mookherjee, Debraj Ray, and Fernando Vega-Redondo (1998). "Evolving Aspirations and Cooperation." *Journal of Economic Theory* 80: 292-331.

Narendra, Kumpati and M.A.L. Thatcher. Learning Automata: An Introduction. Prentice-Hall, Englewood Cliffs, NJ, 1989.

Levitt, Barbara and James G. March (1988). “Organizational Learning.” In Michael D. Cohen and Lee S. Sproull, Eds. (1996). Organizational Learning. Thousand Oaks, CA: Sage, 516-540.

March, James G. (1991). "Exploration and Exploitation in Organizational Learning." *Organization Science* 2: 71-87.

Mosier, C. I. (1951). “Problems and Designs of Cross-Validation.” *Educational and Psychological Measurement*. 11: 5-11.

Mookherjee, Dilip and Barry Sopher (1994) “Learning Behavior in an Experimental Matching Pennies Game.” *Games and Economic Behavior* 7: 62-91.

Narendra, Kumpati S. and Mandayam A.L. Thatcher (1989). Learning Automata: An Introduction. New York: Academic Press.

Payne John W., James R. Bettman, and Eric J. Johnson (1993). The Adaptive Decision Maker. New York: Cambridge University Press.

Pearl, Judea (1984). Heuristics: Intelligent Search Strategies for Computer Problem Solving. Addison-Wesley.

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.

Rotter, Julian B. (1966). “Generalized Expectancies for Internal versus External Control of Reinforcement.” *Psychological Monographs* 80:1-28.

Schunn , Christian D. and Lynne M. Reder (1998). “Strategy Adaptivity and Individual Differences.” In D. L. Medin, Ed. The Psychology of Learning and Motivation. New York: Academic Press.

Selten, Reinhard (1991). “Evolution, Learning and Economic Behavior.” *Games and Economic Behavior* 3: 3-24.

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

Shepard, Roger N. (1987). “Toward a Universal Law of Generalization for Psychological Science.” *Science* 237: 1317-1323.

Simon, Herbert A. (1957). Models of Man. New York: Wiley.

Simon, Herbert A. (1956). “Dynamic Programming under Uncertainly with a Quadratic Criterion Function.” *Econometrica* 24: 19-33.

Simon, Herbert A. (1955). “A Behavioral Model of Rational Choice.” *Quarterly Journal of Economics* 69:99-118.

Staddon, John E.R. and Alliston K. Reid (1990). “On the Dynamics of Generalization.” *Psychological Rev*iew 97: 576-578.

Sutton, Rich S. (1990). “Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming.” *Proceedings of the Seventh International Conference on Machine Learning*, 216-224.

Thorndike, Edward L. (1898). “Animal Intelligence: An Experimental Study Of The Associative Processes in Animals.” *Psychological Review Monograph Supplement.* No. 8.

Thrun, Sebastian B. (1992a). "The Role of Exploration in Learning Control." In David A. White and Donald A. Sofge, Eds. Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches. New York: Van Nostrand Reinhold.

Thrun, Sebastian B. (1992b). "Efficient Exploration in Reinforcement Learning." *Technical Report CMU-CS-92-102*. School of Computer Science, Carnegie Mellon University.

Tinklepaugh, O. L. (1928). "An Experimental Study of Representative Factors in Monkeys." *Journal of Comparative Psychology* 8: 197-236.

Tsetlin, M.L. (1946). "Behavior of Automata in Random Media." Ph.D. Dissertation. Published in Automaton Theory & Modeling of Biological Systems. Translated from the Russian. New York: Academic Press, 1973, pp. 12-83.

Van Huyck, John B., Raymond C. Battalio, and Frederick W. Rankin (1996). "Selection Dynamics and Adaptive Behavior without Much Information." *Mimeo.* Texas A&M Department of Economics.

Vulkan, Nir, and Chris Preist (2000). "Automated Trading in Agents-based Markets for Communication Bandwidth." *Hewlett Packard Technical Report* HPL-2000-24.

Watkins, Chris J.C.H. (1989). "Learning from Delayed Rewards." PhD Dissertation. University of Cambridge, England.

Watson, John B. (1914). Behavior: An Introduction to Comparative Psychology. New York: Holt.

