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

Mikhael Shor[†]
Owen Graduate School of Management
Vanderbilt University
mike.shor@owen.vanderbilt.edu
Phone: 615-292-4355
Fax: 615-292-7177

Revised: June 2004

[*] I would like to thank Ido Erev, Eric Friedman, Stefan Napel, seminar participants at the 2001 Decentralization conference, and at Vanderbilt and Rutgers universities, for useful comments. 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.

[†] Correspondence to 352 Management Hall, 401 21^{st} Avenue South, Nashville, Tennessee 37203

Learning to Respond:
The Use of Heuristics in Dynamic Games

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. This paper evaluates several 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 dynamic 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. Implications for the design of learning models for dynamic, low-information settings such as the Internet are discussed.

JEL Classification Numbers: D83, C91, C73

Keywords: learning, limited information, responsiveness

1 Introduction

Analysis of human behavior demonstrates that people are remarkably responsive to changes in their environment on time scales ranging from millennia (evolution) to milliseconds (reflex). Psychologists have observed that people react quickly in dynamic settings (Payne, Bettman, and Johnson 1993, Schunn and Reder 1998). 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 examined adaptive group behavior, an inquiry applied to organizational decision making in business settings (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.

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, (e.g., Mookherjee and Sopher 1994, Roth and Erev 1995, Van Huyck, Battalio, and Rankin 1996, Erev and Roth 1998), 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. The goal of this paper is to analyze the suitability of some common learning models in low-information dynamic games. A further objective is to determine which heuristics of the various models help capture responsiveness.

Heuristics are "rules of thumb" for deciding among competing alternatives. Incorporating only general principles of behavior, heuristics are tactics for approaching a problem, not fully represented strategies (for a review, see Pearl 1984). While heuristic-based approaches are likely to select relatively better actions among those available, they do not do so necessarily in an optimal fashion. As most can attest from personal experience, people are subject to the same fault. Rather than conduct a horse race among representative learning models, we classify the models according to the heuristics they embody. Specifically, we devise a taxonomy along three dimensions: (i) *history* denotes how much weight individuals place on the fact that an action has previously performed well, (ii) *reference points* provide a performance metric by which to evaluate received payoffs, and (iii) *experimentation* which highlights the tradeoff between acquiring information about one's environment and taking advantage of the information already acquired.

The next section presents an experiment which isolates responsiveness from other behaviors by instituting a 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. However, the learning models we consider, including reinforcement learning, evolving aspirations, satisficing, and responsive automata, differ in how well they capture this adaptability. Models that slow the rate of learning over time are perhaps not suited for non-static environments. The models that do well in both describing and predicting play differ in their quantitative approach but share similar heuristics.

2 Experiments

The data are a subset (the first ten minutes) of those from an experiment by Friedman, Shor, Shenker, and Sopher (2004) on dynamic decision making in low-information environments. Design features of the experiment included very limited information, a dynamic, 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. 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. Also the short length avoids boredom, which may lead to excessive experimentation.[1]

The underlying game was a simple quantity-setting monopoly game with linear demand. In each period, t, a subject selected an action, $q_t \in \{0, 1, \ldots, 100\}$, by moving a slider provided on screen. The resulting payoff was given by $\Pi_t = aq_t - bq_t^2$. Two different treatments were run. In both treatments the game began with the same values of a and b. At seven minutes for the first treatment and five minutes for the second treatment the payoff function changed (Table 1). The two treatments differ in 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 the optimal 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 the second. Subjects were markedly responsive in both treatments (Figure 1). In Treatment 1, the median player recognized the change in the payoff function and learned the new equilibrium within 100 seconds. Play in Treatment 2 was similar except for a slightly longer lag after the payoff change likely caused by one's inability to recognize a change without experimentation.

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 – or initially converging on – 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. (2004) termed a common pattern of experimentation "arrhythmic heartbeat

[1] 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. 2004). Continued participation beyond the game reported here avoids endgame effects.

Table 1: Experimental treatments and changes in underlying payoff functions.
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.

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

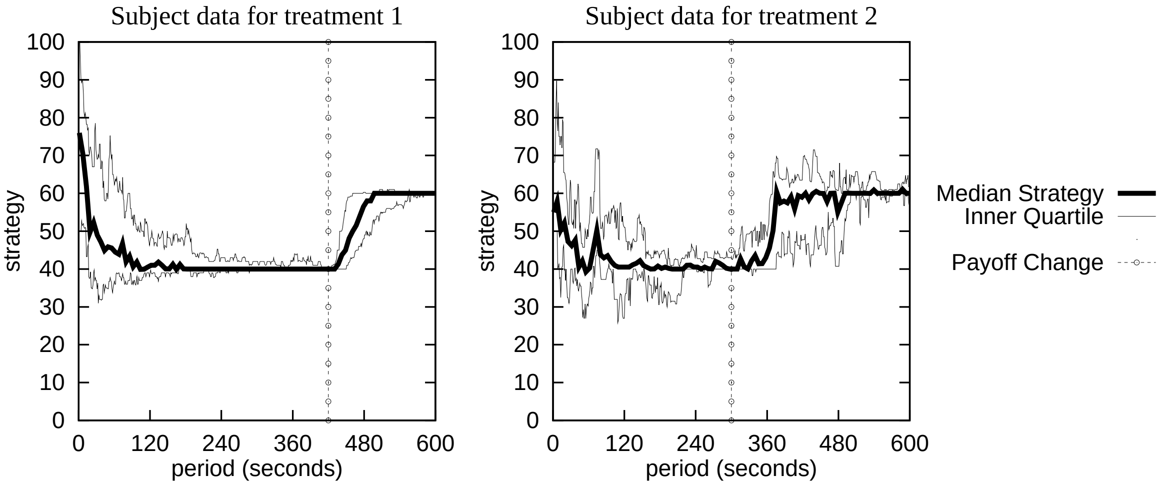


Figure 1: Subject play in experimental treatments

Table 2: Incorporation of heuristics in learning models.

