Action-Reinforcement Learning versus Rule Learning

by

Dale O. Stahl

Malcolm Forsman Centennial Professor
Department of Economics
University of Texas
stahl@eco.utexas.edu

December 2000

ABSTRACT

How well do various learning models predict the dynamics of the population distribution of play in a variety of games? We measure and compare the in-sample and out-of-sample prediction performance of seven action-reinforcement learning models as well as Rule Learning for symmetric normal-form games. The criteria include log-likelihood values and root mean squared error. We conclude that a logit best-reply dynamic with inertia and adaptive expectations (LBRIAE) is best among the action-reinforcement models. We formulate a population Rule Learning model that nests the LBRIAE model and find that it is statistically superior. Sophistication is learned in the sense that random behavior and herd behavior decline substantially, and level-1 and level-2 behaviors increase over time.

---

*Partial funding of this research was provided by grant SBR-9410501 and SBR-9631389 from the National Science Foundation and the Malcolm Forsman Centennial Professorhip, but nothing herein reflects their views. I am especially indebted to Ernan Haruvy who helped design and run the experiments and provided invaluable intellectual stimulation. All errors and omissions are my sole responsibility.

Action-Reinforcement Learning versus Rule Learning

by

Dale O. Stahl

Malcolm Forsman Centennial Professor
Department of Economics
University of Texas
stahl@eco.utexas.edu

February 2002

ABSTRACT

How well do various learning models predict the dynamics of the population distribution of play in a variety of games? We measure and compare the in-sample and out-of-sample prediction performance of seven action-reinforcement learning models as well as Rule Learning for symmetric normal-form games. The criteria include log-likelihood values and root mean squared error. We conclude that a logit best-reply dynamic with inertia and adaptive expectations (LBRIAE) is best among the action-reinforcement models. We formulate a population Rule Learning model that nests the LBRIAE model and find that it is statistically superior. Sophistication is learned in the sense that random behavior and herd behavior decline substantially, and level-1 and level-2 behaviors increase over time.

Word Count = 14,781

---

*Partial funding of this research was provided by grant SBR-9410501 and SBR-9631389 from the National Science Foundation and the Malcolm Forsman Centennial Professorhip, but nothing herein reflects their views. I am especially indebted to Ernan Haruvy who helped design and run the experiments and provided invaluable intellectual stimulation. All errors and omissions are my sole responsibility.

1. Introduction.

One of the major questions for economic game theory in recent years has been "how do humans learn to behave in strategic situations?" Since the demise of the eductive refinement program of game theory, hope has arisen that dynamic learning models might be able to resolve the predictive impotency of equilibrium theories. The approaches to learning in games include: evolutionary game theory and replicator dynamics[1], reinforcement learning and belief learning[2], and Rule Learning[3]. Given the proliferation of learning models in the literature, there is now a critical need to apply model selection methodolgies. There have been a number of studies in the literature that compare models of learning; however, none of these studies is sufficiently comprehensive.[4]

We embrace the instrumentalist position [Friedman (1953)] that the ultimate test of a model is its ability to predict behavior. Accordingly, we present a "horse race" comparing the in-sample and out-of-sample predictive performance of a group of learning models on a large data set. The models are fitted using maximum likelihood methods, and then several measures of performance are computed (maximized likelihood values, and mean squared errors)

Our ultimate goal is to find a learning theory that is applicable to real world situations for which game theory is often applied. For example, consider marketing, which involves multiple stages with substantial information sharing among the participants between stages through trade magazines and word-of-mouth communciation. The pairwise matching protocol in which each player's information is limited to the pair of choices corresponding to his own sequence of matches is not well-suited for these applications. In contrast, this paper focuses on data from experiments in

which the participants receive information about the whole population of players rather than just one coplayer.

This paper also focuses exclusively on models of population dynamics rather than individual player dynamics for several reasons. *First*, in economic applications such as marketing, the payoff function is the discounted present value of revenue minus cost, which is invariant to variations in individual behavior that leave the aggregate payoff unchanged. Thus, a model of population dynamics would suffice, and individual data may not be readily available. Hence, there is a need for a theory of population dynamics. *Second*, Stahl (2000b) shows that aggregating up from individual player dynamics does not provide better predictions of population dynamics.[5] *Third*, when a player faces a sequence of co-players drawn from a large population of potential co-players, decision theory tells us that it is sufficient for the player to have a belief about the potential choices of those co-players: a subjective probability assessment of the likelihood of the potential action choices of the co-players. Further, if the likelihood of these co-player choices changes over time, then it is sufficient to have a consistent method of updating one's subjective probability assessment. It is *not* necessary to be able to track the choices of any specific co-player. In other words, for a rational player, it is sufficient to have a subjective belief about the distribution of choices in the population of co-players and a model of the dynamics of that population distribution.

In section 2, we present the action-reinforcement learning models that are investigated. By "action-reinforcement learning", we refer to the class of dynamic models in which the objects of reinforcement are the actions available to the players in a one-shot game. Specifically, we present population versions of (i) replicator dynamics, (ii) Roth-Erev reinforcement learning,

(iii) Camerer-Ho experience weighted attraction dynamics, and (iv) a Logit-Best-Reply with Inertia and Adaptive Expectations (LBRIAE) model.

In section 3, we describe the data and the pertinent econometric details. We focus on experiments in which the players receive population feedback and the game payoffs are binary lotteries, and we employ standard maximum likelihood techniques for all models, providing widely accepted criteria for model comparisons. In section 4, we present econometric comparisons of the action-reinforcement models and compute three measures of goodness-of-fit both in-sample and out-of-sample. The LBRIAE model emerges as the winner of the horse race among these models, and we explore why.

An obvious shortcoming of action-reinforcement models is that nothing can be learned that the player can transfer to a new but similar situation.[6] For example, the population may converge to action A in the first game, but in the next game action A might be dominated. Hence, when the game changes, the learning model must be revised and reinitialized. Yet, researchers who are university professors know very well that what we teach (and hopefully students learn) are ways of thinking about problems: high level algorithms for recognizing essential features and solving problems.

Also, in contrast to standard game theory, these models assume that players are rather unsophisticated. At one extreme, the players in replicator dynamics could just as well be thoughtless amoeba; learning is a byproduct of natural selection, not a conscious internal process in the players' minds. Most of the learning dynamics ignore much of the information that human players might be assumed to have available: e.q. the history of play and knowledge of the game. Such information would enable them to compute (a) hypothetical payoffs from actions they could have chosen but did not, (b) the best reply to the recent past, and (c) iterates of the best reply mapping.

There is evidence that people do use all available information.[7]

In this spirit, the Rule Learning models of Stahl (1996, 1999, 2000a-b) hypothesize a rich space of *behavioral rules* which players can learn based on performance feedback. These rules span several levels of sophistication (level-n bounded rationality, as well as herd behavior and Nash behavior). There is growing evidence in favor of Rule Learning [Nagel (1995), Ho, *et al* (1996), Duffy and Nagel (1997), Stahl (1996, 1999, 2000a-b), and Sosino, *et al* (1998)].

In section 5, we present a population version of Rule Learning that nests the LBRIAE model which is the best of the action-reinforcement models. Rule Learning adds three features to the LBRIAE model: (i) the use of additional information (as embodied in additional rules), (ii) diversity in the population, and (iii) performance feedback on the probability distribution of rule use in the population. By virtue of this nesting, we can analyze the contribution of each feature. In section 6, we present the econometric analysis of this Rule Learning model, and show that each feature is statistically significant. Conclusions are drawn in section 7.

2. Canonical Action-Reinforcement Dynamics for Finite Human Populations.

2.1. Replicator Dynamics for Human Populations.

Since replicator dynamics [e.g. Hofbauer and Sigmund, 1988] is perhaps the best known model for population dynamics, it is natural to begin our inquiry with it. The basic biological population model deals with a population of J types, with $n_j(t)$ denoting the number of type j's in period t, and a total population of $\bar{n}(t)$ in period t; the proportion of type j's in period t is denoted $p_j(t) \equiv n_j(t)/\bar{n}(t)$. The biological interpretation is that each type

is encoded in some set of genes, and that the population distribution changes because of differential survival (or "fitness") rates. The reproductive rate of a type depends on the population distribution of types and can be represented by a strictly positive J×J matrix U, so the net growth rate of type j is $e_j'Up(t)$, where $e_j \in \Re^J$ is the column vector with 1 in the jth column and 0 elsewhere. Then, the number of type j's at time $t+\Delta t$ is given by

$$n_j(t+\Delta t) = n_j(t)[1 + e_j'Up(t)\Delta t] \quad . \tag{1}$$

Letting the time interval, Δt, go to 0, the continuous-time version of replicator dynamics can be derived as

$$\dot{p}_j(t)/p_j(t) = [e_j - p(t)]'Up(t) \quad . \tag{2}$$

However, the biological interpretation of replicator dynamics hardly seems applicable to human populations playing economic games. The frequency of play in a population, which is how we interpret p(t) in eqs(1-2), is obviously not changing due to a payoff sensitive birth and death process among players (at least not in an economically relevant time frame). However, following Dawkins (1976), we can imagine a mental birth and death process of ideas (or "memes") in an individual player's mind. Under this interpretation, let $\varphi(t)$ denote the *underlying choice probability* for a randomly drawn player from the population; then, eq(2) becomes

$$\dot{\varphi}_j(t)/\varphi_j(t) = \beta[e_j - \varphi(t)]'Up(t) \quad , \tag{3}$$

where we introduce a free scaling parameter, $\beta > 0$, since (unlike biological

models) there is no reason that the game payoffs necessarily translate into realistically scaled growth rates.

While memes offer a more palatable interpretation of replicator dynamics for human populations, we must now face the problem that while we can observe the frequency distribution of choices, $p(t)$, we cannot observe the true underlying probabilities, $\varphi(t)$, in a finite population.[8] It follows that we can only specify the likelihood of the observed frequencies, so maximum likelihood procedures are natural.

To derive a version of eq(3) that can be applied to discrete-time data, we integrate eq(3) over Δt, yielding

$$ln[\varphi_j(t+\Delta t)] \approx ln[\varphi_j(t)] + \beta e_j'Up(t)\Delta t + a_t \ , \tag{3'}$$

where a_t is a scalar (independent of j) which is determined by the requirement that $\Sigma_j \ \varphi_j(t) = 1$; w.l.o.g. we can choose time units so $\Delta t = 1$. Solving eq(3′) gives the "logit form" of discrete-time mental replicator dynamics:

$$\varphi_j(t) = exp[w_j(t)]/\Sigma_k \ exp[w_k(t)] \ , \quad \text{for } t \geq 1, \tag{4a}$$

$$\text{and} \quad w_j(t) = w_j(t-1) + \beta e_j'Up(t-1) \ , \quad \text{for } t > 1. \tag{4b}$$

We interpret $w_j(t)$ as the *strength* of action j in period t. Since $exp\{\beta e_j'Up(t)\} > 0$, the $\varphi_j()$ quantities defined by eq(4a) are always strictly positive - hence proper probabilities.

To complete the mental replicator model we must specify the initial conditions for period 1: $w(1)$. To let $w(1)$ be a free variable would add J-1 parameters to be estimated from the data reducing the degrees of freedom

accordingly, so it is worthwhile considering alternatives. The first is to let w(1) be uniform on the basis of insufficient reason, but this is likely to be a bad fit of the data and could bias the estimate of β. The second alternative is to roll back insufficient reasoning one period, and use eq(4b') for t = 1, taking p(0) to be uniform and w(0) = 0. For notational convenience, we let p^0 denote the uniform distribution over the J actions. Then $w(1) = \beta U p^0$, so $\varphi(1)$ will differ from the uniform distribution in that actions that would perform better against a uniform prior are more likely and vice versa; hence, the first period fit is likely to be much better [see Stahl and Wilson, 1995] with no sacrifice in degrees of freedom.

2.2. Propensity-Based Reinforcment Models.

Propensity is a concept of strength [Bush and Mosteller, 1955; Luce, 1959] such that the probability of choosing an action is monotonic function of the propensity. Letting $w_j(t) > 0$ denote the propensity for action j in period t, the generic specification of choice probabilities is

$$\varphi_j(t) = f(w_j(t))/\Sigma_k f(w_k(t)) \ ,$$

where f() is a positive monotonic function. The two most common functional forms are (i) the *power form* in which $f(x) = x^\lambda$, for some strictly positive scalar λ; and (ii) the *logit form* in which $f(x) = exp(x)$. The only power-form specification considered in this paper will be the Erev-Roth model.[9]

The "Law of Effect" dynamic specifies how propensities evolve:

$$w_j(t) = \theta w_j(t-1) + g_j(t-1) \ , \quad \text{for } t > 1, \tag{5}$$

where $g_j()$ is the "reinforcement function", and $\theta \in [0,1]$ is an *inertia* variable that discounts past propensities. Initial conditions specify $w(1)$. Different versions of this generic propensity-based reinforcment model arise from different specifications of $g()$, such as in the Erev-Roth model and the Camerer-Ho EWA model. Then, letting $\theta = 1$, $g_j(t-1) = \beta e_j'Up(t-1)$, eq(5) becomes identical to replicator dynamics, eq(4). Thus, the propensity model is a generalization of mental replicator dynamics.

Given $\theta < 1$, eq(5) is equivalent to the standard *partial adjustment dynamic* applied to propensities: To see this, define $\gamma(t) \equiv g(t-1)/(1-\theta)$, so eq(5) can be rewritten as

$$\begin{aligned} w(t) &= \theta w(t-1) + (1-\theta)\gamma(t) \\ &= w(t-1) + (1-\theta)[\gamma(t) - w(t)] \end{aligned} \tag{5'}$$

where $\gamma(t)$ is called the "target propensity" towards which a partial adjustment is made.

