Estimation of Learning Models on Experimental Game Data

Hidehiko Ichimura* and Juergen Bracht**
*Department of Economics, University College London. U.K.
**Center for Rationality and Interactive Decision Theory,
The Hebrew University of Jerusalem, Israel*

July 4, 2001

Abstract

The objective of this paper is both to examine the performance and to show properties of statistical techniques used to estimate learning models on experimental game data. We consider a game with unique mixed strategy equilibrium. We discuss identification of a general learning model and its special cases, reinforcement and belief learning, and propose a parameterization of the model. We conduct Monte Carlo simulations to evaluate the finite sample performance of two kinds of estimators of a learning model's parameters: Maximum likelihood estimators of period to period transitions and mean squared deviation estimators of the entire path of play. In addition, we investigate the performance of a log score estimator of the entire path of play and a mean squared deviation estimator of period to period transitions. Finally, we evaluate a mean squared estimator of the entire path of play with observed actions averaged over blocks, instead of behavioral strategies. We propose to estimate the learning model by maximum likelihood estimation as this method performs well on the sample size used in practice if enough cross sectional variation is observed.

There is an extensive and growing literature on learning in experimental economics. It has been observed that quite simple learning models can track aspects of observed behavior on a collection of games, both when the observed behavior conforms to the equilibrium predictions and when it does not Roth and Erev (1995). There is also a growing segment of the literature attempting to compare the empirical performance of learning models (Erev and Roth (1998), Camerer and Ho (1996), Camerer and Ho (1998), Camerer and Ho (1999a), Camerer and Ho (1999b), Feltovich (2000)). Learning models are specified up

*The paper is part of Juergen Bracht's dissertation at the University of Pittsburgh, Department of Economics, U.S.A., (2000). In Spring 2001, the paper was presented at the Empirical Seminar at the Department of Economics of the Hebrew University of Jerusalem, at the Faculty of Industrial Engineering and Management of the Technion in Haifa and at the Department of Economics at Ben-Gurion University in Beersheva.

to some parameters. In order to compare the empirical performance of the learning models, one needs to choose "best performing parameters" for each of the learning models. It turns out this is not a trivial task.

We consider three issues in estimating learning models in a game with unique mixed strategy equilibrium, Mookherjee and Sopher's Matching Pennies (1994), using the framework of the Experience Weighted Attraction (EWA) learning model of Camerer and Ho (1996, 1998, 1999a, 1999b). This learning model provides a convenient framework to address various estimation issues as it includes two representative learning models as special cases: belief learning models and reinforcement learning models. Thus it allows us to address estimation issues for both learning models in a single framework at once.

The first estimation issue we consider is what we mean by "best performing parameters". An estimator is typically defined via some objective function so this is equivalent to considering how we should choose the objective function that defines an estimator. In the literature Camerer and Ho (1999b) uses the maximum likelihood estimator (MLE) and Erev and Roth (1998) defines the minimum distance estimator based on the average prediction error, for example.

The MLE is known as the most efficient estimator under a given specification of a probability model which satisfies some regularity conditions. The regularity conditions are typically satisfied by the learning model we investigate if there are enough cross section observations, that is when enough players are observed. Thus the answer to the question "what objective function" is unambiguously likelihood under the assumption that a learning model's specification and sampling satisfies the regularity conditions. Note that MLE gains efficiency by exploiting the full details of the model specification. However we may not really place too much confidence in the details of the full learning model and instead we may want to just exploit some feature of the model we are confident about. Thus there are two distinct cases to consider when analyzing objective functions that are different from the likelihood: a case where we do not have enough cross sectional observations and another case where we wish to just exploit part of the implications of the specified learning model.

We investigate MLE and other objective functions to examine identification issues. In particular we study the objective functions to understand the relationship between the parameters of the learning models and the variations in the data that are exploited to estimate them.

Clearly it will be desirable if we can estimate parameters with weakest possible information but doing so will result in inefficient (inaccurate) estimator. Hence as the second estimation issue, we investigate via asymptotic approximation and via Monte Carlo simulation, how the alternative methods perform relative to the MLE applied to the case the regularity conditions are satisfied. A Monte Carlo simulation study is conducted using the sample size typically available in the literature to assess (1) whether the alternative estimators can obtain accurate enough information and (2) whether asymptotic distributions approximate finite sample performance of the estimators so that we can carry out inference adequately (asymptotics to be added).

