Individual Learning in Games

Teck H. Ho
Haas School of Business
University of California, Berkeley
California, CA 94720-1900
hoteck@haas.berkeley.edu

August 8, 2006

Forthcoming
The New Palgrave Dictionary of Economics:
Experimental and Behavioral Economics

1 Introduction

Economic experiments on strategic games typically generate data that, in early rounds, violate standard equilibrium predictions. However, subjects normally change their behavior over time in response to experience. The study of learning in games is about how this behavioral change works empirically. This empirical investigation also has a theoretical payoff: If subjects' behavior converges to an equilibrium, the underlying learning model becomes a theory of equilibration. In games with multiple equilibria, this same model can also serve as a theory of equilibrium selection, a long-standing challenge for theorists.

There are two general approaches to studying learning: Population models and individual models.

1. Population models make predictions about how the aggregate behavior in a population will change as a result of aggregate experience. For example, in replicator dynamics, a population's propensity to play a certain strategy will depend on its 'fitness' (payoff) relative to the mixture of strategies played previously (Friedman, 1991; Weibull, 1995). Models like this submerge differences in individual learning paths.

2. Individual learning models allow each person to choose differently, depending on the experiences each person has. For example, in Cournot dynamics, subjects form a belief that other players will always repeat their most recent choice and best-respond accordingly.[1] Since players are matched with different opponents, their best responses vary across the population. Aggregate behavior in the population can be obtained by summing individual paths of learning.

This chapter reviews three major approaches to individual learning in games: experience-weighted attraction (EWA) learning, reinforcement learning, and belief learning (includ-

[1] Another class of learning models involve imitation, where players repeat observed strategies rather than evaluate all strategies (e.g., Schlag, 1999).

ing Cournot and fictitious play).[2] These models of learning strive to explain, for every choice in an experiment, how that choice arose from players' previous behavior and experience. These models assume strategies have numerical evaluations, which are called "attractions." Learning rules are defined by how attractions are updated in response to experience. Attractions are then mapped into predicted choice probabilities for strategies using some well-known statistical rule (such as logit).

The three major approaches to learning assume players are adaptive (i.e., they respond only to their own previous experience and ignore others' payoff information) and that their behavior is not sensitive to the way in which players are matched. Empirical evidence suggests otherwise. There are subjects who can anticipate how others learn and choose actions to influence others' path of learning in order to benefit themselves. So, we describe a generalization of these adaptive learning models to allow this kind of sophisticated behavior. This generalized model assumes that there is a mixture of adaptive learners and sophisticated players. An adaptive learner adjusts his behavior according to one of the above learning rules. A sophisticated player does not learn and rationally best-responds to his forecast of others' learning behavior. This model therefore allows "one-stop shopping" for investigating the various statistical comparisons of learning and equilibrium models.

2 EWA Learning

Denote player i's jth strategy by s_i^j and the other player(s)' strategy by s_{-i}^k. The strategy actually chosen in period t is $s_i(t)$. Player i's payoff for choosing s_i^j in period t is $\pi_i(s_i^j, s_{-i}^k(t))$. Each strategy has a numerical evaluation at time t, called an attraction $A_i^j(t)$. The model also has an experience weight, $N(t)$. The variables $N(t)$ and $A_i^j(t)$ begin with prior values and are updated each period. The rule for updating attraction sets $A_i^j(t)$ to be the sum of a depreciated, experience-weighted previous attraction $A_i^j(t-1)$ plus the (weighted) payoff from period t, normalized by the updated experience weight:

[2]While these learning models are primarily designed for games, they have also been applied to predict product choice at supermarkets. For instance, Ho and Chong (2003) use EWA learning to analyze 130,265 purchase incidences and show that it can predict product choice remarkably well.

$$A_i^j(t) = \frac{\phi \cdot N(t-1) \cdot A_i^j(a, t-1) + [\delta + (1-\delta) \cdot I(s_i^j, s_i(t))] \cdot \pi_i(s_i^j, s_{-i}(t))}{N(t)}. \quad (2.1)$$

where indicator variable $I(x, y)$ is 1 if $x = y$ and 0 otherwise. The experience weight is updated by:

$$N(t) = \rho \cdot N(t-1) + 1. \quad (2.2)$$

Let $\kappa = \frac{\phi - \rho}{\phi}$. Then $\rho = \phi \cdot (1 - \kappa)$ and $N(t)$ approaches the steady-state value of $\frac{1}{1-\phi\cdot(1-\kappa))}$. If $N(0)$ begins below this value, it steadily rises, capturing an increase in the weight placed on previous attractions and a (relative) decrease in the impact of recent observations, so that learning slows down.