Nerlove (1958) showed that when $\gamma(t)$ is a linear function of $p(t-1)$, say $\Gamma[p(t-1)]$, then eq(5′) also describes the dynamics of an *adaptive expectations* model in which the current propensity is a complete adjustment to the target propensity based on the expectation of $p(t)$:

$$w(t) = \Gamma(E[p(t)|t-1]) \ , \quad \text{for } t \geq 1, \tag{6a}$$

and

$$E[p(t)|t-1] = \theta E[p(t-1)|t-2) + (1-\theta)p(t-1), \text{ for } t > 1. \tag{6b}$$

Then, $w(t) = \theta w(t-1) + (1-\theta)\Gamma[p(t-1)]$ for $t > 1$, which is equivalent to eq(5′). Thus, the generic propensity-based reinforcement model (given $\theta < 1$) is equivalent to the standard partial adjustment model, which (given that $\gamma(t)$

is a linear function of p(t-1)) is equivalent to the adaptive expectations model.[10]

This adaptive expectations model is often referred to as "belief learning" [e.g. Cheung and Friedman (1994) and Camerer and Ho (1997)]. For instance, if $\Gamma(p) = \beta Up$, then eq(6a) defines propensities as scaled expected utilities. Further, in the logit-form, the choice probabilities, φ, can be interpreted as logit best-replies to beliefs given by adaptive expectations.

2.2.1 Erev-Roth Reinforcement Learning.

Roth and Erev (1995) put forth a version of propensity-based reinforcement learning that applies to individuals. We aggregate the extended Erev-Roth (1996) model[11] to derive a population version. The aggregate reinforcement function is the expected reinforcement function for a random player conditional on the actual population distribution of play in period t:

$$g_j(t) = (1-J\tilde{\mu})p_j(t)y_j(t) + \tilde{\mu}\bar{y}(t) \ , \tag{7a}$$

where

$$y_j(t) \equiv \left(\frac{n}{n-1} e_j'Up(t) - \frac{U_{jj}}{n-1} - \rho(t) \right) \ , \text{ and } \bar{y}(t) \equiv p(t)\cdot y(t) \ , \tag{7b}$$

and

$$\rho(t) = \rho(t-1) + b^{+}\max\{\bar{y}(t-1),0\} + b^{-}\min\{\bar{y}(t-1),0\} \ , \tag{7c}$$

and $\tilde{\mu} \equiv \mu/(J-1)$ with $\mu \in [0,1)$ is a scalar "spillover" parameter.

Eq(7a) defines a reinforcement function consisting of two components. The first is the "subjective reward", $y_j(t)$, which is the average payoff to any individual who chose j in period t less the average aspiration level for period t, $\rho(t)$, which adjusts according to a potentially asymmetric adaptive

expectation process with $b^+ \leq b^-$. The second component is the population average subjective reward $\bar{y}(t)$. The reinforcement of action j is roughly j's share of the population average subjective reward, plus a spillover portion ($\tilde{\mu}$) of the average subjective reward in the whole population.

Then, the population choice probabilities are defined by

$$\varphi_j(t) = w_j(t)/\Sigma_k w_k(t) \text{ , and} \tag{8a}$$

$$w_j(t) = \max\{\theta w_j(t-1) + g_j(t-1) \text{ , } \eta\} \text{ ,} \tag{8b}$$

where $\eta = 0.0001$ is a fixed parameter to avoid boundary problems.

For initial conditions, Erev-Roth specify the uniform distribution for period 1, and use eqs(7 and 8) for $t > 1$, with $w_j(1) = S\bar{u}/J$ and $\rho(1) = \rho_1\bar{u}$, where $\bar{u}$ is the mean payoff $p^0{}'Up^0$ of the game. This population version of the original Erev-Roth model has 6 parameters (θ, μ, b^-, b^+, S, ρ_1).

An alternative in the spirit of section 2.2 is to apply the principle of insufficient reasoning to period 0, taking $w_j(0) = S\bar{u}/J$, $p(0) = p^0$, and $\rho(0) \equiv \rho_0\bar{u}$; then using eq(7a,b) to determine g(0), and using eq(8) and eq(7c) to determine w(t) and $\rho(t)$ respectively for $t \geq 1$. This version of the Erev-Roth model also has 6 parameters: (θ, μ, b^-, b^+, S, ρ_0).[12]

2.2.2. <u>The Experience Weighted Attraction Model</u>.

Camerer and Ho (1996) put forth an innovative propensity-based reinforcement model that combines several features of the Erev-Roth reinforcement learning and belief learning models. An advantage of the Camerer-Ho model is that by restricting certain parameter values, a number of nested models can be tested. Their EWA model is an individual learning

model, so we aggregate to derive a population version. We also restrict attention to the logit form.[13] The underlying state variables, $w_j(t)$, which Camerer and Ho call "attractions", are updated according to the following dynamic:

$$w_j(t) = \{\theta N(t-1)w_j(t-1) + g_j(t-1)\}/N(t) \ ,$$

where $$g_j(t) \equiv \alpha e_j'\beta Up(t) + (1-\alpha)p_j(t)\beta[e_j'Up(t)n/(n-1) - U_{jj}/(n-1)] \ , \tag{9}$$

$$N(t) = \gamma N(t-1) + 1 \ , \text{ and } N(0) \equiv N_0.$$

The parameter α is the the probability that an individual evaluates the past performnce of all actions; Camerer and Ho call this the "imagination parameter". Notice that when $\alpha = 0$, the last term of eq(9) is similar to the reinforcement function of Erev-Roth, eq(7), sans the aspiration level term $\rho(t)$ and the spillover effect ($\tilde{\mu}$), and when $\alpha = 1$, eq(9) resembles the Luce propensity-based model, eq(5).

The N() term is called the "experience weight". If $N_0 < 1/(1-\gamma)$ and $\gamma \in (0,1)$, then $N(t)/N(t+1)$ declines, putting less weight on new evidence as time passes (the power law of practice), and vice versa. On the other hand, if $\gamma = 0$ or $\gamma = (N_0-1)/N_0$, then N(t) is constant for all t, and since N_0 and β are then not separately identifiable, we can eliminate N(t) entirely from eq(9).

For initial conditions, Camerer and Ho let $\{w_j(1)\}$ be free parameters. Rather than increase the number of free parameters, we take the approach of insufficient reasoning used for the mental replicator and Erev-Roth models, specifying $w_j(0) = 0$ for all j and $p(0) = p^0$, and using eq(9) for period 1. Thus, the logit form of EWA dynamics has five parameters in all: (θ, α, β, γ,

and N_0).

2.3. Logit Best-Reply Dynamics.

As in Kandori, Mailath and Rob (1993), suppose that with probability δ the individual sticks to the action mostly recently played, and with probability $1-\delta$ the individual evaluates the performance of all the actions and switches to the action that appears to be the best reply to the recent past. In other words, the crucial state variable defining the dynamics is the action most recently chosen; if a player chose action j, then we consider him a "type j". A type j player is true to his type or becomes the type corresponding to the apparent best reply. Further, suppose that the apparent best reply is given by the logit best-reply function.[14] These assumptions lead to the following two-parameter *Logit Best-Reply with Inertia* (LBRI) model:

$$\varphi(t) = \delta p(t-1) + (1-\delta) b[p(t-1),\beta] \ , \quad \text{for } t \geq 1,$$

$$\text{where} \quad b_j(p,\beta) = exp(\beta U_j p)/[\Sigma_k \ exp(\beta U_k p)] \tag{10}$$

is the logit best-reply function with precision parameter β, and $p(0) = p^0$. Naturally, the inertia parameter δ is confined to [0,1]. Because of the properties of $b(\cdot,\beta)$ for large values of β, this dynamic can exhibit periodic behavior as well as chaotic behavior. For smaller values of β, $b(,\beta)$ tends to the uniform distribution, which represents a very noisy evaluation process.

An unrealistic feature of this LBRI model is the implicit assumption that last period's frequency distribution of choices is the player's forecast of next period's. Typically, one can reject this static forecasting model in favor of an AR1 model: $E[p(t)|t-1] \equiv$

$$q(t;\theta) = \theta q(t-1;\theta) + (1-\theta)p(t-1), \tag{11}$$

starting with $q(0;\theta) = p(0) = p^0$. As a forecasting model, eq(11) often fits empirical data quite well, and will thereby serve as a useful non-learning (or mindless) benchmark.

Combining adaptive expectations with the logit best-reply dynamic by substituting $q(t-1;\theta)$ for $p(t-1)$ in eq(10), we have a three-parameter Logit Best-Reply with Inertia and Adaptive Expectations model:

$$\varphi(t) = \delta q(t-1;\theta) + (1-\delta)b[q(t-1;\theta),\beta] \ , \text{ for } t \geq 1. \tag{12}$$

For brevity, we will refer to eq(11-12) as the *LBRIAE* model. The adjustment parameters θ and δ are naturally confined to [0,1], while the precision parameter β is assumed to be positive. Note that when $\theta = 0$, eq(12) reduces to eq(10), because adaptive expectations become instantaneous.

This LBRIAE model can be interpreted as a mixture of belief learning and herd behavior with inertia. That is, eq(11) describes the evolution of beliefs, to which a fraction $1-\delta$ of the population choose a logit best reply, while the remaining fraction δ of the population "follows the herd", where following the herd means that with probability θ the player sticks with last period's choice and with probability $(1-\theta)$ the player mimics the recent past play of the population.[15]

When $\delta = 0$, the LBRIAE model reduces to the belief learning dynamic given at the end of section 2.3, which is equivalent to the standard partial adjustment model and the generic propensity-based reinforcement model.

3. The Experimental Data.

As argued in the Introduction, we believe that it is important to test learning theories in environments with substantial population feedback between learning trials. Further, to guard against overfitting the models to unrepresentative samples, we choose data involving a variety of symmetric normal form games: specifically, data from two different experiments, both using binary lottery games.[16]

The first experiment consisted of four sessions, each with two runs of 15 periods each and one (of four) symmetric 5×5 game in each run. The payoff matrices are given in Appendix A. Each matrix was constructed to have a unique (pure-action) Nash equilibrium, and the distinct first-period level-1 and level-2 behaviors, because these features are most likely to "stress-test" the learning theories. There were 22-24 participants in each session (all inexperienced). For more details, see Stahl (2000a). There was convergence to the Nash equilibrium in only two of these eight runs, so this data set has "rich" out-of-equilibrium dynamics.

The second experiment consisted of four sessions, each with two runs of 12 periods each and one symmetric 3×3 game in each run. The payoff matrices are given in Appendix A. The game in the first run of each session had multiple Nash equilibria for the purpose of investigating selection theories [see Haruvy and Stahl, 1998]; two of the second run games have no symmetric pure-strategy Nash equilibria, and so convergent behavior was not anticipated over the 12 period horizon. There were 24-25 participants (all inexperienced) in each session.

For both experiments, in each period each participant was matched with the n-1 other participants and was given the history of choices of those n-1 other players. The binary lotteries were not resolved until all periods of both

runs were completed. The choice data can be obtained at www.eco.utexas.edu/faculty/Stahl/experimental. Plots of the population choice frequencies over time for each game are given in Appendix D.

4. Econometric Comparisons.

Since our goal is a successful and parsimonious model of learning, the results reported here are based on pooling the data sets; in other words, one set of parameters is estimated for the entire pool of data. Generally, in our experience, there is considerably more heterogeneity in the participant population than any of these models can possibly capture, so not surprizingly this homogeneity assumption can be statistically rejected. Notwithstanding mis-specifications regarding homogeneity, we are keenly interested in the relative performance of the alternative parsimonious reinforcement models. Moreover, the comparative results reported here still hold when the analysis is done separately on the 5×5 and the 3×3 data sets.

The models of Section 2 define a likelihood function that can be maximized with respect to the free parameters. For a finite population of size n, let $n_r(t)$ denote the J×1 integer-valued vector of observed choices of each action at time t in "run r". Then, $n_r(t)$ is multinomially distributed with probability vector $\varphi_r(t)$. Hence, the likelihood of $n_r(t)$ given $\varphi(t)$ is proportional to

$$L[n_r(t)|\varphi_r(t)] \equiv \Pi_j \, \varphi_{rj}(t)^{n_{rj}(t)} \quad . \tag{13}$$

Further, since the probabilities $\varphi_r(t)$ are conditional on $n_r(t-1)$, the likelihood of the entire dynamic path is simply the product of these period-by-period likelihoods; and the likelihood of the entire data set is obtained

by taking the product over all 16 runs. Let LL denote the logarithm of this total likelihood value.

Before estimating the models of Section 2, it is worthwhile presenting several benchmarks for the log-likelihood (LL) values on our data. The completely uninformative model (equal probable outcomes) has LL = -7086.77, while a one-parameter logit Nash equilibrium model has LL = -6660.24. We will see that all the learning models do substantially better. At the other extreme, the best a model could do is predict $\varphi(t) = p(t)$, in which case LL = -4295.95.

For each model, the log-likelihood function for the entire pooled data set is maximized with respect to the parameter values. As a way of separating out the first period effects from the dynamic effects, we also compute the LL value for all but the first period of each run. As another measure of fit, we compute the Root Mean Squared Error (RMSE): the Euclidean distance between the empirical choice frequences, $p(t)$, and the predicted choice probabilities, $\varphi(t)$, summed over all periods. This is a "one-period-ahead" quadratic scoring measure:[17] i.e. the period t forecast is based on the actual data from period t-1 in contrast to a simulated data path from period 1 to period t-1.

We considered using the Pearson Chi-square measure of goodness-of-fit (PCS) given by

$$\Sigma_t \Sigma_j [n_j(t) - n\varphi_j(t)]^2/n\varphi_j(t) \quad . \tag{14a}$$

However, PCS can be unreliable for small sample sizes. A rule of thumb is that each cell should have at least 5 observations. However, in many of our games there are many periods in which this criterion is not satisfied: e.g. when the players converge to an equilibrium, there will be many actions that

are seldem chosen in later periods. Therefore, we do not report PCS measures here.

