

SELECTION DYNAMICS AND ADAPTIVE BEHAVIOR WITHOUT MUCH INFORMATION

John B. Van Huyck, Raymond C. Battalio, Frederick W. Rankin

October 2001

Comments to john.vanhuyck@tamu.edu
Related research available at <http://erl.tamu.edu>



Abstract: This paper investigates whether behavior in a coordination game changes when subjects are limited to the information used by reinforcement learning algorithms. In the experiment subjects converge to an absorbing state at rates that are orders of magnitude faster than reinforcement learning algorithms, but slower than under complete information. Usually, this state is very close to a mutual best response outcome. All of the subjects are within a dime of giving a best response and 82.5 percent of the subjects gave a best response to the behavior of the other subjects in their cohorts without any information about their own best response function. The stability conditions derived from the best response dynamic are to conservative both under complete information and reinforcement information.

Key Words: Stability, Equilibrium Selection, Information, Reinforcement Learning, Adaptive Behavior.

JEL classification: c72, c92.

Acknowledgments: Rajiv Sarin found an error in our reinforcement learning simulations. Eric Battalio programmed the graphical user interface. The National Science Foundation and Texas Advanced Research Program provided financial support. Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not necessarily reflect the views of the National Science Foundation or the Texas Advanced Research Program.

© 1996-2001 by the authors. All rights reserved.

I. INTRODUCTION

Economists usually assume that decision makers know the consequences of their actions, form rational expectations, and possess internally consistent preferences. We do this in order to predict how people will behave in novel situations. When the situation is strategic, the knowledge assumptions needed to deduce a prediction are even stronger. Yet, people often have to make decisions without much information and at least as often ignore information that is readily available.

For example, it is difficult to communicate a common knowledge description of a game to undergraduate students. Moreover, we sometimes doubt that subjects in an experiment are actually using this information to deduce a mutually consistent way to behave. As Vernon Smith (1990, p12) wrote, “Many years of experimental research have made it plain that real people do not solve decision problems by thinking about them in the way we do as economic theorists. Only academics learn primarily by reading and thinking. Those who run the world, and support us financially, tend to learn by watching, listening, and doing. ... When experiments approximate the predictions of theory, it is because subjects experience the choices of others and then choose based on what they have learned to expect.”

Van Huyck, Cook, and Battalio (1994) report an experiment rejecting the stability conditions derived from the myopic best response dynamic and find that stability conditions based on relaxation algorithms with inertia make more accurate predictions. They were careful to communicate the best response function to their subjects, since this information seemed central to the theory being tested. Smith’s observation suggests that the subjects’ knowledge of the best response function doesn’t explain why the theory made accurate predictions. Our research hypothesis is that taking away information only used in a deductive analysis of the situation, like the best response function, will not influence behavior since subjects don’t use it anyway.

Reinforcement learning algorithms only require that players know their feasible actions and respond to the consequences of their actual choices. These algorithms can also be used as selection theories, although convergence to a mutually consistent outcome is not guaranteed in general. This paper reports an experiment in which subjects are limited to the information used by reinforcement learning algorithms and contrasts the results with Van Huyck, Cook, and Battalio (1994).

The reinforcement learning algorithms considered below do not accurately predict observed behavior under reinforcement information conditions. Humans converge to the interior equilibrium at speeds that are orders of magnitude faster than the models. Observed behavior under

complete information and reinforcement information treatments do differ in the length of time it takes to converge to a mutual best response outcome. However, all cohorts converged to the market statistic predicted by the interior equilibrium regardless of the information conditions or the stability conditions derived from the myopic best response dynamic. Average subject behavior did increase in exactly the way predicted by a conventional comparative static analysis.

The paper is organized as follows: Section II reviews Van Huyck, Cook, and Battalio's (1994) analytical framework and the Cross Dynamic studied in Borghers, Morales, and Sarin (2001); Section III reports the experimental design; Section IV reports the experimental results; Section V estimates an empirical model of satisfaction; Section VI compares the complete and reinforcement information treatments; Section VII concludes with a discussion of low cognitive game theory and the problem of modeling imagination in light of our results.

II. ANALYTICAL FRAMEWORK

The following analytical framework is from Van Huyck, Cook, and Battalio (1994). Let e^1, \dots, e^n , denote the actions taken by n players, where n is odd and greater than one. Let e denote this action combination, and let $M(e)$ denote the median of e . The game $\Gamma(\omega)$ is defined by the following payoff function and action space for each of these n players who are indexed by i :

$$\pi(e^i, e^{-i}) = 0.5 - |e^i - \omega M(e)(1 - M(e))|, \quad (1)$$

where $\omega \in (1, 4]$, $e^i \in \mathbf{E} = [0, 1]$, e^{-i} denotes $\{e^1, \dots, e^{i-1}, e^{i+1}, \dots, e^n\}$, and $|\cdot|$ is the absolute value function. Assume that the payoff functions and feasible actions are common knowledge.

In game $\Gamma(\omega)$ a player's best response to a given median M is $b(M) = \omega M(1 - M)$, which is a best response function that has been widely studied in the literature on nonlinear dynamics. The parameter ω "tunes" $b(M)$. Figures 1a and 1b graph $b(M)$ for $\Gamma(2.44)$ and $\Gamma(3.85)$.¹

{insert Figures 1a and 1b about here.}

Common knowledge that players are individually rational requires the serial deletion of dominated strategies. The set of serially undominated

¹ The exact values are 2.439024 and 3.846154, which are rounded to 2.44 and 3.85 for convenience in the text.

action combinations in game $\Gamma(\omega)$ is $[0, 0.25\omega]^n$ when $\omega \geq 2$ and $[0, 1-1/\omega]^n$ otherwise. Let $\mathbf{U}(\omega)$ denote the two dimensional space of serially undominated action combinations. Figures 1a and 1b graph $\mathbf{U}(2.44)$ and $\mathbf{U}(3.85)$ respectively, where the set $\mathbf{U}(\omega)$ is indicated by grey shading.²

The principle of individual rationality does not make very precise predictions in game $\Gamma(\omega)$. Requiring a mutual consistency condition allows one to make more precise predictions. An action combination e^* constitutes a strict equilibrium, if it satisfies the following mutual best response condition:

$$\pi(e^i, e^{-i*}) < \pi(e^{i*}, e^{-i*}) \quad (2)$$

for all $e^i \in \mathbf{E}$ and for all i . An observed action combination is a mutual best response outcome if it satisfies (2).

A symmetric strict equilibrium satisfies condition (2) when all of the players use the same action. Both of the strict equilibria of $\Gamma(\omega)$ are symmetric. Hence, it is convenient to denote the equilibria by the ordered pair (e, M) . The requirement that $\omega \in (1, 4]$ results in two strict equilibria: a corner equilibrium $(0, 0)$ and an interior equilibrium $(1 - 1/\omega, 1 - 1/\omega)$. Figures 1a and 1b indicate these equilibria with 'NE'. The equilibria occur at the intersection of $b(M)$ and the 45° line.

While imposing the mutual consistency requirement has significantly increased the precision of our prediction, it still leaves an equilibrium selection problem. Moreover, deductive selection principles, like payoff-dominance and symmetry, fail to reduce the set of equilibria, since both of the strict equilibria are efficient and symmetric. Hence, even if subjects are individually rational, giving them common information about $\Gamma(\omega)$ is not likely to produce mutually consistent behavior.

A. Selection Dynamics and Asymptotic Stability

If an equilibrium point is viewed as a potential convention that might arise amongst the players when they interact repeatedly, then some equilibria can be ruled unstable and, hence, unlikely conventions. An equilibrium point is unstable if it does not correspond to an asymptotically stable fixed point of some selection dynamic. In this section we focus on the myopic best response dynamic as a selection dynamic: see Van Huyck, Cook and Battalio (1994) for a general discussion of "relaxation algorithms," which include the myopic best response dynamic, the partial

² See Van Huyck, Cook, and Battalio (1994) for derivation.

adjustment dynamic, and least squares learning.³

Suppose that players choose their current action as a best response to last period's median action, then the selection dynamic can be reduced to the following difference equation:

$$M_{t+1} = \omega M_t(1 - M_t) \quad (3)$$

This difference equation has been studied in Baumol and Benhabib (1989), Boldrin and Woodford (1990), and Eckalbar (1993).

Figure 2a illustrates the use of the myopic best response dynamic (3) as a selection dynamic for $\Gamma(2.44)$. Notice that for an initial condition close to the corner equilibrium the dynamic diverges towards the interior equilibrium. However, initial conditions close to the interior equilibrium converge to the interior equilibrium. Hence, the myopic best response dynamic implies that the corner equilibrium is unstable and the interior equilibrium is stable.

This distinction between unstable and stable equilibria in $\Gamma(\omega)$ is useful to an economist conducting *a priori* comparative static analysis. Suppose one wants to know if increasing ω will increase e^i and M . The unstable corner equilibrium (0,0) is not a function of ω , but the stable interior equilibrium $(1 - 1/\omega, 1 - 1/\omega)$ is an increasing function of ω . Hence, appealing to something like Samuelson's correspondence principle would allow the analyst to predict that increasing ω will increase e^i and M .

However, increasing ω globally leads to complex dynamics in game $\Gamma(\omega)$ under the myopic best response dynamic. Let $\mathbf{A}(\omega)$ denote the attractor set under the myopic best response dynamic. For ω in (1,3) $\mathbf{A}(\omega)$ contains the single element $\{1 - 1/\omega\}$; but for ω in [3,4] the dynamics get complicated. For ω in [3, 3.449499) attracting two cycles appear. This bifurcation continues until $\mathbf{A}(4) = \mathbf{E}$.

⟨insert Figures 2a and 2b about here.⟩

Figure 2b illustrates the path for game $\Gamma(3.85)$ under the myopic best response dynamic starting at the same initial condition used in figure 2a. While the path still diverges from the corner equilibrium, it does not converge to the interior equilibrium. Instead it converges to a six cycle. Neither the corner nor the interior equilibrium are stable under the myopic best response dynamic.

³ See also Sargent (1993).

B. Learning without much information

In this section we restrict attention to learning models in which players know only their feasible actions and the historical payoffs associated with actions they have played. Moreover, we restrict attention to a class of reinforcement learning models called stochastic learning models in the literature. In these models players choose strategies according to a vector of probabilities rather than according to a reasoned search based on some belief about the underlying stochastic process generating the payoffs.