Finally we further examine the finite sample variance covariance matrix to

Table 1: Mookheree and Soher's Matching Pennies

	action R	action L
action R	4.00,0.00	0.00,4.00
action L	0.00,4.00	4.00,0.00

understand what kind of data we should collect. Specifically we examine the marginal efficiency gain of having longer time series observation holding number of players constant and compare it with that of having more players holding the repetition of games constant. This information should help us to design an appropriate experiment under a budget constraint.

0.1 Sections

The organization of the paper is as follows: Section 1 describes the matching pennies game and the EWA model. Section 2 describes the estimation methods we investigate. Section 3 discusses identification of the EWA model and specialize the result to the reinforcement and belief learning models. We discusses what variation in the data is exploited to estimate which parameter. Section 4 reports Monte Carlo results. Section 5 concludes with comments and outlines future research.

1 Data Generating Process

The data generating process is specified by the learning model and the game environment. In this section, we describe both.

1.1 Game

The game studied in this paper is the Matching Pennies. It is a 2x2 constant sum game. It has a unique mixed strategy equilibrium involving 50% mixing of each strategy. The game in normal form is shown in table 1.

The game was used in Mookherjee and Sopher's (1994) experimental study, for example. Players are indexed by i ($i = 1, 2$). The strategy space is the same for both players and consists of 2 discrete choices, R and L. It is denoted as S, and thus $S = \{R, L\}$. An element s_i in S denotes a strategy of player i. The scalar valued payoff function of player i is $\pi_i(s_1, s_2)$. We denote the actual strategy chosen by player i in period t by $s_i(t)$ and his/her opponent's strategy by $s_{-i}(t)$. In this situation, with some abuse of notation we denote ith player's payoff as $\pi_i(s_i(t), s_{-i}(t))$.

1.2 EWA, Belief and Reinforcement Learning Models

At the core of the EWA learning model are three state variables, $N_i(t)$ and $A_i^R(t)$ and $A_i^L(t)$, for each i in period t. Here $N_i(t)$ is referred to as "observation-

equivalent" and controls the speed of learning. $A_i^s(t)$ is an indicator of player i's "attraction" to strategy s after period t has taken place. These state variables jointly determine individual choice probabilities in the learning model. The EWA model specifies initial "observation equivalent" and "attractions", how $N_i(t)$ and $A_i^s(t)$ ($s = R$ and L) are updated, and how attractions determine choice probabilities.

At the beginning of the first round of play "observation equivalent" is given by a parameter $N_i(0) = N(0)$ and an initial attraction of both strategies are also left as some parameters: i.e.

$$A_i^R(0) \text{ and } A_i^L(0) \text{ for } i = 1, 2.$$

After each period, attractions and observation equivalents of player i are updated. The updating rule of observation equivalents is:

$$N_i(t) = \rho \cdot N_i(t-1) + 1, \quad t \geq 1,$$

where ρ denotes the discount rate of observation-equivalents. Note that since we assume the same initial condition for both players and ρ is the same, $N_i(t)$ is the same for both players. Hence we drop the i subscript from $N_i(t)$ from now on.

Let $I\{A\} = 1$ if statement A is true and 0 otherwise. The updating rules of attractions are specified as

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

where ϕ denotes a discount factor of attractions and δ denotes "imagination". A key component of the updating rule is the payoff that a strategy either yielded, or would have yielded, in a period. The model weights hypothetical payoffs that unchosen strategies would have earned by a parameter δ and weights payoffs actually received from a chosen strategy $s_i(t)$ by an additional $1-\delta$. Attractions are then a weighted average of past discounted attractions and those actual or imagined payoffs, normalized by observation equivalents.