Similarly, T-period-ahead likelihood measures are unreliable for small sample sizes. For instance, we could compute the likelihood of the actual data in the final period (T) of run r conditional on only the estimated parameters by integrating out the history of play for periods 1 to T-1:

$$L_r[n_r(T)] = \int \ldots \int \Pi_{t=1}^{T} L_r(n_r) dn_r(T-1) \times \ldots \times dn_r(1) \quad , \qquad (14)$$

approximating this integration by simulating paths of play. With J actions and N participants, there are $M \equiv (N+J)!/(N!J!)$ possible realizations of $n_r(T)$; e.g. M=3276 when N=25 and J=3. Thus, eq(14) defines a probability distribution with M-1 degrees of freedom. Clearly, there is no valid way that one or two observations (as in our case) can be used to assess the goodness-of-fit of this probability distribution. Therefore, we do not report T-period-ahead likelihood measures here.

4.1 The Models of Section 2.

The results of maximum likelihood estimation of the seven base models of section 2 are given in Table I. Note that the LL values are roughly increasing as one moves down the Table. The logit best-reply dynamic with inertia and adaptive expectations (LBRIAE) has the largest LL value for the entire pooled data as well as the data excluding period 1 of each run "LL(-1)", and the highest T-period-ahead goodness-of-prediction value. EWA dynamics is a close second in this initial heat of the horse race. It is also noteworthy that the mindless AR1 model, eq(11), does better than mental replicator dynamics. However, model comparisons are of dubious value unless

all the models being compared are on an equal footing with regard to errors/deviations.

4.2 Models with Trembles and/or Mutations.

Trembles are transient departures from the underlying choice probabilities, $\varphi(t)$, that affect only the current period choices and not the underlying propensities, $w(t)$. To model trembles, we keep the dynamics on $w_j()$ and $\varphi_j()$ unchanged from section 2, and we define post-tremble choice probabilities, $\varphi_j^*(t)$, as

$$\varphi_j^*(t) \equiv (1-\varepsilon)\varphi_j(t) + \varepsilon/J \ , \tag{15}$$

where $\varepsilon \in [0,1]$ is a free parameter.

Mutations, on the other hand, are permanent departures and hence can be modeled as changes in the underlying propensities, $w(t)$. For the logit form of the EWA model, mutations are already incorporated by virtue of the θ and β parameters, which attenuate the log-likelihood ratios among the available actions. Similarly, the δ and β parameters of the LBRI and LBRIAE models implicitly incorporate mutations, so no further modification is needed. However, the logit form of the mental replicator dynamics, eq(4), does require some modification to incorporate mutations. In the spirit of the other logit forms, it is natural to do this by introducing an inertia parameter θ, so

$$w_j(t) = \theta w_j(t-1) + \beta e_j' Up(t-1) \ . \tag{4'}$$

This modification makes mental replicator dynamics equivalent to the logit form of propensity-based reinforcement, eq(5), with $g(t) = \beta Up(t-1)$. Eq(4')

is also equivalent to the logit quantal response learning model adopted by Anderson, et. al. (1997). It is also a special case of EWA dynamics, eq(9), when $\alpha = 1$, $\gamma = 0$, and $n = \infty$.

The Erev-Roth model involves $\max\{\cdot, \eta\}$, with η fixed at 0.0001, to ensure that the predicted choice probabilities are positive. A natural way to model mutations is to allow η to be a free parameter to be estimated.

Obviously, we can also consider both mutations and trembles. Indeed, since the logit form of EWA dynamics implicitly entails mutations, using eq(15) effectively gives a model with both mutations and trembles. It would be unfair not to allow both forms of deviations/errors for all the models being considered.

We therefore considered both forms of deviations for all the models, and tested the significance of each additional parameter (using standard chi-square tests). We report in Table II, the "best" version from each model class in the sense that it is statistically superior to the base model and no other additions make a statistically significant improvement. These best versions are (i) both trembles and mutations for mental replicator dynamics, (ii) only mutations for both versions of the Erev-Roth model, (iii) trembles for the EWA model (which already implicitly contains mutations), and (iv) trembles for the AR1 and the logit best-reply models (which already include mutations). *Without exception, all learning models in Table II outperform the mindless AR1 model with trembles - the LBRIAE model dramatically so.*

We observe that the three parameter LBRI model performs substantially better than replicator dynamics and both versions of the Erev-Roth model. Also the four parameter LBRIAE model performs better than the LBRI model and the EWA dynamic.

We also report in Table II a four-parameter version of the logit EWA model

"w/o EW" in which the experience weights $N(t) = 1$ for all t, which reduces the set of free parameters to (θ, α, β, and ε).[18] This restriction leads to a decrease in the maximized log-likelihood value of only 0.43, which is clearly statistically insignificant, and, therefore the restricted version is much preferred.

4.3 Out-of-Sample Prediction.

A requirement of a successful empirical model is that it be capable of predicting behavior both in-sample (Table II) and out of sample. With a large data set, one can select a holdout sample to use for forecasting in numerous ways. To provide a robust test, we selected four different holdout samples: the first (second, third and fourth) sessions of each data set. Thus each holdout sample contains 25% of the 5×5 game data and 25% of the 3×3 game data. The residual 75% of the data is used to estimate the model and those parameter estimates are used to forecast the population behavior for the holdout sample. Each holdout sample not only involves completely different subjects than the residual data but also different games or order of games (see Appendix A).

Three measures of prediction success were computed: (i) the log-likelihood of the holdout data given the parameter estimates, (ii) the log-likelihood excluding the first period of each run, and (iii) the root mean squared error of the forecast. Table III gives the measures aggregated for the four holdout samples. The log-likelihood values are simple sums, while the RMSE measure is an average over holdout samples. These results reinforce those of Table II.

4.4 Model Selection.

As a basic model selection principle, if model A has a higher likelihood value both in-sample and out-of-sample and entails no more parameters than

model B, then we prefer model A over model B. We note that the preference sometimes encountered for a structual model over a reduced-form model is justified by the potential of the former to out perform the latter out-of-sample because of the ability to accomodate structural changes. However, all the models investigated in this paper use all the structural information available (the game and the history of play), and the premise of our selection principle is that A performs better than B out-of-sample, so a preference for B over A cannot be justified.

From Tables II and III, we can readily reject mental replicator dynamics, on the grounds that by the LL values, it is astronomically less likely to have produced the data than the simple three-parameter LBRI dynamic. We can also reject both Erev-Roth models on the grounds that despite having 7 parameters, it does not fit the data as well as the three-parameter LBRI model on every measure of fitness.

While the EWA model and the restricted EWA model perform better than the LBRI model, the former have respectively three and one more parameters. A fairer comparison is between the four-parameter EWA model and the four-parameter LBRIAE model, which is statistically superior to its nested cousin (the LBRI model). The pseudo-R^2 (the improvement in LL as a percentage of the maximum possible improvement over the uninformative model) of the LBRIAE is a respectible 0.829. Both in-sample and out-of-sample, the LBRIAE model performs better than the four-parameter EWA model on all measures. Therefore, according to our model selection principles, we declare the winner of the horse race to be the LBRIAE model.

4.5 Parameter Estimates and Hypothesis Tests.

Table IV contains the parameter estimates and bootstrapped standard

errors for the models of Table II. We note that most all the parameters are estimated fairly accurately; the one exception is b^+ in Erev-Roth reinforcement dynamics. With b^+ insignificantly different from zero, the aspiration level can never increase, and (given $b^- \approx 1$) it will decrease whenever actual payoffs fall below the aspiration level, which it does in many sessions since the initial asperation level is roughly the average payoff of the game.

The fact that the experience weights were found to be insignificant in the six-parameter EWA model but that the imagination parameter α is strongly significant reveals that the real contribution of the EWA model is this imagination parameter. Without the experience weights, the EWA model is equivalent to a logit propensity-based model with imagination.

We tested the hypothesis that $\alpha = 0$ (i.e. that only the most recently chosen action is reinforced), and found that the LL value decreased by 271.59; since twice this difference is distributed chi-square with one degree of freedom, we overwhelmingly reject that hypothesis. In other words, we strongly reject the Erev-Roth assumption that actions not chosen are not reinforced.

To test the hypothesis that $\alpha = 1$ (i.e. that all actions are reinforced), we merely need to compare the LL value for the EWA model with the LL value for logit mental replicator dynamics. The latter is worse by 208.66; so we overwhelmingly reject that hypothesis. In other words, the simple belief-learning model with logit best-replys and the quantal-response learning model [Anderson, et. al.] are strongly rejected in favor of the EWA model.

We also see that setting $\alpha = 1$ results in an increase in the estimate of the inertia parameter θ from 0.3269 (in the EWA model) to 0.6757 (in the logit replicator model). This suggests that a substantive effect of the α parameter

is to increase the impact of recent history on current propensities by adding $(1-\alpha)p(t-1)$ to the law of motion, eq(9). In contrast, recent history enters the logit replicator model only through $\theta w(t-1)$ which is proportional to $\theta\varphi(t-1)$ and correlated with $\theta p(t-1)$. Evidently, the EWA model does at better job of incorporating history than the logit replicator model. Moreover, this interpretation of $\alpha < 1$ is a plausible alternative to the interpretation as "lack of imagination."

For logit best-reply dynamics with inertia (LBRI and LBRIAE), we tested and strongly rejected $\beta = \infty$; in other words, the exact best-reply function is rejected in favor of the logit best-reply function in these partial adjustment models. The estimate $\hat{\beta}$ is of sufficient magnitude to put over 95% probability on the best reply when the best reply is 10 or more payoff points better; however, payoff differences of a few points do not have a significant impact on the choice probabilities. The estimate $\hat{\delta} = 0.615$ (and 0.605 respectively) indicates substantial inertia in actions.

To test for no herd behavior in the LBRIAE model (i.e. $\delta = 0$), we merely need to compare the LL value for the LBRIAE model with the LL value for logit mental replicator dynamics. The latter is worse by 220.21; so we overwhelmingly reject that hypothesis. By comparing the β estimates, we can infer that replicator dynamics tries (rather unsuccessfully) to make up for the absense of specific herd behavior by lowering the responsiveness to recent expected utilities, $\beta Up(t-1)$.

To test $\delta = 1$ in the LBRIAE model, we merely compare the LL value with that of the AR1 model. The latter is worse by 297.20; so we overwhelmingly reject that hypothesis. In other words, *both herd behavior and logit best-replies make very statistically significant contributions* to the fit of the data.

The utility scaling parameter β in the LBRI and LBRIAE models is substantially larger than the analogous parameter in the EWA model and the logit replicator model; evidently, by incorporating trends in the data specifically as additive herd behavior, the remaining population appears to be fairly accurate best-repliers to that trend. Further, the trend (or herd) part of the LBRIAE model responds rather quickly to new data: the weight on current observations is $(1-\hat{\theta}) = 0.7493$. In contrast, the mindless AR1 model is less responsive: $(1-\hat{\theta}) = 0.5980$.

Recalling that LBRIAE reduces to LBRI when $\theta = 0$, the estimate $\hat{\theta} = 0.251$ with a standard error of 0.028 demonstrates that θ is strictly positive. Further, the likelihood ratio test strongly rejects the LBRI model in favor of the LBRIAE model. The estimated tremble parameter $\hat{\varepsilon} = 0.076$ (and 0.053) is modest but vital in preventing erroneous near-zero probability predictions.

5. An Improved Population Rule Learning Model.

In this section we present a revision of the population Rule Learning model of Stahl (2000b) that encompasses the LBRIAE model and fits the data better than the former model. First, instead of incorporating herd behavior as an evidence-based rule, herd behavior is incorporated in the partial adjustment framework of the LBRIAE model. Second, since (as reported in section 4) we found that modelling trembles makes a very significant contribution to the fit of the data, we introduce an additional parameter for the probability of using the tremble rule in the first period.

5.1 The Rule Learning Framework.

A *behavioral rule* is a mapping from information Ω^t to $\Delta(A)$, the set of probability measures on the actions A. For the purposes of presenting the

abstract model, let $\rho \in R$ denote a generic behavioral rule in a space of behavioral rules R; $\rho(\Omega^t)$ is the mixed strategy generated by rule ρ given information Ω^t.

The second element in the model is a probability measure over the rules: $\varphi(\rho,t)$ denotes the probability of using rule ρ in period t. Because of the non-negativity restriction on probability measures, it is more convenient to specify the learning dynamics in terms of a transformation of φ that is unrestricted in sign. To this end, we define $w(\rho,t)$ as the *log-propensity* to use rule ρ in period t, such that

$$\varphi(\rho,t) \equiv \exp(w(\rho,t)/[\int \exp(w(x,t)dx] \ . \tag{16}$$

Given a space of behavioral rules R and probabilities φ, the induced probability distribution over actions for period t is

$$\hat{p}(t) \equiv \int_R \rho(\Omega^t)d\varphi(\rho,t) \ . \tag{17}$$

Computing this integral is the major computational burden of this model.

The third element of the model is the equation of motion. The Law of Effect states that rules which perform well are more likely to be used in the future. This law is captured by the following dynamic on log-propensities:

$$w(\rho,t+1) = \beta_0 w(\rho,t) + \beta_1 \rho(\Omega^t)Up^t \ , \text{ for } t \geq 1 \ , \tag{18}$$

where $\beta_1 > 0$ is a scaling parameter.

Given a space of rules R and initial conditions $w(\cdot,1)$, the law of motion, eq(18), completely determines the behavior of the system for all t >

1. The remaining operational questions are (i) how to specify R, and (ii) how to specify $w(\cdot,1)$.

An attractive feature of this general model is that it encompasses a wide variety of learning theories. For instance, to obtain replicator dynamics, we can simply let R be the set of J constant rules that always choose one unique action in A for all information states. Fictitious play and Cournot dynamics can be seen as very special cases in which R is a singleton rule which chooses a (possibly noisy) best-response to a belief that is a deterministic function of the history of play. Moreover, the general model can include these constant rules, best-response rules and other rules.