Attractions are mapped into choice probabilities using a logit rule (other functional forms fit about equally well; Camerer and Ho, 1998):

$$P_i^j(t+1) = \frac{e^{\lambda \cdot A_i^j(t)}}{\sum_k e^{\lambda \cdot A_i^k(t)}}. \quad (2.3)$$

where λ is the payoff sensitivity parameter. The key parameters are δ, ϕ, and κ (which are generally assumed to be in the [0,1] interval).

The most important parameter, δ, is the weight on foregone payoffs relative to realized payoffs. It can be interpreted as a kind of "imagination" of foregone payoffs, or responsiveness to foregone payoffs (when δ is larger players move more strongly toward ex post best responses). We call it "consideration" of foregone payoffs. The weight on foregone payoff δ is also an intuitive way to formalize the "learning direction" theory of Selten and Stoecker (1986). Their theory consists of an appealing property of learning: Subject move in the direction of ex-post best-response. Broad applicability of the theory has been hindered by defining "direction" only in terms of numerical properties of ordered strategies (e.g., choosing 'higher prices' if the ex-post best response is a higher price than the chosen price). The parameter δ defines the "direction" of learning set-theoretically

by shifting probability toward the set of strategies with higher payoffs than the chosen ones.

The parameter ϕ is naturally interpreted as depreciation of past attractions, $A_i^j(t-1)$. In a game-theoretic context, ϕ will be affected by the degree to which players realize other players are adapting, so that old observations on what others did become less and less useful. So we can interpret ϕ as an index of (perceived) "change" in the environment.

The parameter κ determines the growth rate of attractions, which in turn affects how sharply players converge. When $\kappa = 0$, the attractions are weighted averages of lagged attractions and payoff reinforcements (with weights $\phi \cdot N(t-1)/(\phi \cdot N(t-1)+1)$ and $1/(\phi \cdot N(t-1)+1)$. When $\kappa = 1$ and $N(t) = 1$, the attractions are cumulations of previous reinforcements rather than averages (i.e., $A_i^j(t) = \phi \cdot A_i^j(t-1) + [\delta + (1-\delta) \cdot I(s_i^j, s_i(t))] \cdot \pi_i(s_i^j, s_{-i}(t))$). In the logit model, the *differences* in strategy attractions determine their choice probabilities. When κ is high the attractions can grow furthest apart over time, making choice probabilities closer to zero and one. We therefore interpret κ as an index of "commitment".

3 Reinforcement Learning

In cumulative reinforcement learning (Harley, 1981; Roth and Erev, 1995), strategies have levels of attraction which are incremented by only received payoffs. The initial reinforcement level of strategy j of player i, s_i^j, is $R_i^j(0)$. Reinforcements are updated as follows:

$$R_i^j(t) = \begin{cases} \phi \cdot R_i^j(t-1) + \pi_i(s_i^j, s_{-i}(t)) & \text{if } s_i^j = s_i(t), \\ \phi \cdot R_i^j(t-1) & \text{if } s_i^j \neq s_i(t). \end{cases} \tag{3.1}$$

Using the indicator function, the two equations can be reduced to one:

$$R_i^j(t) = \phi \cdot R_i^j(t-1) + I(s_i^j, s_i(t)) \cdot \pi_i(s_i^j, s_{-i}(t)). \tag{3.2}$$

This updating formula is a special case of the EWA rule, when $\delta = 0$, $N(0) = 1$, and $\kappa = 1$.