Figure 1. *Subject data for experimental treatments*

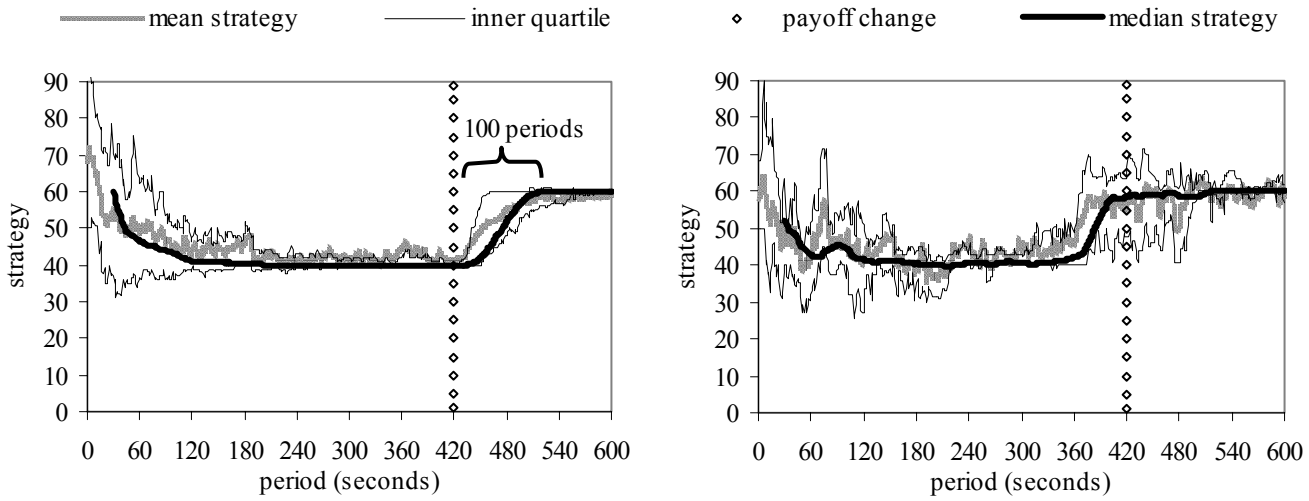


Figure 2. *Comparison of reinforcement learning models and the contribution of certain parameters to the models' descriptive power.*
Dashed lines and boxes represent comparable models.