Within the class of linear stochastic learning rules, Borgers, Morales, and Sarin (2001) prove that the Cross (1973) learning rule and certain affine transformations of it are the only rules satisfying certain properties of which the principle requirement is absolute expediency, see also Schlag (1994). Given a stationary environment, a learning rule is absolutely expedient if expected payoffs are on average strictly higher each period.

The Cross dynamic is defined for a discrete action space and Van Huyck, Cook, and Battalio (1994,p.984-85) use a discrete approximation of the continuous action space, \mathbf{E} , so we digress briefly to introduce a discrete action space. Let $\Phi = \{0,1,\dots,100\}$ denote the subjects' finite set of actions. The function $f:\Phi\rightarrow\mathbf{E}$ mapping subjects' actions into the unit interval is $f(x) = (100 - x)/100$. Notice that this flips the best-response function and the 45° line. This is not particularly important for the reinforcement information experiment since subjects are not provided with information on the best-response function. We framed the problem in this way simply for comparability with Van Huyck, Cook, and Battalio (1994). Let $G(\omega)$ denote game $\Gamma(\omega)$ when actions are restricted to the finite grid Φ and are mapped to \mathbf{E} by $f(\bullet)$.

The strict equilibria for $G(2.44)$ are (41,41) and (100,100). The strict equilibria for $G(3.85)$ are (26,26) and (100,100). The values for ω were chosen to insure that the interior equilibria existed in pure strategies and, hence, remained strict given the 101×101 grid. Using f to map e back to the unit interval implies that effort in the interior equilibrium increases from approximately 0.59 to 0.74 as ω goes from 2.439024 to 3.846154. Both equilibria have a payoff of \$0.50 per player per period.

When the action space is Φ , $\mathbf{U}(2.44) = \{40,41,\dots,100\}$ and $\mathbf{U}(3.85) = \{4,5,\dots,100\}$. Hence, individual rationality or behavior consistent with adaptive learning (in the Milgrom/Roberts (1991) sense) imply that subjects will not choose $e^i \in \{0,1,2,\dots,39\}$ in $G(2.44)$ and will not choose $e^i \in \{0,1,2,3\}$ in $G(3.86957)$ either initially or after behavior has converged.

Having introduced the finite action space we can continue with our exposition of the Cross Dynamic. Let p_t^i denote player i 's mixed strategy defined on Φ , where $p_t^i(j)$ denotes the probability assigned to action j in

period t . If at time t player i 's state is p_t^i and he plays action j and receives payoff π_t^i , then the change in the state, Δp_t^i , is given by the following equations:

$$\begin{aligned}\Delta p_t^i(j) &= r(\pi_t^i)(1 - p_t^i(j)) \\ \Delta p_t^i(k) &= -r(\pi_t^i)p_t^i(k) \quad k \neq j,\end{aligned}\tag{4}$$

where $r(\pi)$ denotes the reinforcement strength of payoff π . There is little theory to inform the choice of $r(\bullet)$ other than the restriction to $(0,1)$ and the reasonable proposition that $r(\bullet)$ is increasing in π . Here we normalize the best feasible payoff to 1 and the worst feasible payoff to 0 and assume that the reinforcement strength of payoff π is proportional to the normalized payoff $u(\pi)$. Let $r(\pi) = \alpha u(\pi)$, where α is a step size parameter.

Figure 3 graphs two realizations of the Cross Dynamic for α equal to 0.05 and a uniform initial condition. We have mapped actions back into \mathbf{E} to avoid confusing the reader and promote comparison with previous analysis. Figure 3a is for $G(2.44)$ and figure 3b is for $G(3.85)$. In both cases the dynamic traces out clockwise cycles around the equilibrium point. Unlike the myopic best response dynamic, the cycles are stochastic. Occasionally, the cycle is broken by a switch back, but then the clockwise cycling resumes. Neither process has converged to a stationary outcome within 75 periods.

⟨insert figure 3a and 3b about here.⟩

It takes about 750 periods for most realizations of the Cross dynamic to converge to a stationary outcome with α equal to 0.05. Figure 4a graphs the simulated distribution function of the median in period 750 for α equal to 0.05 and a uniform initial condition. Each treatment is simulated 25 times. In period 750, the simulated medians ranged from 0.49 to 0.69 for the $G(2.44)$ and from 0.70 to 0.81 for $G(3.85)$. So the simulations converge to outcomes near the interior equilibrium. As figure 4a illustrates, increasing ω shifts the simulated distribution function for M to the right.

The outcomes were not mutual best response outcomes. To give some feel for how close the distributions are to mutual best response outcomes we map the outcome back into the payoff space. In a mutual best response outcome all players earn 0.50. The simulated payoffs ranged from 0.23 to 0.50 in $G(2.44)$ and from 0.19 to 0.50 in $G(3.85)$. These represent fairly substantial deviations from best responses.

A second problem with the Cross Dynamic with α equal to 0.05 is that it has failed to eliminate the play of strictly dominated actions. The set of

serially undominated actions for $\Gamma(2.44)$, $U(2.44)$, is $[0, 0.6091]$. Twenty-four percent of the medians lie outside of $U(2.44)$.

As McAllister (1991) emphasized in his application of reinforcement learning algorithms to Van Huyck, Battalio, and Beil's (1990) coordination game, one confronts a tradeoff between more rapid adaptation and the possibility that the algorithm converges to a bad outcome. This tradeoff amounts to the difference between using reinforcement learning algorithms as models of human behavior or as equilibrium selection dynamics.

Figure 4*b* graphs simulated distribution functions for the period 1500 medians with α equal to 0.01. The simulated medians ranged from 0.51 to 0.74 for $G(2.44)$ and from 0.62 to 0.8 for $G(3.85)$. Now twenty percent of the medians lie outside of $U(2.44)$. The simulated distribution function has shifted to the right, but there is now a positive probability⁴ of observing a violation of the basic comparative static result that increasing ω increases e^i and M . However, comparing figure 4*a* and 4*b* all in all one is struck more by the similarity than the differences between the two sets of simulations.⁵

⟨insert figure 4*a* and 4*b* about here.⟩

III. EXPERIMENTAL DESIGN

Our experiment consists of two treatments: cohorts 1-4, which consist of five subjects playing $G(2.44)$, and cohorts 5-8, which consist of five subjects playing $G(3.85)$. The sessions repeated $G(\omega)$ seventy-five periods and this was announced at the beginning of the session. Payoffs were in dollars rounded to the nearest ten-thousandths of a dollar.

Figure 5*a* is a half-tone image of the main screen used to communicate $G(\omega)$ to our subjects. On the computer the slider is blue. Once the subject has clicked on the slider, he can slide the mouse on the mouse pad to change the displayed action. Clicking the mouse restores the cursor and clicking on proceed submits the choice. The main screen also displays the history of choices and earnings associated with that choice ordered by periods. Notice that a subject does not have any information about the best response to an action, the security of an action, or the historical median.

Our reinforcement information graphical user interface is inspired by Van Huyck, Cook, and Battalio's (1994, p.984) graphical user interface.

⁴Specifically, about 3 percent of the time.

⁵ See Roth and Erev (1995) on the importance of intermediate run results in organizing experimental data.

They used a ‘blue box’ to communicate a complete information description of the game, see figure 5*b*. Notice also that Van Huyek, Cook, and Battalio restrict the action space to integers between 1 and 90. Here we use the integers 0 to 100, which seems more natural. In order to preserve the strict interior equilibrium we had to adjust the tuning parameter ω slightly.

⟨insert figure 5*a* and 5*b* about here.⟩

The only information subjects had beyond that used in reinforcement learning algorithms was the number of periods to be played and the fact that the rule mapping the actions of the five players in their cohort to their payoffs was stationary and deterministic, see the instructions in appendix A.⁶ We provided this information in order to minimize the number of questions we would have to answer during the instruction phase of a session. The instructions were read aloud while the subjects followed along on their monitors. The instructions covered the general information about the experiment as well as the use of the graphical user interface.

The experiment was conducted in the TAMU economics laboratory. Seating at the terminals was determined by lot. Forty subjects participated in the experiment, five in each cohort. All were recruited from undergraduate economics courses at Texas A&M University. The sessions take about one hour to conduct. If the subjects coordinate on either the corner or the interior equilibrium for all seventy-five periods, they would each earn \$37.50.

IV. EXPERIMENTAL RESULTS

A. Initial Behavior

Subjects have no information on the different payoff surfaces and we attempted to keep all else constant, so one would hope that the initial behavior in the two treatments was similar. Figure 6 reports the initial distribution of individual actions by treatment. The solid line denotes the empirical distribution function for the $G(2.44)$ cohorts and the dashed line denotes the empirical distribution function for the $G(3.85)$ cohorts. The 45° line denotes uniform play. Formal statistical tests fail to reject the null hypothesis of uniform play and fail to reject the null hypothesis that both

⁶ The subjects did not know the maximum and minimum feasible payoff, which is needed to normalize payoffs for the Cross model.

samples were drawn from the same population.⁷ The largest Kolmogorov T statistic is 0.18 and the critical value at the 5 percent level of statistical significance is 0.29. The Smirnov T statistic is 0.15 and the critical value at the 5 percent level of statistical significance is 0.8.

⟨insert figure 6 here.⟩

B. Adaptive Behavior

Figure 7 graphs the observed medians for cohorts 1 to 4 in the phase space. Recall that cohorts 1 to 4 played $G(3.85)$. The initial median is denoted by m in the graphs. The most striking feature of the figures is how much more quickly human behavior converges to the intersection of the best response function and the 45° line than did the Cross dynamic, compare figures 7 and 3.

<Insert Figure 7 about here>

Specifically, Cohort 1 coordinates on the interior equilibrium median in period 10 and never leaves this state after period 14. Cohort 2 coordinates on the interior equilibrium median in period 21 and never leaves this state after period 38. Cohort 3 coordinates on the interior equilibrium median in period 32 and never leaves this state after period 55. Cohort 4 coordinates on the interior equilibrium median in period 31 and never leaves this state after period 45, see Table 1.

Table 1: Period of Convergence to Equilibrium Median

Cohort	G(2.44)		G(3.85)	
	First	Absorbed	First	Absorbed
1 or 5	10	24	10	14
2 or 6	18	25	21	38
3 or 7	18	18	32	55
4 or 8	15	25	31	45

⁷ These results motivate our use of the uniform distribution as the initial distribution in section IIB.

Figure 8 graphs the observed medians for cohorts 5 to 8 in the phase space. Recall that cohorts 5 to 8 played $G(2.44)$. Again, convergence to the interior equilibrium median is surprisingly fast. Cohort 5 coordinates on the interior equilibrium median in period 10 and never leaves this state after period 24. Cohort 6 coordinates on the interior equilibrium median in period 18 and never leaves this state after period 25. Cohort 7 coordinates on the interior equilibrium median in period 18 and never leaves this state. Cohort 8 coordinates on the interior equilibrium median in period 15 and never leaves this state after period 25.

