Replicating a Model:

Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria

By Ido Erev AND Alvin E. Roth

Presented by James Paine
February 15, 2019


Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria

By Ido Erev and Alvin E. Roth*

We examine learning in all experiments we could locate involving 100 periods or more of games with a unique equilibrium in mixed strategies, and in a new experiment. We study both the ex post ("best fit") descriptive power of learning models, and their ex ante predictive power, by simulating each experiment using parameters estimated from the other experiments. Even a one-parameter reinforcement learning model robustly outperforms the equilibrium predictions. Predictive power is improved by adding "forgetting" and "experimentation," or by allowing greater rationality as in probabilistic fictitious play. Implications for developing a low-rationality, cognitive game theory are discussed. (JEL C72, C92)

Game theory has traditionally been developed as a theory of strategic interaction among players who are perfectly rational, and who (consequently) exhibit equilibrium behavior. This approach has been complemented by evolutionary game theory, which, motivated by biological evolution, seeks to understand how equilibria could arise in the long term by selection among generations of players who need not be rational or even conscious decision makers. Somewhere in between are models of learning, which consider the adaptive behavior of goal-oriented players who may not be highly rational, both to provide foundations for theories of equilibrium and to model empirically observed behavior.

The present paper considers how well simple learning models, motivated by the psychology of learning, can model the interaction of players who must learn about the game and each other in the course of playing the game, over time spans that may not be long enough to lead to equilibrium. Our goal will be to model observed behavior, starting with behavior observed in experimental settings. (In the conclusion we will also consider the implications of this approach for applied economics in naturally occurring, nonexperimental settings.) We will show that a wide range of experimental data can be both well described ex post and robustly predicted ex ante by a very simple family of learning theories.

Economists have traditionally avoided explaining behavior as less than rational for fear of developing many fragmented theories of mistakes. Part of the attraction of highly rational models is the idea that there may be many ways to be less than rational, but only one way (or in light of the equilibrium refinement literature perhaps only a few ways) of being highly rational. In this view, the success in economics of the assumptions of utility maximization and equilibrium behavior is in large part due to the prospect that they may provide a useful approximation of great generality, even if they are not precisely correct models of human behavior (cf., Roth, 1996a).

* Erev: Faculty of Industrial Engineering and Management, Technion, Haifa, Israel 32000, and Department of Economics, University of Pittsburgh, Pittsburgh, PA 15260 (e-mail: erev@tx.technion.ac.il); Roth: Department of Economics, Harvard University, Cambridge, MA 02138, and Harvard Business School, Boston, MA 02163 (e-mail: aroth@hbs.edu). The work of both authors is partially supported by grants from the National Science Foundation. We have benefited from helpful conversations with Yaelle Bereby-Meyer, Nick Feltovich, Daniel Gopher, Joachim Meyer, Ayala Cohen, Dan Hanuremesh, and Shmuel Zamir. Yaelle Bereby-Meyer also contributed to the design and programming of the new experiment. We are indebted to Barry O'Neill, Jack Ochs, and Antoon Rapoport for access to unpublished parts of their data. The present version reflects numerous comments by three anonymous referees on several earlier drafts. This work was completed while Roth was at the University of Pittsburgh.

$$q_{n,k,1}=q_{n,j,1} \text{ for all } k,j$$

$$p_{n,k,t}=\frac{q_{n,k,t}}{\sum_j q_{n,j,t}}$$

$$R(x)=x-x\_min$$

$$q_{n,j,t+1}=\begin{cases} q_{n,j,t}+R(x_n) & \text{if } j=i \\ q_{n,j,t} & \text{otherwise} \end{cases}$$ *i.e. minimum payout by player*

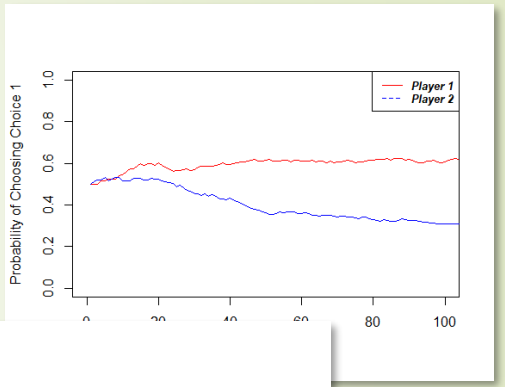
$$p_{n,j,1}=\frac{1}{M_n} \quad \text{where } M_n \text{ is the number of player } n\text{'s choices}$$

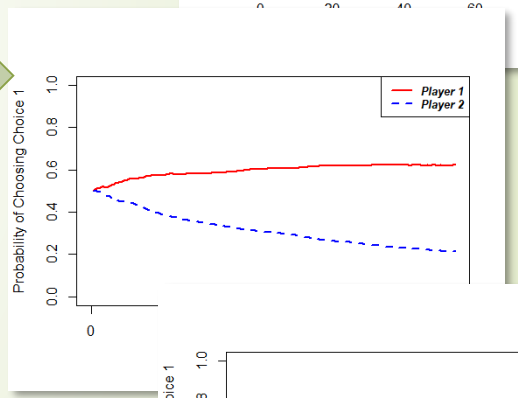
$$s_n=\frac{\sum_j q_{n,j,1}}{X_n}$$

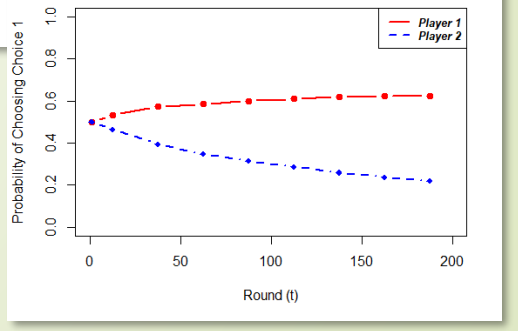
$$q_{n,j,t+1}=(1-\phi)\,q_{n,j,t}+f(R,j,\varepsilon,M)$$

$$\text{where } f(R,j,\varepsilon,M)=\begin{cases} R(x)(1-\varepsilon) & \text{if } j=k \\ R(x)\varepsilon/2 & \text{if } j=k\pm 1 \\ 0 & \text{otherwise} \end{cases}$$

```
for (i in 1:Iterations) {
cat("\014")
print(paste0("Iteration: ", i, " of ", Iterations))

#initalize the probability and q array space for this iteratio
p = array(data = NA, dim = c(NumPlayer,max(as.numeric(M[1:Num
q=p

#initialize the choice array k for this iteration
k = array(data = NA, dim = c(NumPlayer,Rounds))

#Initialization round
#set t=1 values for the probabilities and propensities
for (n in 1:NumPlayer) {
  for (j in 1:as.numeric(M[n])) {

    p[n,j,1]= 1/as.numeric(M[n])
    q[n,j,1]= p[n,j,1]*S1*as.numeric(X[n])
  }
}

#Loop through subsequent rounds
for(t in 2:Rounds) {

  for(n in 1:NumPlayer){
    #Probabilistically get player n's choice
    k[n,t]=sample(c(1:as.numeric(M[n])),1,replace=FALSE,prob=p
  }

  #Look at the payout matrix to determine which row matches t
  FullPayout = Payout[apply(Payout[,1:NumPlayer], 1, function
  R = FullPayout[,(NumPlayer+1):ncol(Payout)]-x_min

  for(n in 1:NumPlayer){

    #for each strategy option, see if a reward was observed
    for(j in 1:as.numeric(M[n])){

      q[n,j,t] = q[n,j,(t-1)]

      if (as.numeric(k[n,t]) == j) {
        #update the propensities based on the reward observed
        q[n,j,t] =  q[n,j,(t-1)]+as.numeric(R[n])
      }
    } #next j in the observation of choice and propentsity up

    #Update probabilities based on newly updated propensities
    for(j in 1:NumPlayer){
      p[n,j,t] = q[n,j,t]/sum(q[n,1:as.numeric(M[n]),t])
    } #next j in the update of probability loop
```