Model	History	Reference Points	Experimentation
Reinforcement Learning (RL)	Yes	No	Undirected
Basic			
Forgetfulness			
Experimentation			
Full model			
RL with Reference Points	Yes		Recency
Fixed reference		Fixed	
Evolving reference		Evolving	
World Resetting	Resettable	No	Recency Sampling
Responsive Learning Automata	No	No	Undirected
Aspirations Models	No	Evolving	Sampling
Satisficing			
Evolving Aspirations			

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

Given the low-information design of the experiment and the nature of the 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 be applicable, and learning should occur through some adaptive, or backward-looking mechanism. Further, in light of experimental support for such "myopic" learning, 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 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, learning models generally incorporate Thorndike's classic law of effect (Thorndike 1898, Broadbent 1961), capturing the notion that people learn from the rewards, or payoffs, received and attributed to a particular action. An action which performs well, or results in high payoffs, is more likely to be used in the future, while an action which performs poorly will be used with less frequency. 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.

3.1 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. The implication for responsiveness is that 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 the most simple formulation of Roth and Erev's (1995) 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. Strategies are played with probabilities proportional to their propensities. Suppose that a player has two strategies, A and B, and that payoffs fall in the range $[0, 1]$. After a few periods, if the propensity of strategy A is 4 and of 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.

History dependence implies that the transition of the probability distribution over actions from the last period to the current one depends not only on the last payoff received but on time-dependent parameters. Conversely, 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. In general, a memoryless formulation may be expressed as $p_t = f(p_{t-1}, a_t, \pi_t; \theta)$, with p, a,and π representing the probability distribution, action taken, and payoff received, and θ a set of parameters. This implies a Markovian property with only last period's play entering explicitly. If a model depends on time explicitly (for example, incorporating a learning term which diminishes over time), or implicitly, by having probabilities reflect the whole of past experience, then such models are history dependent. Roughly, a history-dependent model implies that behavior is updated by ever smaller increments as time passes. Reinforcement learning models are history dependent due to the construction of propensities. 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.

3.2 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. Entry-level 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 people may have little information about possible payoffs, Simon suggests that people may develop aspirations by which they evaluate a strategy based upon whether it yields payoffs higher or lower than the satisficing level.[2] Satisficing

[2] For a discussion of satisficing, see Gigerenzer and Todd (1999).

was formalized into a model of learning by Karandikar, Mookherjee, Ray, and Vega-Redondo (1998) who suppose that an action which yields a payoff above a player's aspirations will continue to be played. The likelihood of repeating a certain action decreases if its payoffs fall short of one's aspiration.

Reference points are incorporated more broadly than satisficing models. Roth and Erev allow for reference points to scale payoffs downward so that *relatively* low payoffs may serve as negative reinforcements and also consider variable reference points, which evolve as the game progresses. Similarly Karandikar, et al. (1998) allow aspiration levels to evolve in the direction of realized payoffs. This captures the idea first put forward by Tinklepaugh (1928, in an experimental study of monkeys), that "individuals" not only learn about the payoff implications of various actions as the game progresses but also learn what defines a "good" payoff.

3.3 Experimentation

At the heart of learning lies a struggles between using strategies currently believed to be optimal and intentionally playing sub-optimally to gain valuable information about one's environment. This exploration-versus-exploitation dilemma appears in any dynamic environment from control theory (Thrun 1992b) to organizational learning (March 1991). While all learning models considered incorporate experimentation in some form (since "trial and error" is a fundamental learning method), they differ in how it is modeled.

Computer science and artificial intelligence researchers draw a distinction between undirected and directed experimentation. Undirected experimentation involves superimposing random strategy selection on the learning process. Such selection may be from a uniform distribution, implying an equal chance of trying any strategy, or may be utility-driven, selecting strategies proportional to their expected utilities. Conversely, directed exploration (for a survey, see Thrun 1992a,b) implies using strategies that contribute most to the estimates of the underlying payoff function. Directed exploration incorporates purposeful experimentation in order to gain particular knowledge about the environment. 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 sampling.[3] Recency-based exploration (Sutton 1990) assumes that knowledge about the world decays, or decreases in informational value with 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.[4] Sampling implies that decision makers initiate an experimentation phase in which they explore enough of the action space to gain good estimates of the payoff function. The estimates are then used to exploit the environment, or play a "best" strategy, until commencing another period of experimentation. This is reflected in the analysis of Friedman, et al. (2004) who find that many subjects embarked on periods of experimentation in which a broad portion of the strategy space was sampled.

Though no model in economics appears to incorporate directed experimentation explicitly, some learning models contain the flavor of these exploration techniques through appropriate parameterization. For example, consider reinforcement learning models with reference points in which the propensity to play a strategy

[3] 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.

[4] Barto, Sutton, and Watkins (1989) and Watkins (1989) propose methods for estimating the "exploration bonus" of a strategy based on recency. 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. More recently used strategies possess less informative value in exploration.

increases if its payoff exceeds the reference point and decreases if it falls short. High initial propensities and high initial reference points can lead to recency exploration. Initially as actions are tried, corresponding propensities decrease making unexplored strategies more attractive. 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 utility maximization may be achieved through occasional upward shocks to one's aspiration level. "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 sampling.