Attractions determine how frequently players choose a particular strategy. The probability that player i chooses a strategy s in period $t+1$ is given by the following logit choice rule:

$$P_i^s(t+1) = \frac{e^{\lambda \cdot A_i^s(t)}}{e^{\lambda \cdot A_i^R(t)} + e^{\lambda \cdot A_i^L(t)}}, \quad t \geq 0,$$

where λ denotes sensitivity of players to attractions[1]. An implicit assumption in the logit model is that disturbances are added to attractions that have a double exponential form (McFadden (1974), Yellot (1977)).

[1] Recall that we defined $A_i^s(t)$ to be an indicator of player i's attraction to strategy s *after* period t has taken place. Alternatively, $A_i^s(t)$ could be interpreted as denoting player i's attraction to strategy s at the beginning of period t. Then, the choice rule would be

$$P_i^s(t) = \frac{e^{\lambda \cdot A_i^s(t)}}{e^{\lambda \cdot A_i^R(t)} + e^{\lambda \cdot A_i^L(t)}}, \quad t \geq 0$$

The model is a reinforcement learning model if $\delta = 0$, $N(0) = 1$ and $\rho = 0$. This implies that individuals react only to the actual reward and that $N(t) = 1$. Attractions are called propensities to choose strategies in the context of the reinforcement learning model. Propensities are stock variables of past actual payoffs and initial propensities. Let $Q_i^s(t)$ denote the propensity of strategy s of player i. The updating rule of propensities is:

$$Q_i^s(t) = \phi \cdot Q_i^s(t-1) + I(s_i(t) = s) \cdot \pi_i(s, s_{-i}(t)), \quad t \geq 1.$$

One can verify that this is a special case of the EWA specification when $\delta = 0$, $N(0) = 1$, and $\rho = 0$. The autoregressive form of the updating equation is

$$Q_i^s(t) = \phi^t \cdot Q_i^s(0) + \sum_{\tau=0}^{t-1} \phi^\tau \cdot I(s_i(t-\tau) = s) \cdot \pi_i(s, s_{-i}(t-\tau)), \quad t \geq 1.$$

The model is a belief learning model if $\delta = 1$, $\phi = \rho$ and initial attractions are equal to expected payoffs given initial beliefs. Attractions are expected payoffs of strategies in the context of the belief learning model. Let $E_i^s(t)$ denote player i's expected payoff of strategy s. The autoregressive form of the updating equation is[2]

$$E_i^s(t) = \frac{\phi^t \cdot E^s(0) N(0) + \sum_{\tau=0}^{t-1} \phi^\tau \cdot \delta \cdot \pi_i(s, s_{-i}(t-\tau))}{1 + \phi + \cdots + \phi^{t-1} + \phi^t N(0)}, \quad t \geq 1.$$

Note that, in the EWA model, the parameter δ measures the relative weight given to foregone payoffs, compared to actual payoffs, in updating attractions. In reinforcement learning, foregone payoffs do not count towards updating propensities. In belief learning, actual and foregone payoffs do count with equal weight towards updating expected payoffs of chosen and unchosen strategies, respectively.

1.2.1 Simulation

Once the values of the parameters of the learning model are chosen, the corresponding model can be simulated. Each sample is a panel of $I \times T$ observations where I is the size of the subject pool and T is the number of rounds. In the next section, we will describe estimation methods for the learning models on the simulated data.

2 Estimation techniques

One way to view learning models is as forecasting rules that, given information from previous rounds, predict (possibly probabilistically) a subject's choices in the current round. Another way to view learning theories is as predictors of

[2] See section 3 for the derivation of the general form of the updating equation.

typical behavior in all rounds, given only some initial conditions. The learning model can be estimated in either way.

First, we describe estimation procedures that minimizes the error of the period to period transitions. These transitions are based on the observed data on payoffs and actions up to the current round. We will use two measures of closeness of predictions to actual choices.

Then we describe estimation procedures that minimize the error between the entire simulated path of play and the observed choices. We will use two measures of closeness of the predicted trajectory to observed actions. In addition, we describe an estimator of the entire path of play with observed actions averaged over blocks, instead of behavioral strategies, as used by, among others, Erev and Roth (1998) and Feltovich (2000).