Paper Background

- Published in The American Economic Review in September 1998
- Based, in part, on earlier work by Erev and Roth from 1995
- Conducts a review of 12 previous studies of two player economic games
- Presents three reinforcement learning models and compares the model to the empirical data from the previous studies
- Concludes that the base, 1 parameter, model and the intermediary 3-parameter model, are able to replicate empirical data better (in general and when tuned) than equilibrium estimates
- Base model has been generalized for other RL models and follows similar choice-theory work

Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria

By Ido Erev and Alvin E. Roth*

We examine learning in all experiments we could locate involving 100 periods or more of games with a unique equilibrium in mixed strategies, and in a new experiment. We study both the ex post ("best fit") descriptive power of learning models, and their ex ante predictive power, by simulating each experiment using parameters estimated from the other experiments. Even a one-parameter reinforcement learning model robustly outperforms the equilibrium predictions. Predictive power is improved by adding "forgetting" and "experimentation," or by allowing greater rationality as in probabilistic fictitious play. Implications for developing a low-rationality, cognitive game theory are discussed. (JEL C72, C92)

Game theory has traditionally been developed as a theory of strategic interaction among players who are perfectly rational, and who (consequently) exhibit equilibrium behavior. This approach has been complemented by evolutionary game theory, which, motivated by biological evolution, seeks to understand how equilibria could arise in the long term by selection among generations of players who need not be rational or even conscious decision makers. Somewhere in between are models of learning, which consider the adaptive behavior of goal-oriented players who may not be highly rational, both to provide foundations for theories of equilibrium and to model empirically observed behavior.

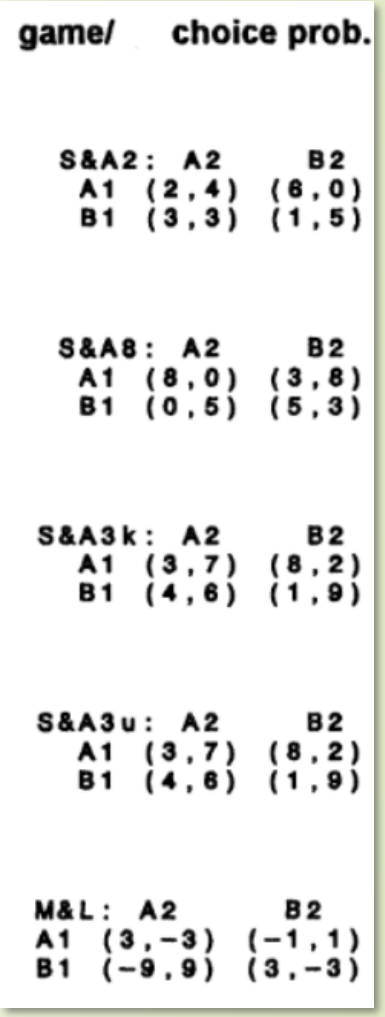
The present paper considers how well simple learning models, motivated by the psychology of learning, can model the interaction of players who must learn about the game and each other in the course of playing the game, over time spans that may not be long enough to lead to equilibrium. Our goal will be to model observed behavior, starting with behavior observed in experimental settings. (In the conclusion we will also consider the implications of this approach for applied economics in naturally occurring, nonexperimental settings.) We will show that a wide range of experimental data can be both well described *ex post* and robustly predicted *ex ante* by a very simple family of learning theories.

Economists have traditionally avoided explaining behavior as less than rational for fear of developing many fragmented theories of mistakes. Part of the attraction of highly rational models is the idea that there may be many ways to be less than rational, but only one way (or in light of the equilibrium refinement literature perhaps only a few ways) of being highly rational. In this view, the success in economics of the assumptions of utility maximization and equilibrium behavior is in large part due to the prospect that they may provide a useful approximation of great generality, even if they are not precisely correct models of human behavior (cf., Roth, 1996a).

* Erev: Faculty of Industrial Engineering and Management, Technion, Haifa, Israel 32000, and Department of Economics, University of Pittsburgh, Pittsburgh, PA 15260 (e-mail: erev@technion.technion.ac.il); Roth: Department of Economics, Harvard University, Cambridge, MA 02138, and Harvard Business School, Boston, MA 02163 (e-mail: aroth@hbs.edu; http://www.economics.harvard.edu/faculty/roth/roth.html). The work of both authors is partially supported by grants from the National Science Foundation. We have benefitted from helpful conversations with Yoella Bereby-Meyer, Nick Feltovich, Daniel Gopher, Joachim Meyer, Ayala Cohen, Dan Hamermesh, and Shmuel Zamir. Yoella Bereby-Meyer also contributed to the design and programming of the new experiment. We are indebted to Barry O'Neill, Jack Ochs, and Amnon Rapoport for access to unpublished parts of their data. The present version reflects numerous comments by three anonymous referees on several earlier drafts. This work was completed while Roth was at the University of Pittsburgh.

Model Details

- Assumes players only see the results of their choices, i.e. have no knowledge of payouts a priori
- Based on 'propensities' that change as players receive a reward for their actions
- Each player probabilistically chooses based on the weighted value of the propensity of any choice relative to the other choices
- Each model version adds elements to modify the propensity update routine
- General routine:
  i. Each player randomly picks an action based on their propensities
  ii. Payoffs by player are determined by the payout matrix
  iii. Players update their propensities based on the observed reward
  iv. Repeat above



Model Details

- In general, all propensities (q) start equal (at t=1 or t=0)

$$q_{n,k,1}=q_{n,j,1}\ for\ all\ k,j$$

- The probability (p) of any choice (j or k) is the relative weight of that propensity
  - Therefore, all choices have an equal chance at the first time step

$$p_{n,k,t} = \frac{q_{n,k,t}}{\sum_j q_{n,j,t}}$$

- The reward received is *relative to the minimum reward possible*

$$R(x) = x - x\_min$$

- The propensity of a choice is updated *if and only if* that choice was chosen (k) and a reward was observed

$$q_{n,j,t+1} = \begin{cases} q_{n,j,t} + R(x)\ if\ j = k \\ q_{n,j,t} \quad otherwise \end{cases}$$

Model Details

- 1-Parameter Model
  - Used parameter called ‘strength’ to determine the initial propensities
  - Value scales the rate at which the curves change (ie the ‘speed of learning’)

$$X_n = |\mathrm{avg}(\mathrm{x_n})| \quad i.e. minimum\ payout\ by\ player$$

$$p_{n,j,1} = \frac{1}{M_n} \quad where\ M_n\ is\ the\ number\ of\ player\ n's\ choices$$

$$s_n = \frac{\sum_j q_{n,j,1}}{X_n}$$

$$\therefore q_{n,j,1} = (p_{n,j,1})(s_n)(X_n)$$

  - The above is used to initialize the model at the first time step
  - The value of this strength term was tuned in the paper to fit all 12 referenced datasets