4 Models

4.1 Reinforcement Learning

The initial formulation of Roth and Erev (1995) provided a simple learning model incorporating both the law of effect and power law of practice. Proposed variants of the model (Roth and Erev 1995, Erev and Roth 1998) additionally incorporate 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

$$\begin{aligned} \rho_t(i) &= (1-\gamma)\rho_{t-1}(i) + (1-\varepsilon)(\pi_t(i) - \alpha_{t-1}) \\ \rho_t(j) &= (1-\gamma)\rho_{t-1}(j) + (\varepsilon/S)(\pi_t(i) - \alpha_{t-1}), \quad j \neq i \end{aligned}$$

where S is the number of pure strategies, γ is a forgetfulness or recency parameter, ε is the experimentation probability, and α_t is a reference point.[5] The reference points, α_t, evolve according to the following rule:

$$\alpha_t = \lambda\alpha_{t-1} + (1-\lambda)\pi_t(i) \qquad (1)$$

where λ is a persistence parameter. When $\lambda = 1$, reference points remain constant, while $\lambda = 0$ implies that next period's point of reference is simply this period's payoff. A strategy is played with probability proportional to its propensity:

$$p_t(i) = \rho_t(i) / \sum\nolimits_j \rho_t(j)$$

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

[5] Roth and Erev conjecture that payoffs might only need be generalized to the nearest strategies, such that $\rho_t(j) = (1-\gamma)\rho_{t-1}(j) + (\varepsilon/2)(\pi_t(i) - \alpha_{t-1}), j = i \pm 1$. In preliminary simulations, this formulation had little effect on the models' predictions.

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	$\rho_0, \gamma, \varepsilon$	$\alpha_t = \lambda = 0$
Fixed Reference	ρ_0, α	$\varepsilon = \gamma = 0, \alpha_t = \alpha$
Evolving Reference	$\rho_0, \alpha_0, \lambda$	$\varepsilon = \gamma = 0$

4.2 World Resetting

The world resetting model extends the basic reinforcement learning procedure by discarding history when past performance no longer appears relevant. Its construction is motivated by subjects informing the experimentalist that upon recognizing a change in the environment, they "reset" their learning, or ignore historical payoffs. A hypothesized subject maintains in memory not only propensities for each strategy but also estimates of the payoff function. When payoffs begin to differ substantially from these estimates, the model resets propensities to their initial value, ρ_0. While a number of criteria exist for such resetting (e.g., Vulkan and Preist 2003), the experimental setting does not provide the ability to discrimination between them. 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.

In general, such a test of model fitness consists of three components. First is a mechanism for forming payoff expectations. For example, the expected payoff from a given strategy may be simply a weighted average of all payoffs received from that strategy. Second is some measure by which a subject notes in each period the distance between the realized and predicted payoffs. A world is deemed *understood* if the distance is sufficiently small for a reasonable period of time. Third is a criteria for determining when one's estimates fail to provide an accurate description of payoffs. This triggers a "resetting" in which the learning model's parameters revert to initial values as if the person begins learning a different task.

Without any noise in the payoffs the first two components are trivial. In our experimental setting, any weighted average of a strategy's past payoffs would yield the same result since the payoffs are deterministic. Similarly, any reasonable measure of distance between expected and realized payoffs would equal 0 except following the singular change in the payoff function. The only remaining issue is how many periods after the change in payoffs does a subject require before re-initializing learning. In the simulations reported, values of between 1 and 30 periods yielded almost identical results in terms of model fitting and estimation. The results presented in the next section are for five periods.

The world resetting model is the basic reinforcement learning model described above but without the need to amass history indefinitely. While the model has a single parameter for our purposes, ρ_0, it is not the intention to suggest that the model is "simple." For most purposeful applications, a criterion for environmental stability requires a number of parameters on top of the parameters inherent in the learning process itself. However, for the purposes of determining the role of history or memory in learning in dynamic settings, the world resetting model provides a useful benchmark for the reinforcement learning models.

4.3 Responsive Learning Automata

Learning automata were originally simple, one-period memory systems modeled after biological processes, and designed for solving control problems (for a survey, see Narendra and Thatcher 1989). Unlike reinforcement learning in which a dependence on history exists in the updating of propensities, the responsive learning automata (Friedman and Shhenker 1996) maintains only a probability distribution over strategies. If in period t the player uses strategy i and receives payoff $\pi_t(i)$, probability updating is governed by

$$\begin{aligned} p_{t+1}(i) &= p_t(i) + \varepsilon\beta\pi_t(i)\sum_{j\neq i}\omega_t(j)p_t(j) \\ p_{t+1}(j) &= p_t(j) - \varepsilon\beta\pi_t(i)\omega_t(j)p_t(j), \quad j \neq i \end{aligned}$$

where

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

and β captures the speed of learning. Again, ε captures the probability of experimentation and S is the number of pure strategies. The probability of playing the same strategy as in the previous period increases with the payoff received while all other strategies decrease in probability proportionally. However, no strategy may drop below some threshold, ε/S, guaranteeing the persistence of experimentation.

4.4 Aspirations

Aspiration models incorporate the no-memory property of responsive learning automata with a reference-point-based behavioral assumption. In light of little information about the game and its attainable payoffs, people may develop aspirations. A strategy is played more often if the resulting payoff exceeds the 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 players repeat a strategy as long as payoffs exceed aspirations. If payoffs fall short of aspirations, the likelihood of repeating the strategy decreases in proportion to the magnitude of the disappointment – the difference between the aspiration and received payoff. Further, aspirations are subject to occasional trembles which are the source of experimentation in the model.

With probability $(1-\varepsilon)$ aspirations evolve according to (1), equivalent to the updating of reference points in reinforcement learning. In each period, aspirations may also "tremble" with probability ε, in which case they are randomly drawn from some distribution over the payoff space. In what follows, the distribution is assumed to be uniform over $[0, 100]$. When a realized payoff falls short of the aspiration level, probability updating is governed by (if strategy i played at time t):

$$\text{If } \alpha_t > \pi_t, \quad \begin{aligned} p_{t+1}(i) &= \tfrac{p_t(i)}{[1+\beta(\alpha_t-\pi_t)]} \\ p_{t+1}(j) &= \tfrac{p_t(j)}{1-p_t(i)}\left(1 - \tfrac{p_t(i)}{1+\beta(\alpha_t-\pi_t)}\right), \quad j \neq i \end{aligned} \tag{2}$$

When the learner is satisfied, having received a payoff exceeding her 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, one repeats the previous action:

$$\text{If } \pi_t \geq \alpha_t, \quad p_{t+1}(i) = 1, \quad p_{t+1}(j) = 0, \quad j \neq i \tag{3}$$