(Insert figures 8 here)

There appears to be a tendency to overshoot the interior equilibrium median in $G(3.85)$ and to avoid overshooting in $G(2.44)$. This is vaguely similar to the treatment difference predicted by the myopic best response dynamic, see figure 2. In order to formalize this treatment difference we construct a contingency table for the median in which we group observations by whether they are less than, equal to, or greater than the interior equilibrium median, M^* . Given the median is above M^* the odds that the response is in $G(3.85)$ rather than $G(2.44)$ is 4.5, see table 2. The Fisher's Exact Test (2-tail) rejects the null hypothesis of no treatment effect at all conventional significance levels. So there is evidence of overshooting in the data, but this overshooting does not destabilize the convergence to M^* .

Table 2: Contingency Table for Median.

	$G(3.85)$	$G(2.44)$	Total
$M_t < M^*$	59 9.83 50.86 19.67	57 9.50 49.14 19.00	116 19.33
$M_t = M^*$	187 31.17 44.74 62.33	231 38.50 55.26 77.00	418 69.67
$M_t > M^*$	54 9.00 81.82 18.00	12 2.00 18.18 4.00	66 11.00
Total	300 50.00	300 50.00	600 100.00
Fisher's Exact Test (2-tail) Prob. 0.000			

C. Terminal Behavior

It is possible to test the null hypothesis that the empirical and simulated distribution functions of the medians were derived from the same population. Doing so requires viewing the Cross Dynamic as something more than a selection dynamic. For treatment $G(2.44)$, the Smirnov statistic for the maximum vertical distance between the empirical distribution function in period 75 and the simulated distribution function in period 75 with $\alpha = 0.05$ is 0.8, which exceeds the critical value of approximately 0.73 at the five percent level of statistical significance. Hence, we can reject the null hypothesis that the empirical and simulated distributions come from the same population distribution.

For treatment $G(3.85)$, things are not so simple. The Smirnov statistic is 0.48 so one fails to reject. One suspects this is due to a lack of power, but the Characteristic Function Test Statistic of 7.27 has a probability value of 0.12 and, hence, also fails to reject at conventional significance levels.

The distribution of individual actions in period 75 tells a more dramatic story. Figure 9 graphs the period 75 empirical and simulated distribution functions of actions for treatments $G(2.44)$ and $G(3.85)$. The Smirnov statistic for the null hypothesis that the empirical distributions $e[2.44,75]$ and $e[3.85,75]$ were drawn from the same population is 1.0 and, hence, is

easily rejected at all conventional significance levels. The Characteristic Function Statistic for the null hypothesis that $e[2.44,75]$ and the simulated distribution $s[2.44,0.05,75]$ were drawn from the same population is 28.93 and, easily rejects the null hypothesis at all conventional significance levels. The Characteristic Function Statistic for the null hypothesis that $e[3.85,75]$ and $s[3.85,0.05,75]$ were drawn from the same population is 37.93 and easily rejects the null hypothesis at all conventional significance levels. The Cross Dynamic does not predict observed behavior under the reinforcement information treatments accurately.

A hypothesis that can not be rejected is that the empirical distribution functions $e[2.44,75]$ and $e[3.85,75]$ were drawn from the theoretical distribution predicted by the respective interior equilibrium. The Kolmogorov statistic for $e[2.44,75]$ is 0.15 and for $e[3.85,75]$ is 0.20. The critical value at the five percent significance level is 0.294. Paradoxically, it is the complete information theory that more accurately predicts observed behavior under the reinforcement information treatment.

It is interesting to note that the human learning process was consistent with adaptive learning in the sense of Milgrom and Roberts (1991). No subject played a dominated action in period 75 even though they could not deduce what $U(\omega)$ contains. In contrast, 37 percent of the mass in the simulated distribution $s[2.44,0.05,75]$ is on dominated actions in period 75 and in period 750 36 percent of the mass in the simulated distribution $s[2.44,0.05,750]$ remains on dominated actions.

Cohort 7 actually converges to a mutual best response outcome. However, the other cohorts have either one or two subjects who are not giving a best response ex post. These subjects have stopped exploring and are playing the same action repeatedly. For treatment $G(3.85)$, this satisficing behavior costs one subject 10¢, three subjects 3¢, and two subject 2¢ per period. For treatment $G(2.44)$, this satisficing behavior costs one subject 9¢, one subject 4¢, and one subject 1¢ per period. All of the subjects are within a dime of giving a best response and 82.5 percent of the subjects gave a best response to the behavior of the other subjects in their cohorts without any information about their best response function, see Table 3.⁸

⁸ Recall that all of the automata in the simulations with $\alpha = 0.05$ are within 31¢ of a best response in period 750. They are within 65¢ of a best response in period 75. The maximum deviation possible given an interior equilibrium median is 59¢ in $G(2.44)$ and 74¢ in $G(3.85)$.

Table 3: Frequency Distribution of the Distance from a Best Response.

Distance from best response (cents)	G(2.44)		G(3.85)		Total	
	Count	Percent	Count	Percent	Count	Percent
0	17	85%	14	70%	31	82.5%
1	1	5%	0	0%	1	2.5%
2	0	0%	2	10%	2	5.0%
3	0	0%	3	15%	3	2.5%
4	1	5%	0	0%	1	2.5%
9	1	5%	0	0%	1	2.5%
10	0	0%	1	5%	1	2.5%
	20	100	20	100	40	100

VI. COMPARISON WITH VAN HUYCK, COOK, AND BATTALIO.

Van Huyck, Cook, and Battalio (1994) attempted to provide a complete information description of the game to their subjects: call this treatment the complete information treatment. Here we have limited the information to that used by reinforcement learning algorithms: call this treatment the reinforcement information treatment. How does observed behavior under complete and reinforcement information differ?

If one is only concerned with the long run, the answer is not much. All treatments converge to the interior equilibrium median, M^* , regardless of the predicted stability or information treatment, see figure 10. But inspecting figure 10 does suggest that it takes longer for the median to converge to the interior equilibrium median and that the process is noisier under the reinforcement information treatment.

Since this is a fairly subtle distinction, we resort to econometric procedures to characterize the difference. Specifically, we use the statespace procedure in *SAS* version 6.11 to fit an ARMA model to the median time series. The procedure employs canonical correlation analysis for the automatic identification of the state space model.

Since the series exhibit an extreme form of heteroskedasticity we truncate them at period 25. In order to estimate the model by treatment we append session series by treatment to form one time series per treatment, which, of course, introduces an additional heteroskedasticity problem. Finally, M^* was subtracted from the observed median time series so the estimated model is for deviations from the interior equilibrium median.

Table 4 reports the estimated ARMA models. The estimated complete information treatment models are ARMA(1,1) and the estimated reinforcement information treatment models are ARMA(2,2). Moreover, the estimated autoregressive parameters reveal much more persistence under the reinforcement information treatments. So there is a statistically significant difference between the two information conditions. Observed behavior under complete information and reinforcement information treatments does differ in the time it takes to converge and the estimated variance of the moving average shock is larger for reinforcement information treatments than for the complete information treatments.

The tuning parameter ω also influences the estimated ARMA models in a systematic way. Increasing ω increases the estimated variance of the moving average shock and reduces the persistence of the autoregressive component, that is, increasing ω makes the time series less predictable.

Table 4: ARMA model of median times series by treatment

Treatment	AR process		MA Process	
	M_{t-1}	M_{t-2}	Order	ϵ^2
comp G(3.87)	0	.	1	0.0036
comp G(2.47)	0.35*	.	1	0.0003
reinf G(3.85)	0.57*	0.17	2	0.0114
reinf G(2.44)	0.46*	0.32*	2	0.0086

* denotes significantly different from 0 at the five percent level.

One last difference is worth noting. Under complete information only one subject out of forty failed to give an ex post best response in the terminal period, while under reinforcement information nine subjects out of forty failed to best respond. Under reinforcement information all of the subjects are within a dime of giving a best response, while under complete information all subjects are within three cents of giving a best response.

VII. SUMMARY AND CONCLUSIONS

In the experiment subjects converge to an absorbing state at rates that

are orders of magnitude faster than reinforcement learning algorithms. These states are always very close to a mutual best response outcome and the terminal median is always exactly equal to the interior equilibrium median. The interior equilibrium is behaviorally stable. Given a theory that selects the interior equilibrium, standard comparative static arguments accurately predict that increasing the tuning parameter, ω , increases individual effort, e^i , and median effort, M .⁹

All of this is true under both the complete and reinforcement information treatments. So there is a sense in which the hypothesis that taking away information only used in a deductive analysis of the situation will not influence behavior since subjects don't use it anyway is not contradicted by the experiment. However, the information treatment does influence behavior in a subtle, but statistically significant way. It takes longer for the median to converge to the interior equilibrium median, the process is noisier under the reinforcement information treatment, and more subjects fail to give a best response to the median in the terminal period.

A careful analysis of the data reveals that random search models of reinforcement learning, like Erev and Roth (1995) or the closely related Cross Dynamic, do not accurately describe behavior even when subjects are restricted to reinforcement information. Specifically, our subjects are able to search the action space much more efficiently than the random-search-reinforcement-learning analysis allows.¹⁰ Our subjects do better even under information conditions that favor the reinforcement learning algorithm. It appears to us that human cognition is not well described by either the choice-theoretic analysis or the random-search-reinforcement-learning analysis.¹¹

Sarin and Vahid (2001) propose a payoff assessment model that takes account of the similarity amongst strategies to explain the data reported in

⁹ Van Huyck, Cook, and Battalio (1994) demonstrate that for some "relaxation algorithms" the interior equilibrium is stable. The myopic best response dynamic is not one of them. Smith (1990, p.13) cites several cobweb market, public goods, and oligopoly experiments that also find the myopic best response dynamic to be too conservative. None of these limit subjects to reinforcement information as done here.

¹⁰ Note that when you speed up these algorithms they begin to get stuck in absorbing states that are not even close to being mutually consistent, something our subjects don't do, and if one continually smears probability to prevent this then the process doesn't converge, which is also something our subjects don't do.

¹¹ See Possajennikov (1997) for an analysis of convergence of reinforcement learning algorithms.

this paper. Their payoff assessment model fits remarkably well and is more successful than the original Cross model and versions that include similarity or declining step size. Chen and Khoroshilov (2001) use the data reported in this paper to compare the payoff assessment model to a version of experience weighted attraction learning model and a version of the relative payoff sum model. They also find the payoff assessment model fits the data best.