2.1 Estimation of Behavior in the current round given the history of t plays up to the current round

In this subsection, we will describe estimation procedures which use a measure of accuracy of the forecasts of the model. The individual level predictions of the model in particular situations will be compared with the decisions made by players in those situations. The predictions of individual decisions are made given histories of the play up to the current round. Hence, in the reinforcement learning model, we assume that the propensity for playing action R in round t is equal to the (discounted) sum of payoffs received in rounds up to $t-1$ (plus the initial propensity). Given their propensities, players' predicted probabilities are obtained as discussed in Section 2.2.

We discuss two estimation methods that correspond to using two different measures of closeness of predictions to actual choices: mean squared deviation (MSD) and log likelihood. Both criteria are derived by pairing the predicted probability of R being chosen by player i in round t, $P_i^R(t)$, according to the model and the actual probability that R was chosen – which is either zero or one – for each choice made by either type of player. We first describe maximum likelihood estimation and then mean squared deviation estimation.

Each player has two actions. Let $D_i^R(t) = 1$ if action R is chosen by player i in period t and let $D_i^R(t) = 0$ if action L is chosen by player i in period t. Let T denote the length of the repeated game. Let I denote the number of players.

The likelihood function, the formula for the joint probability distribution, of the Matching Pennies data is

$$L = \prod_{i=1}^{I} \prod_{t=1}^{T} P_i^R(t)^{D_i^R(t)} \cdot (1 - P_i^R(t))^{1 - D_i^R(t)}.$$

The log likelihood function for Matching Pennies data is

$$\ln L = \sum_{i=1}^{I} \sum_{t=1}^{T} D_i^R(t) \cdot \ln P_i^R(t) + (1 - D_i^R(t)) \cdot \ln(1 - P_i^R(t)).$$

The probabilities, $P_i^R(t)$, are given by the choice rule. The maximum likelihood estimates of the parameters are the values that give the greatest probability of obtaining the observed data.

The objective function of the mean squared deviation estimator is

$$MSD = \frac{1}{I \bullet T} \sum_{i=1}^{I} \sum_{t=1}^{T} [D_i^R(t) - P_i^R(t)]^2$$

Note that $I \bullet T$ gives the total number of observations, the sample size.

The MSD statistic is a measure of how closely the probabilistic predictions of a learning model conform to observed events. The MSD statistic used here is equivalent to the "quadratic scoring rule", whose theoretical properties are examined in Selten (1998). It is proposed in Brier (1956) and used in Tang (1996), Chen and Tang (1998) and Feltovich (2000), among others.

From a statistical point of view, the MSD objective function ignores the presence of heteroskedasticity in the error term. This leads to an efficiency loss. The efficiency loss is larger when there is more variation in $P_i^R(t)$.

2.2 Estimation of Behavior in all rounds, given some initial conditions

In this subsection, we will describe estimation procedures using measures of closeness of the entire simulated path of play and the observed choices during the play. The predictions of the entire aggregate play are made by running *additional* simulations given some parameters. We then see how close simulation trajectories track either observed aggregate experimental trajectories or observed individual choices over time. Players' predicted probabilities are obtained as discussed above and always aggregated over simulated players.

First, we use two measures of closeness of predictions to actual behavioral strategies: mean squared deviation, MSD_{RA}, and a log score, $\log score_{RA}$. Second, we use a measure of closeness of predictions averaged over blocks to observed choices averaged over blocks: MSD_{RA_ave}.

The former criteria are derived by pairing the predicted probability of R being chosen by a representative agent, $P_{RA}^R(t) = K^{-1} \sum_{i=1}^{K} P_i^R(t)$, in round t, according to the model (and its parameters) and the actual choice of, $D_i^R(t)$, which is either zero or one, for each choice made by a player. Hence, we perform K sets of simulations to predict $P_{RA}^A(t)$.[3] The latter criterion is derived by pairing the predicted probability of R being chosen, $P_{RA}^R(t)$, averaged over blocks to the actual probability that R was chosen, $D_i^R(t)$, averaged over blocks and players. Let b designate B different blocks of time. Then, for T/B blocks, $T = 40$ and $B = 4$, we have $b = 1$ for periods 1 to T/B, $b = 2$ for periods $2 \cdot T/B + 1$ to $3 \cdot T/B$ ... and $b = B$ for periods $(B-1) \cdot T/B + 1$ to T. The