Since the decision maker does not care about the magnitude of payoffs when payoffs exceed aspirations, the trembling aspirations model given by (2) and (3) will be referred to as the *satisficing* model hereafter. Borgers and Sarin (2000) 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:

$$\text{If } \pi_t \geq \alpha_t, \quad \begin{array}{rcl} p_{t+1}(i) & = & \frac{p_t(i)+\beta(\pi_t-\alpha_t)}{1+\beta(\pi_t-\alpha_t)} \\ p_{t+1}(j) & = & \frac{p_t(j)}{1+\beta(\pi_t-\alpha_t)}, \quad j \neq i \end{array} \tag{4}$$

Decision makers do not satisfice since they react to ever greater payoffs. We term the formulation consisting of (2) and (4) *evolving aspirations* hereafter. Similar in spirit to reinforcement learning with reference points, aspirations determine whether a received payoff drives a strategy's probability to be revised upward or downward. However, the model is substantially distinct from reinforcement learning models since it has no memory and with the inclusion of trembles in aspirations, it allows for sampling experimentation.

5 Model Performance

The experimental design accommodated 101 possible actions in each period. Since the payoff function is continuous in strategies, subjects learn to generalize, associating realized payoffs with strategies close to the one actually used. It is not apparent, however, how this generalization occurs.[6] The models considered do not generalize instead treating each strategy as entirely distinct. Accordingly, the game is reduced to ten strategies, $\{10, 20, \ldots, 100\}$ for the purpose of fitting and simulating the models and the experimental data are aggregated by mapping players' strategies into the next highest among the ten available.[7] The best parameters for each model are obtained using the mean squared deviation criterion (MSD, Simon 1956). Selten (1998) provides an axiomatic explanation of MSD and a discussion of its desirable properties.[8] Simulations were run using the updating rules prescribed by each model and the payoff functions from Treatment 1 of the experiment. The simulated paths of play were compared to the actual play of each subject. Parameters were chosen using an iterative procedure. Initially, a broad grid was chosen to permit

[6] For a description 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.

[7] 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. The reduction to strategies which are multiples of ten reflects people's affinity for round numbers (e.g., Higgins, Rholes, and Jones 1977), a schema documented in retail (Vanhuele and Drèze 2002) and financial (Hasbrouck 1999) settings.

[8] Denote the probability distribution over actions by a vector p. If strategy i was used, MSD $= (1-p_i)^2 + \sum_{j\neq i} p_j^2$. A lower MSD implies a better fit. The percentage-of-inaccuracies criterion (Erev and Roth 1998) yielded similar qualitative comparisons across models. Unfortunately, obtaining best-fitting parameters using maximum likelihood is computationally infeasible. For example, consider the evolving aspirations model in which aspirations "tremble" with some probability in every period. The likelihood of a given path of play depends on when these trembles occurred. With 600 periods, one must average over 2^{600} possible combinations of periods in which aspirations trembled, a value exceeding the number of atoms in the known universe.

250 sets of parameter values. For logically-constrained parameters – 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. The inner quartile of the parameter values which minimized subjects' MSD was used for the new bounds on the grid in the next iteration. A total of eight iterations were used for each model, for a total of 2000 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 20,000 simulations were used per subject (1,120,000 total) for identifying the best individual-level parameters. MSD scores are also obtained for two benchmark models: the equilibrium prediction and a random choice model, which assumes that players select each action with equal probability in every period.

Each model was fit both to play of individual subjects as well as to aggregated data. While individually fitting a model's parameters to each subject perhaps holds 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.[9] Recognizing if learning models perform well on an individual level allows us to determine if the models represents heuristics common to most subjects, even if the exact parameter values representing that heuristic differ from person to 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.

5.1 Individual Fit

All of the models perform substantially better than either the equilibrium or random choice benchmark (Table 3). Aspirations-based models appear to do quite well followed by responsive learning automata and two variants of reinforcement learning: the full model and the world resetting learner. To compare models' performance, we inquire about the proportion of subjects for whom one model performs better than another (Table 4). Each entry in Table 4 represents the fraction of subjects for whom the model in the row results in a lower MSD than the model in the column. All learning models describe play at least as well as the random choice benchmark for every subject and most models substantially outperform the equilibrium prediction. Even one-parameter models outperform the equilibrium for over 2/3 of the subjects.

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, each of the top four models outperforms each of the lower six; the lower six all incorporate propensities while none of the top four feature history dependence. This suggests that history dependence is not an applicable heuristic for some real-time changing environments. The ordering of the models is strict – each listed model predicts better for a majority of subjects than every model below it (above the diagonal, all entries are greater than 0.5). In fact, satisficing results in a lower MSD for at least three-fourths of all subjects, when compared to any non-aspirations-based

[9] 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.

Table 3: Performance of models in describing individual path of play in Treatment 1.

Model	Params	MSD
Reinforcement Learning (RL)		
Basic	ρ_0	0.817
Forgetfulness	ρ_0, γ	0.675
Experimentation	ρ_0, ε	0.725
Full model	$\rho_0, \varepsilon, \gamma$	0.612
RL with Reference Points		
Fixed reference	ρ_0, α_0	0.692
Evolving reference	$\rho_0, \alpha_0, \lambda$	0.672
World Resetting	ρ_0	0.618
Responsive Learning Automata	ε, β	0.606
Aspirations Models		
Satisficing	$\lambda, \beta, \varepsilon$	0.545
Evolving Aspirations	$\lambda, \beta, \varepsilon$	0.567
Benchmark Models		
Random Choice		0.900
Equilibrium		1.026

Table 4: Relative performance of model fits to individual data.