REFERENCES

- Aumann, Robert and Adam Brandenberger, "Epistemic Conditions for Nash Equilibrium" *Econometrica* 63(5), September 1995, 1161-80.
- Baumol, William J. and Jess Benhabib. "Chaos: Significance, Mechanism, and Economic Applications." *Journal of Economic Perspectives* 3(1) Winter 1989, 77-105.
- Boldrin, Michele and Michael Woodford. "Equilibrium Models Displaying Endogenous Fluctuations and Chaos: A survey." *Journal of Monetary Economics* 25 1990, 189-222.
- Borgers, Tilman, Antonio J. Morales, and Rajiv Sarin, "Expedient and Monotone Learning Rules," laser-script, July 2001.
- Bray, M. "Learning, Estimation, and the Stability of Rational Expectations." *Journal of Economic Theory* 26, 1982, 318-39.
- Bush, Robert and Frederick Mosteller, *Stochastic Models for Learning* (New York, NY: Wiley, 1955).
- Conover, W.J. *Practical Nonparametric Statistics*. 2nd ed. (New York, NY: John Wiley & Sons, 1980).
- Chen, Yan, and Yuri Khoroshilov, "Learning Under Limited Information," laser-script, October 2001.
- Cross, J.G., "A Stochastic Learning Model of Economic Behavior," *Quarterly Journal of Economics*, 87, 1973, 239-66.
- Eckalbar, John C. "Economic Dynamics." In *Economic and Financial Modeling with Mathematica*, edited by H. Varian. (New York, Springer-Verlag, 1993).
- Erev, Ido and Alvin E. Roth, "On the need for low rationality, cognitive game theory: Reinforcement learning in experimental games with unique, mixed strategy equilibria", laser-script August 1995.
- Lucas, Robert E., Jr. "Adaptive Behavior and Economic Behavior." In *Rational Choice: the contrast between economics and psychology*, edited by R. Hogarth and M. Reder. (Chicago, University of Chicago Press, 1987).
- McAllister, Patrick H., "Adaptive Approaches to Stochastic Programming," *Annals of Operations Research*, 30, 1991, 45-62.
- Milgrom, Paul and John Roberts. "Adaptive and Sophisticated Learning in Normal Form Games." *Games and Economic Behavior* 3(1) February 1991, 82-100.
- Possajennikov, Alexandre, "An Analysis of a Simple Reinforcing Dynamics: Learning to Play an 'Egalitarian' Equilibrium," laser-script, January 1997.
- Rassenti, Stephen, Stanley S. Reynolds, Vernon L. Smith, and Ferenc Szidarovszky, "Learning and Adaptive Behavior in Repeated

- Experimental Cournot Games," laser-script, October 1993.
- Roth, Alvin E., and Ido Erev, "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the intermediate term," *Game and Economic Behavior* 8(1), January 1995, 164-212.
- Sargent, Thomas J. *Bounded Rationality in Macroeconomics*. (Oxford, Clarendon Press, 1993).
- Sarin, Rajiv, "Learning Through Reinforcement: The Cross Model," laser-script March 1995.
- Sarin, Rajiv, and Farshid Vahid, "Strategic Similarity and Coordination," laser-script July 2001.
- Schlag, K., "A Note on Efficient Linear Learning Rules," laser-script 1994.
- Selten, Reinhard, "The Chain Store Paradox," *Theory and Decision* 1978; reprinted in *Models of Strategic Rationality* (Dordrecht, Kluwer Academic Publishers, 1988).
- Smith, Vernon L., "Experimental Economics: Behavioral Lessons for Microeconomic Theory and Policy," 1990 Nancy L. Schwartz memorial lecture, Kellogg Graduate School of Management, Northwestern University.
- Van Huyck, John B., Raymond C. Battalio, and Richard O. Beil, "Tacit Coordination Games, Strategic Uncertainty, and Coordination Failure," *American Economic Review*, 80, 1990, 234-48.
- Van Huyck, John B., Raymond C. Battalio, and Richard O. Beil. "Strategic Uncertainty, Equilibrium Selection, and Coordination Failure in Average Opinion Games." *The Quarterly Journal of Economics* Vol. CVI, No. 426, August 1991: 885-910.
- Van Huyck, John B., Joseph P. Cook and Raymond C. Battalio. "Adaptive Behavior and Coordination Failure," *Journal of Economic Behavior and Organization*, 32, 1997, 483-503.
- Van Huyck, John B., Joseph P. Cook and Raymond C. Battalio, "Selection Dynamics and Adaptive Behavior," *Journal of Political Economy* 102(5), 1994, 975-1005.

Appendix A: Instruction Text File for Graphical User Interface

^HP^CPAINSTRUCTIONS

This is an experiment in the economics of strategic decision making. Various agencies have provided funds for this research. If you follow the instructions and make appropriate decisions, you can earn an appreciable amount of money. At the end of today's session, you will be paid in private and in cash.

It is important that you remain silent and do not look at other people's work. If you have any questions, or need assistance of any kind, please raise your hand and an experimenter will come to you. If you talk, laugh, exclaim out loud, etc., you will be asked to leave and you will not be paid. We expect and appreciate your cooperation.

You will be making choices on a Logitech mouse, which is located on the mouse pad in the middle of your table. You may move the pad to the right or left if this would be more comfortable. Hold the mouse in a relaxed manner with your thumb and little finger on either side of the mouse. Rest your wrist naturally on the table surface. When you move the mouse, let your hand pivot from the wrist. Use a light touch. Your cursor (a white arrow on your screen) should move when you slide the mouse on the mouse pad. If it does not, raise your hand.

To participate, you must be able to move the cursor onto an object and click any one of the mouse buttons. We will call pointing at an object and then clicking your mouse "clicking on" an object displayed on the screen. Click on the page down icon located below to display the next page.

^HP^CPAGENERAL

In this experiment you will participate in a market of five people. At the beginning of period one, each of the participants in this room will be randomly assigned to a group of size five and will remain in the same group for seventy-five periods. That is, you will remain grouped with the same four other participants for the next seventy-five periods.

In each period, every participant will pick a value of X. The values of X you may choose are any one of the 101 integers 0, 1, 2, . . . , 98, 99 or 100. The value you pick for X and the value of X picked by the other four participants in your group will determine the payoff you receive for that period. The rule used to determine your earnings will remain the same for the next seventy-five periods. That is, if you pick the same value for X for two different periods AND the other four participants in your group pick the same values for X for the same two periods, then your earnings will be the same in both periods. However, if you pick the same value for X for two different periods AND the other four participants in your group do not pick the same values for X for the same two periods, then your earnings in the two periods may be different.

^HR5^HP^CPAMAIN SCREEN

We will now view the main screen. You will use the main screen to make your choices each period. While you view the main screen I will read the description of the main screen contained in the next two pages. You can review the text that I am reading at any time during the experiment by returning to the instructions. Click on the word "MAIN" located on the second line down from the top of the screen now. (The second line is the light blue line on your screen).

The top line of the main screen displays the current period number, the title of the screen and your current balance. The second line has the word "PROCEED", the abbreviation "INSTR" and the word "RECORD" on it. During the session you will be able to return to these instructions by clicking on "INSTR." You will also be able to view the history of play by clicking on "RECORD", which I will explain in a moment.

The remainder of the screen contains: (1) A blue vertical bar that you will use to make your choice each period, and (2) A historical record of your past choices and your earnings

for each period of this session.

Please look at the monitor at the front of the room while I demonstrate how to use the vertical blue bar to enter your choice each period.

[^]HP[^]CPA Now click on the blue bar labelled YOUR CHOICE. Your mouse cursor is replaced by a green horizontal line. Immediately to the right of the blue bar your current choice appears in green. By moving your mouse up and down you can pick any value of X, (0, 1, 2, ..., 98, 99, 100), for your choice of X for the current period. Click your mouse a second time to restore the cursor.

When you are ready to enter a choice for a period you do so by selecting a value for X and clicking a second time to restore the cursor. Next click on "PROCEED", located on the second line of the main screen. Click on "PROCEED" now and notice that the message 'DO YOU WANT TO PROCEED' in appears in yellow. To proceed you click on the word "YES", in green at the right side of the line. If you want to change your choice at this point you would click on the word "NO", in red. Click on "NO" now and notice that the choice you had entered is canceled and you must now make another choice to proceed.

Now click on the blue bar again. Click the mouse a second time to select a value of X for the current period and to restore the cursor. Click on "PROCEED" now. If you click on "YES" your choice for the period will be entered. Please click on "YES" now to return to the instructions.

[^]HR6[^]HP[^]CPAWAITING SCREEN

During a session a waiting screen will appear after you have made a choice. NOTICE: the second line DOES NOT contain PROCEED. While you are waiting for all of the other participants to pick a value for X for the current period you may view the instructions and the record screen by clicking on "INSTR" or "RECORD." When all participants have made a choice for the current period you will be automatically switched to the outcome screen. The choice displayed on the WAITING SCREEN is the choice that you made during the demonstration of the main screen. You will automatically return to the instructions in twenty seconds. Click on "WAITING" now.

[^]HR7[^]HP[^]CPAOUTCOME SCREEN

During a session, after everyone has made their choices, the outcome screen will appear. The outcome screen summarizes what happened each period for ten seconds. Your choice and period earnings will be highlighted in [^]CKA^{green}[^]CPA.

During the experiment, after the period 1 outcome screen has been displayed for ten seconds, you will automatically advance to period 2. Your main screen for period 2 will appear and you may then make a choice for period 2 whenever you are ready. After the outcome screen for period 2 has been displayed for ten seconds you will automatically advance to period 3. This will continue for seventy-five periods.

The outcome screen is not active and, therefore, your mouse cursor will not be present while the outcome screen is displayed. Click on "OUTCOME" now. The value displayed on the outcome screen for YOUR CHOICE is the selection that you made earlier during these instructions. A series of question marks [^]CKA??????[^]CPA are displayed for earnings. During the experiment your period earnings will be displayed in the location where the question marks are currently displayed. You will automatically return to the instructions in twenty seconds.

[^]HR8[^]HP[^]CPARECORD SCREEN

The record screen records the period outcomes and updates your earnings balance. The record screen contains all of the information contained in the past history on the MAIN SCREEN and the WAITING SCREEN plus an additional column labelled Balance that has your balance at the end of each period. At the beginning of the first period your balance is zero. At the end of each period your current period earnings will be added to your balance.