Model Details

- 3-Parameter Model
  - Keeps the 'strength' parameter to initialize the model
  - Adds two additional terms: ε (or 'experimentation') and φ (or 'forgetting')
  - φ essentially acts as a discount factor, reducing the importance of the current value of the propensity relative to the new reward
  - ε updates propensities of 'nearby' choice after a reward is realized

$$q_{n,j,t+1} = (1 - \varphi)\, q_{n,j,t} + f(R, \mathrm{j}, \varepsilon, \mathrm{M})$$

$$where\ f(R, \mathrm{j}, \varepsilon, \mathrm{M}) = \begin{cases} R(x)(1-\varepsilon) & if\ j = k \\ R(x)\varepsilon/2 & if\ j = k \pm 1 \\ 0 & otherwise \end{cases}$$

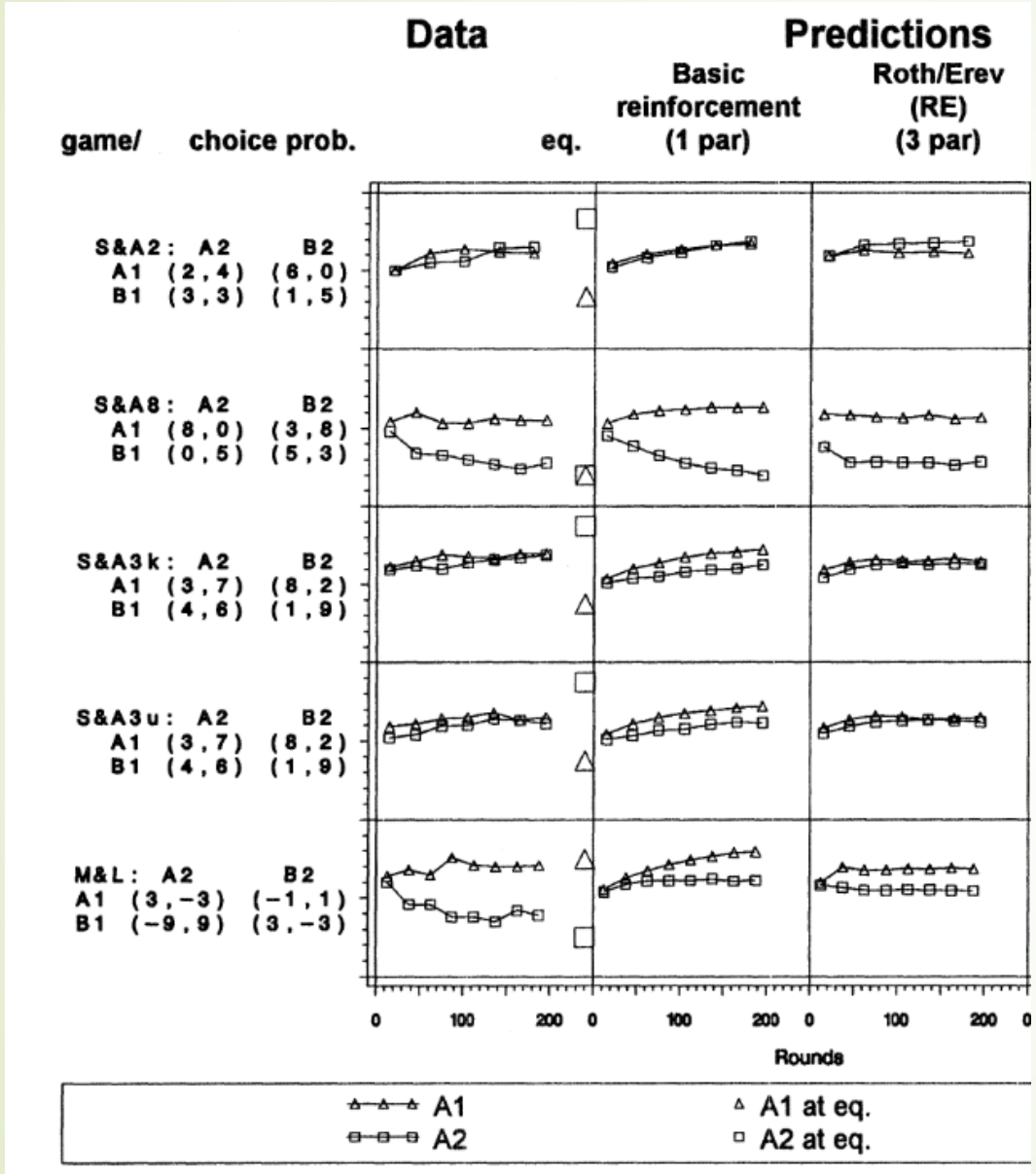
  - *Note: there's a slightly more complicated f() expression when the number of choices is greater than 2. This is incorporated into the R script*