#	Model	1	2	3	4	5	6	7	8	9	10	R	E	best
1	Satisficing	–	.61	.75	.75	.75	.80	.79	.80	.82	.86	.95	1.0	9
2	Evolving Aspirations	.36	–	.77	.66	.64	.75	.80	.84	.89	.93	.96	.98	8
3	World Resetting	.25	.23	–	.54	.54	.71	.77	.80	.91	.96	1.0	.91	7
4	Responsive LA	.25	.34	.46	–	.59	.75	.89	.88	.98	.98	1.0	.86	6
5	RL Full model	.25	.36	.46	.41	–	.71	.73	.84	.95	.98	1.0	.91	5
6	RL Evolving reference	.20	.25	.29	.25	.29	–	.57	.54	.86	1.0	1.0	.80	4
7	RL Forgetfulness	.21	.20	.23	.11	.00	.43	–	.63	.77	.89	1.0	.79	3
8	RL Fixed reference	.20	.16	.20	.13	.14	.00	.29	–	.64	.91	1.0	.80	2
9	RL Experimentation	.18	.11	.09	.02	.00	.14	.14	.29	–	.86	1.0	.73	1
10	RL Basic	.14	.07	.04	.02	.00	.00	.00	.00	.00	–	1.0	.68	0
R	Random Choice	.00	.00	.00	.00	.00	.00	.00	.00	.00	.00	–	.59	0
E	Equilibrium	.00	.02	.09	.14	.09	.20	.21	.20	.27	.32	.41	–	0

Each cell entry represents the proportion of subjects for whom the model in the row predicted strictly better (lower MSD) than the model in the column. Models 1 and 2 did at least as well as Random Choice for every subject, but *strictly* better for 95% and 96% of subjects, respectively. 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: Performance of models in describing aggregate path of play in Treatment 1.

Model	Params	Mean Squared Deviation Overall	 Pre change	 Post change
Reinforcement Learning				
Basic	ρ_0	0.852	0.828	0.909
Forgetfulness	ρ_0, γ	0.849	0.840	0.871
Experimentation	ρ_0, ε	0.838	0.818	0.883
Full model	$\rho_0, \varepsilon, \gamma$	0.836	0.828	0.854
RL with Reference Points				
Fixed reference	ρ_0, α_0	0.824	0.759	0.976
Evolving reference	ρ_0, α_0, γ	0.794	0.725	0.955
World Resetting	ρ_0	0.829	0.817	0.856
Responsive Learning Automata	ε, β	0.820	0.800	0.868
Aspirations Models				
Satisficing	$\gamma, \beta, \varepsilon$	0.750	0.703	0.857
Evolving Aspirations	$\gamma, \beta, \varepsilon$	0.788	0.772	0.825
Benchmark Models				
Random Choice		0.900	0.900	0.900
Equilibrium		1.026	1.050	0.969

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

model. Binary Wilcoxon signed-rank matched-pairs tests allow us to reject the hypothesis (at 1%) that the MSD scores of the satisficing and evolving aspirations models are generated from the same population as any of the other models. In fact, of the 66 pairwise comparisons of the twelve models presented in the table, the Wilcoxan tests suggest that 62 are significantly different at 5% and 60 at 1%.

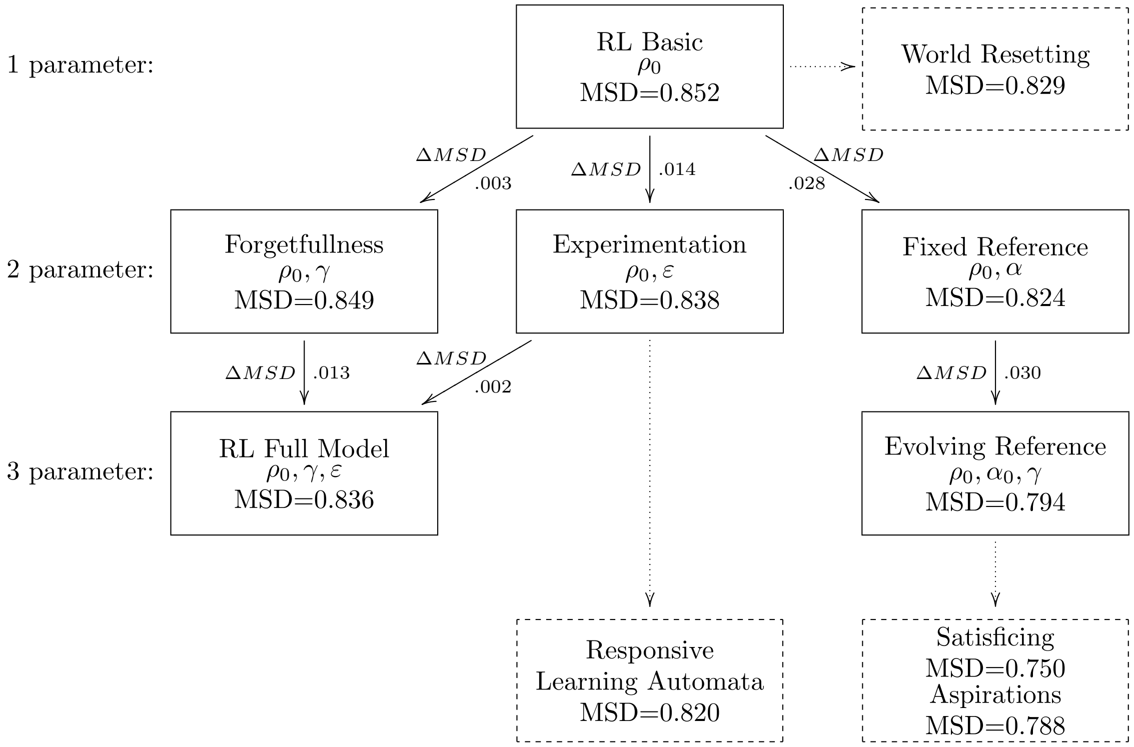
5.2 Aggregate Fit

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 provides a single set of parameters that may encapsulate general behavior. A model calibrated with a particular set of parameters may then be able to explain play in a variety of experiments (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. Nevertheless, each model surpasses both benchmarks and aspiration-based models continue to outperform (5).[10] Reinforcement learning models are the worst performing but benefit from the inclusion of reference points. Again the satisficing model performs best. It incorporates reference points, allows for experimentation in stages by occasional trembles in aspiration levels, and is not history dependent. Both individually fit and aggregate data suggests that history-dependent learning models do not explain the data well.