At the end of this experiment you will be paid your ending balance, (the sum of all of your period earnings), in cash.

Click on the word "RECORD" located on the second line down from the top of your screen now. As the experiment proceeds the records for the earlier periods will scroll off the top of the record screen. You may review the earlier records by clicking on the page up, page down, line up and line down icons located at the bottom of the record screen. Click on RETURN to leave the RECORD SCREEN.

^HP^CPASUMMARY

^E1***^E0 At the beginning of period one, each of the participants in this room will be randomly assigned to a group of size five and will remain in the same group for seventy-five periods.

^E1***^E0 In each period, every participant will pick a value of X. The values of X you may choose are any one of the 101 integers 0, 1, 2, . . . , 98, 99 or 100. The value you pick for X and the value of X picked by the other four participants in your group will determine the payoff you receive for that period.

^E1***^E0 The rule used to determine your earnings will remain the same for the next seventy-five periods. That is, if you pick the same value for X for two different periods AND the other four participants in your group pick the same values for X for the same two periods, then your earnings will be the same in both periods. However, if you pick the same value for X for two different periods AND the other four participants in your group do not pick the same values for X for the same two periods, then your earnings in the two periods may be different.

^HP^CPA^E1***^E0 You make a choice by (i) selecting a value between 0 and 100 for X using the blue bar, (ii) clicking the mouse a second time, to select your choice and restore your cursor and then (iii) clicking on "PROCEED" and "YES" to enter and confirm your choice for the current period.

^E1***^E0 Remember that you can view the instructions or the record screen by clicking on the appropriate word on the light blue bar.

^E1***^E0 Your balance at the end of the session, if positive, will be paid to you in private and in cash. If your balance is negative you will be paid zero.

If you have a question, please raise your hand, and an experimenter will come to assist you. If there are no questions, period one of the experiment will begin.

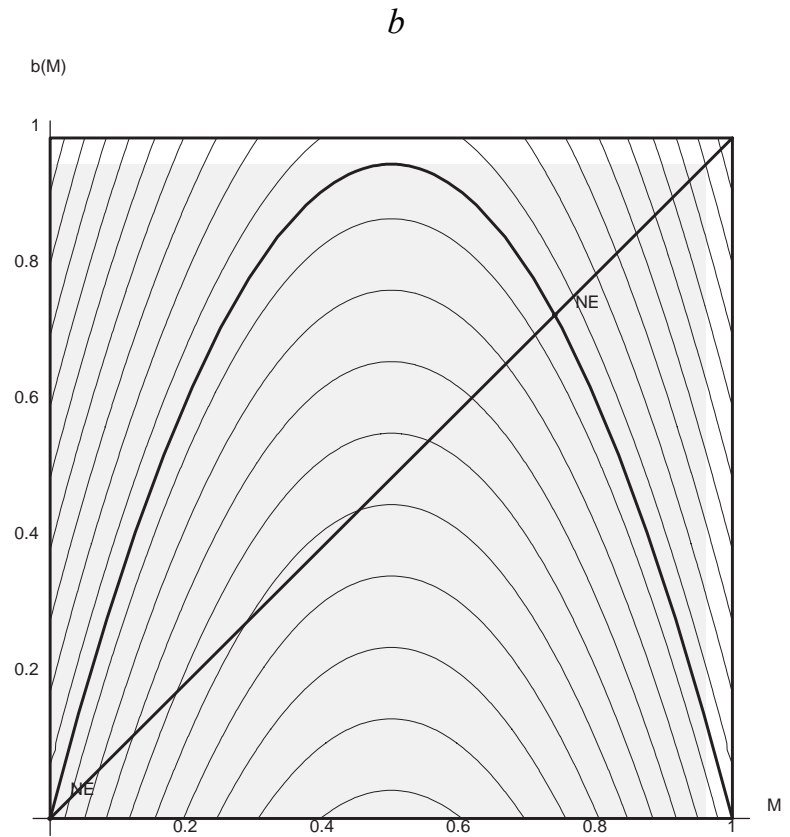
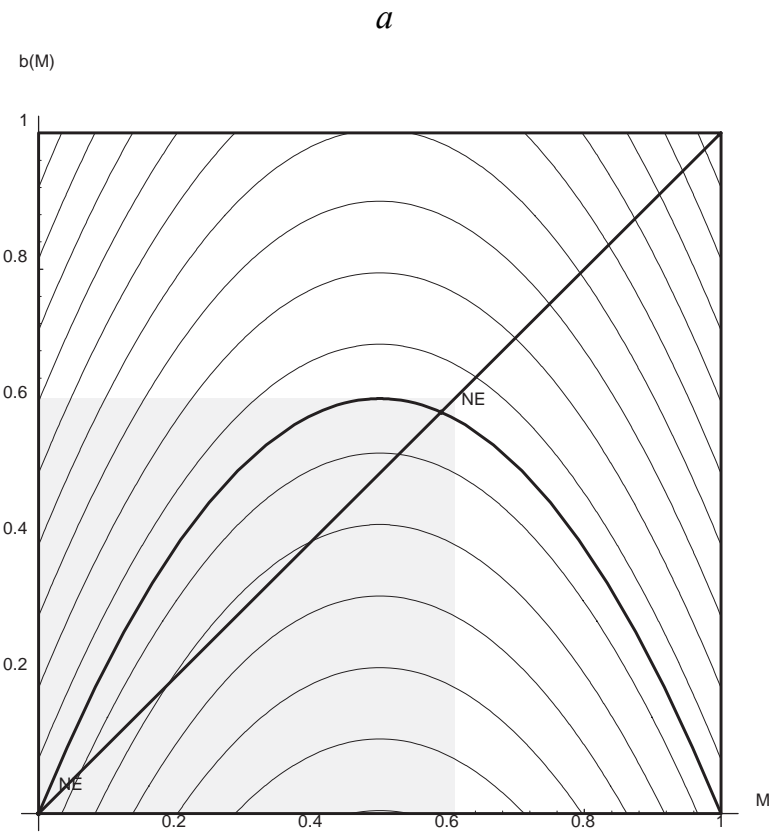


Figure 1: *a*, Graph of $b(M)$ and $U(2.44)$ for $\Gamma(2.44)$. The intersection of $b(M)$ and the 45° line is a strict equilibrium, NE; *b*, Graph of $b(M)$ for $\Gamma(3.85)$ and $U(3.85)$, which is denoted by grey shading. (The 10 unit contours are denoted by thin curves.)

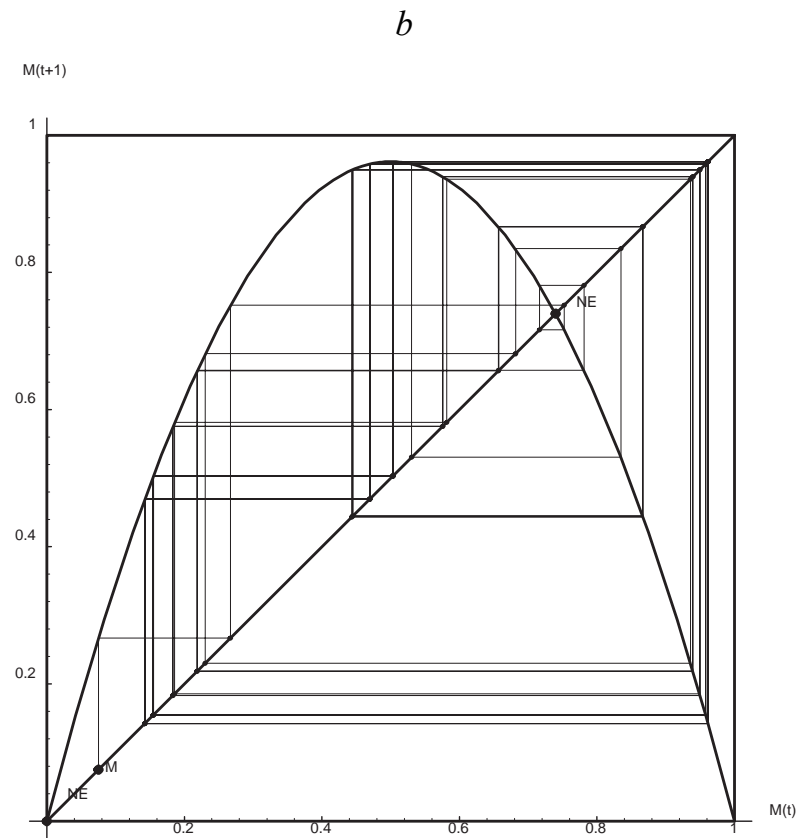
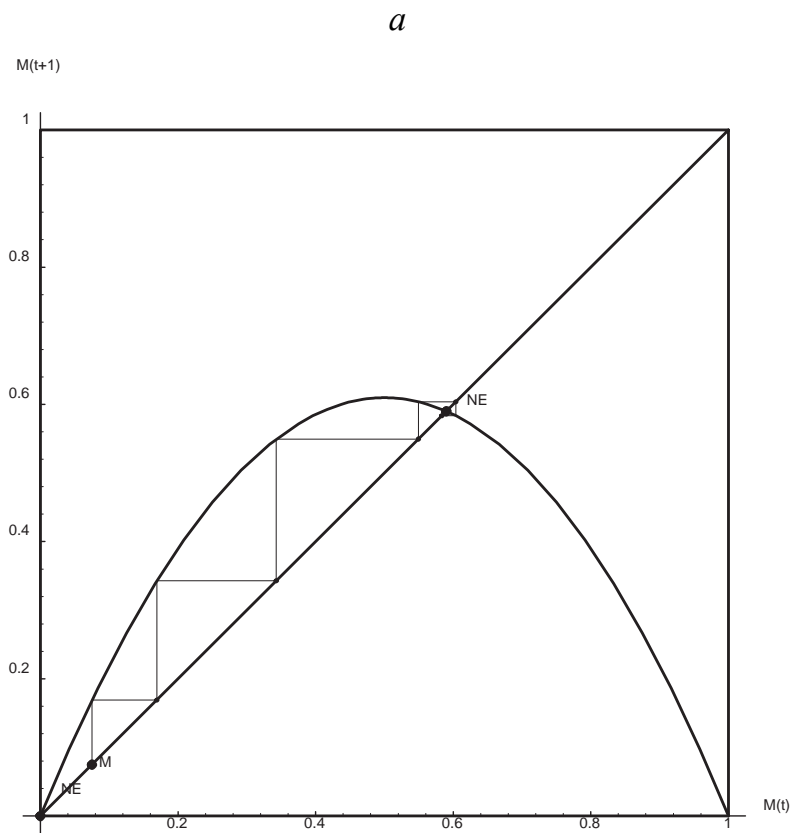


Figure 2: *a*, Path of a selection dynamic for $\Gamma(2.44)$ with the property that the interior equilibrium is stable and the corner equilibrium is unstable. M denotes the initial condition; *b*, Path for $\Gamma(3.85)$ under the same selection dynamic as *a*. Neither equilibrium is stable.