In average reinforcement learning, updated attractions are averages of previous attractions and received payoffs (e.g. Mookerjhee and Sopher, 1994, 1997; Erev and Roth, 1998). For example

$$R_i^j(t) = \phi \cdot R_i^j(t-1) + (1-\phi) \cdot I(s_i^j, s_i(t)) \cdot \pi_i(s_i^j, s_{-i}(t)). \tag{3.3}$$

A little algebra shows that this updating formula is also a special case of the EWA rule, when $\delta = 0$, $N(0) = \frac{1}{1-\phi}$, and $\kappa = 0$. Since the two reinforcement models are special cases of EWA learning, their predictive adequacy can be tested empirically by setting the appropriate EWA parameters to their restricted values and seeing how much fit is compromised (adjusting, of course, for degrees of freedom).

4 Belief Learning

In belief-based models, adaptive players base their responses on beliefs formed by observing their opponents' past plays. While there are many ways of forming beliefs, we consider a fairly general "weighted fictitious play" model, which includes fictitious play (Brown, 1951; Fudenberg and Levine, 1998) and Cournot best-response (Cournot, 1960) as special cases. It corresponds to bayesian learning if players have a Dirichlet prior belief.[3]

In weighted fictitious play, players begin with prior beliefs about what the other players will do, which are expressed as ratios of strategy choice counts to the total experience. Denote total experience by $N(t) = \sum_k N_{-i}^k(t)$. Express the belief that others will play strategy k as $B_{-i}^k(t) = \frac{N_{-i}^k(t)}{N(t)}$, with $N_{-i}^k(t) \geq 0$ and $N(t) > 0$.

Beliefs are updated by depreciating the previous counts by ϕ, and adding one for the strategy combination actually chosen by the other players. That is,

$$B_{-i}^k(t) = \frac{\phi \cdot N_{-i}^k(t-1) + I(s_{-i}^k, s_{-i}(t))}{\sum_h [\phi \cdot N_{-i}^h(t-1) + I(s_{-i}^h, s_{-i}(t))]}. \tag{4.1}$$

[3]This class of models however does not include the belief-type learning approach proposed by Crawford (1995).

This form of belief updating weights the belief from one period ago ϕ times as much as the most recent observation, so ϕ can be interpreted as how quickly previous experience is discarded. When $\phi = 0$ players weight only the most recent observation (Cournot dynamics); when $\phi = 1$ all previous observations count equally (fictitious play).

Given these beliefs, we can compute expected payoffs in each period t,

$$E_i^j(t) = \sum_k B_{-i}^k(t)\pi(s_i^j, s_{-i}^k). \tag{4.2}$$

The crucial step is to express period t expected payoffs as a function of period $t-1$ expected payoffs. This yields:

$$E_i^j(t) = \frac{\phi \cdot N(t-1) \cdot E_i^j(t-1) + \pi(s_i^j, s_{-i}(t))}{\phi \cdot N(t-1) + 1}. \tag{4.3}$$

By expressing expected payoffs as a function of lagged expected payoffs, we make the belief terms disappear. This is because the beliefs are only used to compute expected payoffs, and when beliefs are formed according to weighted fictitious play, the expected payoffs which result can also be generated by generalized reinforcement according to previous payoffs. More precisely, if the initial attractions in the EWA model are expected payoffs given some initial beliefs (i.e., $A_i^j(0) = E_i^j(0)$), $\kappa = 0$ (or $\phi = \rho$), and foregone payoffs are weighted as strongly as received payoffs ($\delta = 1$), then EWA attractions are exactly the same as expected payoffs. Put differently, belief learning is "mathematically equivalent" or "observationally equivalent" to EWA learning with $\delta = 1$, $\kappa = 0$ and $A_i^j(0) = E_i^j(0)$.

This demonstrates a close kinship between reinforcement and belief approaches. Belief learning is nothing more than generalized attraction learning in which strategies are reinforced equally strongly by actual payoffs and foregone payoffs and attractions are weighted averages of past attractions and reinforcements.[4]

[4]Hopkins (2002) compares the convergence properties of reinforcement and fictitious play and finds that they are quite similar in nature and that they will in many cases have the same asymptotic behavior.