By comparing nested formulations of learning models, we assess the value of certain parameters and the behavioral traits they embody. Figure 2 exhibits the fit of reinforcement learning models and the contribu-

[10] That the same models seem to perform best when fit to both individual and aggregate data may suggest that people are rather idiosyncratic, but employ similar heuristics at different rates.

Figure 2: The impact of history on describing play. Comparison of reinforcement learning models and the contribution of certain parameters to the models' descriptive power. Dashed lines and boxes represent comparable memoryless models with the same number of parameters.



tion of additional parameters to the models' explanatory power. Beginning with the basic (one-parameter) model, solid arrows in Figure 2 show the improvement in mean squared deviation from the inclusion of additional parameters. The parameter γ, representing "forgetfulness," contributes little explanatory power. Its addition to the basic model decreases MSD by about 0.3% and its addition 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. A fixed reference point is the parameter which contributes the most explanatory power to the basic model, 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 reference points to evolve. Finally, incorporating disappointment-based satisficing behavior causes the largest change in MSD, from 0.794 to 0.750.

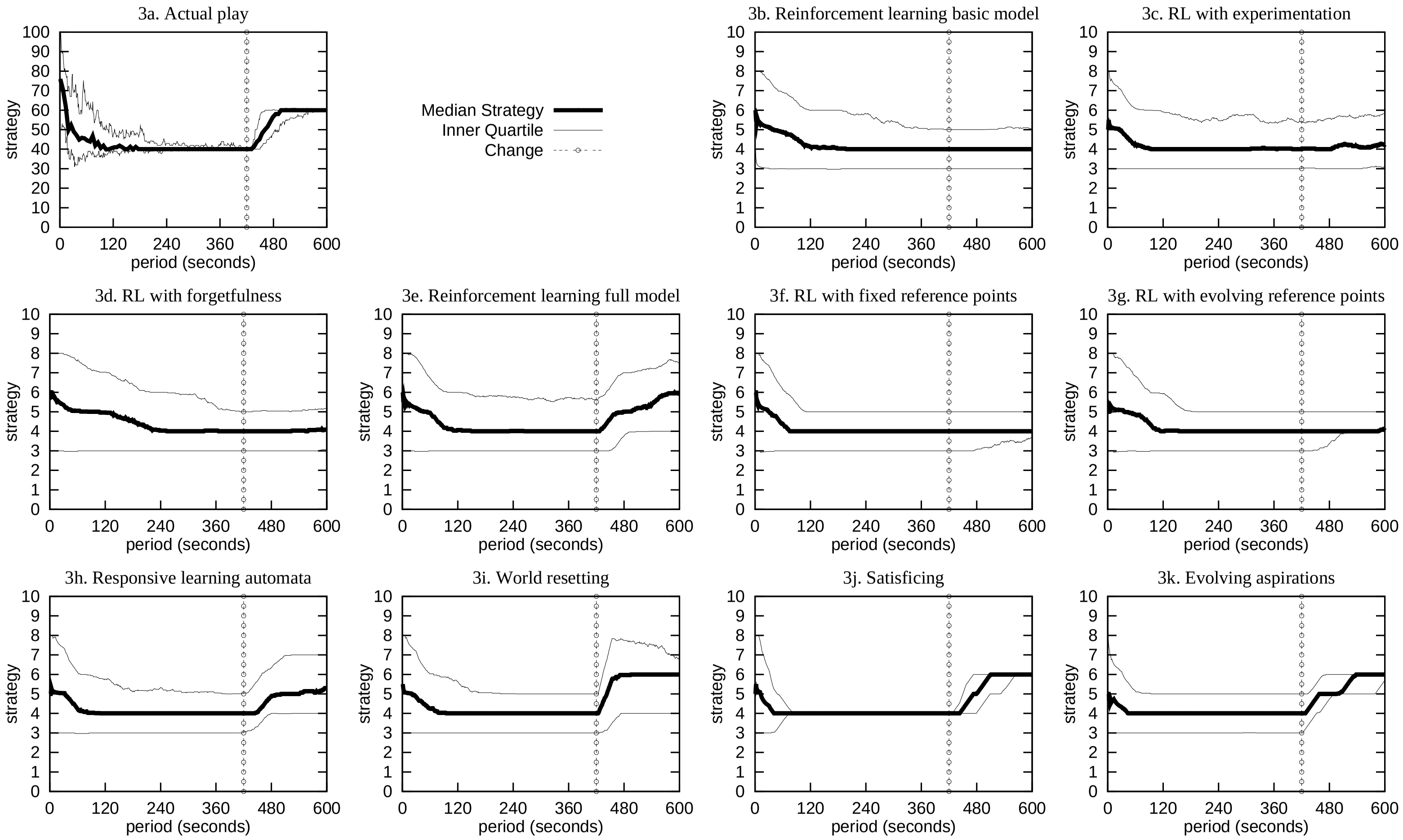
We can compare the history-dependent models with models of equal complexity but without memory or history-dependence. The dotted lines in Figure 2 point to memoryless models which have the same number of parameters yet all explain the data better. For example, the two-parameter reinforcement learning 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 one representing the speed of learning. Despite an equal number of parameters, the responsive learning automata model generates a superior fit.

Numerical comparisons of models may not provide an accurate picture of how well the models describe behavior and responsiveness qualitatively. The path of play predicted by each learning model was simulated 5,000 times using the parameters that minimized MSD (Figure 3). Numerically, both reinforcement learning 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 is comparable to the responsiveness observed in 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, models with reference points quickly converge and exhibit tight bounds while the full reinforcement learning model is affected by constant experimentation. 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.