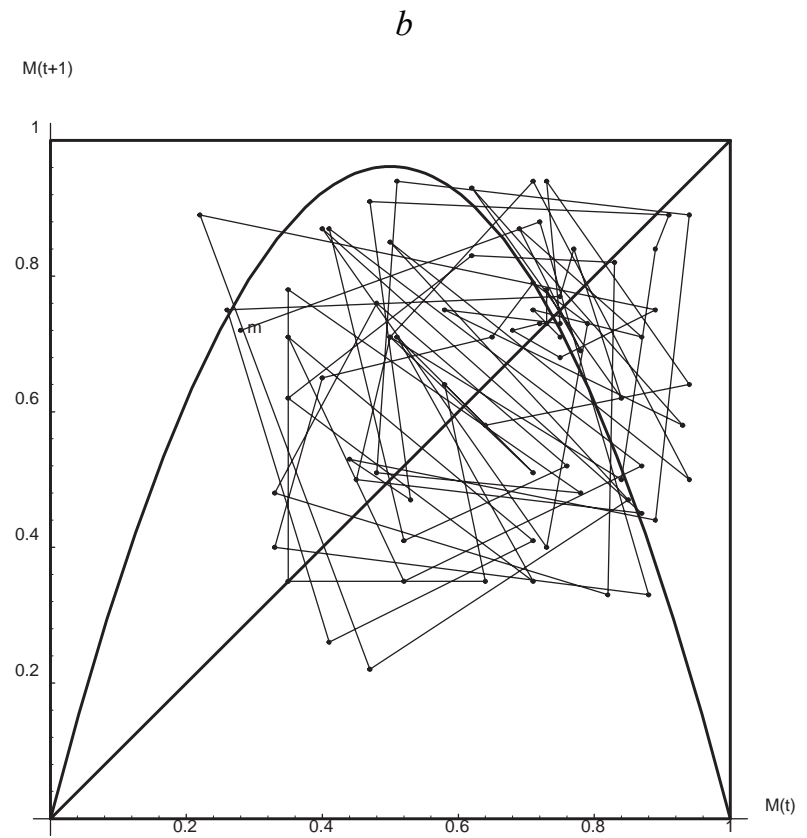
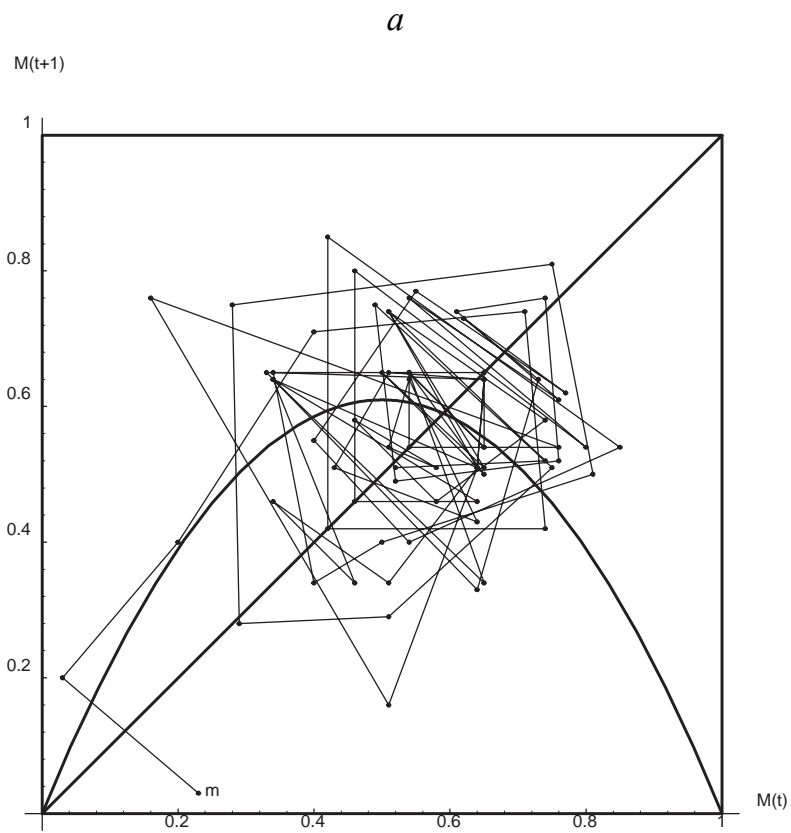


Figure 3: *a*, First 75 periods of a simulated realization of the Cross Dynamic for $\Gamma(2.44)$; m denotes initial state; α was 0.05.
b, First 75 periods of a simulated realization of the Cross dynamic for $\Gamma(3.85)$; α was 0.05.

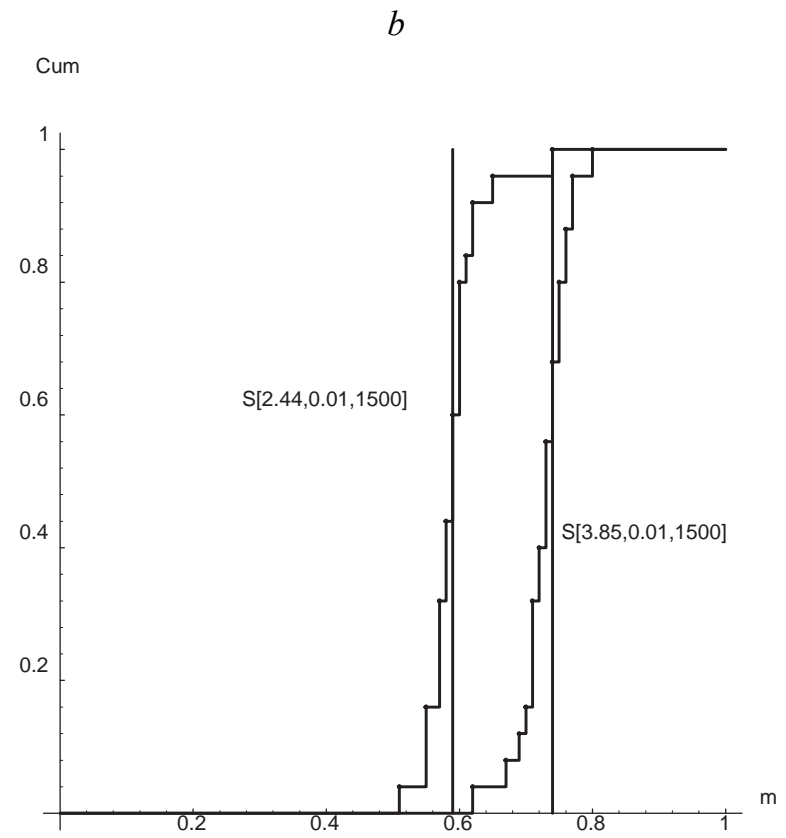
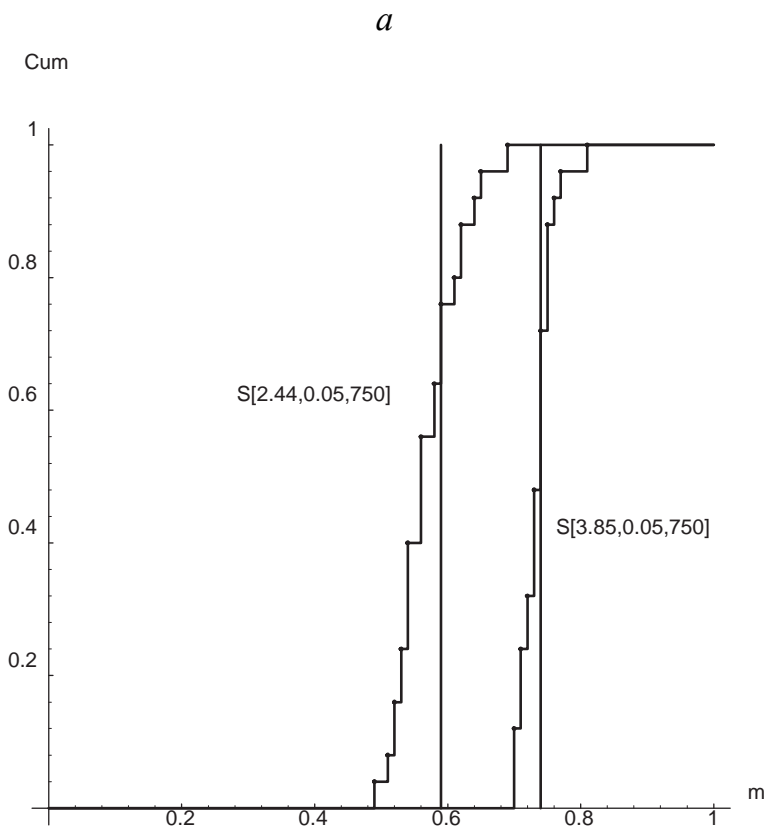


Figure 4: *a*, Simulated distribution functions of median in period 750 for $\Gamma(2.44)$ and $\Gamma(3.85)$ under the Cross Dynamic with α equal to 0.05. *b*, Simulated distribution functions of median in period 1500 for $\Gamma(2.44)$ and $\Gamma(3.85)$ under the Cross Dynamic with α equal to 0.01.