5 A Graphical Representation

Since reinforcement and belief learning are special cases of EWA learning, it is possible to represent all three learning models in a three-dimensional EWA cube (see Figure 1). The vertex $\delta = 1$ and $\kappa = 0$ corresponds, to weighted fictitious play models. The corners $\phi = 0$ and $\phi = 1$ correspond to Cournot best-response dynamics and fictitious play, respectively. Reinforcement models in which only chosen strategies are reinforced according to their payoffs correspond to vertices in which $\delta = 0$, and $\kappa = 1$ (cumulative reinforcement) or $\kappa = 0$ (averaged reinforcement). Interior configurations of parameter values incorporate both the intuition behind reinforcement learning, that realized payoffs weigh most heavily ($\delta < 1$), and the intuition implicit in belief learning, that foregone payoffs matter too ($\delta > 0$).

The cube shows that contrary to popular belief for many decades, reinforcement and belief learning are simply two extreme configurations on opposite edges of a three-dimensional cube, rather than fundamentally unrelated models. Figure 1 also shows estimates of the three parameters in 20 different studies (Camerer, Ho, and Chong, 2002). Each point is a triple of estimates.[5] Most points are sprinkled throughout the cube, rather than at the extreme vertices mentioned in the previous paragraph, although some (generally from games with mixed-strategy equilibria) are near the averaged reinforcement corner $\delta = 0$ and $\kappa = \phi = 1$.[6] Parameter estimates are generally significantly inside the interior of the cube, rather than near the vertices. Thus, we may conclude that subjects' behavior is often neither belief nor reinforcement learning.[7] The cube shows which games these nested approaches fail and why.

[5]These parameter estimates were typically obtained by the maximum likelihood method. Initial attractions could be either estimated using data or set to plausible values using the cognitive hierarchy model of one-shot games (see Camerer, Ho, and Chong, 2004).

[6]Ho, Camerer, and Chong (in press) provide an explanation for how δ and ϕ vary across games by endogenizing them as functions of game experience. This self-tuning approach provides a one-parameter model, which makes EWA learning more amenable to field applications.

[7]Players are assumed to have the same learning parameters. Allowing for heterogeneity often improves fit. However, the population is not simply a mixture of reinforcement and belief learning. In fact, such a mixture is significantly worse than a mixture of two types of EWA learners in predictive performance (Camerer and Ho, 1998).

6 Linking Learning and Equilibrium Models

The adaptive learning models presented above do not permit players to anticipate learning by others. Omitting anticipation logically implies that players do not use information about the payoffs of other players, and that whether players are matched together repeatedly or are randomly re-matched should not matter. Both of the latter implications are unintuitive and experiments with experienced subjects have provided evidence to show otherwise.

In Camerer, Ho, and Chong (2002) and Chong, Camerer, and Ho (2006), we proposed a simple way to include "sophisticated" anticipation by some players that others are learning, using two additional parameters. We assume a fraction α of players are sophisticated. Sophisticated players think that a fraction $(1-\alpha')$ of players are adaptive and the remaining fraction α' of players are sophisticated like themselves. They use the EWA model (which nests reinforcement and belief learning as special cases) to forecast what the adaptive players will do, and choose strategies with high expected payoffs given their forecast.

All the adaptive models discussed above (EWA, reinforcement, belief learning) are special cases of this generalized model with $\alpha = 0$. The assumption that sophisticated players think some others are sophisticated, creates a small whirlpool of recursive thinking which implies that quantal response equilibrium (QRE; McKelvey and Palfrey, 1995) and Nash equilibrium, are special cases of this generalized model. Our specification also shows that equilibrium concepts combine two features which are empirically and psychologically separable: "social calibration" (accurate guesses about the fraction of players who are sophisticated, $\alpha = \alpha'$); and full sophistication ($\alpha = 1$). Psychologists have identified systematic departures from social calibration called "false uniqueness" or overconfidence ($\alpha > \alpha'$) and "false consensus" or curse of knowledge ($\alpha < \alpha'$).