All memoryless models display both convergence to the equilibrium and responsiveness to the change in the payoff function (Figures 3h-3k). We examine the tradeoff between strong convergence to the optimal strategy and experimentation significant enough to recognize environmental changes by decomposing the MSD score into two time intervals (Table 5). A low MSD score before the change in payoffs indicates accurate tracking of subjects' convergence while a low MSD after the change reflects a good fit to subjects' responsiveness. The basic (one parameter) model of reinforcement learning is identical in performance to the full (three parameter) model prior to the change. Post-change, the two are quite distinct as the basic model fails to predict better than even the random choice benchmark. Similarly, the addition of reference points to reinforcement learning, whether fixed or evolving, leads to strong convergence initially, but the poorest responsiveness of any of the models considered.

The performance of the basic reinforcement learning model benefits most not from the inclusion of experimentation or forgetfulness but from a notion of world-resetting. If the model "throws out" built-up propensities when underlying payoffs change, its descriptive power improves considerably both post-change and even before the change in payoffs. The latter may sound counterintuitive; why should the 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? The value of initial propensities largely drive experimentation in the basic model and experimentation is central to recognizing environmental variations. If appreciable experimentation must persist even after a good amount of time has elapsed then the best-fitting estimated initial propensities will be large (ρ_0=310). The tradeoff is that large initial propensities will lead to slower convergence. Since the 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 world resetting model is a fair competitor. 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. Ironically, while the model was developed after the author was privy to the data and subjects' sentiments, it does not perform well when compared with aspiration-based learning models (Figures 3j-3k). Both aspiration-based models track the data well throughout but only the satisficing model mimics 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 equally important for many applications especially if the efficiency or total payoffs in a game fall off substantially as players move away from equilibrium. Both



3a. Actual play

3b. Reinforcement learning basic model

3c. RL with experimentation

3d. RL with forgetfulness

3e. Reinforcement learning full model

3f. RL with fixed reference points

3g. RL with evolving reference points

3h. Responsive learning automata

3i. World resetting

3j. Satisficing

3k. Evolving aspirations

Figure 3: Simulated play for Treatment 1

actual play and simulations with the satisficing model exhibit an inner quartile of play which is quite broad initially but quickly converges on the equilibrium strategy. Shortly after the change in payoffs, simulated and actual play once more become more volatile but again soon converge on the new equilibrium.

5.3 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 (or sometimes post hoc or cross-validation; 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 faced with 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.[11]

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. We adopt an out of sample prediction approach, comparing simulated play using the models fit to Treatment 1 to actual data from Treatment 2 which incorporates a different experimental design. The primary distinction is that a subject playing the optimal action in Treatment 1 will instantly note a change in payoffs 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, balancing the relative weighing of initial convergence and responsiveness in the MSD scores while in Treatment 1, seventy percent of play occurred before the change.

The measure of how well the models predicted play in Treatment 2 was again 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, 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, world resetting, evolving aspirations, and satisficing models increase less than one percent in mean squared deviation. This suggests that they predict the initial convergence of

[11] 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.

Table 6: Predictive power of learning models in describing path of play in Treatment 2.

Model	Params	Mean Squared Deviation Overall	Pre change	Post change
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	$\rho_0, \varepsilon, \gamma$	0.859	0.849	0.868
RL with Reference Points				
Fixed reference	ρ_0, α_0	0.900	0.804	0.996
Evolving reference	ρ_0, α_0, γ	0.914	0.805	1.024
World Resetting	ρ_0	0.859	0.857	0.860
Responsive Learning Automata	ε, β	0.873	0.849	0.896
Aspirations Models				
Satisficing	$\gamma, \beta, \varepsilon$	0.809	0.808	0.811
Evolving Aspirations	$\gamma, \beta, \varepsilon$	0.822	0.821	0.823
Benchmark Models				
Random Choice		0.900	0.900	0.900
Equilibrium		1.360	1.346	1.373

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

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. For the remaining models, we again simulate 5,000 paths of play to provide a graphical comparison (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. 4e) displays comparable convergent behavior.

The responsive learning automata model (Figure 4c) appears to require the longest time to react to the change in the payoff function. The fitted value of β, representing the speed of learning, is 0.0005, indicating slower adjustment times. After the change, play initially stabilizes on strategy 5 and only much later begins to move towards the post-change equilibrium strategy of 6. Similar stepwise movement towards the new equilibrium appears in the evolving aspirations model (Figure 4f) and, to a lesser degree, in the full reinforcement learning model (Figure 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.

Aspirations-based models continue to perform well 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.