5.2. The Family of Evidence-Based Rules.

Our approach to specifying the space of rules is to specify a finite number of empirically relevant discrete rules that can be combined to span a much larger space of rules. In Stahl (1999), the family of "evidence-based" rules was introduced as an extension of the SW95 level-n rules. Evidence-based rules are derived from the notion that a player considers evidence for and against the available actions and tends to choose the action which has the most net favorable evidence based on the available information.

The first kind of evidence comes from a "null" model of the other players. The null model provides no reason for the other players to choose any particular strategy, so for the first period of play by virtue of insufficient reason, the belief is that all strategies are equally likely. The expected utility payoff to each available action given the null model is $y_1(\Omega^1) \equiv Up^0$. We interpret y_{1j} as "evidence" in favor of action j stemming from the null model and no prior history.

For later periods ($t > 1$), the players have empirical data about the past

choices of the other players. It is reasonable for a player to use simple distributed-lag forecasting. In particular, we specify adaptive expectations given by eq(11). The expected utility payoff given this belief is $y_1(\Omega^t;\theta) \equiv Uq^t(\theta)$. We can interpret $y_{1j}(\Omega^t;\theta)$ as "level-1" evidence in favor of action j stemming from the null model and prior history h^t.

The second kind of evidence is based on the SW95 "level-2" player who believes all other players are level-1 players, and hence believes that the distribution of play will be $b(q^t(\theta))$, where $b(q^t(\theta)) \in \Delta(A)$ puts equal probability on all best responses to $q^t(\theta)$ and zero probability on all inferior responses. The expected utility conditional on this belief is $y_2(\Omega^t;\theta) \equiv Ub(q^t(\theta))$. We can interpret $y_{2j}(\Omega^t;\theta)$ as "level-2" evidence in favor of action j.

The third kind of evidence incorporates Nash equilibrium theory within the model. Letting p^{NE} denote a Nash equilibrium of G, $y_3 \equiv Up^{NE}$ provides yet another kind of evidence on the available actions. Based on Haruvy and Stahl (1999), for games with multiple Nash equilibria, we define the evidence to be the average of y_3 over all the Nash equilibria equally weighted.

So far we have defined three kinds of evidence: $Y \equiv \{y_1, y_2, y_3\}$. The next step is to weigh this evidence and specify a probabilistic choice function. Let $\nu_k \geq 0$ denote a scalar weight associated with evidence y_k. We define the weighted evidence vector:

$$\bar{y}(\Omega^t;\nu,\theta) \equiv Y(\Omega^t;\theta)\nu \ , \tag{19}$$

where $\nu \equiv (\nu_1, \nu_2, \nu_3)'$. Note that the rules spanned by $\bar{y}(\Omega^t;\nu,\theta)$ can be viewed as a four-dimensional space of SW95 "worldly" rules.

We assume that the player assesses the weighted evidence with

some error or idiosyncratic additions, and chooses the action which from his/her perspective has the greatest net favorable evidence. There are many ways to go from such a weighted evidence measure to a probabilistic choice function. We opt for the multinomial logit specification because of its computational advantages when it comes to empirical estimation. Hence, the probability of choosing action j is

$$\hat{p}_j(\Omega^t;\nu,\theta) \equiv \exp[\bar{y}_j(\Omega^t;\nu,\theta)]/\Sigma_\ell \exp[\bar{y}_\ell(\Omega^t;\nu,\theta)] \ . \tag{20}$$

Note that, given the four-dimensional parameter vector (ν,θ), eq(20) defines a mapping from Ω^t to $\Delta(A)$, and hence is a *behavior rule* as defined abstractly above. By putting zero weight on all but one rule, eq(20) defines an archetypal rule - one for each kind of evidence corresponding to the underlying model of other players. Eq(19) generates the space of rules spanned by these archetypal rules.

5.3 Herd Behavior

As explained in section 2.3, following the herd does not mean exactly replicating the most recent past, but rather following the past with perhaps some inertia as represented by $q^t(\theta)$. Eq(11) can be interpreted as a behavioral rule: specifically, with probability θ the player repeats his/her choice propensities from last period q^{t-1}, and with probability $(1-\theta)$ the player mimics the population distribution of choices, p^t. Letting θ range over [0,1] generates a one-dimensional subspace of herd rules, and experience can affect the distribution of θ in the population. Let $\rho_h(\theta)$ denote a herd behavioral rule, so when using eq(17) the probability of using this rule is denoted by $\varphi(\rho_h(\theta),t)$.

5.4 Initial Propensities

The initial log-propensity function for the evidence-based rules conditional on the class of evidence-based rules is specified as

$$\tilde{w}(\nu,\theta,1) = -0.5\|(\nu,\theta)-(\bar{\nu},\bar{\theta})\|^2/\sigma^2 , \tag{21a}$$

where $\|(\nu,\theta)-(\bar{\nu},\bar{\theta})\|$ denotes the distance between rule (ν,θ) and the mean of the distribution $(\bar{\nu},\bar{\theta})$, and σ^2 is the variance around the mean. Hence, the initial conditional probability distribution over these evidence-based rules is

$$\tilde{\varphi}(\nu,\theta,1) \equiv \exp(\tilde{w}(\nu,\theta,1)/[\int\exp(\tilde{w}(x,1)dx], \tag{21b}$$

where the integration in the denominator is over the evidence-based rules (i.e. excluding herd behavior). Hence, the initial conditional propensity over the evidence-based rules is essentially a normal distribution with mean $(\bar{\nu},\bar{\theta})$ and standard deviation σ.

Similarly, the log-propensity function for the herd rules is conditionally specified as

$$\tilde{w}(\rho_h(\theta),1) = -0.5\|\theta - \bar{\theta}\|^2/\sigma^2 , \tag{22a}$$

using the same $\bar{\theta}$ and σ as for the evidence-based rules. Then the initial conditional probability distribution over the herd rules is

$$\tilde{\varphi}(\rho_h(\theta),1) \equiv \exp(\tilde{w}(\rho_h(\theta),1)/[\int\exp(\tilde{w}(\rho_h(x),1)dx]. \tag{22b}$$

Further, in keeping with the notation of the LBRIAE model, we introduce a single variable δ to denote the initial *unconditional* probability of using a herd rule, so

$$\varphi(\rho_h(\theta),1) = \delta\tilde{\varphi}(\rho_h(\theta),1). \tag{23}$$

Finally, we introduce another parameter ε to augment the initial probability of trembles (i.e. $\nu \equiv 0$). With zero weight on any evidence, the resulting choice probabilities are uniformly random: p^0.

Given δ probability of a herd rule and ε probability of a tremble, the probability of an evidence-based rule is $1-\delta-\varepsilon$; hence, the initial unconditional probability distribution over evidence-based rules is

$$\varphi(\nu,\theta,1) \equiv (1-\delta-\varepsilon)\ \tilde{\varphi}(\nu,\theta,1)\ . \tag{24}$$

Then the initial unconditional log-propensity for rule ρ (whether an evidence-based rule, herd rule, or tremble) is

$$w(\rho,1) \equiv ln[\tilde{\varphi}(\rho,1)]\ . \tag{25}$$

With the initial log-propensities so specified, the dynamic eq(17) fully determines $w(\rho,t)$ for all $t > 1$.

5.5 Transferrence.

Since this theory is about rules that use information about the game as input, we should be able to predict behavior in a temporal sequence that involves a variety of games. For instance, suppose an experiment consists of

one run with one game for T periods, followed by a second run with another game for T periods. How is what is learned about the rules during the first run transfered to the second run with the new game? A natural assumption would be that the log-propensities at the end of the first game are simply carried forward to the new game. Another extreme assumption would be that the new game is perceived as a totally different situation so the log-propensities revert to their initial state. We opt for a convex combination:

$$w(\rho,T+1) = (1-\tau)w(\rho,1) + \tau w(\rho,T*) \ , \tag{26}$$

where T* indicates the update after period T of the first run, and τ is the *transference* parameter. If $\tau = 0$, there is no transferance, so period T+1 has the same initial log-propensity as period 1; and if $\tau = 1$, there is complete transference, so the first period of the second run has the log-propensity that would prevail if it were period T+1 of the first run (with no change of game). This specification extends the model to any number of runs with different games without requiring additional parameters.

5.6. The Likelihood Function.

The theoretical model involves 10 parameters: $\xi \equiv (\delta, \varepsilon, \bar{\nu}_1, \bar{\nu}_2, \bar{\nu}_3, \bar{\theta}, \sigma, \beta_0, \beta_1, \tau)$. The quadruple of parameters $(\bar{\nu}_1, \bar{\nu}_2, \bar{\nu}_3, \bar{\theta})$ represents the mean of the participant's initial probability distribution φ over evidence-based rules and σ is the standard deviation of that probability distribution; β_0 and β_1 are the learning parameters of eq(18); and τ is the transference parameter in eq(26) for the initial propensity of the subsequent runs. This revised model has one additional parameter over the model of Stahl (2000b): namely ε.[19]

Let $s^h \equiv (s^{h1},\ldots,s^{h2T}) \in A^{2T}$ denote the choices of participant h for an experiment consisting of two runs of T periods each with action set A, and let n_j^t denote the number of participants who choose action j in period t.

The rule propensities and law of motion yield population choice probabilities:

$$p_j^{e\tilde{t}}(\xi) \equiv \int_R \hat{P}_j(\Omega^t;\rho)\varphi(\rho,t|\xi)d\rho \ . \tag{27}$$

Then the log of the joint probability of the data conditional on ξ is

$$LL(\xi) \equiv \Sigma_t \ \Sigma_j \ n_j^t \ \log[p_j^{e\tilde{t}}(\xi)] \ . \tag{28}$$

To perform the integration required by eq(27), we impose a finite grid on the (ν,θ)-space [see Appendix B for details].

5.7 Maximum Likelihood Estimates

The log-likelihood function, eq(28), for the entire pooled data set was maximized with respect to the parameter values. Compared to the 9-parameter model of Stahl (2000b) for this exact same data set,[20] the revised model increases the LL by 17.31. We would like to know whether this improvement is due to the extra parameter or the different specification of herd behavior. To address this question, we restricted the tremble parameter (ε) to be 0 and re-estimated the model. The restricted LL value decreased by 9.05. Thus, we can decomposs the 17.31 improvement into a 9.05 improvement due to the addition of the ε parameter, and the remaining 8.26 improvement due to the different specification of herd behavior.[21]

The parameter estimates for the revised model are displayed in Table V.

Note that the estimates of three parameters fall on the boundary of the parameter space: $\bar{\nu}_3 = 0$, $\beta_0 = 1$, and $\tau = 1$.[22] We interpret this finding as indicating that a model with just 7 "interior" parameters can fit the data as well as the full 10 parameter model.

6. Performance and Testing of Rule Learning

Three alternative models are nested within our revised Rule Learning model. First, by restricting the learning parameter $\beta_1 = 0$, we eliminate rule learning completely, leaving the "Diverse Worldly LBRIAE" model. Second, by further restricting the variance parameter $\sigma^2 = 0$,[23] we eliminate diversity in the population, leaving the "Worldly LBRIAE" model. Third, by further eliminating all evidence except level-1 evidence ($\bar{\nu}_2 = \bar{\nu}_3 = 0$), we obtain the original LBRIAE model. The corresponding in-sample and out-of-sample performance of each of these restricted models is reported in Tables VI and VII. A fourth alternative model, Rule Learning without diversity, which applies Rule Learning to just the three rules of the Worldly LBRIAE model, is also tested.

6.1 Diverse Worldly LBRIAE

The maximum LL value with β_1 restricted to 0 decreases from -4720.09 to -4730.79. Twice this difference (21.40) is asymptotically distributed chi-square with one degree of freedom and has a p-value of 3.7×10^{-6}. Thus, we can strongly reject *no rule learning*.

6.2 Worldly LBRIAE

The maximum likelihood estimates with β_1 and σ restricted to 0 are displayed in Table V. The LL value decreases to -4756.04. Compared to the

Diverse Worldly LBRIAE model, twice the difference (50.50) is asymptotically distributed chi-square with one degree of freedom and has a p-value of 1.2×10^{-12}. Thus, we can strongly reject *homogeneity* of the population. Compared to the revised Rule Learning model, twice the difference (71.90) is distributed chi-square with 2 degrees of freedom and has a p-value of 2.4×10^{-16}.

6.3 LBRIAE

The maximum likelihood estimates with β_1, σ, and $\overline{\nu}_2$ and $\overline{\nu}_3$ restricted to 0 are displayed in Table V. The LL value decreases to -4772.52. Compared to the Worldly LBRIAE model, twice the difference (32.96) is asymptotically distributed chi-square with one degree of freedom and has a p-value of 9.4×10^{-9}. Thus, we can strongly reject the unsophisticated LBRIAE model in favor of more sophisticated learning that takes account of level-2 evidence. The other compound hypotheses tests are rejected also at extremely "low" confidence levels.

6.4 Rule Learning without Diversity

To evaluate the relative contribution of Rule Learning (β_1) versus diversity (σ), we also considered the alternative Rule Learning model with σ restricted to 0. This version of Rule Learning contains just three rules: trembling (p^0), herd ($q^t(\overline{\theta})$), and a single evidence-based (or SW95 "worldly") rule with fixed weights ($\overline{\nu}_1$, $\overline{\nu}_2$, $\overline{\nu}_3$, $\overline{\theta}$). Since this model does not require integration over the rule space as in eq(26), it is computationally much simpler than the full Rule Learning model and the Diverse Worldly LBRIAE model. The maximized LL value of Rule Learning without diversity is -4739.34, a decrease of 19.25. Twice this difference is distributed chi-square with one

degree of freedom ($\sigma = 0$) and has a p-value of 5×10^{-10}. Thus, we strongly reject this restriction.

As with the full Rule Learning model, the estimates of $\bar{\nu}_3$ and τ fall on the boundary of the parameter space (see Table V), leaving seven interior parameters. The further restriction of $\beta_1 = 0$ results in the Worldly LBRIAE model, and a further decrease of 16.70 in the maximized LL value, which with two degrees of freedom (β_0 and β_1) has a p-value of 6×10^{-8}. Thus, we strongly reject this restriction. In other words, adding rule learning without diversity to the Worldly LBRIAE model makes a statistically significant contribution.