Formally, adaptive learners follow the EWA updating equations given above (i.e., (2.1) and (2.2)). Sophisticated players have attractions $B_i^j(t)$ and choice probabilities $Q_i^j(t+1)$ specified as follows:

$$B_i^j(t) = \sum_k [(1-\alpha') \cdot P_{-i}^k(t+1) + \alpha' Q_{-i}^k(t+1)] \cdot \pi_i(s_i^j, s_{-i}^k), \quad (6.1)$$

$$Q_i^j(t+1) = \frac{e^{\lambda \cdot B_i^j(t)}}{\sum_k e^{\lambda \cdot B_i^k(t)}}. \quad (6.2)$$

The generalized model has been applied to experimental data from 10-period p-beauty contest games (Ho, Camerer, and Weigelt, 1998). In these games, seven subjects choose numbers in [0,100] simultaneously. The subject whose number is closest to p times the average (where $p = .7$ or .9) wins a fixed prize. Subjects playing for the first time are called "inexperienced"; those playing another 10-period game (with a different p) are called "experienced".

Table 1 reports results and parameter estimates. For inexperienced subjects, adding sophistication to adaptive EWA improves log likelihood (LL) substantially both in- and out-of-sample. The estimated fraction of sophisticated players is $\hat{\alpha} = .236$ and their estimated perception $\hat{\alpha}' = 0$. The consideration parameter δ is estimated to be .781.

Experienced subjects show a larger improved fit from sophistication, and a larger estimated proportion, $\hat{\alpha} = .752$. (Their perceptions are again too low, $\hat{\alpha'} = .413$, showing a degree of overconfidence.) The increase in sophistication due to experience reflects a kind of "learning about learning," which is similar to rule learning (i.e., subjects switch their learning rule over time) (Stahl, 2000).[8]

Figure 2a shows actual choice frequencies for experienced subjects across the ten periods. Figures 2b-e show predicted frequencies for cumulative reinforcement, weighted fictitious play, EWA, and the generalized model. Figure 2b shows that cumulative reinforcement model learns far too slowly because only one player wins each period and the losers get no reinforcement. Figure 2c shows that belief models with low values of ϕ update beliefs very quickly but do not capture anticipatory learning, in which some

[8]There is also an empirical question of whether learning parameters should be the same across games. For prediction, it is best if parameters are stable and universal across games, but that is unlikely to be true if the parameters reflect behavioral tendencies that respond to changes in games. The challenge is then to model how the "self-tuning" change occurs (see Ho, Camerer, and Chong, in press).

Table 1. Parameter Estimates for p-beauty Contest Game

	inexperienced subjects		experienced subjects	
	Generalized Model	EWA Learning	Generalized Model	EWA Learning
ϕ	**0.436**	0.000	**0.287**	0.220
δ	**0.781**	0.900	**0.672**	0.991
κ	**1.000**	1.000	**0.927**	1.000
N(O)	**0.253**	0.000	**0.000**	0.887
α	**0.236**	0.000	**0.752**	0.000
α'	**0.000**	0.000	**0.412**	0.000
LL (in sample)	**-2095.32**	-2155.09	**-1908.48**	-2128.88
LL (out of sample)	**-968.24**	-992.47	**-710.28**	-925.09

subjects anticipate others are learning, best-respond, and leapfrog ahead. As a result, the frequency of low choices (1-10) predicted by belief learning only grows from 20% in period 5 to 35% in period 10, while the actual frequencies grow from 40% to 55%. Similarly, EWA learning is not able to capture the fast equilibration of experienced subjects (Figure 2d). Adding sophistication (Figure 2e) captures the actual frequencies quite closely.

7 Conclusions

We describe three major approaches of adaptive learning models. We show that EWA learning is a generalization of reinforcement and belief learning and that the latter two nested models are intimately related. Specifically, they differ mainly in the way they treat foregone payoffs; reinforcement learning ignores them and belief learning treats them the same as actual payoffs. Estimation results from dozens of studies show that the emergence of behavior is neither reinforcement nor belief learning in most games. The

EWA cube provides a simple way for detecting how these simpler models fail and why.