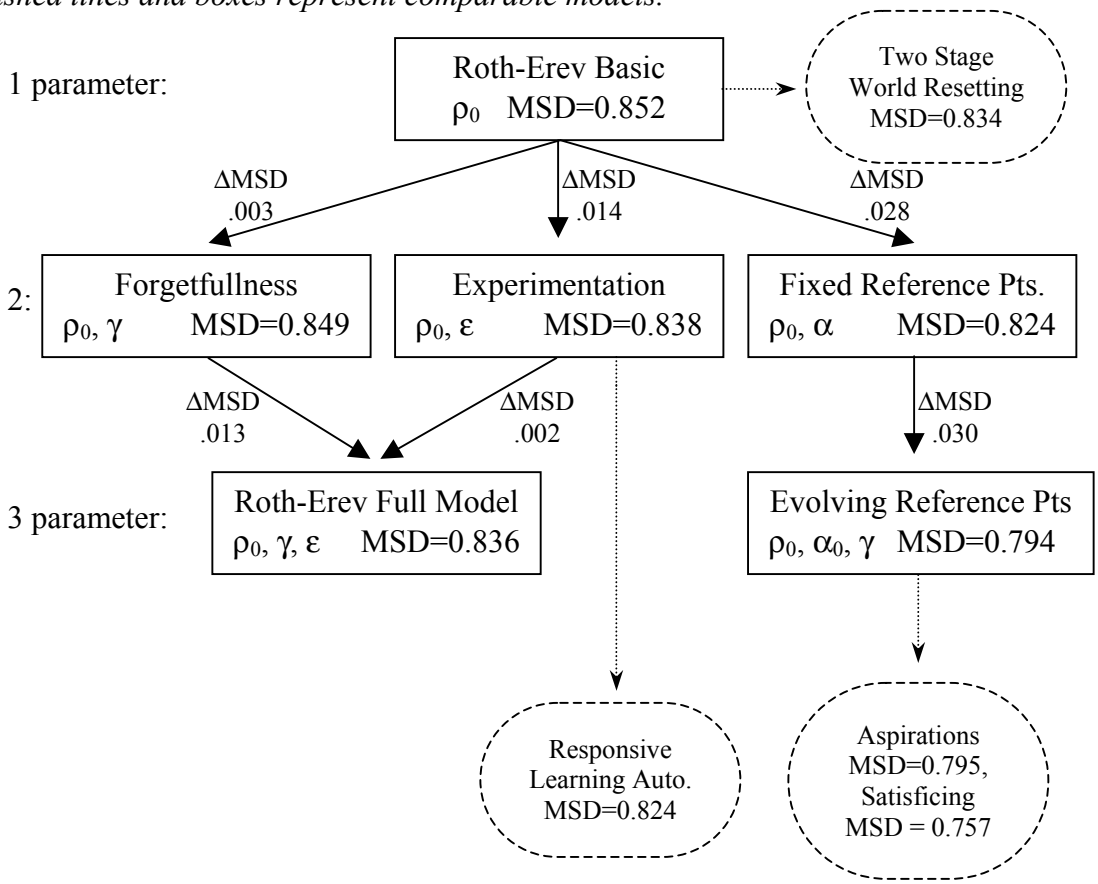
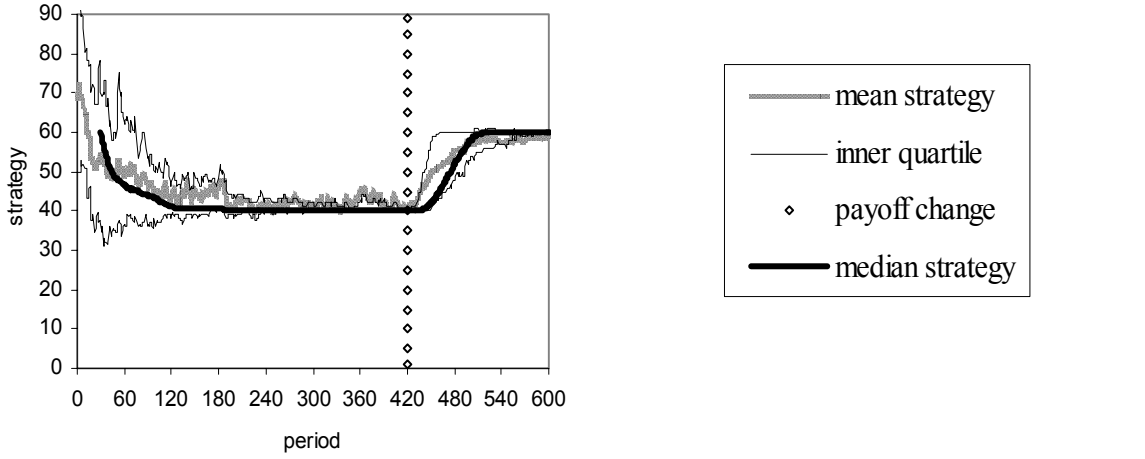
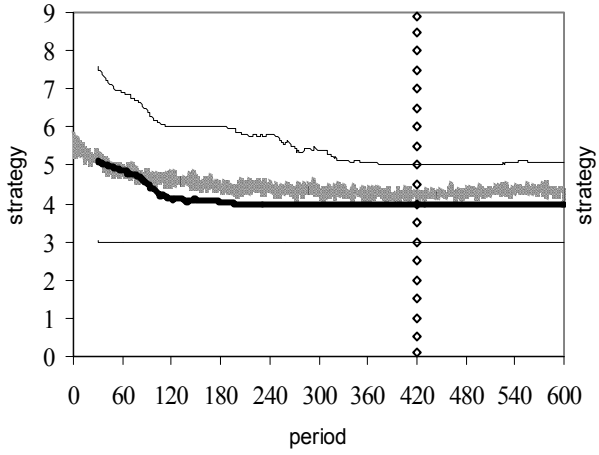


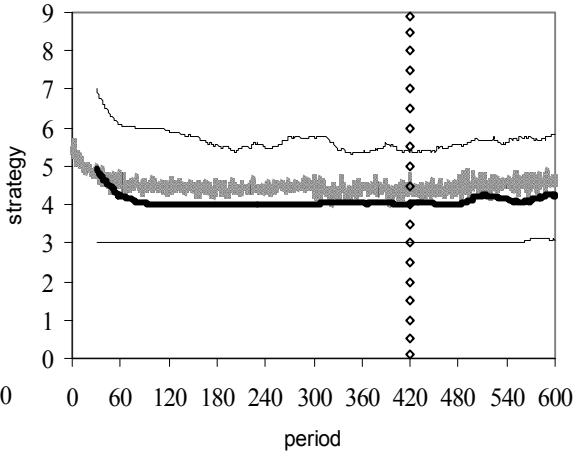
Figure 3. *Simulated Play for Treatment 1.*



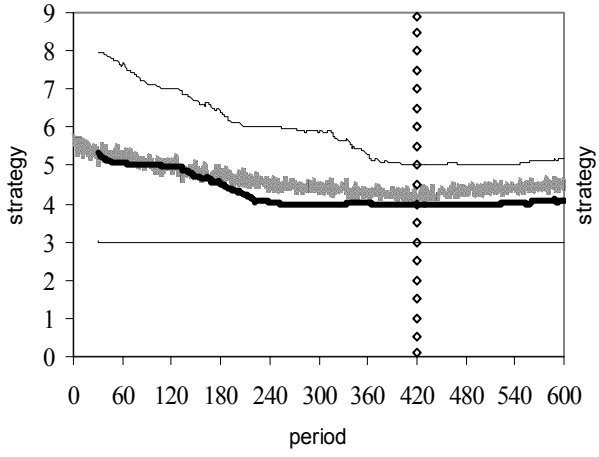
3a. Experimental Data for Treatment 1.