[3] Following common practise, we set K equal to some number. For the Monte Carlo study, we choose $K = 50$.

formula below is given for this kind of aggregation.[4]

The three objective functions[5] of these three estimators are

$$MSD_{RA} = \frac{1}{I \cdot T} \sum_{i=1}^{I} \sum_{t=1}^{T} \left(P_{RA}^{R}(t) - D_{i}^{R}(t) \right)^{2}$$

$$\log score_{RA} = \frac{1}{I \cdot T} \sum_{i=1}^{I} \sum_{t=1}^{T} \left(D_{i}^{R}(t) \cdot \log P_{RA}^{R}(t) + (1 - D_{i}^{R}(t)) \cdot \log(1 - P_{RA}^{R}(t)) \right)$$

$$MSD_{RA\_ave} = \frac{1}{B} \left(\sum_{b=1}^{B} \left(\frac{1}{T/B} \sum_{t=(b-1)T/B+1}^{bT/B} P_{RA}^{R}(t) \right) - \left(\frac{1}{T/B} \frac{1}{I} \sum_{i=1}^{I} \sum_{t=(b-1)\cdot T/B+1}^{b\cdot T/B} D_{i}^{R}(t) \right) \right)^{2}.$$

Note that $I \bullet T$ gives the total number of observations and that B gives the total number of blocks.

To estimate the parameters we minimize the error between the entire simulated path of play and the observed choices. Erev and Roth (1998), among others, compare the predictions of different learning models by computing the mean-squared deviation (MSD_{RA} and MSD_{RA_ave}) of the predicted and observed behavior, period by period, for each experimental game, both for all subjects and individual pairs (when individual level data are available). Roth, Erev, and Slonim (1998) propose the log scoring rule for estimation ($\log score_{RA}$), but evaluate the closeness of predictions to the data using both MSD scores.

3 Identification of Parameters and Model Diagnostics

In this section, we discuss identification of Camerer and Ho's (1999b) general EWA learning model and its special case, the reinforcement learning model.

Recall that

$$N(t) = 1 + \rho + \cdots + \rho^{t-1} + \rho^{t} N(0)$$

and that the updating rules of attractions for $t \geq 1$ is specified is

$$A_{i}^{s}(t) = \frac{\phi \cdot N(t-1) \cdot A_{i}^{s}(t-1) + [\delta + (1-\delta) \cdot I\{s_{i}(t) = s\}] \cdot \pi_{i}(s, s_{-i}(t))}{N(t)}.$$

Defining $X_{i}^{s}(t) = A_{i}^{s}(t) N(t)$, we have

$$X_{i}^{s}(t) = \phi \cdot X_{i}^{s}(t-1) + [\delta + (1-\delta) \cdot I\{s_{i}(t) = s\}] \cdot \pi_{i}(s, s_{-i}(t)),$$

[4] For the Monte Carlo, we set (T/B) equal to 25 for sample (a) and (T/B) equal to 10 for sample (b). This is common practice.

[5] To be precise, the measures have to be calculated separetly for either type (row or column player) of player and then averaged. We leave that out for simplicity.

so that for $t \geq 1$

$$X_i^s(t) = \phi^t \cdot X_i^s(0) + \sum_{\tau=0}^{t-1} \phi^\tau \cdot [\delta + (1-\delta) \cdot I\{s_i(t-\tau) = s\}] \cdot \pi_i(s, s_{-i}(t-\tau)),$$

and that for $t \geq 1$

$$A_i^s(t) = \frac{\phi^t \cdot A^s(0) N(0) + \sum_{\tau=0}^{t-1} \phi^\tau \cdot [\delta + (1-\delta) \cdot I\{s_i(t-\tau) = s\}] \cdot \pi_i(s, s_{-i}(t-\tau))}{1 + \rho + \cdots + \rho^{t-1} + \rho^t N(0)}.$$

Camerer and Ho (1999b) specifies the link between attractions and probability as the logit choice rule:

$$P_i^s(t+1) = \frac{e^{\lambda \cdot A_i^s(t)}}{e^{\lambda \cdot A_i^R(t)} + e^{\lambda \cdot A_i^L(t)}}, \quad t \geq 0.$$