Along this path of model changes, we have decomposed the total LL increase of 35.95 (= -4720.09 + 4756.04) into (i) a increase of 16.39 due to rule learning without diversity, and (ii) a increase of 19.56 due to diversity (with rule learning). The alternative path yielded a decomposition into (iii) a increase of 25.25 due to diversity without rule learning, and (iv) a increase of 10.70 due to rule learning (with diversity). Either way, both features add significantly to the Worldly LBRIAE model.

6.5 In-sample and Out-of-sample Performance

Table VI displays our three measures of in-sample goodness-of-fit for the revised population Rule Learning model and its nested alternatives, and Table VII gives the parallel out-of-sample measures aggregated over the four holdout samples. Compared with the LBRIAE model the improved performance of the revised Rule Learning model is substantial for all measures.

6.6 What Rules are Learned

The grid for (ν_1, ν_2, ν_3 and θ) used to perform the integration required

by eq(27) consists of 175 points, which is constructed from 5 values for θ and 35 values for (ν_1, ν_2, ν_3); see Appendix B for details. In addition, there are 5 points corresponding to the herd rules for the five θ values. We also add one point corresponding to the mean ($\bar{\nu}$, $\bar{\theta}$). With only 181 discrete rules[24] at which the φ distribution is evaluated, it is feasible to examine all these φ values and assess which rules are "important" in the sense of being a significant part of the initial distribution and/or later distributions. By this criteria, of the 35 values for (ν_1, ν_2, ν_3), at most five of them are important, and all of the herd rules and the mean point ($\bar{\nu}$, $\bar{\theta}$) are important.

The initial distribution, $\varphi(\nu,\theta;1)$, is depicted in Figure 1. The probability, $\varphi(\nu,\theta;1)$, is given along the vertical axis. The values of (ν_1, ν_2, ν_3) are given along the horizontal axis, and the values for θ are given along the "Theta" axis. "Other" denotes the aggregate of $\varphi(\nu,\theta;1)$ over the 150 rules comprised by the 30 "unimportant" values of (ν_1, ν_2, ν_3) times the five θ values. The rule depicted second from the left with (ν_1, ν_2, ν_3) = (0.80, 0.07, 0) is the rule corresponding to the mean ($\bar{\nu}$, $\bar{\theta}$); it is shown with all probability mass in the cell corresponding to θ = 0.25 (since that value is closest to $\bar{\theta}$ = 0.35). The herd rules are at the far right side of the graph. The cell in the near right corner combines the tremble rule and the herd rule with θ = 1 (since they are observationally equivalent). Clearly the tremble and herd rules are the dominant rules in the initial period, followed by the level-1 rules (the left side of the graph). The Other group has a total probability of only 2.9%. Only one rule that gives weight to Nash evidence was found to be important (0.81, 0, 0.27); it is displayed just to the left of the Other category and has a total probability mass of 3.4%.

Using graphs like Figure 1, we examined how $\varphi(\nu,\theta;t)$ changes over time for

each experimental session. The change was continuous over time and remarkably similar over sessions. Figure 2 displays the total change in φ from the initial period to the final (30th) period, averaged over the four sessions with 5x5 games. The analogous figure for the sessions with 3x3 games is virtually the same. From Figure 2, we see that (i) trembles and herd behavior decline substantially over time (-16.1%), (ii) the level-1 rules and the rules that combine level-1 and level-2 evidence increase in importance over time (+13.9%), (iii) the Other rules are roughly unchanged over time (+0.8%), and the one "initially important" rule that gives weight to Nash evidence (0.81, 0, 0.27) increases by only (+1.4%).

Thus, we can infer that the major phenomenon that the LBRIAE model misses but Rule Learning picks up is the decline of trembling and simplistic herd behavior and the rise of more sophisticated level-1 and level-2 behavior. Indeed, trembling is virtually extinguished after 30 periods. Herd behavior, however, does survive (37.2%) because as a trend emerges (in particular, if play converges to a Nash equilibrium), then all the rules including following the herd perform equally well and are therefore equally reinforced.

7. Conclusions.

The literature on learning does not adequately deal with how one might choose among the rapidly growing number of models. Part of the reason for this lack of model comparison is the lack of agreed upon criteria for comparison. There are many imperfect yardsticks (e.g. likelihood, mean squared error, T-period-ahead forecasts, etc., each in-sample and out-of-sample), and it is tempting to use the yardstick that favors one's own model.

We consider a rich variety of action-reinforcement learning models, and numerous yardsticks for comparison. After putting the models on equal

footings, such as deriving population versions of the models where necessary, and allowing for mutations and trembles, the winner of the "horse race" among the seven reinforcement models considered is the four-parameter LBRIAE model (logit best-reply with inertia and adaptive expectations), and this winner dramatically outperforms the mindless AR1 model, the mental replicator model, and the Erev-Roth models. Second place in the horse race goes to the logit EWA model without experience weighting.

Further, we formulate a population Rule Learning model that nests the LBRIAE model, thereby enabling standard hypothesis testing. We strongly reject the three implied restrictions. We reject the hypothesis that a best-reply dynamic using only level-1 evidence (LBRIAE) is as good as a best-reply dynamic using level-2 evidence as well (the Worldly LBRIAE). We reject the hypothesis that the population is homogeneous in favor of a heterogeneous model (the Diverse Worldly LBRIAE) that admits a distribution over the evidence weights. Finally, we reject that this distribution is constant over time in favor the Rule Learning model in which the distribution changes according to the Law of Effect. These findings are reinforced by alternative performance measures. The measure of the log-likelihood functions for all periods except the first period demonstrates that the improvement in log-likelihood is not just a first-period phenomena. While the root-mean-square-error measures are not substantially different, they are consistent with the other measures. Finally, the dominance of Rule Learning in-sample is replicated for out-of-sample forecasting.

These findings do not mean that the LBRIAE model and Rule Learning will outperform all the alternatives for a specific game and data set, nor for individual data. Rather, if you want a model (and a single set of estimated parameters) that predicts population frequencies well across a variety of

symmetric normal-form games and environments in which players receive population feedback (in contrast to limited pairwise experience), then you should pick the horse-race winner: Rule Learning.

REFERENCES

Anderson, S., J. Goeree, and C. Holt (1997), "Stochastic Game Theory: Adjustment to Equilibrium Under Bounded Rationality," Department of Economics, University of Virginia, mimeo.

Barron, G., and I. Erev (2000), "Toward a General Descriptive Model of One Shot and Repeated Decision Making Under Risk and Uncertainty," Technion Working Paper.

Bush, R., and F. Mosteller (1955), *Stochastic Models of Learning*, Wiley.

Camerer, C, and T. Ho (1997), "EWA Learning in Games: Preliminary Estimates from Weak-Link Games," in D. Budescu, I. Erev, and R. Zwick (ed), *Games and Human Behavior: Essays in Honor of Amnon Rapoport*.

_____ (1999), "Experience-Weighted Attraction Learning in Normal-Form Games, *Econometrica*, forthcoming.

Camerer, C., T. Ho, and J. Chong (2000), "Sophisticated EWA Learning and Strategic Teaching in Repeated Games," mimeo.

Chen, Y., and F-F Tang (1998), "Learning and Incentive-Compatible Mechanisms for Public Goods Provision: An Experimental Study," *J. of Pol. Econ.*, 106, 633-662.

Cheung, Y-W, and D. Friedman (1997), "Individual Learning in Games: Some Laboratory Results," *Games and Economic Behavior*, 19, 46-76.

_____ (1998), "A Comparison of Learning and Replicator Dynamics Using Experimental Data," *J. of Economic Behavior and Organization*, 35, 263-280.

Crawford, V. (1994), "Adaptive Dynamics in Coordination Games," *Econometrica*, 63, 103-143.

Dawkins, R. (1976), *The Selfish Gene*, Oxford University Press.

Duffy, J., and R. Nagel (1997), "On the Robustness of Behavior in Experimental Beauty contest Games," *Economic Journal*, 107, 1684-1700.

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

_____ (1998), "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique Mixed Strategy Equilibria," *American Economic Review*, 88, 848-881.

Feltovich, N. (2000), "Reinforcement-Based vs. Beliefs-Based Learning in Experimental Asymmetric-Information Games," *Econometrica*, 60, 605-641.

Friedman, D., D. Massaro and M. Cohen (1995), "A Comparison of Learning Models," *J. of Mathematical Psychology*, 39, 164-178.

Friedman, M. (1953), *Essays in Positive Economics*, Univ. of Chicago Press.

Fudenberg, D., and D. Levine (1997), *The Theory of Learning in Games*, MIT Press.

Haruvy, E. and D. Stahl (1998), "Can Adaptive Dynamics Resolve Multiplicity," Presentation to ESA Meeting in Tucson, AZ.

Haruvy, E. and D. Stahl (1999), "An Empirical Theory of Equilibrium Selection based on Heterogeneity," mimeo.

Ho, T., K. Weigelt, and C. Camerer (1996), "Iterated Dominance and Iterated Best-Response in Experimental 'p-Beauty' Contests", *Amer. Econ. Rev.*, forthcoming.

Hofbauer, J., and K. Sigmund (1988), *The Theory of Evolution and Dynamical Systems*, Cambridge University Press.

Kandori, M., Mailath, G., and Rob, R. (1993), "Learning, Mutation, and Long-Run Equilibria in Games," *Econometrica*, 61, 29-56.

Luce, R. Duncan (1959), *Individual Choice Behavior*, Wiley.

Mookherjee, D. and B. Sopher (1994), "Learning Behavior in an Experimental Matching Pennies Game," *Games and Economic Behavior*, 7, 62-91.

_____ (1997), "Learning and Decision Costs in Experimental Constant Sum Games," *Games and Economic Behavior*, 19, 97-132.

Nagel, R. (1995), "Unraveling in Guessing Games: An Experimental Study," *Amer. Econ. Rev.*. 85, December, 1313-1326.

Nagel, R., and F. Tang (1998), "Experimental Results on the Centipede Game in Normal Form: An Investigation on Learning," *J. of Math. Psych.* 42, 356-84.

Nerlove, M. (1958), *Distributed Lags and Demand Analysis*, USDA Handbook No. 141, Government Printing Office.

Rapoport, A., I. Erev, E. Abraham, and D. Olson (1997), "Randomization and Adaptive Learning in a Simplified Poker Game," *Organizational Behavior and Human Decision Processes*, 69, 31-49.

Roth, A. and I. Erev (1995), "Learning n Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term," *Games and Economic Behavior*, 8, 164-212.

Selten, R. (1990), "Anticipatory Learning in Games," in *Game Equilibrium Models. Vol. I: Evolution and Game Dynamics* (R. Selten, ed), Springer-Verlag.

Selten, R. (1991), "Evolution, Learning, and Economic Behavior," *Games and Economic Behavior*, 3, 3-24.

Sonsino, D., I. Erev, S. Gilat, and G. Shabtai (1998), "On the Likelihood of Repeated Zero-sum Betting by Adaptive Human Agents," working paper, Faculty of Industrial Engineering and Management, Technion, Israel.

Stahl, D. (1996), "Boundedly Rational Rule Learning in a Guessing Game," *Games and Economic Behavior*, **16**, 303-330.

Stahl, D. (1999), "Evidence Based Rules and Learning in Symmetric Normal-Form Games, *Int'l J. of Game Theory*, **28**, 111-130.

Stahl, D. (2000a), "Rule Learning in Symmetric Normal-Form Games: Theory and Evidence, *Games and Economic Behavior*, **32**, 105-138.

Stahl, D. (2000b), "Population Rule Learning in Symmetric Normal-Form Games: Theory and Evidence," *J. of Econ. Org. and Behavior*, forthcoming.

Stahl, D. and P. Wilson (1995), "On Players Models of Other Players: Theory and Experimental Evidence," *Games and Economic Behavior*, **10**, 218-254.

Tang, F-F. (1998), *Anticipatory Learning in Two-Person Games: An Experimental Study*, Lecture Series in Mathematical and Economic Systems, Springer-Verlag (in press).

Van Huyck, J., R. Battalio and R. Beil (1991), "Strategic Uncertainty, Equilibrium Selection Principles, and Coordination Failures in Average Opinion Games," *Quart. J. of Economics*, **106**, 885-911.

Van Huyck, J., J. Cook, and R. Battalio (1994), "Selection Dynamics, Asymptotic Stability, and Adaptive Behavior," *J. of Political Economy*, **102**, 975-1005.

Van Huyck, J., R. Battalio and F. Rankin (1996), "Evidence on Learning in Coordination Games," mimeo.

ENDNOTES

[1]For example, Selten (1990, 1991), Kandori, *et al* (1993), Crawford (1994), Van Huyck, *et al* (1994).

[2]For example, Mookherjee and Sopher (1994, 1997), Friedman, *et al* (1995), Roth and Erev (1995), Anderson, *et al* (1997), Camerer and Ho (1997, 1999), Cheung and Friedman (1997, 1998), Fudenberg and Levine (1997), Rapoport, *et al* (1997), Erev and Roth (1998), and Feltovich (2000).

[3]For example, Stahl (1996, 1999, 2000a-b), and Barron and Erev (2000).

[4]Cheung and Friedman (1994, 1997, 1998) study a clever three-parameter individual belief-learning model as well as replicator dynamics in a variety of games and under a variety of information conditions. Mookherjee and Sopher (1997) formulate a generalized quantal response learning model that nests individual belief-learning and stimulus-learning along with six versions of "perceived payoff", and they fit pairwise-matched experimental data for several 4×4 and 6×6 constant-sum games with a unique mixed-strategy Nash equilibrium. Camerer and Ho (1997) fit their model to a variety of experimental data and test a number of nested hypotheses (including the models considered by Cheung and Friedman and Erev and Roth). Erev and Roth (1998) study a variety of versions of their individual reinforcement model as well as a version of Camerer and Ho's model. Stahl (1996, 1999, 2000a) studies individual Rule Learning and tests a number of nested alternatives. Chen and Tang (1998) focus on public goods games, Nagel and Tang (1998) focus on a normal-form of the centipede game, and Tang (1998) focuses on 2×2 games with a unique mixed-strategy

equilibrium. Barron and Erev (2000) investigate reinforcement dynamics applied to three rules: loss avoidance, fictitious play, and best-reply to last observation. Feltovich (2000) focuses on asymmetric-information games and several simple learning models.