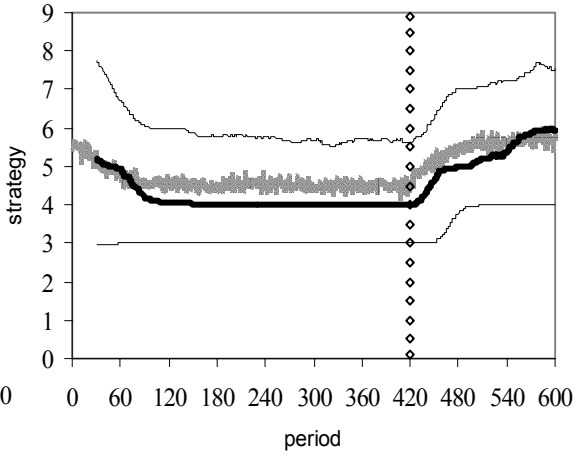
3b. Roth-Erev RL Basic (1 parameter) Model.



3c. RL Basic Model with Experimentation.

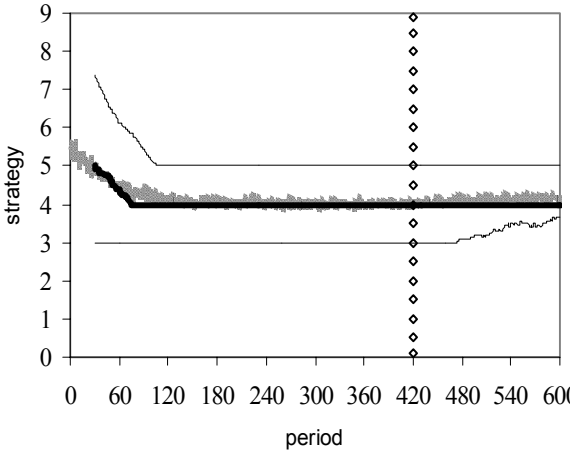


3d. RL Basic Model with Forgetfulness.

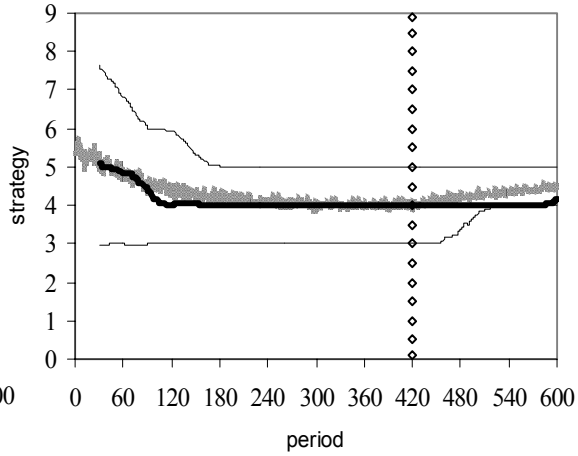


3e. Roth-Erev Full (3 Parameter) Model.

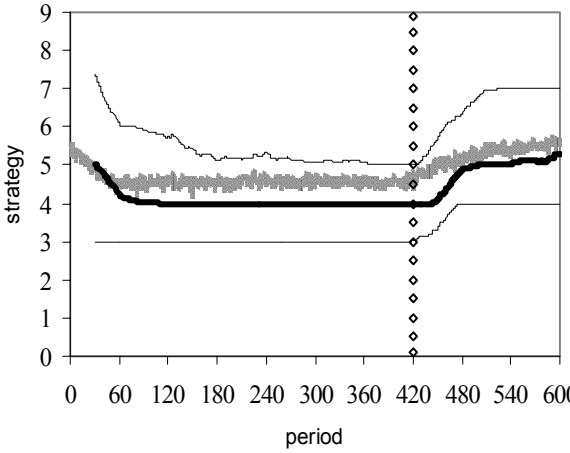
Figure 3 (cont). *Simulated Play for Treatment 1.*



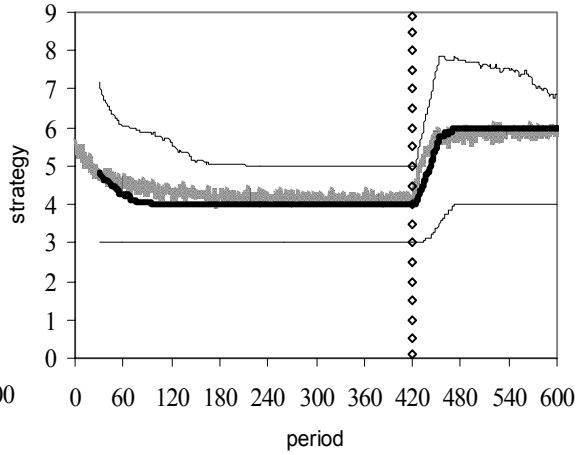
3f. RL Basic Model with Fixed Reference Pts.