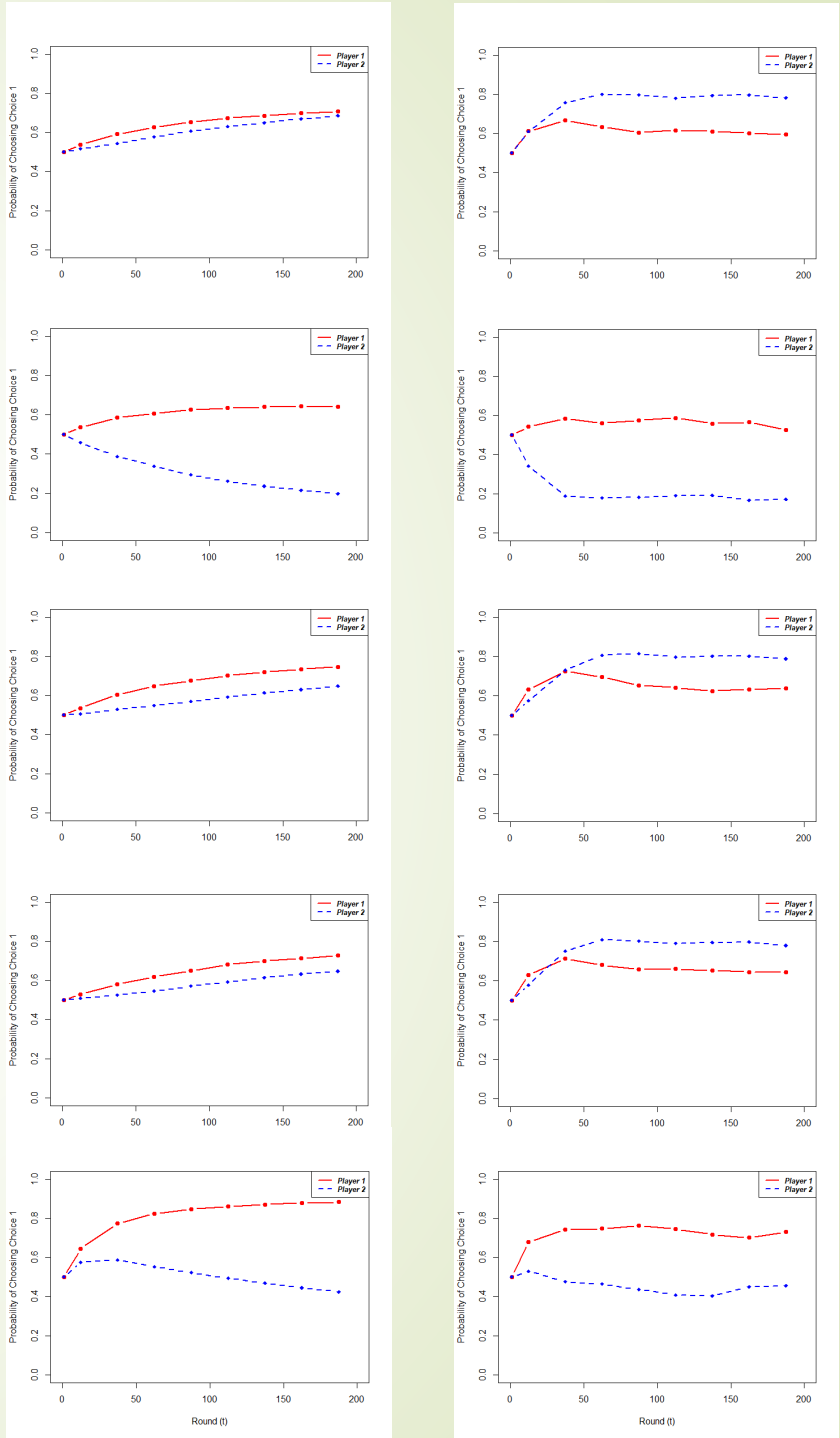
Replication

- Recreated the 1 parameter and 3 parameter models
  - R as the programming language
  - Standardized format for payout matrices for the referenced games
  - Code written to be flexible with any number of player choice combinations (so long as every player-choice combos is non-empty)
- Matched the model terminology and structure from Erev and Roth
  - As opposed to later multiarm bandit, RL, choice models, or those of a similar structure
- Used the published parameter values and visually compared my results to Erev and Roth

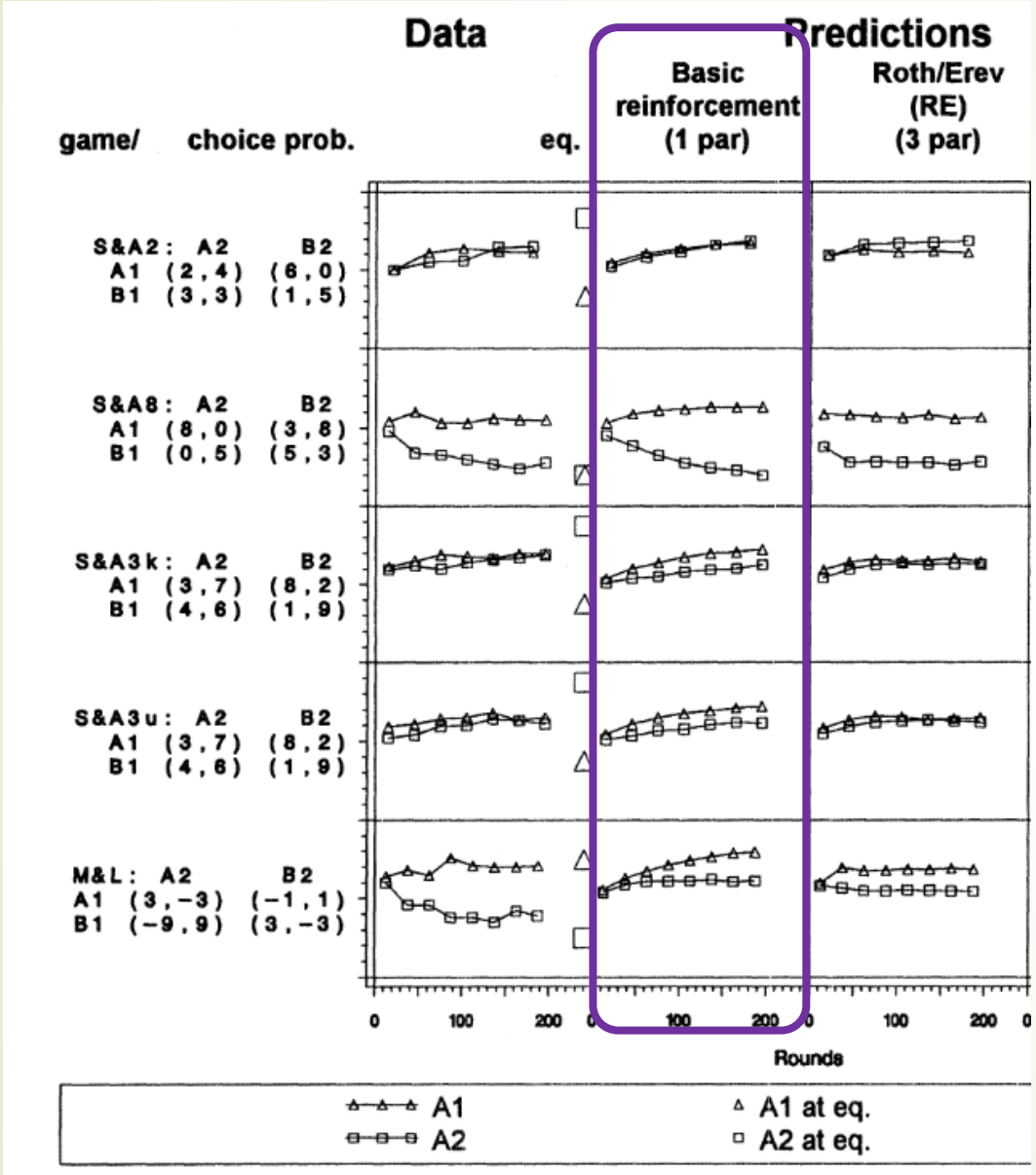
```
for (i in 1:Iterations) {
cat("\014")
print(paste0("Iteration: ", i, " of ", Iterations))


#initalize the probability and q array space for this i
p = array(data = NA, dim = c(NumPlayer,max(as.numeric(M
q=p

#initialize the choice array k for this iteration
k = array(data = NA, dim = c(NumPlayer,Rounds))


#Initialization round
#set t=1 values for the probabilities and propensities
for (n in 1:NumPlayer) {
  for (j in 1:as.numeric(M[n])) {

    p[n,j,1]= 1/as.numeric(M[n])
    q[n,j,1]= p[n,j,1]*S1*as.numeric(X[n])
  }
}

#Loop through subsequent rounds
for(t in 2:Rounds) {

  for(n in 1:NumPlayer){
    #Probabilistically get player n's choice
    k[n,t]=sample(c(1:as.numeric(M[n])),1,replace=FALSE
  }

  #Look at the payout matrix to determine which row mat
  FullPayout = Payout[apply(Payout[,1:NumPlayer], 1, fu
  R = FullPayout[,(NumPlayer+1):ncol(Payout)]-x_min

  for(n in 1:NumPlayer){

    #for each strategy option, see if a reward was obse
    for(j in 1:as.numeric(M[n])){

      q[n,j,t] = q[n,j,(t-1)]

      if (as.numeric(k[n,t]) == j) {
        #update the propensities based on the reward ob
        q[n,j,t] =  q[n,j,(t-1)]+as.numeric(R[n])
      }
    } #next j in the observation of choice and propents

    #Update probabilities based on newly updated propen
    for(j in 1:NumPlayer){
      p[n,j,t] = q[n,j,t]/sum(q[n,1:as.numeric(M[n]),t]
    } #next j in the update of probability loop
```