We also describe a generalization of these adaptive models to study anticipation by some players that others are learning. This generalized model nests equilibrium and the adaptive learning models as special cases and is a powerful framework for analyzing both equilibrium and learning simultaneously. We show that it can improve the predictive performance of the adaptive learning models when players are experienced and able to anticipate how others learn.

References

[1] Brown, G. "Iterative Solution of Games by Fictitious Play," In Activity Analysis of Production and Allocation, New York: John Wiley & Sons, 1951.

[2] Camerer, Colin F. and Teck-Hua Ho. "EWA learning in normal-form games: Probability rules, heterogeneity and time-variation," Journal of Mathematical Psychology, 42, 1998.

[3] Camerer, Colin F. and Teck-Hua Ho. "Experience-weighted attraction learning in normal-form games," Econometrica, July 1999, 67, 827-874.

[4] Camerer, C. Ho, T-H., and Chong, J-K., "Sophisticated Learning and Strategic Teaching," Journal of Economic Theory, 104 (2002), 137-18

[5] Camerer, C.F., Ho, T-H, Chong, J-K. "A Cognitive Hierarchy Model of One-Shot Games," Quarterly Journal of Economics, August 2004, 119(3), 861-898.

[6] Chong J-K, Camerer, C., and Ho, T-H. "A Learning-based Model of Repeated Games with Incomplete Information," Games and Economic Behavior, Vol. 55, no. 2, pp. 340-371, 2006.

[7] Cournot, A. *Recherches sur les principes mathematiques de la theorie des richesses.* Translated into English by N. Bacon as Researches in the Mathematical Principles of the Theory of Wealth. London: Haffner, 1960.

[8] Crawford, V. P. "Adaptive Dynamics in Coordination Games," Econometrica, 63, 103-143, 1995.

[9] Erev, Ido and Roth, Alvin. (1998) "Modelling Predicting How People Play Games: Reinforcement learning in experimental games with unique, mixed-strategy equilibria," American Economic Review, vol. 88, no. 4, pp. 848-881 1998.

[10] Friedman, D. "Evolutionary Games in Economics," Econometrica, Vol. 59, No. 3 (May 1991), 637-666.

[11] Fudenberg, Drew and Levine, David. The Theory of Learning in Games. Cambridge: MIT Press, 1998.

[12] Harley, Calvin, "Learning the Evolutionary Stable Strategies," Journal of Theoretical Biology, 89, (1981), pp. 611-633.

[13] Ho, Teck-Hua, Camerer, Colin, and Weigelt, Keith, "Iterated Dominance and Iterated Best-response in p-Beauty Contests," American Economic Review, 1998, 88:4, 947-969.

[14] Ho, T-H., Camerer, C. and Chong, J-K. "Self-tuning Experience-Weighted Attraction Learning in Games," Journal of Economic Theory, in press.

[15] Ho, T-H. and Chong, J-K., "A Parsimonious Model of SKU Choice," Journal of Marketing Research, Vol. XL (Aug 2003), 351-365.

[16] Hopkins, Edward, "Two Competing Models of how People Learn in Games," *Econometrica*, 70(6), (2002), pp. 2141-2166.

[17] McKelvey, Richard and Thomas Palfrey, "Quantal Response Equilibria for Normal Form Games," *Games and Economic Behavior*, 10, (1995), pp. 6-38.

[18] Mookerjhee, Dilip, and Barry Sopher, "Learning Behavior in an Experimental Matching Pennies Game," Games and Economic Behavior, 7, (1994), pp. 62-91.

[19] Roth, Alvin E. and Ido Erev, "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Models in the Intermediate Term," Games and Economic Behavior, 8(1), (1995), pp. 164-212.

[20] Schlag, Karl. (1999). "Which One Should I Imitate?" Journal of Mathematical Economics, 31(4), 493522.

[21] Selten, Reinhard and Rolf Stoecker. "End behavior in sequences of finite prisoner's dilemma supergames: A learning theory approach." Journal of Economic Behavior and Organization. 7, 1986, 47-70.