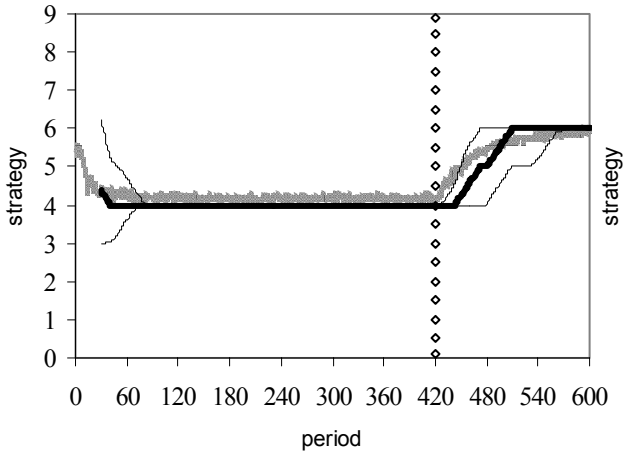
3g. RL Basic Model with Evolving Ref. Pts.



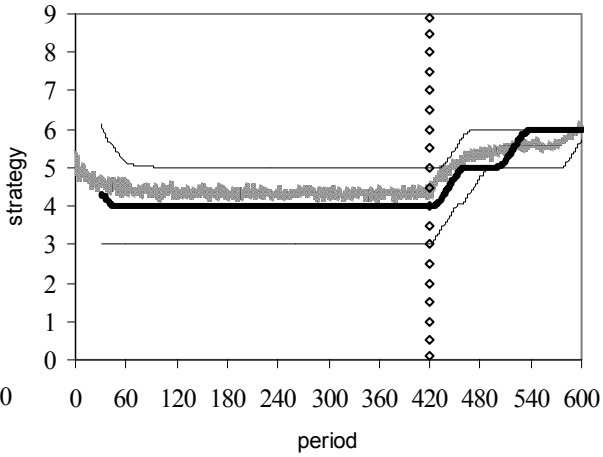
3h. Responsive Learning Automata.



3i. Two-Stage World Resetting.