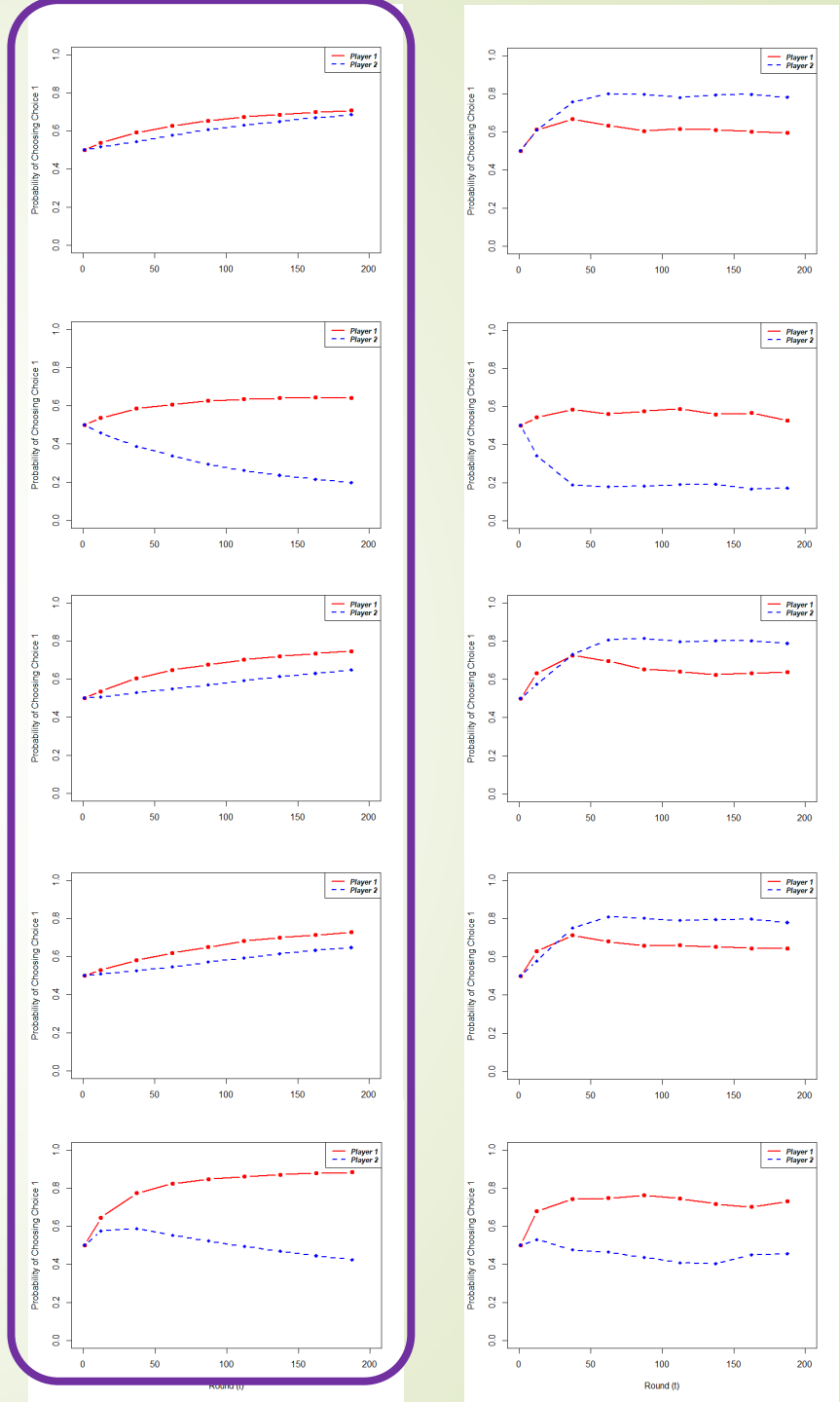
Model Recreation – Directional Results



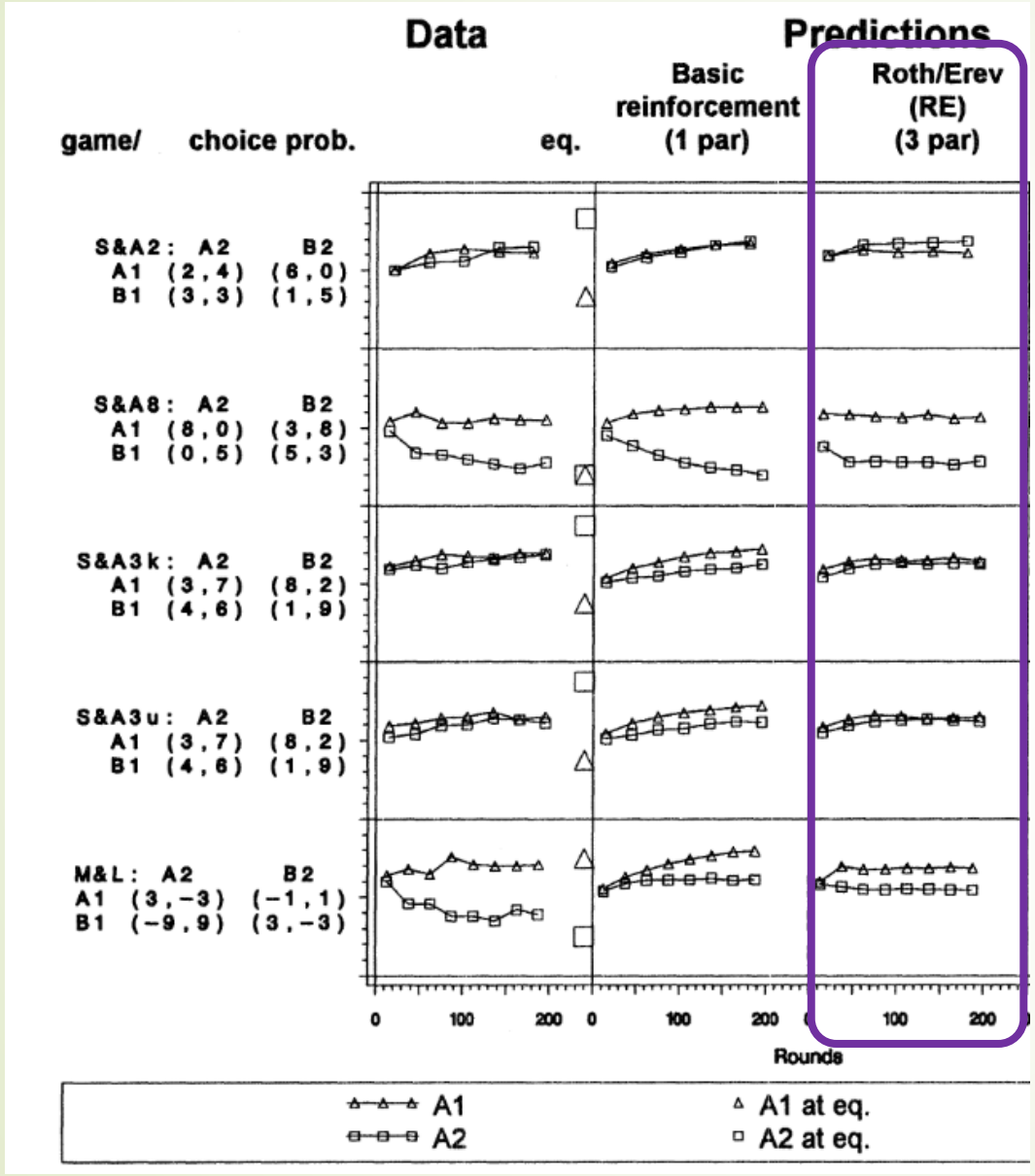


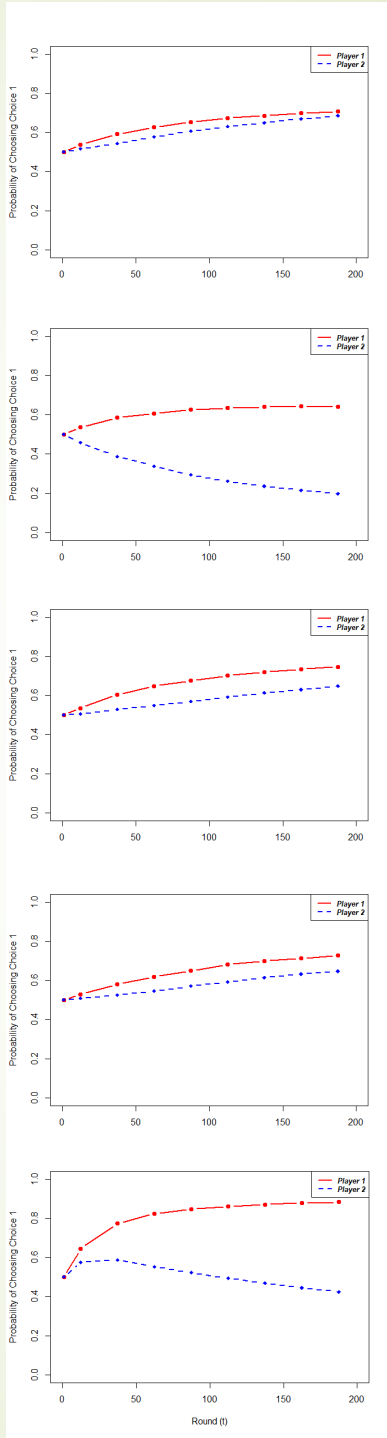
Model Recreation – Directional Results

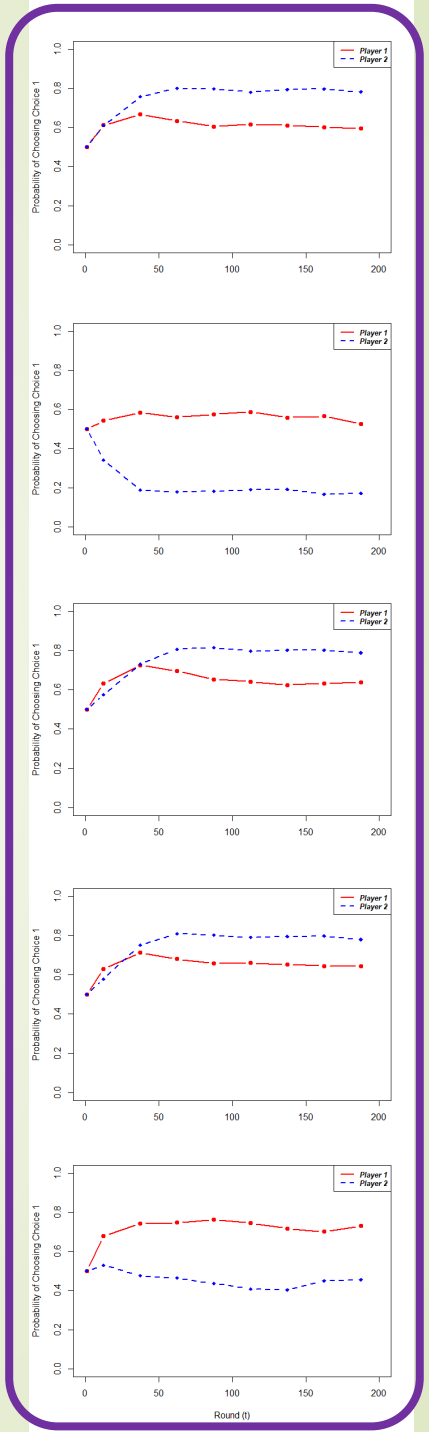




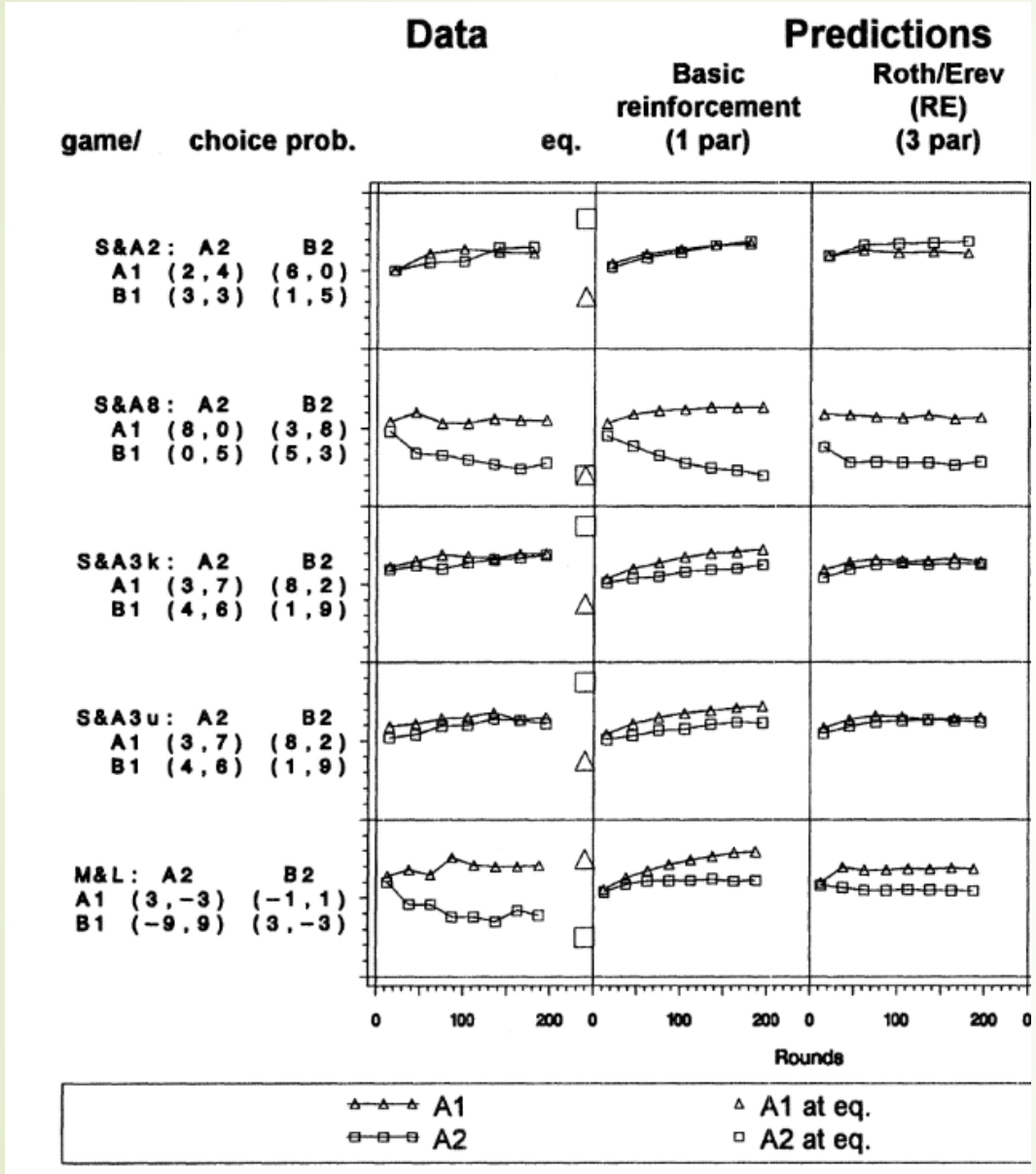
Model Recreation – Directional Results

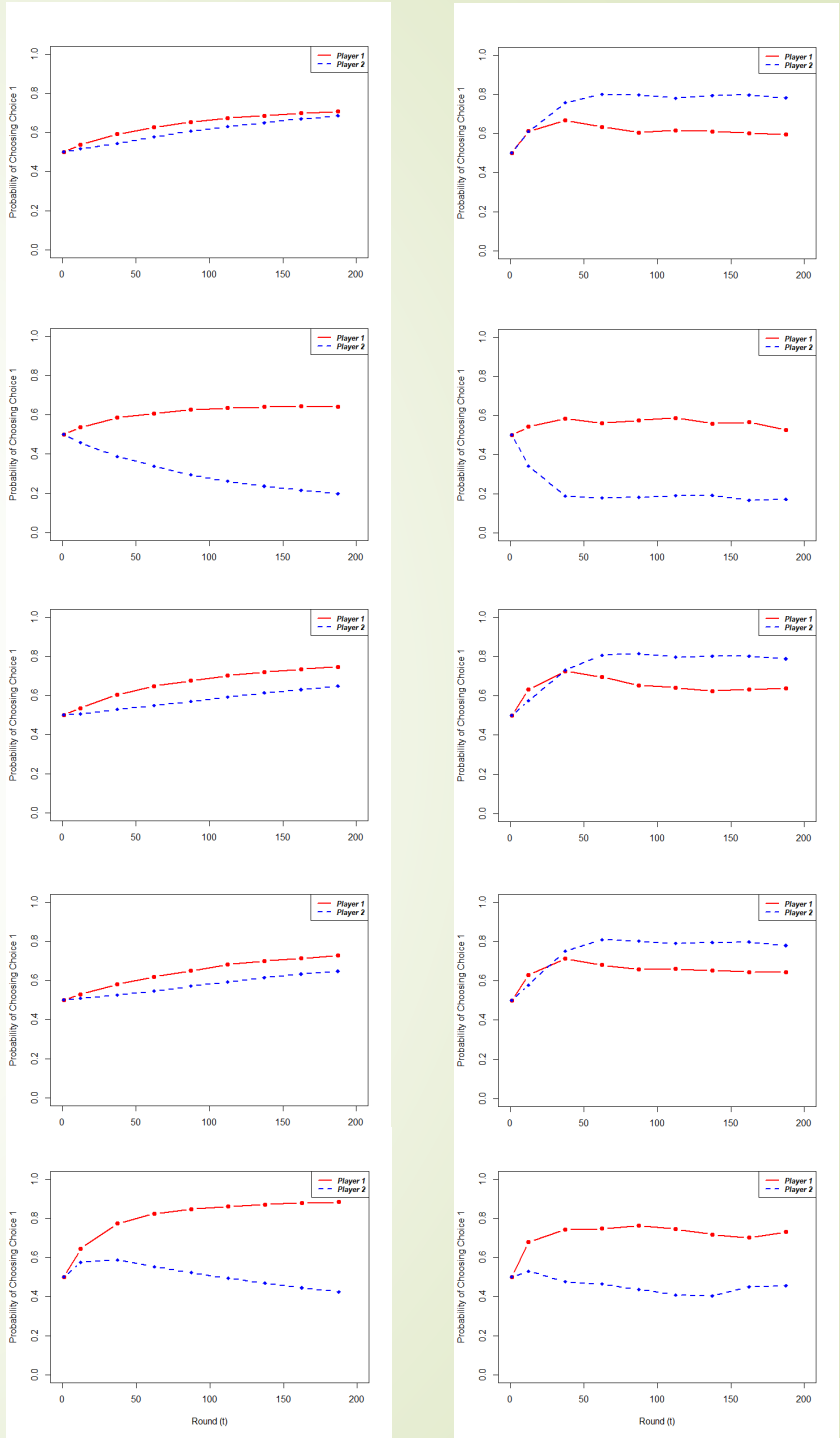






Model Recreation – Directional Results





R Script Demo

1-BasicReinforcement-ErevModeling.R

2-3Param-ErevModeling.R