[5]While the *ex post* likelihood of the data is always higher given individual parameter estimates, since there is no way to identify individual parameters *ex ante*, the appropriate comparison is between the likelihood for a population model and the likelihood for a model in which the population is simulated by random draws with replacement from an urn filled with the *ex post* individual parameter estimates. Based on this comparison, there is no advantage to the latter individual-parameter urn model.

[6]Recently, Camerer, Ho and Chong (2000) have proposed an extension of the EWA model that includes a sophisticated player and the possibility of transferance.

[7]e.g. Mookerjhee and Sopher, 1994; Van Huyck, Battalio and Rankin (199?); Erev and Rappoport, 1998; Costa-Gomes, Broseta and Crawford, 1998; Barron and Erev, 2000; and Camerer, Ho and Chong, 2000.

[8]Note that p(t) remains to the right of the payoff matrix U, in eq(3), since it is the payoff against the actual frequency of play in the population that matters.

[9]We investigated power-form specifications, but have consistently found the logit form to fit better [as do Camerer and Ho]. Also, an inconvenience of the

power form is boundary requirement that $w_j(t) \geq 0$ for all t.

[10]We implicitly take the initial conditions, w(1), to be the same for all three models. However, this could be problematic if the range of $\Gamma()$ is not the full domain for w(1), since eq(6a) requires that $w(1) = \Gamma(E[p(1)|0])$.

[11]This is the extended model alluded to in section VI of Erev and Roth (1998).

[12]The Erev-Roth (1998) 3-parameter model is nested within these 6 parameter models with $b^- = b^+ = 0$ and $\rho_1 = \rho_0 = \min\{U_{jk}\}/\bar{u}$. These restrictions are strongly rejected on our data with p-values of 10^{-14} or less.

[13]The power form was found to perform worse, consistent with Camerer and Ho' own findings.

[14]Obviously one could consider alternative specifications of noisy best-reply functions such as the probit form; however, there is a limit to the amount of data mining we will engage in.

[15]To extend to asymmetric n-player games, we would use an n-population model with separate AR1 processes for each population. The own-population AR1 process would represent "following one's herd", while the other AR1 processes would describe beliefs about the other players (populations).

[16]There is very little other suitable data available. Cheung and Friedman (1997) ran experiments using mean-matching and population feedback, but this treatment occured in a sequence of different information treatments that would

confound our analysis. Van Huyck, Battalio and Beil (1991) provide feedback of a summary statistic for the population but not the whole distribution. Nagel (1995) made public an anonymous list of all choices, but the action set was infinite [see Stahl (1996) for a similar analysis of that data.] There is, of course, a wealth of data using pairwise matching; however, at the very least we would expect a slower rate of learning in such environments, which would necessitate separate parameter estimation from the mean-matching data. We leave that task for future research.

[17]We also computed the Mean Absolute Deviation, $\Sigma_j \, |p_j(t) - \varphi_j(t)|$, but we do not report this here because it does not add anything to the model comparisons over the RMSE.

[18]For the full six-parameter EWA, we found γ and N_0 to be insignificantly different from zero, which motivated our consideration of the restricted model without experience weighting.

[19]Herd behavior still entails one parameter (now δ instead of $\bar{\nu}_0$).

[20]The code for the latter model can be downloaded from www.eco.utexas.edu/faculty/Stahl/experimental/Rule_learning.

[21]In the old specification, herd evidence was defined as $ln[q^t(\theta)]$, then weighted by a free parameter ν_0 and combined with other evidence as in eq(19). Thus, pure herd behavior yielded probabilistic choice functions: $p_j(\nu_0,\theta) = q_j^t(\theta)^{\nu_0}/[\Sigma_k \, q_k^t(\theta)^{\nu_0}]$, which distort the $q^t(\theta)$ probabilities in a non-linear way, as opposed to the linear nature of the partial adjustment model. In

retrospect, the stylistic benefit of expressing all behavioral rules in the form of evidence-based rules is outweighed by the loss in goodness-of-fit.

[22]This is also true of the original model (Stahl, 2000b).

[23]In the limit, as $\sigma \to 0$, the initial probability distribution over evidence-based rules, $\tilde{\varphi}(\nu,\theta,1)$, converges to a degenerate distribution with all probability mass on the rule characterized by the mean $(\bar{\nu},\bar{\theta})$. Thus, instead of using eq(21a) with $\sigma = 0$, we instead estimate the limit model with a single evidence-based rule $(\bar{\nu},\bar{\theta})$, a single herd rule $\rho_h(\bar{\theta})$, and a tremble rule.

[24]Decreasing the number of divisions per dimension of the grid definitely worsens the fit of the model, while doubling the number of divisions (which entails 415 points) makes a negligible improvement. Therefore, the 181 point grid is quite adequate.

[25]This admittedly ad hoc value came about by considering the case of J=5, and an upper bound on r corresponding to 0.9999 of the interval [1/J,1]. Since the results are robust to this number and to J, this value become "grandfathered" in our code.

Table I. Base Models on GX4 & GX6 Data (optional)

Model	Eqs	np	LL	LL(-1)	RMSE
Logit Mental Replicator	4	1	-5460.50	-4985.22	.160
AR1	11	1	-5080.96	-4566.79	.124
Erev-Roth					
uniform prior:	7,8	6	-4935.20	-4421.02	.110
insuff. reason:	"	6	-4867.77	-4424.57	.101
LBRI	10	2	-4927.92	-4489.15	.098
Logit EWA	9	5	-4822.30	-4398.97	.091
LBRIAE	11,12	3	-4803.45	-4371.71	.088

Table II. Best Versions - GX4 & GX6 Data

Model	Deviations	np	LL	LL(-1)	RMSE
AR1	trembles*	2	-5069.72	-4555.54	.123
Logit Mental Replicator	both	3	-4992.73	-4563.06	.118
Erev-Roth					
uniform prior:	mutations	7	-4916.93	-4402.76	.107
insuff. reason:	mutations	7	-4828.06	-4405.53	.094
LBRI	trembles*	3	-4805.13	-4369.35	.088
Logit EWA					
	trembles*	6	-4783.64	-4353.29	.089
w/o EW:	trembles*	4	-4784.07	-4353.28	.089
LBRIAE	trembles*	4	-4772.52	-4338.17	.086

*These models implicitly contain a form of mutations.
"both" = mutations and trembles

Table III. Sum of Best Version Forecasts of Sessions

Model	LL	LL(-1)	RMSE
AR1	-5094.54	-4580.11	.125
Logit Mental Replicator	-5024.31	-4594.62	.120
Erev-Roth			
uniform prior:	-4967.04	-4449.86	.111
insuff. reason:	-4868.41	-4441.80	.099
LBRI	-4834.45	-4397.58	.091
Logit EWA	-4808.68	-4375.82	.091
w/o EW:	-4803.73	-4372.83	.091
LBRIAE	-4794.42	-4359.11	.088

Table IV. Parameter Estimates and Standard Errors for Best Versions

Parm	AR1	Logit Repl	LBRI	LBRIAE	Logit EWA w/o EW	Erev-Roth Uniform	Erev-Roth Insuf. R
θ	.4020 (.0418)	.6757 (.0139)		.2507 (.0281)	.3269 (.0417)	.3701 (.0520)	.3835 (.0396)
β		.0868 (.0036)	.3650 (.0222)	.3955 (.0391)	.0975 (.0047)		
ε	.0365 (.0221)	.1182 (.0106)	.0764 (.0047)	.0530 (.0093)	.0684 (.0128)		
η						.8091 (.1052)	.7513 (.0633)
δ			.6154 (.0102)	.6053 (.0202)			
α					.6659 (.0280)		
S						.1696 (.0918)	.2325 (.1004)
μ						.0966 (.0113)	.1006 (.0107)
ρ_0, ρ_1						.7392 (.0313)	1.049 (.0373)
b^-						.9998 (.0261)	1.000 (.0375)
b^+						.0100 (.0105)	2.3E-8 (4.5E-6)

Table V. Maximum Likelihood Parameter Estimates for Revised Populations Rule Learning and Nested Hypotheses

Parameter	Revised Pop Rule Learning	Diverse Worldly LBRIAE	Rule Lrn w/o Div.	Worldly LBRIAE	LBRIAE
δ	0.533 (0.030)	0.516* (0.030)	0.574 (0.041)	0.532* (0.019)	0.573* (0.019)
ε	0.090 (0.023)	0.033 (0.010)	0.118 (0.030)	0.059 (0.009)	0.053 (0.009)
$\bar{\nu}_1$	0.797 (0.204)	0.968 (0.124)	0.375 (0.027)	0.358 (0.030)	0.395 (0.039)
$\bar{\nu}_2$	0.070 (0.020)	0.092 (0.016)	0.027 (0.003)	0.022 (0.003)	
$\bar{\nu}_3$	0.000 (3×10^{-9})	0.000 (4×10^{-7})	0.000 (1×10^{-8})	0.000 (1×10^{-12})	
$\bar{\theta}$	0.353 (0.027)	0.359 (0.027)	0.127 (0.019)	0.184 (0.020)	0.251 (0.028)
σ	0.767 (0.089)	0.757 (0.079)			
β_0	1.000 (0.019)		0.974 (0.016)		
β_1	0.008 (0.0014)		0.011 (0.0025)		
τ	1.000 (0.008)		1.000 (2×10^{-5})		
LL	-4720.09	-4730.79	-4739.34	-4756.04	-4772.52
No. of interior parmeters	7	6	7	5	4

*We report here the combined estimate of $\delta(1-\varepsilon)$.

Table VI. In-Sample Performance Measures for Rule Learning

Model	LL	LL(-1)	RMSE
LBRIAE	-4772.52	-4338.17	0.086
Worldly LBRIAE	-4756.04	-4329.20	0.083
Rule Lrn w/o Diversity	-4739.34	-4308.68	0.081
Diverse Worldly LBRIAE	-4730.79	-4305.20	0.080
Pop Rule Learning	-4720.09	-4292.34	0.079

Table VII. Out-of-Sample Performance Measures for Rule Learning

Model	LL	LL(-1)	RMSE
LBRIAE	-4794.42	-4359.11	0.088
Worldly LBRIAE	-4793.51	-4368.02	0.087
Rule Lrn w/o Diversity	-4769.55	-4337.63	0.085
Diverse Worldly LBRIAE	-4760.96	-4335.62	0.083
Pop Rule Learning	-4754.17	-4325.98	0.083

APPENDIX A: Payoff Matrices

Session

Run 1

6/27/95

68	10	76	33	75
73	4	59	0	8
3	92	16	15	99
86	54	25	41	6
72	98	92	8	52

6/29/95

2	31	0	99	6
6	10	97	40	24
98	96	38	48	19
42	40	80	51	48
97	46	5	68	49

8/10/95

22	79	35	56	75
22	38	78	55	99
27	58	1	11	0
70	1	34	59	37
56	84	60	23	2

8/15/95

19	43	96	85	85
28	62	88	74	24
67	21	38	48	38
40	58	0	15	92
16	15	86	99	79

2/17/98

20	0	60
0	60	0
10	25	25

2/19/98

80	60	50
60	70	90
0	0	100

4/07/98

50	0	0
70	35	35
0	25	55

4/09/98

60	60	30
30	70	20
70	25	35

Run 2

6/27/95

19	43	96	85	85
28	62	88	74	24
67	21	38	48	38
40	58	0	15	92
16	15	86	99	79

6/29/95

22	79	35	56	75
22	38	78	55	99
27	58	1	11	0
70	1	34	59	37
56	84	60	23	2

8/10/95

2	31	0	99	6
6	10	97	40	24
98	96	38	48	19
42	40	80	51	48
97	46	5	68	49

8/15/95

68	10	76	33	75
73	4	59	0	8
3	92	16	15	99
86	54	25	41	6
72	98	92	8	52

2/17/98

30	50	100
40	45	10
35	60	0

2/19/98

30	100	22
35	0	45
51	50	20

4/07/98

70	60	90
60	80	50
40	20	100

4/09/98

68	4	49
86	41	4
72	25	39

APPENDIX B. Computational Methods and Issues.

To perform the integration required by eq(15), we impose a finite grid on the (ν,θ)-space. To ensure that the results are robust to the specification of the grid, we experimented with a variety of finite grids. We have settled on a computationally efficient grid consisting of $5\times35 = 175$ points generated as follows. First, for θ, we specify 5 uniformly spaced points: {0, 0.25, 0.5, 0.75, 1.0}.

The ν-subspace is represented by 35 points. To describe these 35 points, first recall eq(20) which defines the probabilistic choice function for each rule, and consider a rule with one and only one $\nu_k > 0$. In games with binary lottery payoffs (0 to 100%), how responsive is the choice behavior of such a rule to, say, a 10 percentage point difference in payoff? In other words, suppose the evidence for one action, say j, is 10 points higher than any other action. How much more likely is action j to be chosen? Using eq(20), the probability of choosing action j out of J actions would be

$$r = \exp(10\nu_k)/[J-1+\exp(10\nu_k)]. \tag{B1}$$

Clearly, for large (small) values of ν_k, this probability r is close to 1 (1/J), and is a non-linear function of ν_k. For the many symmetric normal form games we have used, payoff differences of 10% are typical. A smaller difference, say 5%, is on the margin of what we generally consider statistically significant, while a larger difference, say 20%, seems too crude relative to human discriminating abilities. Thus, a 10 percentage point payoff difference is a reasonable standard by which to assess the behavioral

impact of the ν_k weights.