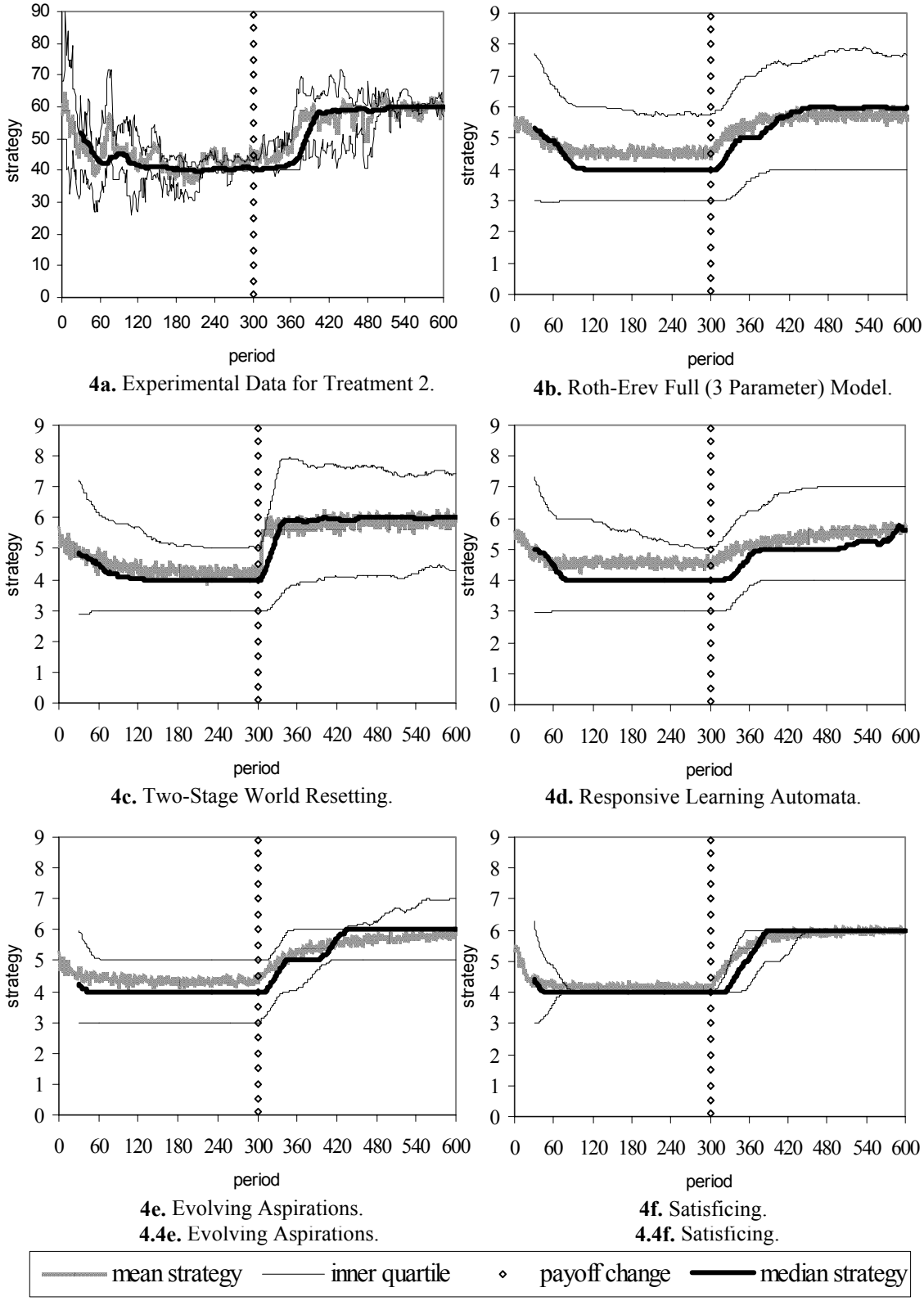


3j. Satisficing.



3k. Evolving Aspirations.

Figure 4. *Simulated Play for Treatment 2.*



4a. Experimental Data for Treatment 2.

4b. Roth-Erev Full (3 Parameter) Model.

4c. Two-Stage World Resetting.

4d. Responsive Learning Automata.

4e. Evolving Aspirations.

4.4e. Evolving Aspirations.

4f. Satisficing.

4.4f. Satisficing.

Figure 5. *Comparison of Learning Paths. Upper two graphs represent post-parameter change play (right axis). Lower two graphs represent pre-parameter change play (left axis). All paths represent the average strategy choice across all subjects.*