Critiques on Model Formulation

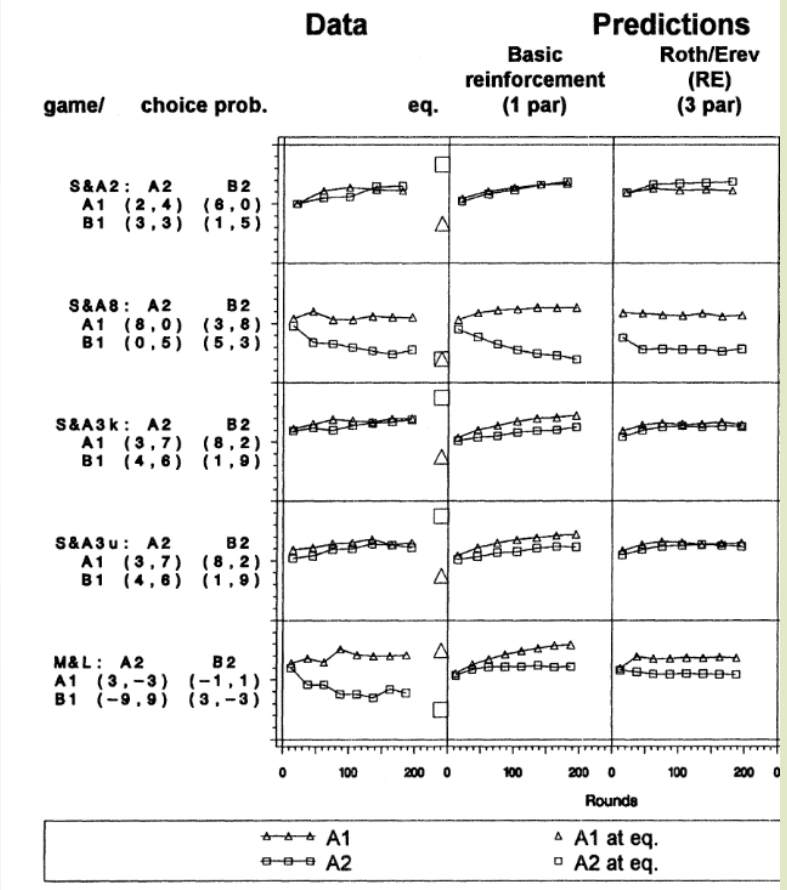
- Model is probabilistic but no confidence ranges or iteration details given
- 'Forgetting' and 'Experimentation' terms
  - Can be manipulated to get non-physical results: very high forgetting relative to experimentation can quickly invert reward structure
  - Forgetting is extremely myopic, looking only at current value
  - Experimentation assumes that choices are ordered (i.e., 1 then 2 then 3 ….)
- Implicit assumptions arise from modeling choices
  - Reward for each player is relative to the lowest reward possible
    - Loss of any amount may have greater impact
    - Solutions to different payouts that are multiples of each other are the same
      - In equilibrium, but the rate of change for larger rewards is faster
    - This implicitly adds the assumption that the worst case scenario, at minimum, is known a priori
  - Speed of learning relies on knowing the average payout
    - Implies participants know the scale of the payouts and adjust accordingly