We would like the range of ν_k weights in our grid to span fairly the range of choice behaviors. In other words, we would like our grid points to correspond to equally spaced behaviors (probabilities). Solving eq(B1) for ν_k in terms of r gives ν_k equal to

$$0.1 \ln[(J-1)r/(1-r)] \equiv f(r,J) . \tag{B2}$$

For any probability $r \in [1/J,1)$, $f(r,J)$ gives the magnitude of the ν_k weight such that an action with a 10 point payoff advantage would have a choice probability of r. We note, however, that $f(1,J) = \infty$, and from eq(B1), we can see that the choice probability an action that has a 10 point higher payoff than any other action will be quite close to one for all values of $\nu_k > 1$. Thus, bounding the ν_k weights from above by 1 would not induce any significant loss in the behavior that could be represented. We chose $\bar{v} \equiv 0.1 \ln(49996) \simeq 1.082$, as our upper bound.[26] This corresponds to a choice probability of 49996/(49995+J), which when J=5 is 0.99992, and is clearly quite insensitive to J for all practical values of J that would be used in experimental games. Now, take four equally spaced points from $[1/J,(J-1)/J]$: r_h ($[1 + 0.25h(J-1)]/J$, for $h = 0,\ldots,3$, This partition yields 5 values for (k: $\{f(r_h,J)$, for $h = 0,\ldots,3\}$, plus $\bar{v}$. While these values are unevenly spaced in ν-space, they induce evenly spaced behaviors (probabilistic choices) for the reference 10 point payoff difference.

---

[1]This admittedly ad hoc value came about by considering the case of J=5, and an upper bound on r corresponding to 0.9999 of the interval [1/J,1]. Since the results are robust to this number and to J, this value become "grandfathered" in our code.

Next we require that the sum of the weights over the three dimensions of ν-space, ν_k not exceed $\bar{v}$. This restriction effectively creates competition amoung the 3 kinds of evidence (level-1, level-2 and Nash): when some ν_k is increased, the weight on some other kind of evidence has to be decreased. Without this competition, the MLE procedure will produce the following spurious results. Suppose that after 7 or 8 periods, choice behavior has converged to the extend that the level-1 and level-2 rules put high probability on the best-response. The MLE procedure is likely to drive both ν_1 and ν_2 as high as possible, creating only a slight increase in the log-likelihood value but obscuring the relative importance of level-1 and level-2 evidence. Our restriction forces the MLE to assess the relative importance of each kind of evidence. To explore the effect of this restriction, we experimented with a variety of values for the upper bound (some higher and some lower than $\bar{v}$). Typically, an increase in the upper bound will increase the likelihood function only slightly, and have no significant effect on the parameter estimates (other than better identifying the relative ν weights). Thus, the results are robust to the specification of the upper bound.

Given this restriction on the sum of the weights, it is natural to specify the other points of the grid so the sum of the weights equals a value in $\{f(r_h,J)\ ,\ h = 0,\ldots,3\}$. To do this, define the 35 point "triangular" grid:

$$T \equiv \{(i_1,\ldots,i_3) \in \{0,1,2,3,4\}^3 \mid \text{isum}\ \Sigma_{k=1}^{3}\ i_k \leq 4\}.$$

Then for any $\underline{i} \in T$, define the weight for dimension k as $\nu_k(\underline{i}) \equiv f(r_{isum},J)(i_k/\text{isum})$. The "distance", $\|\cdot\|$, used in eq(9a and 10a) is the Euclidean distance in this index grid.

The entire grid is then the Cartesian product of the 5 θ points and these

35 ν points. In addition to this fixed grid of 175 points, we add the mean $(\overline{\nu},\overline{\theta})$ as a variable grid point, making a total of 176 points in all.

To find a ξ vector that maximizes LL(ξ), eq(28), we use a simulated annealing algorithm [Goffe (1994)] for high (but declining) temperatures, and then feed the result into the Nelder and Mead (1965) algorithm. We find the simulated annealing algorithm to be effective in exploring the parameter space, but very slow to converge once it settled in on a local maximum, while the latter algorithm converges much faster locally.

APPENDIX D. Empirical Frequencies of Data

Figures 4a-h for the "GX4" data set are already published in Stahl (2000b).

APPENDIX C. Empirical Frequencies of Data

Figures 4a-h for the "GX4" data set are already published in Stahl (2000a).

Figure 4a: Distribution of Choices
Session 627 Run 1 Game II

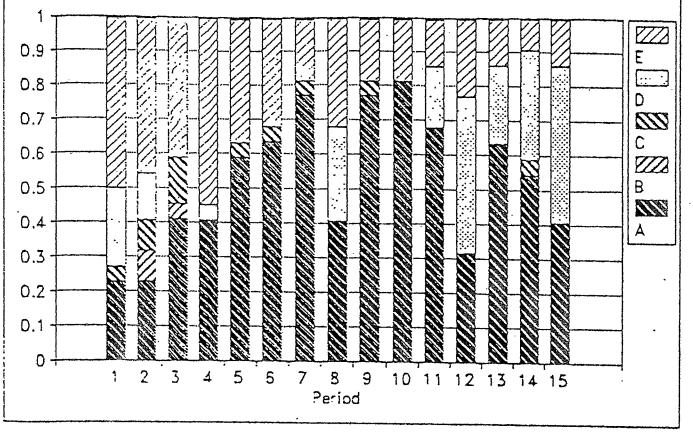


Figure 4b: Distribution of Choices
Session 627 Run 2 Game I

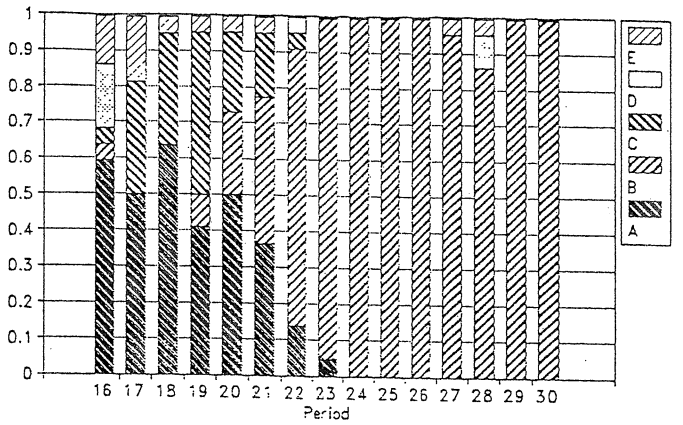


Figure 4c: Distribution of Choices
Session 629 Run 1 Game III

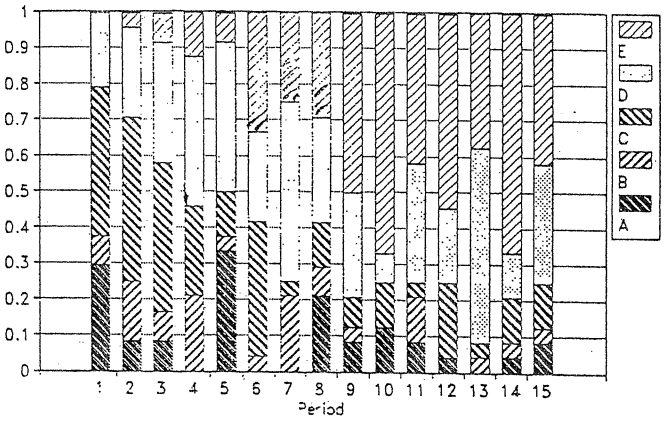


Figure 4d: Distribution of Choices
Session 629 Run 2 Game IV

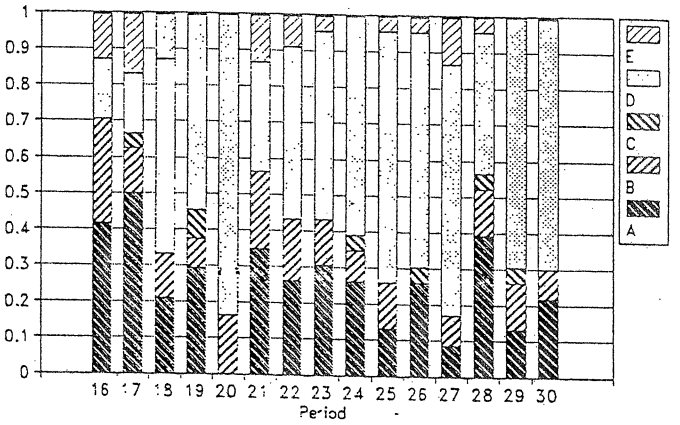


Figure 4e: Distribution of Choices
Session 810 Run 1 Game IV

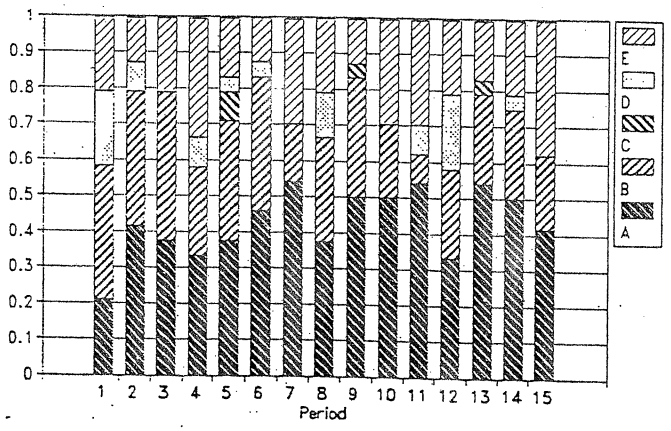


Figure 4f: Distribution of Choices
Session 810 Run 2 Game III

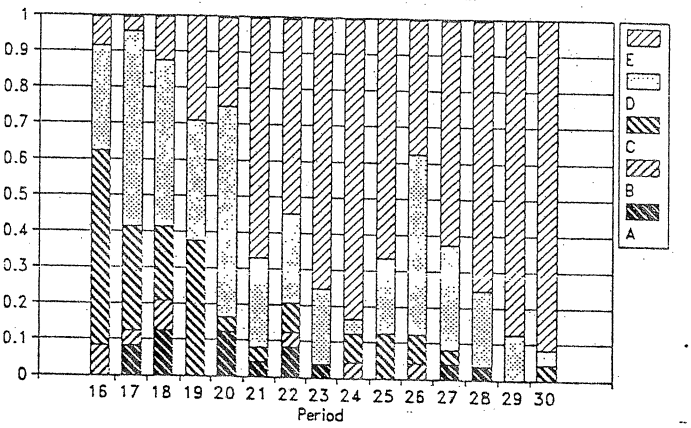


Figure 4g: Distribution of Choices
Session 815 Run 1 Game I

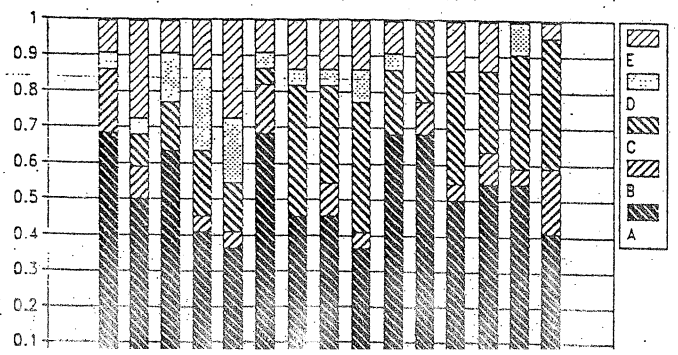
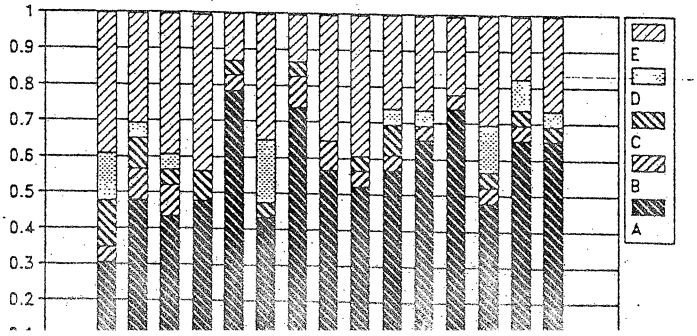
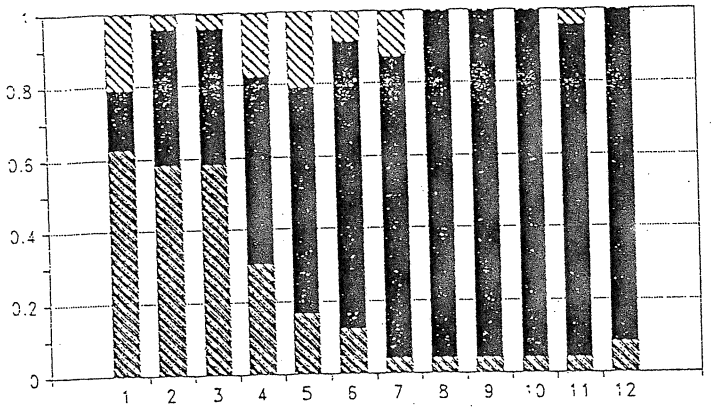