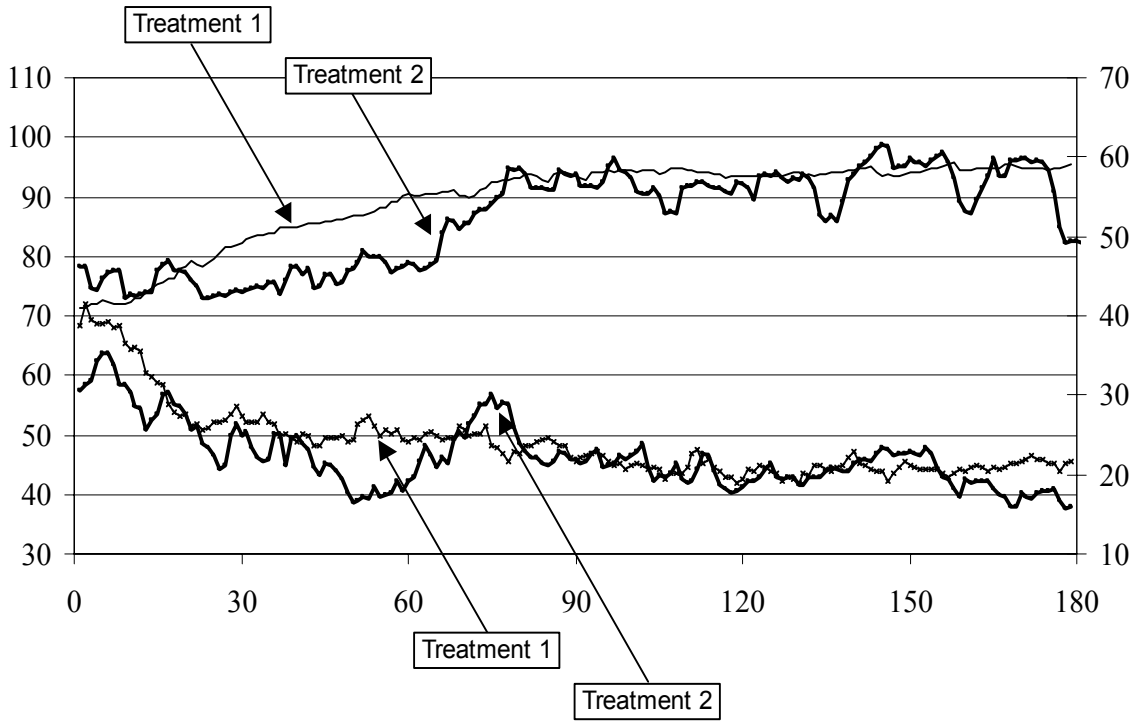


Figure 6. *Distribution of Tremble Probabilities.*

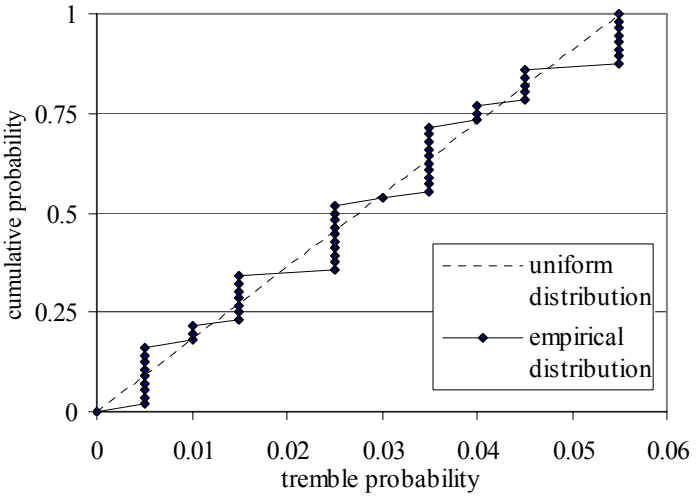


Figure 7. *Median Play of Subjects in Treatment 1 by Level of Aspiration Tremble Probability.*

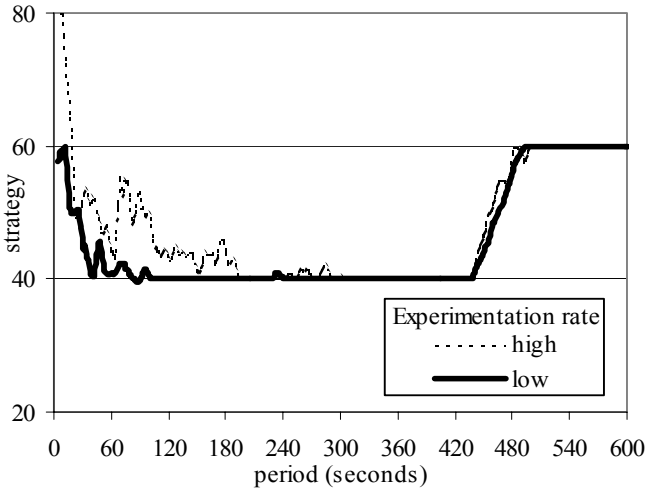


Table 1. *Treatments and changes in underlying payoff functions.*

Time	Treatment 1	Treatment 2
0	$\Pi_t = 3q_t - \frac{3}{80}q_t^2 \qquad q^* = 40, \Pi^* = 60$	$\Pi_t = 3q_t - \frac{3}{80}q_t^2 \qquad q^* = 40, \Pi^* = 60$
5		$\Pi_t = \begin{cases} 3q_t - \frac{3}{80}q_t^2 & q_t < 45 \quad q^* = 60, \\ \frac{8}{3}q_t - \frac{1}{45}q_t^2 & q_t \geq 45 \quad \Pi^* = 80 \end{cases}$
7	$\Pi_t = \frac{10}{3}q_t - \frac{1}{36}q_t^2 \qquad q^* = 60, \Pi^* = 100$	
10		

Both treatments began with the same payoff functions. At seven minutes for Treatment 1 and five minutes for Treatment 2, payoffs changed. Starred variables represent equilibrium strategies and payoffs.

Table 2. *Incorporation of heuristics in learning models.*

Model	History	Reference Points	Experimentation	
			Period or Stage	Directed or Undirected
Roth-Erev Reinforcement Learning Basic Forgetfulness Experimentation Full model	Yes	No	Period	Undirected
R-E Reinforcement with Reference Points Fixed reference Evolving reference	Yes	 Fixed Evolving	Period	Recency
Two-stage World Resetting	Resettable	No	Stage	Recency, Full sampling
Responsive Learning Automata	No	No	Period	Undirected
Aspirations Models Satisficing Evolving Aspirations	No	Evolving	Stage	Full sampling

Table 3. *Mean squared deviations (MSD) for best fitting parameters fit to treatment 1.*