This implies that the choice probability depends only on $\lambda\left[A_i^R(t) - A_i^L(t)\right]$ and that for $t \geq 1$:

$$\begin{aligned}
&\lambda\left[A_i^R(t) - A_i^L(t)\right] \\
&= \lambda\left[1 + \rho + \cdots + \rho^{t-1} + \rho^t N(0)\right]\left\{\phi^t \cdot \left[A^R(0) - A^L(0)\right] N(0)\right. \\
&+ \sum_{\tau=0}^{t-1} \phi^\tau \delta\left[\pi_i(R, s_{-i}(t-\tau)) - \pi_i(L, s_{-i}(t-\tau))\right] \\
&+ \left.\phi^\tau(1-\delta)\left[I\{s_i(t-\tau) = R\}\pi_i(R, s_{-i}(t-\tau)) - I\{s_i(t-\tau) = L\}\pi_i(L, s_{-i}(t-\tau))\right]\right\}.
\end{aligned}$$

By further exploiting the structure of the payoff we obtain for $t \geq 1$

$$\begin{aligned}
&\lambda\left[A_i^R(t) - A_i^L(t)\right] \\
&= \lambda\left[1 + \rho + \cdots + \rho^{t-1} + \rho^t N(0)\right]\left\{\phi^t \cdot \left[A^R(0) - A^L(0)\right] N(0)\right. \\
&+ \left.4\sum_{\tau=0}^{t-1} \phi^\tau \left(\delta \cdot \left[I\{s_i(t-\tau) \neq s_{-i}(t-\tau)\} + I\{s_i(t-\tau) = s_{-i}(t-\tau)\}\right]\right)(-1)^{I\{L = s_{-i}(t-\tau)\}}\right\}.
\end{aligned}$$

As Camerer and Ho note the choice probabilities are the same if $A^R(0) - A^L(0)$ takes the same value, other things equal, so without a loss of generality, we set $A^L(0) = 0$. With this normalization they estimate

$$\lambda,\ \rho,\ N(0),\ \phi,\ A^R(0), \text{ and } \delta.$$

With this specification when λ is 0, none of the other parameters of the model is identifiable. Note that $\lambda = 0$ represents the equilibrium prediction. We have shown elsewhere that the matching pennies experiment might very well conform with the equilibrium prediction. This implies that the specification does not allow accurate estimation of all of the rest of the parameters.

We examine the likelihood using the following parametrization:

$$\rho,\ N(0),\ \lambda A^R(0)N(0),\ \phi,\ \lambda\delta, \text{ and } \lambda.$$

That is, we study for $t \geq 1$

$$\begin{aligned}
&\lambda\left[A_i^R(t) - A_i^L(t)\right] \\
&= \left[1 + \rho + \cdots + \rho^{t-1} + \rho^t N(0)\right]\left\{\phi^t \cdot \lambda A^R(0)N(0)\right. \\
&\left. + 4\sum_{\tau=0}^{t-1}\phi^\tau\left(\lambda\delta \cdot \left[I\left\{s_i(t-\tau) \neq s_{-i}(t-\tau)\right\} + \lambda \cdot I\left\{s_i(t-\tau) = s_{-i}(t-\tau)\right\}\right]\right)(-1)^{I\{L=s_{-i}(t-\tau)\}}\right\}.
\end{aligned}$$

The idea is to separate out the contribution of the effect of sensitivity parameter λ on initial condition and on δ without excluding the equilibrium prediction. While the reparametrization is not going to resolve the identification of δ, when λ is close to zero, parameters $\lambda A^R(0)N(0)$, $\lambda\delta$, and λ may well be estimated well while $A^R(0)$ and δ will not be so that the likelihood is easier to optimize. In fact this is consistent with our experience. More concretely our inspection of the original likelihood function clearly shows a ridge over the area where $\lambda\delta$ is constant, holding other parameter values at the true values. This explains the difficulty we faced in optimizing the original likelihood. The reparametrization is important in carrying out the Monte Carlo simulation study where repeated optimization is required.