Critiques on Paper Reproducibility

- Direct verification of results is difficult
  - Original data from the 12 referenced games were not readily available for verification
  - As a result, difficult to confirm the MSE values in the paper
- Model recreation limitations
  - Probabilistic algorithm but no seed value referenced
  - Data presented as pre-batched into average groups of 10-40 (typically 20) time steps
  - Unclear how many iterations were run to obtain the average results seen

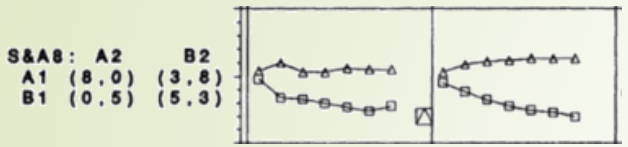
Extensions and Suggestions

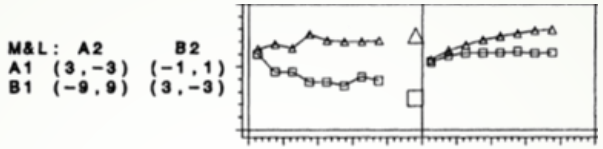
- Include ‘No Choice’ in a more general propensity weighting function
- Replace full sequence batching of trials with single trail using Monte Carlo at each time step
- Run the simulations longer to see landscape of outcomes
  - Pure equilibrium assumes rational play at every time step
  - Implied surface of surface of end states based on previous choices, both rational and not
  - Interesting to see conditions under which traditional equilibrium is reached or not and what other equilibria may exists

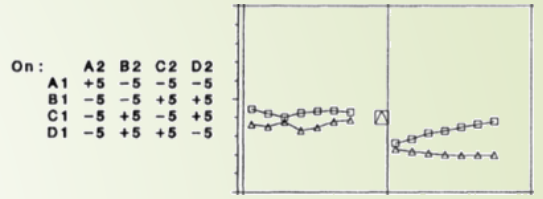


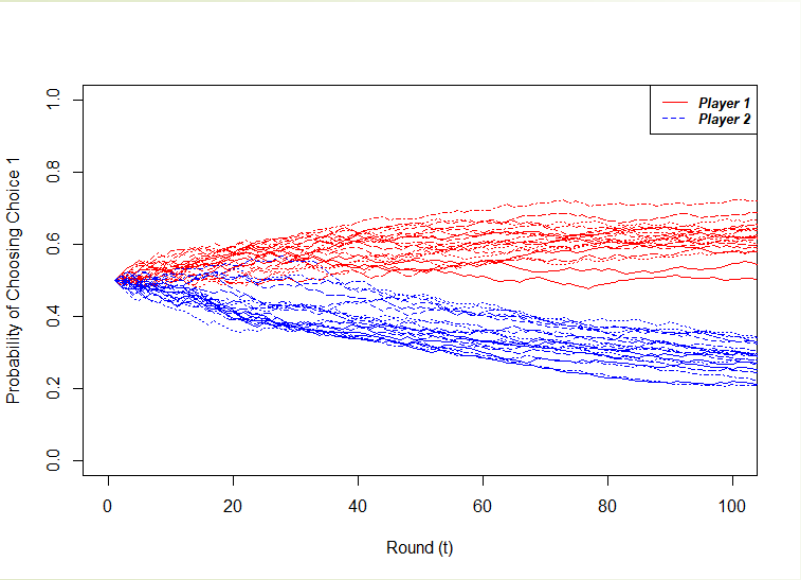
Extensions and Suggestions

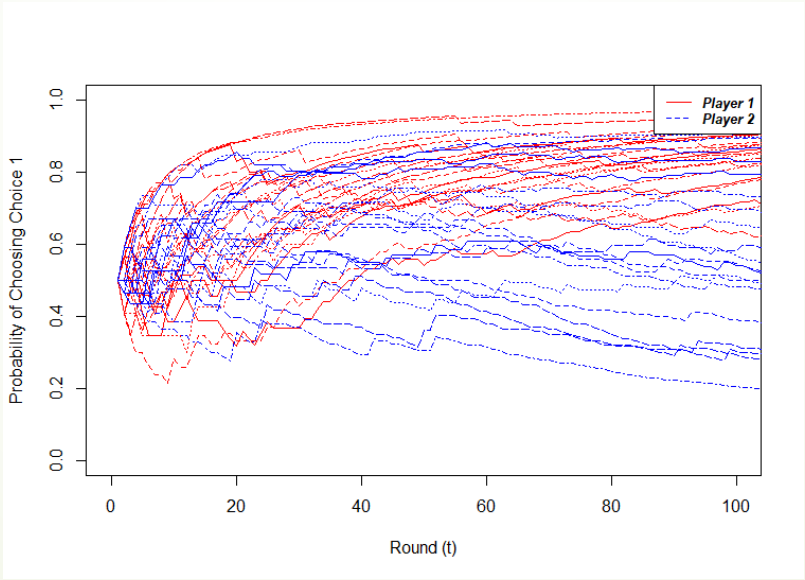
- Run the simulations longer to see landscape of outcomes

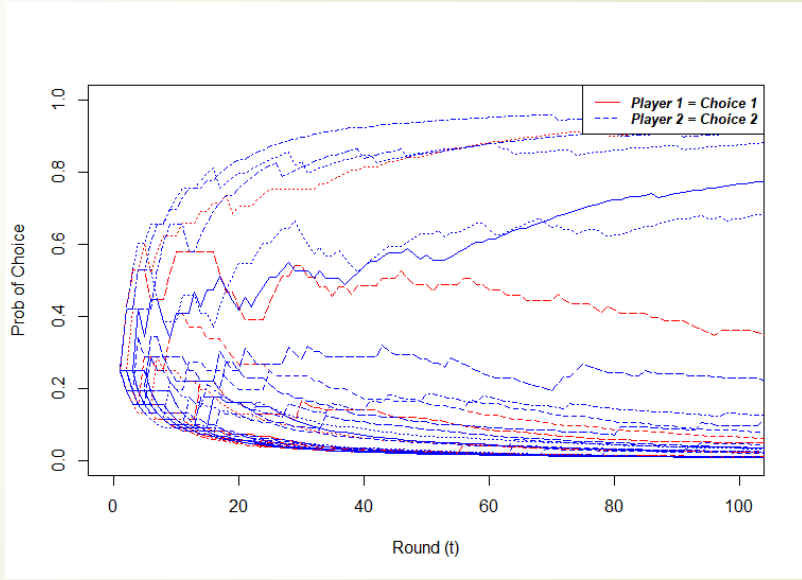












Some simulations generally converge to the aggregate results, but some do not!