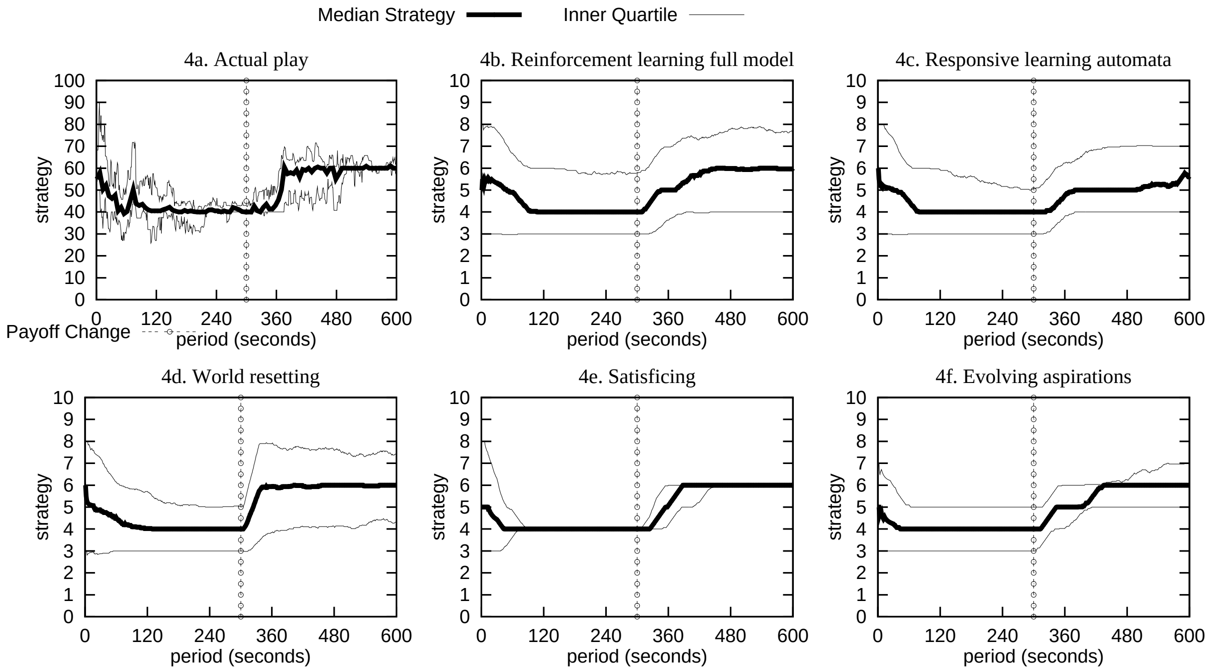


Figure 4: Simulated play for Treatment 2

6 Conclusion

History-based models of learning have performed quite well, explaining data from a variety of repeated games experiments (Roth and Erev 1995). However, they appear ill suited to real-time dynamic games. 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.[12]

The incorporation of reference points provides the greatest contribution to explanatory and predictive power of the models considered. However, fixed reference points may hinder responsiveness in dynamic environments. Fixed reference points do not permit 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

[12] 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.

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 neither independent from one period to the next nor as rare as often assumed in the literature was shown by Friedman, et al. (2004). 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 that some other strategies may outperform what currently seems best.

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 appear to be 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 and economics.

References

Audia, Pino G., Edwin A. Locke, and Kenneth G. Smith (2000): "The Paradox of Success: An Archival and Laboratory Study of Strategic Persitence Following a Radical Environmental Change," *Academy of Management Journal*, 43(5), 837–854.

Barto, Andrew G., Richard S. Sutton, and Chris J.C.H. Watkins (1989): "Learning and Sequential Decision Making," in *Learning and Computational Neuroscience: Foundations of Adaptive Networks*, ed. by M. Gabriel, and J.W. Moore, pp. 539–602. MIT Press.

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

Borgers, Tilman, and Rajiv Sarin (2000): "Naive Reinforcement Learning with Endogenous Aspirations," *International Economic Review*, 41(4), 921–950.

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

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.

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, and Scott Shhenker (1996): "Synchronous and Asynchronous Learning by Responsive Learning Automata," Mimeo, Cornell University School of Operations Research and Industrial Engineering.

Friedman, Eric, Mikhael Shor, Scott Shenker, and Barry Sopher (2004): "An Experiment on Learning with Limited Information: Nonconvergence, Experimentation Cascades, and the Advantage of Being Slow," *Games and Economic Behavior*, 47, 325–352.

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.

Hasbrouck, Joel (1999): "Security Bid/Ask Dynamics with Discreteness and Clustering," *Journal of Financial Markets*, 2(1), 1–28.

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

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

Higgins, E. Tory, William S. Rholes, and Carl R. Jones (1977): "Category Accessibility and Impression Formation," *Journal of Experimental Social Psychology*, 13, 141–154.

Kaelbling, Leslie P. (1993): "Learning in Embedded Systems," Ph.D. thesis, Massachusetts Institute of Technology.

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

Levitt, Barbara, and James G. March (1988): "Organizational Learning," in *Organizational Learning*, ed. by Michael D. Cohen, and Lee S. Sproull, pp. 516–540. Thousand Oaks, CA: Sage.

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

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

Mosier, Charles I. (1951): "Problems and Designs of Cross-Validation," *Educational and Psychological Measurement*, 11, 5–11.

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

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.* Reading, MA: 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 *The Psychology of Learning and Motivation*, ed. by D. L. Medin. New York: Academic Press.

Selten, Reinhard (1991): "Evolution, Learning and Economic Behavior," *Games and Economic Behavior*, 3, 3–24.

——— (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. (1955): "A Behavioral Model of Rational Choice," *Quarterly Journal of Economics*, 69, 99–118.

——— (1956): "Dynamic Programming under Uncertainly with a Quadratic Criterion Function," *Econometrica*, 24.

——— (1957): *Models of Man.* New York: Wiley.

Staddon, John E.R., and Alliston K. Reid (1990): "On the Dynamics of Generalization," *Psychological Review*, 97, 576–578.

Sutton, Rich S. (1990): "Integrated Architectures for Learning, Planning, and Reacting Based on Approximating Dynamic Programming," in *Proceedings of the Seventh International Conference on Machine Learning*, pp. 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): "Efficient Exploration in Reinforcement Learning," Discussion Paper CMU-CS-92-102, School of Computer Science, Carnegie Mellon University, Pittsburgh, Pennsylvania.

——— (1992b): "The Role of Exploration in Learning Control with Neural Networks," in *Handbook of Intelligent Control: Neural, Fuzzy and Adaptive Approaches*, ed. by D. A. White, and D. A. Sofge, Florence, Kentucky. Van Nostrand Reinhold.

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

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.

Vanhuele, Marc, and Xavier Drèze (2002): "Measuring the Price Knowledge Shoppers Bring to the Store," *Journal of Marketing*, 66(4), 72–85.

Vulkan, Nir, and Chris Preist (2003): "Automated Trading in Agents-based Markets for Communication Bandwidth," *International Journal of Electronic Commerce*, fortcoming.

Watkins, Chris J.C.H. (1989): "Learning from Delayed Rewards," Ph.D. thesis, University of Cambridge.

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