Another point this derivation implies is the importance of cross section variation in estimating $N(0)$ and $\lambda A^R(0)N(0)$ when $|\rho| < 1$ and $|\phi| < 1$. In these cases as T goes to infinity, the impact of the initial conditions declines and hence large T does not help to improve efficiency of estimators of $N(0)$ and $\lambda A^R(0)N(0)$. In fact for consistency of the estimators of $N(0)$ and $\lambda A^R(0)N(0)$, theoretically the cross sectional sample size needs to diverge to infinity. On the other hand, ρ, ϕ, $\lambda\delta$, and λ may be estimated consistently with only time series variation. This is indeed so even holding the cross sectional variation constant. As we discussed earlier, large T observations are not affected by initial conditions and hence inconsistency of the initial condition parameters do not impact consistent estimation of the rest of the parameters.

Our observation that the parameter δ is not be identified when $\lambda = 0$ is not unimportant. As Camerer and Ho note, the parameter δ "is the most important in EWA because it shows most clearly the different ways in which EWA, reinforcement and belief learning capture two basic principles of learning – the actual law of actual effect and the law of simulated effect." In EWA, if $\delta = 0$ then only chosen but not unchosen strategies receive reward. If $\delta \neq 0$, like in belief learning (where $\delta = 1$) then unchosen strategies that would have yielded high payoffs are more likely to be chosen subsequently. One of the conclusions Camerer and Ho reach after examining the matching pennies data is that the data do not distinguish the two types of learning models.

We agree with their observation with some qualification. It seems to us that if we are to use EWA as an encompassing model of belief based and reinforcement based learning, then matching penny may not be an appropriate game to

attempt to distinguish the two learning models. This is so because if we use the EWA model, in the neighborhood of equilibrium play of Matching Pennies, these two models are hard to distinguish. On the other hand if there is some other type of encompassing model that does not have the identification problem we discussed or if there are matching pennies games in which play is not close to equilibrium play, then it seems to us that Matching Pennies may well be a suitable game to study learning behavior.

4 Monte Carlo study

In this section, we report on a Monte Carlo simulation study to evaluate the performance of the maximum likelihood estimator and four alternative estimators which are described above.

The data generating process is specified by Matching Pennies and the reinforcement learning model. The EWA model is a reinforcement learning model if $\rho = 0$, $\delta = 0$, and $N(0) = 1$.[6] The parameters of the reinforcement learning model are $\omega = \lambda A^R(0), \lambda$ and ϕ.[7]

Each sample is a panel of $I \times T$ observations where I is the population size and T is the number of rounds. We use MATLAB 5.2's simplex procedure ("fmins") to obtain the estimates for 100 simulations.

The performance of the estimators is evaluated on two kinds of data sets: sample (a) and sample (b). In sample (a), we hold the number of subject pairs constant and vary the length of play. We chose to investigate a sample with 1 subject pair and 50, 125, 200, and 500 rounds of play.[8] In sample (b), we hold the length of play constant and vary the number of subject pairs. We chose to investigate a sample with 40 rounds of play and 5, 10, 20 and 40 subject pairs.[9]

For the maximum likelihood estimator, we report on simulation results for eight sets of true values of the three parameters. For the alternative estimators, we specialize to four sets of true values of the three parameters. The sets of true values used in the study are reported in table 2.

4.1 Finite Sample Results for Maximum Likelihood Estimator

We report on the performance of the maximum likelihood estimator for two kinds of data sets: sample (a) and sample (b).

4.1.1 Sample (a): 1 pair of player and a varying number of rounds

The first part of the experiment was carried out for $I = 2$ and four values of $T : T = 50, 125, 200$ and 500. The results are summarized in tables 1 and 2

[6] See section 2.2.

[7] See section 4 for identification of the model.

[8] Roth et al (1998) had fixed pairs of subjects play games with unique mixed strategy equilibrium for 500 rounds.

[9] Mookherjee and Sopher had 10 subject pairs playing a Matching Pennies for 40 rounds.