Figure 4h: Distribution of Choices
Session 815 Run 2 Game II

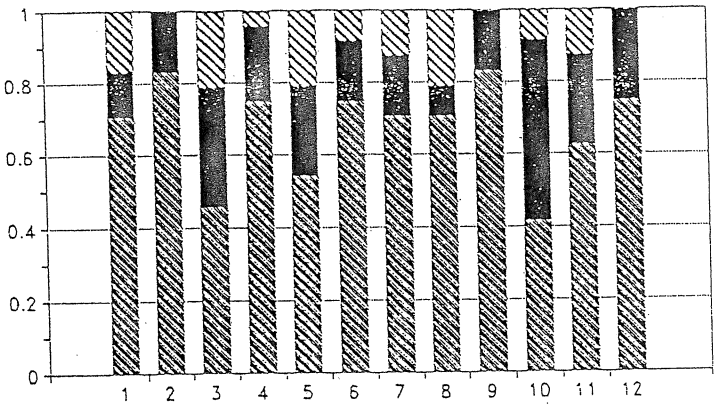


APPENDIX Choice Frequencies for GX6 Data

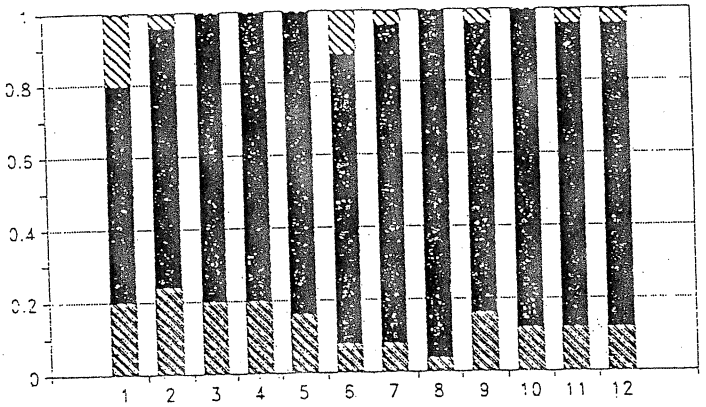
Session 1 Run 1



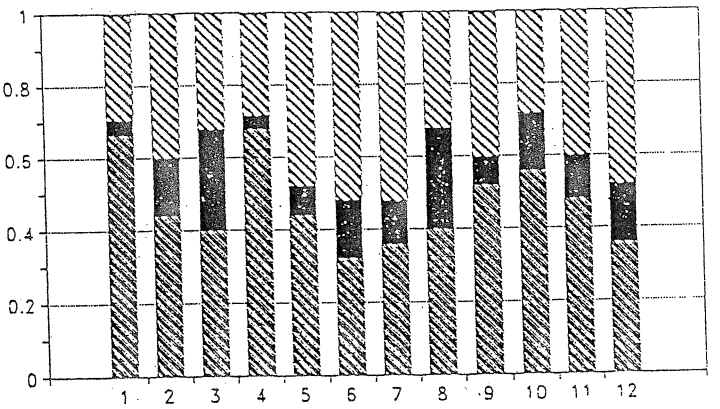
Session 1 Run 2



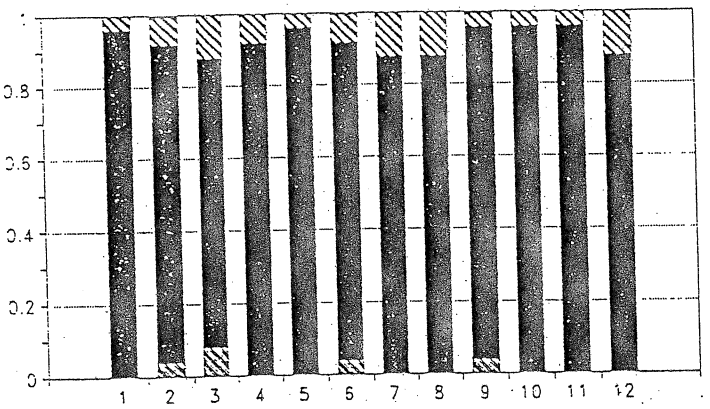
Session 2 Run 1



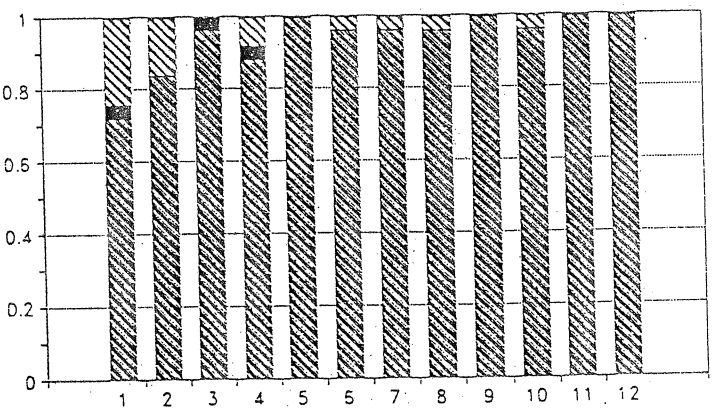
Session 2 Run 2



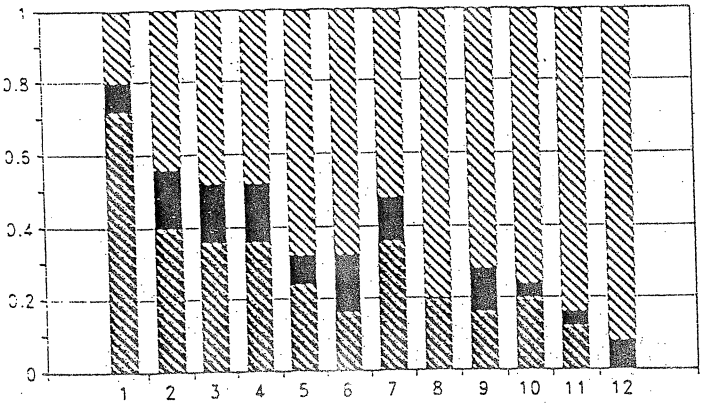
Session 3 Run 1



Session 3 Run 2



Session 4 Run 1



Session 4 Run 2

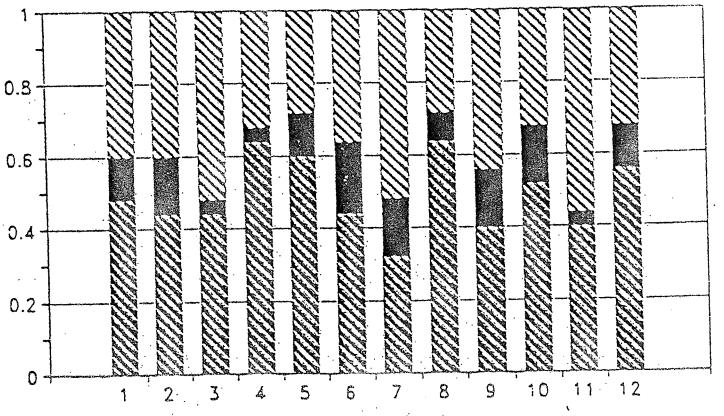


Figure 1. Initial Distribution over Rules

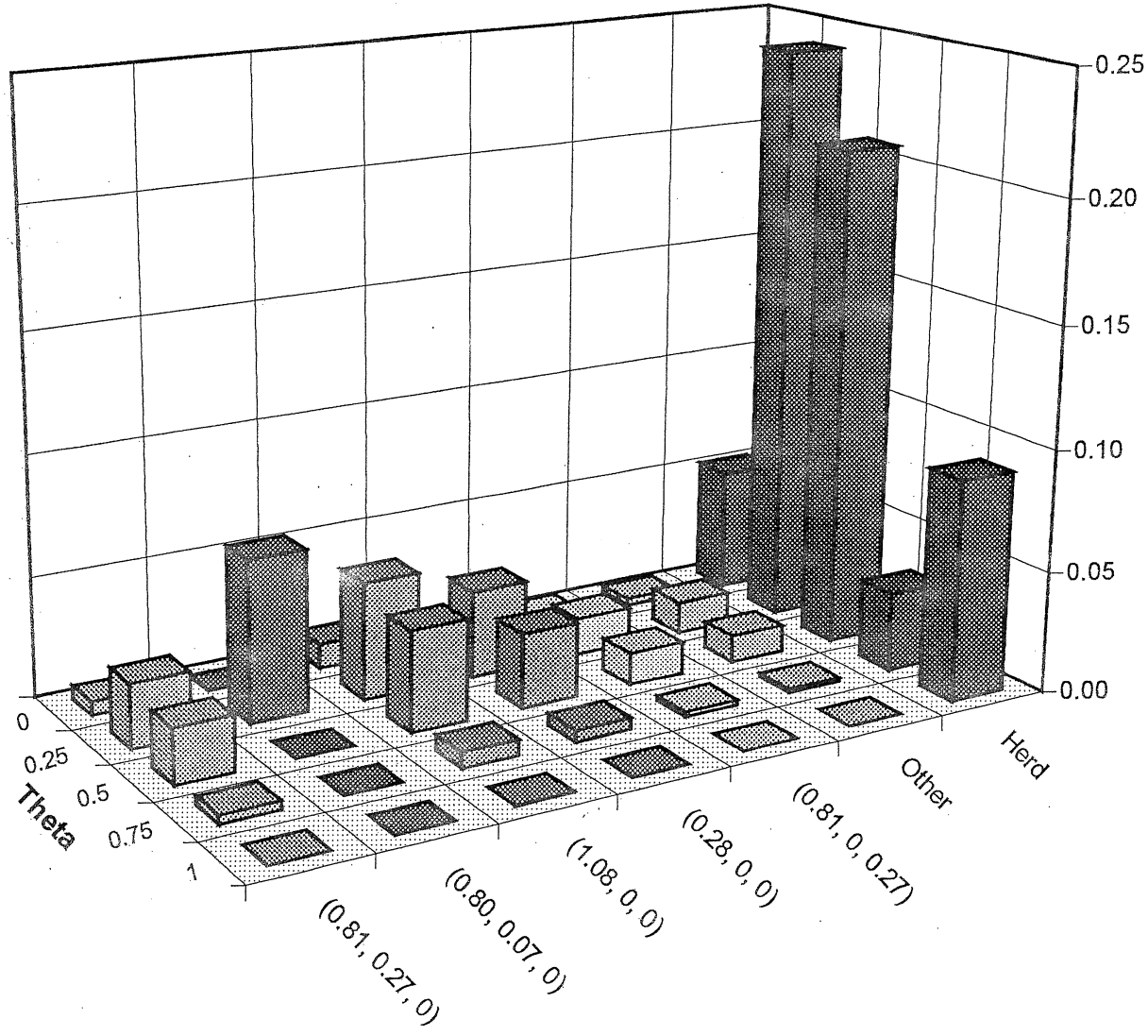
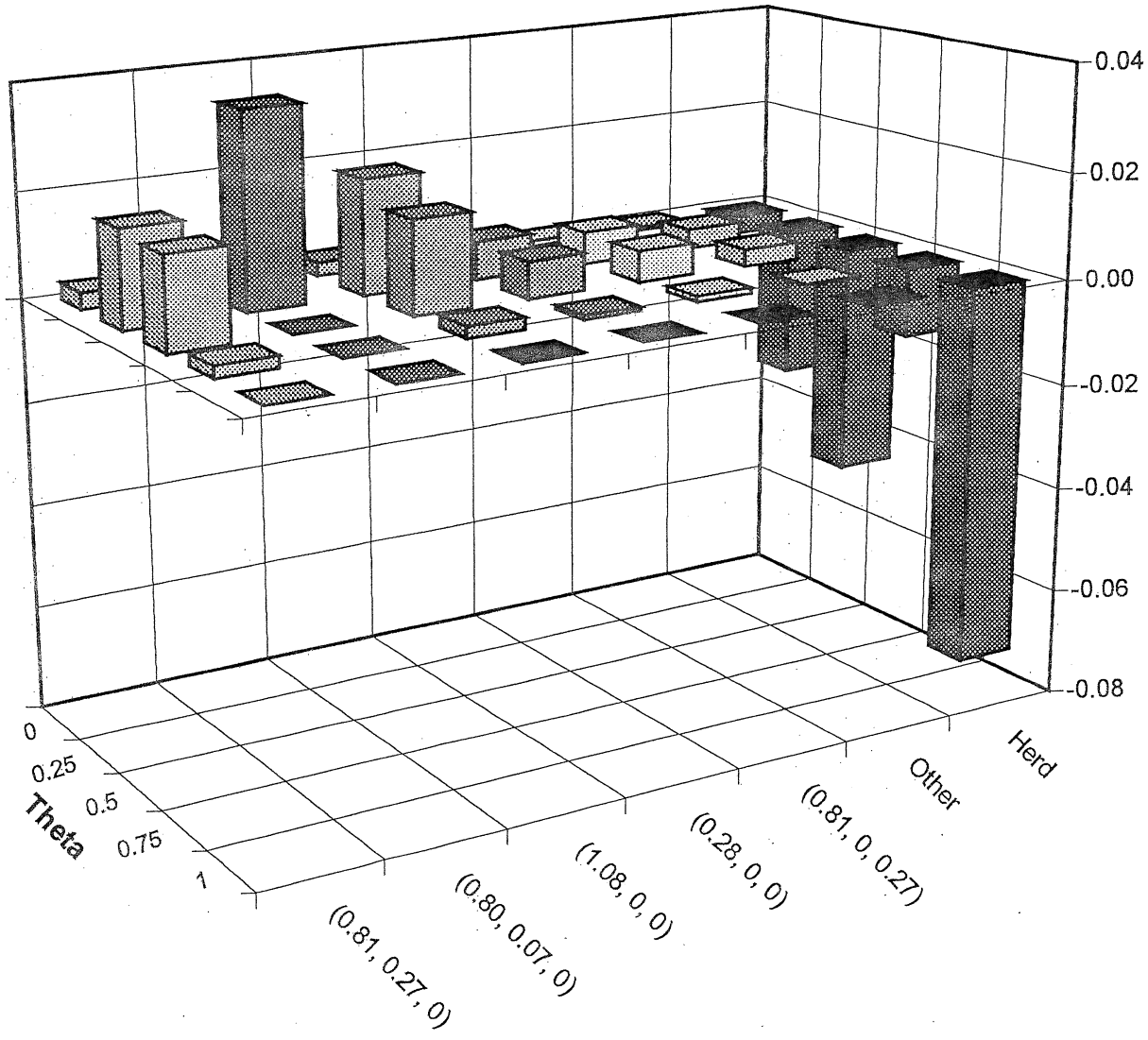



Figure 2. Change in Distribution over Rules: Period 1-30