[22] Stahl, D. "Rule Learning in Symmetric Normal-Form Games," Games and Economic Behavior: Theory and Evidence, 32, 2000, 105-138.

[23] Weibull, J. Evolutionary Game Theory, MIT Press, 1995.

Figure 1: EWA's Model Parametric Space

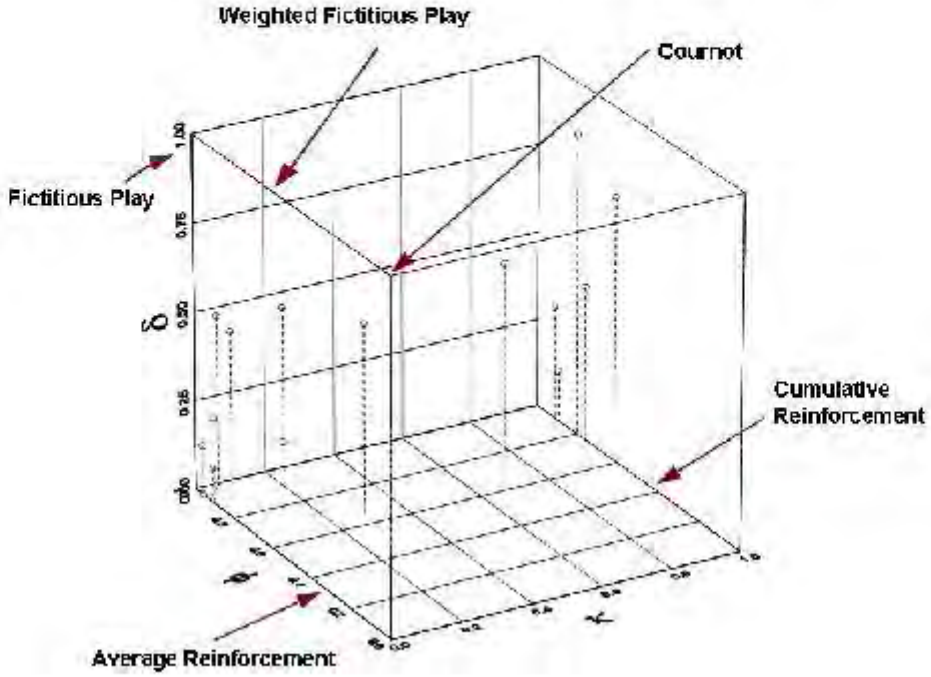

Figure 2a: Actual Choice Frequencies for Experienced Subjects

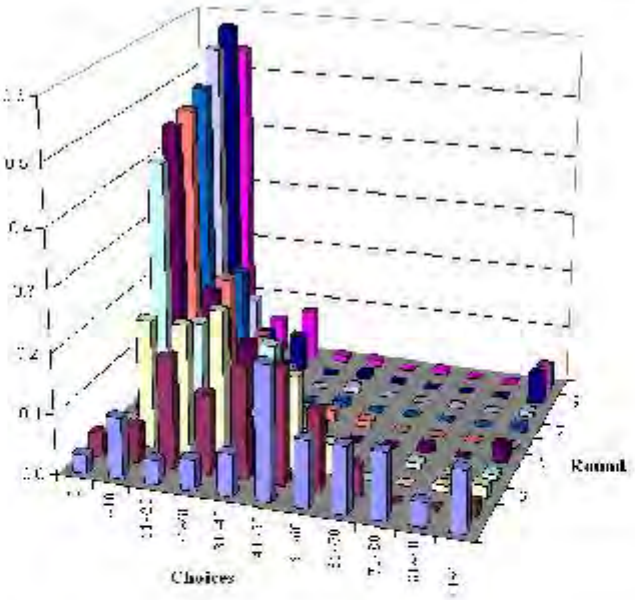

Figure 2b: Cumulative Reinforcement Learning Predicted Frequencies for Experienced Subjects