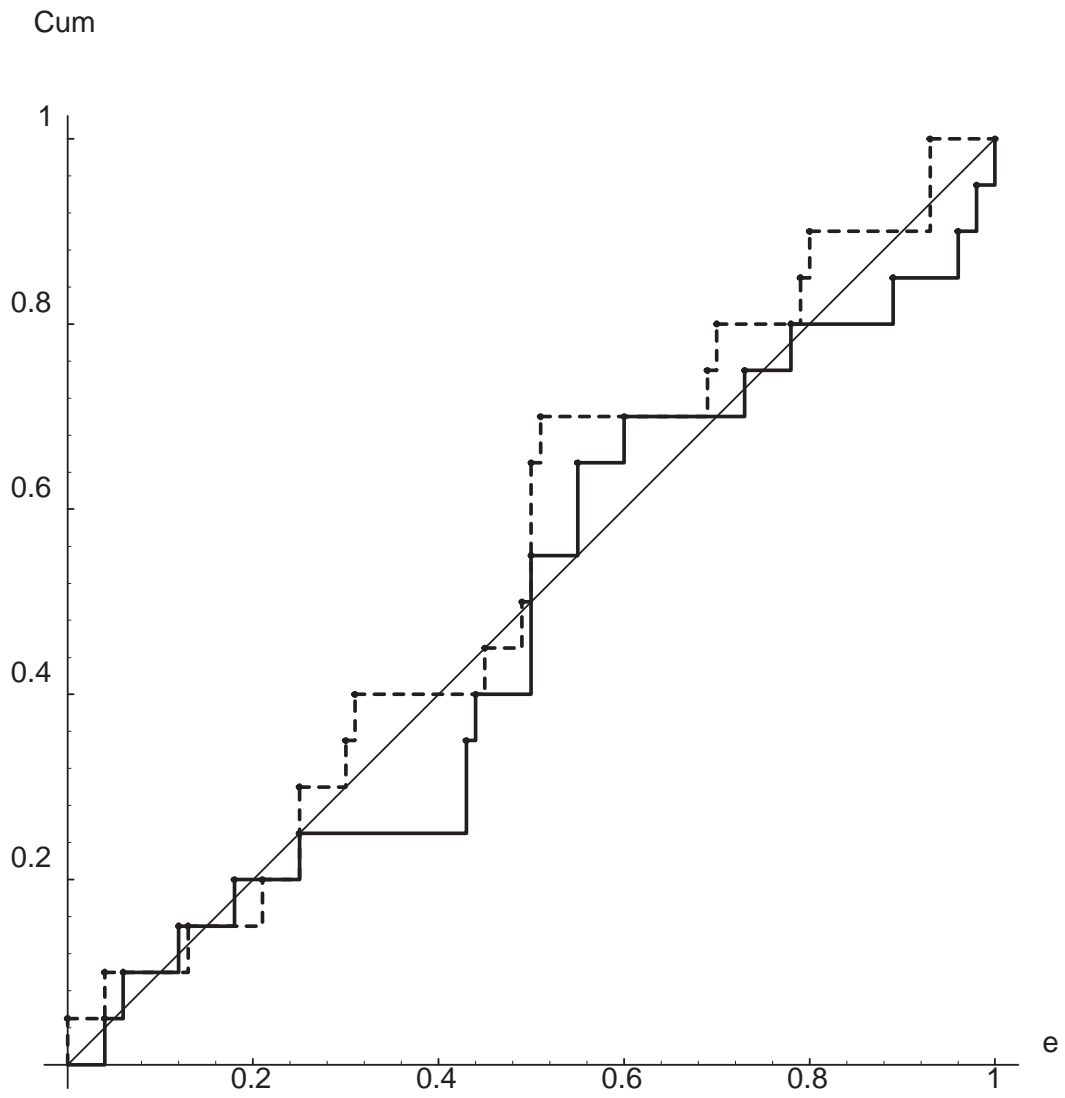


Figure 6: Period 1 empirical distribution functions for treatments $G(2.44)$, solid line, and $G(3.85)$, dashed line.

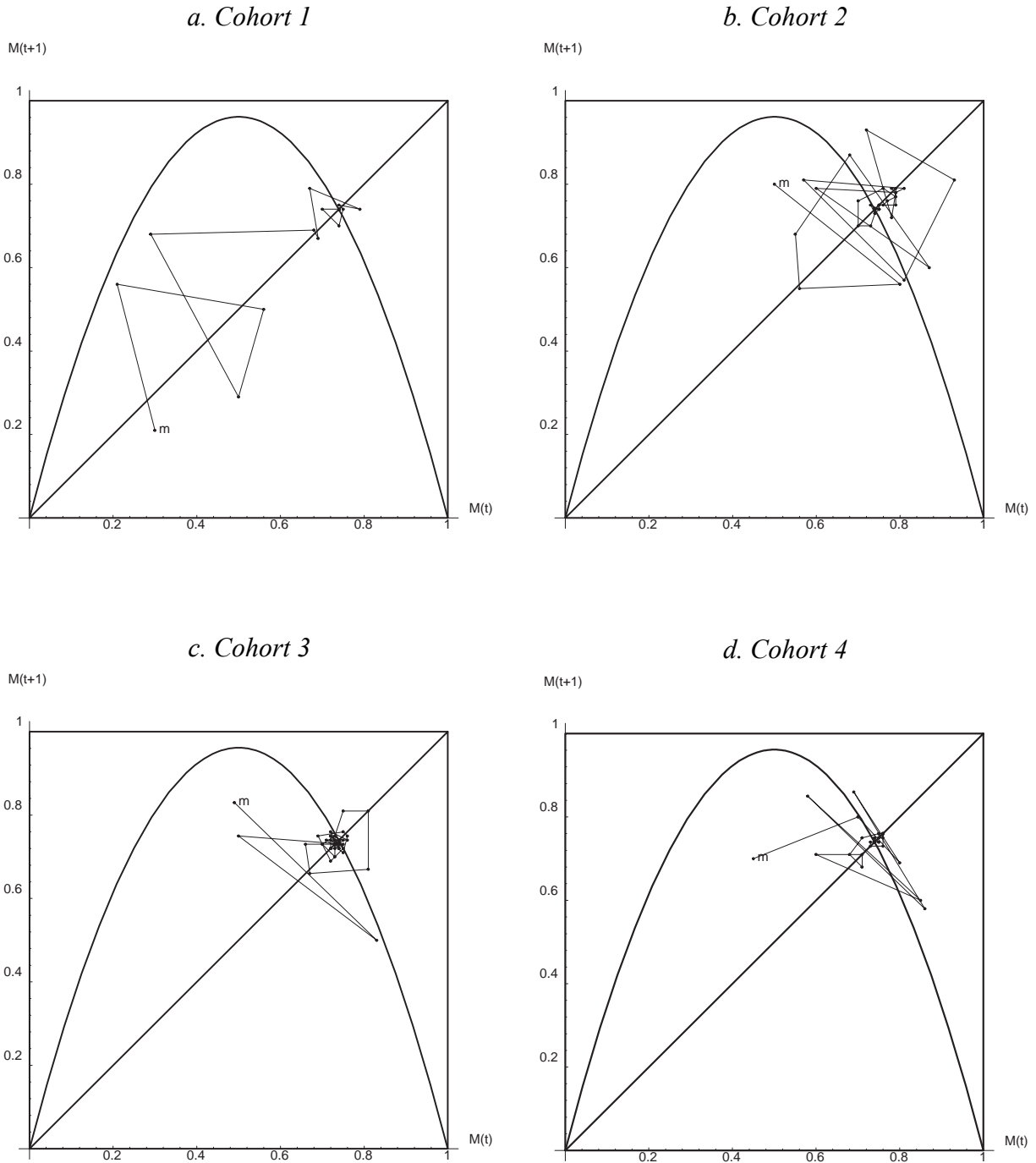


Figure 7: Observed medians for $G(3.85)$ graphed in the phase space; *a.* Cohort 1, *b.* Cohort 2, *c.* Cohort 3, *d.* Cohort 4.

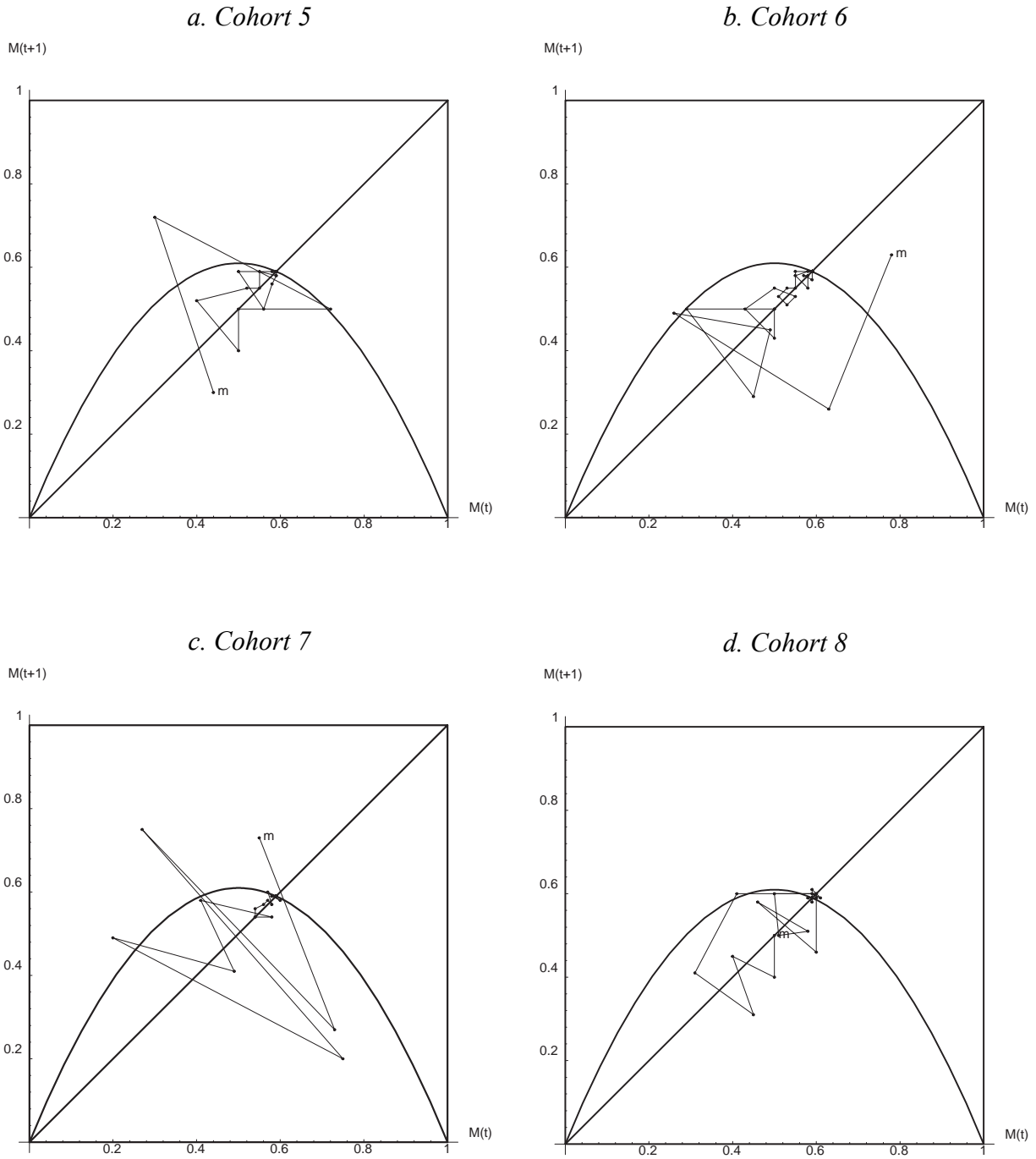


Figure 8: Observed medians for $G(2.44)$ graphed in the phase space; *a.* Cohort 5, *b.* Cohort 6, *c.* Cohort 7, *d.* Cohort 8.