Model	Params	A Fit to Individual Data	B Fit to Aggregate Data
Roth-Erev Reinforcement Learning			
Basic	ρ_0	0.817	0.852
Forgetfulness	ρ_0, γ	0.676	0.849
Experimentation	ρ_0, ε	0.725	0.838
Full model	ρ_0, γ,ε	0.620	0.836
R-E Reinforcement with Reference Points			
Fixed reference	ρ_0, α	0.693	0.824
Evolving reference	ρ_0, α_0, γ	0.695	0.794
Two-stage World Resetting	α_0	0.624	0.834
Responsive Learning Automata	ε,β	0.615	0.824
Aspirations Models			
Satisficing	γ,β, ε	0.554	0.757
Evolving Aspirations	γ,β, ε	0.575	0.795
Benchmark Models			
Random Choice		0.900	0.900
Equilibrium		0.970	0.970

Models were fit to each individual separately (and MSDs averaged across subjects) as well as to the entire data set, assuming a common set of parameters for all subjects. MSD scores shown are for the parameters that minimized MSD in each case. Lower MSD implies better fit.

Table 4. *Comparison of model fits to individual data.*

#	Model	1	2	3	4	5	6	7	8	9	10	11	12	best
1	Satisficing	—	.64	.75	.75	.77	.79	.82	.80	.82	.86	1.0	1.0	9
2	Evolving Aspirations	.39	—	.68	.77	.68	.82	.79	.86	.89	.93	.98	1.0	8
3	Responsive LA	.25	.32	—	.46	.59	.80	.80	.82	.96	.98	.84	1.0	6
4	Two-stage World Resetting	.25	.23	.54	—	.50	.75	.77	.80	.89	.95	.91	1.0	6
5	R-E Full model	.23	.32	.41	.50	—	.71	.82	.80	.93	1.0	.91	1.0	5
6	R-E Forgetfulness	.21	.18	.20	.25	.29	—	.50	.68	.77	.89	.79	1.0	3
7	R-E Evolving reference	.18	.21	.20	.23	.18	.50	—	.54	.79	1.0	.75	1.0	3
8	R-E Fixed reference	.20	.14	.18	.20	.20	.32	.46	—	.64	.91	.80	1.0	2
9	R-E Experimentation	.18	.11	.04	.11	.14	.23	.21	.36	—	1.0	.73	1.0	1
10	R-E Basic	.14	.07	.02	.05	.07	.11	.00	.09	.14	—	.68	1.0	0
11	Equilibrium	.00	.02	.16	.09	.09	.21	.25	.20	.27	.32	—	.41	0
12	Random Choice	.07	.04	.00	.00	.00	.00	.00	.00	.00	.00	.59	—	0

Each number reflects the proportion of subjects for whom the model in the row predicted at least as well as the model in the column. The final column, labeled "best" reflects the number of other models the model in the row "beat," in the sense of predicting better for a majority of individuals.

Table 5. *Decomposed mean squared deviations (MSD) for parameters best fitting treatment 1 data.*

Model	Params	Pre change	Post change
Roth-Erev Reinforcement Learning			
Basic	ρ_0	0.829	0.906
Forgetfulness	ρ_0, γ	0.840	0.871
Experimentation	ρ_0, ε	0.818	0.883
Full model	ρ_0, γ,ε	0.826	0.859
R-E Reinforcement with Reference Points			
Fixed reference	ρ_0, α	0.759	0.976
Evolving reference	ρ_0, α_0, γ	0.725	0.955
Two-stage World Resetting	α_0	0.822	0.861
Responsive Learning Automata	ε,β	0.802	0.876
Aspirations Models			
Satisficing	γ,β, ε	0.705	0.879
Evolving Aspirations	γ,β, ε	0.776	0.838
Benchmark Models			
Random Choice		0.900	0.900
Equilibrium		0.938	1.044

MSD is reported for the first seven minutes (pre change) and last three minutes (post change).

Table 6. *Decomposed mean squared deviations (MSD) for prediction of treatment 2 data.*

Model	Params	Overall	Pre change	Post change
Roth-Erev Reinforcement Learning				
Basic	ρ_0	0.879	0.841	0.918
Forgetfulness	ρ_0, γ	0.876	0.849	0.904
Experimentation	ρ_0, ε	0.878	0.846	0.911
Full model	ρ_0, γ,ε	0.861	0.851	0.871
R-E Reinforcement with Reference Points				
Fixed reference	ρ_0, α	0.900	0.804	0.996
Evolving reference	ρ_0, α_0, γ	0.914	0.805	1.024
Two-stage World Resetting	α_0	0.863	0.861	0.865
Responsive Learning Automata	ε,β	0.880	0.861	0.899
Aspirations Models				
Satisficing	γ,β, ε	0.816	0.815	0.818
Evolving Aspirations	γ,β, ε	0.829	0.828	0.829
Benchmark Models				
Random Choice		0.900	0.900	0.900
Equilibrium		1.388	1.441	1.336

MSD is reported for the first five minutes(pre change) and last five minutes (post change).