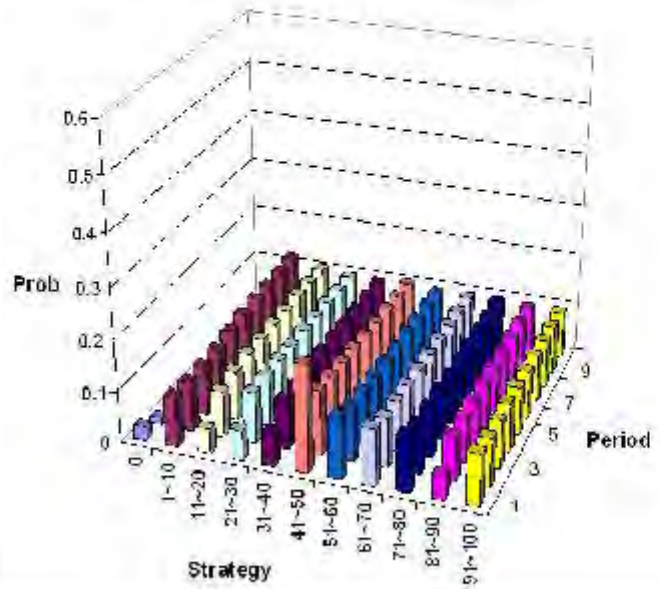

Figure 2c: Belef Learning Predicted Frequencies for Experienced Subjects

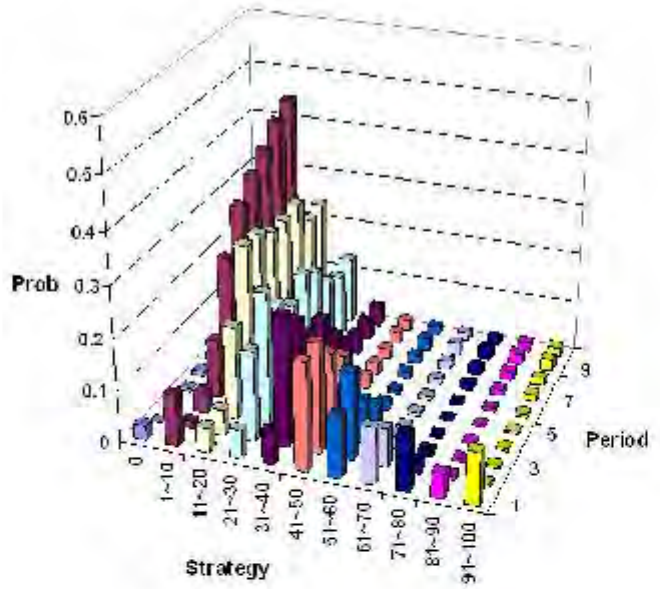

Figure 2d: EWA Learning Predicted Frequencies for Experienced Subjects

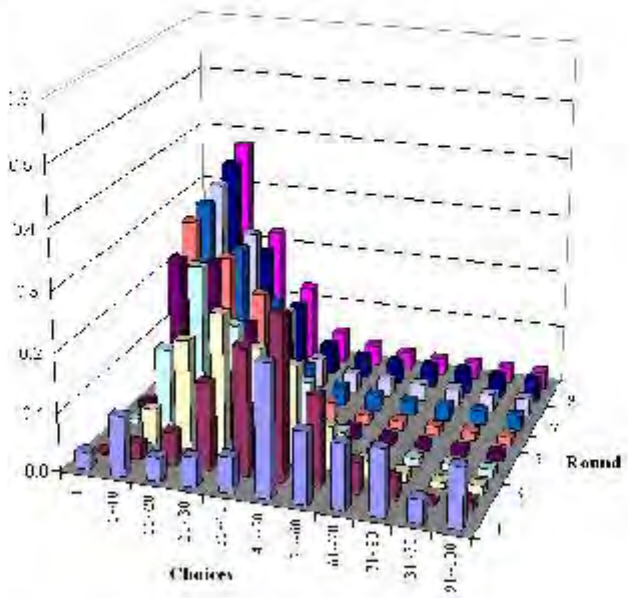

Figure 2e: Generalized Modal Predicted Frequencies for Experienced Subjects