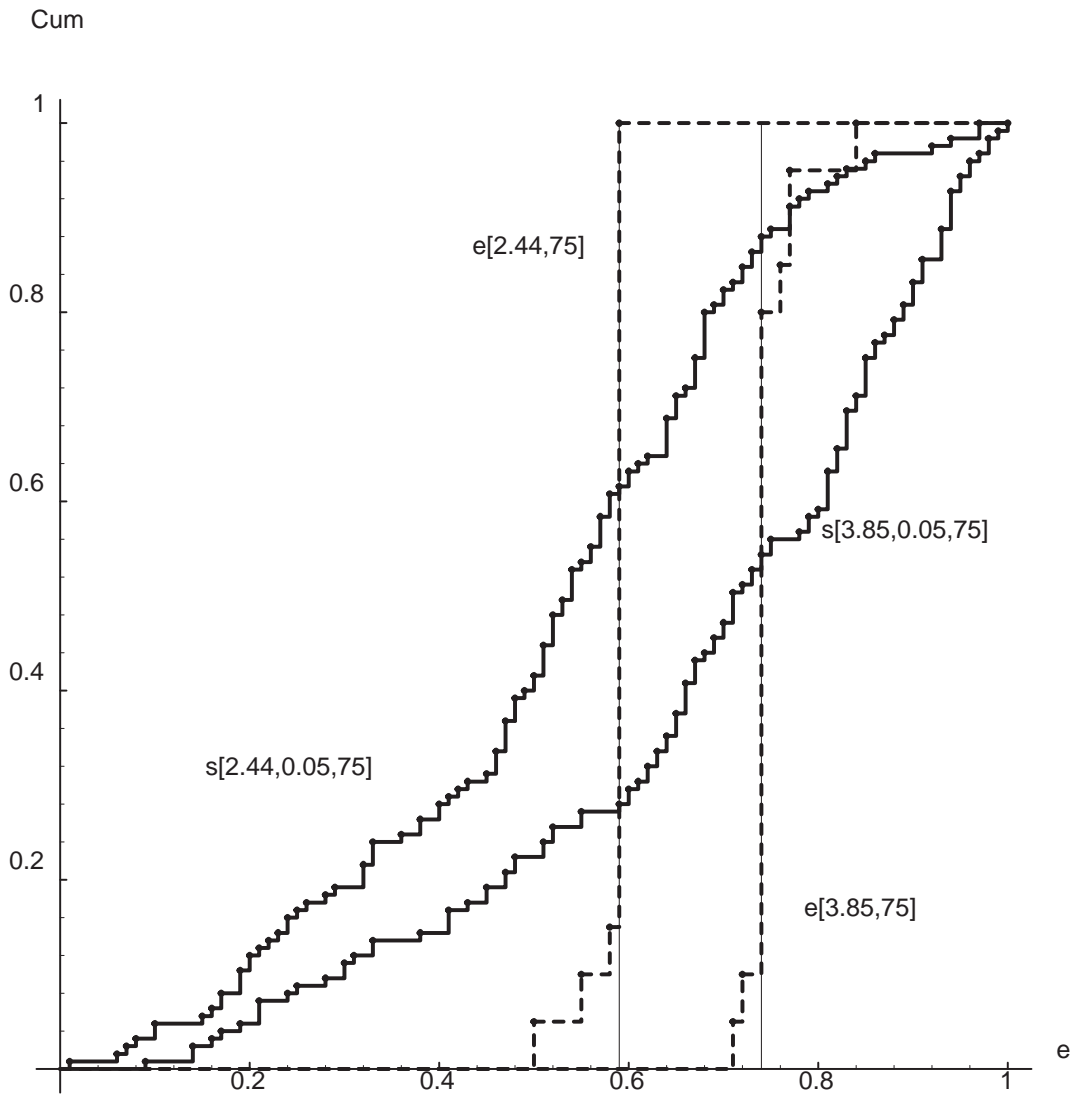
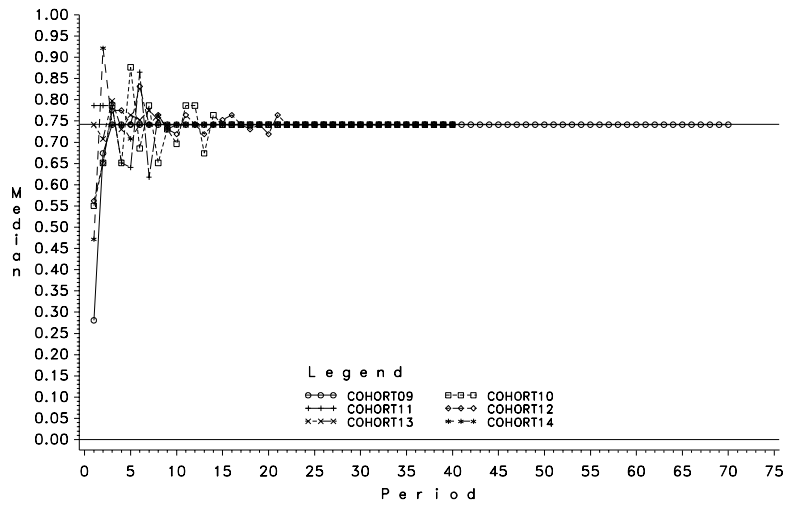
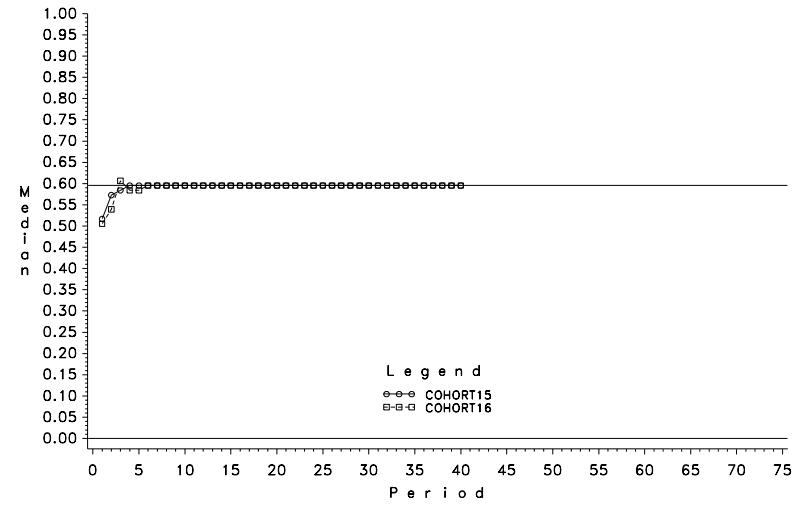


Figure 9: Period 75 empirical and simulated distribution functions of actions for treatments $G(2.44)$ and $G(3.85)$. Dashed lines are the empirical distribution functions labeled $e[\omega, t]$. Thick solid lines are the simulated distribution functions labeled $s[\omega, \alpha, t]$. Thin vertical lines denote the interior equilibrium prediction.

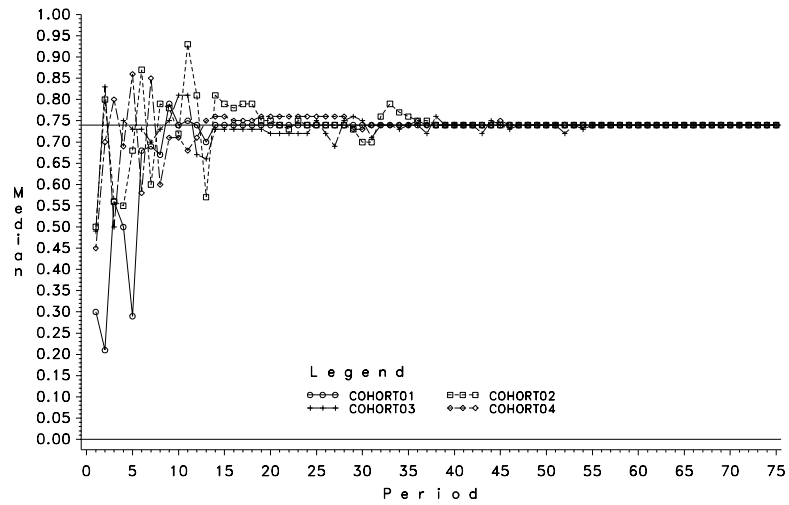
a. $G(3.87)$ and best response information.



b. $G(2.47)$ and best response information.



c. $G(3.85)$ and reinforcement information.



d. $G(2.44)$ and reinforcement information.

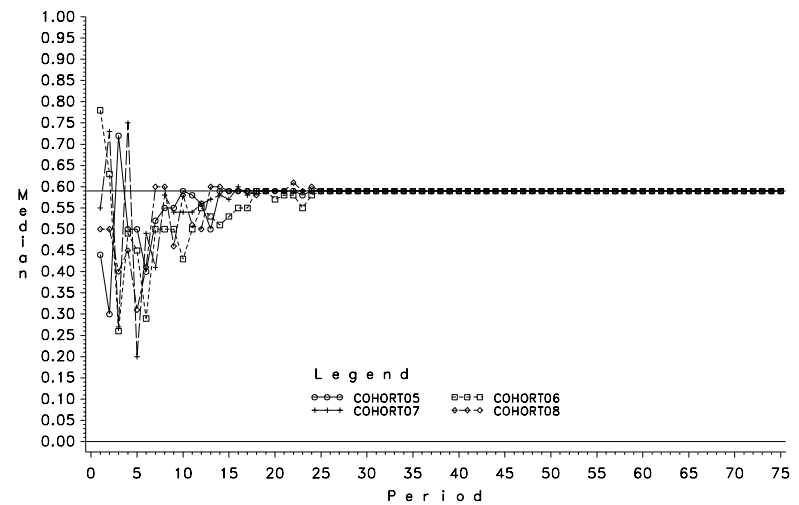


Figure 11: Median time series by treatment. Horizontal lines denote the interior and corner equilibrium.