Table 2: 8 sets of true values of the 3 parameters of the reparameterized reinforcement learning model

	Q^0	ω	ϕ	λ
Set 1	0.00	0.00	0.20	0.20
Set 2	0.00	0.00	0.20	-0.20
Set 3	0.00	0.00	0.80	0.20
Set 4	0.00	0.00	0.80	-0.20
Set 5	4.00	0.80	0.20	0.20
Set 6	4.00	-0.80	0.20	-0.20
Set 7	4.00	0.80	0.80	0.20
Set 8	4.00	-0.80	0.80	-0.20

in the appendix. Each cell in that table reports on a summary statistic of the empirical distribution of the MLE for a set of true values and for a particular sample size.

For instance, the mean, standard deviation and median of the empirical distribution of the MLE of ω on data generated by 1 pair of players playing 50 rounds and by set 1 of true values (i.e $\omega_0 = 0.00$, $\phi_0 = 0.20$ and $\lambda_0 = 0.20$) are -73.644, 225.48 and -0.0034 respectively.

The mean, standard deviation and median of the empirical distribution of the MLE of ω on the data generated by 1 pair of players playing 125 rounds and by set 1 of true values (i.e $\omega_0 = 0.00$, $\phi_0 = 0.20$ and $\lambda_0 = 0.20$) are -41.194, 170.45, and -0.0015 respectively.

For the true values $\omega_0 = 0.00$, $\phi_0 = 0.20$, and $\lambda_0 = 0.20$, the kernel density estimates of the sampling distribution of $\hat{\omega}$, $\hat{\phi}$, and $\hat{\lambda}$ are shown in figure 1 in the appendix.

The salient features of the performance of the MLE for one pair of players are as follows. For the entire set of true values, the distributions of parameter estimates of ω display a large variation. This is consistent with our discussion in the identification section. Consistency of the MLE of ω requires $I \to \infty$ and with one pair of players data is just not informative about ω_0. This is especially so when ϕ_0 is small. When ϕ_0 is small, the information about ω_0 decreases quickly toward zero over time. Reflecting this the variation is less pronounced when the true value of the discount parameter, ϕ_0, is 0.80. However the variation does not substantially decrease while the time series dimension is increased. As discussed, the estimator is inconsistent even when $T \to \infty$ when I is fixed. There is a huge bias in the estimated means of ω. The estimation of ϕ is much better. It is more accurate for a true value $\phi_0 = 0.8$ than for a value $\phi_0 = 0.2$. This is also consistent with the theoretical consideration that smaller ϕ_0 implies less time series information about it. The estimation of λ is also more accurate when ϕ_0 is larger.

Summary 1 *The performance of the maximum likelihood estimator applied to a pair of players on the 8 sets of true values is not satisfactory for the initial*

choice propensity ω. This is consistent with the theoretical prediction that cross sectional variation is needed to consistently estimate ω. The smaller the discount factor ϕ, the harder it is to obtain information from time series observation and that effect shows up on all estimators.

4.1.2 Sample (b): 40 rounds of play and a varying number of pairs

The second part of the experiment was carried out for four values of $I : I = 10, 20, 40$ and 80 and $T = 40$. The means, standard deviations, and medians of the empirical distributions of the estimators $\hat{\omega}$, $\hat{\phi}$, and $\hat{\lambda}$ are shown in tables 3 and 4 in the appendix. For the true values $\omega_0 = 0.00$, $\phi_0 = 0.20$ and $\lambda_0 = 0.20$ (that is *set* 1), the kernel density estimates of the sampling distributions of $\hat{\omega}$, $\hat{\phi}$, $\hat{\lambda}$ are shown in the figure 2 in the appendix.

The salient features of the performance of the MLE are as follows. For the entire set of true values, the distribution of parameter estimates of ω displays some variation. The variation does decrease while increasing the sample size. The estimation of ω is more accurate when $\phi_0 = 0.80$. The estimation of parameter ϕ is more reliable. It is more accurate for a true value $\phi_0 = 0.8$ than for $\phi_0 = 0.2$. The estimation of λ is very accurate across the entire set of true values.

Summary 2 *On sample (b), the performance of the maximum likelihood estimator of the three parameters on the 8 sets of true values is good.*

Finally, we compare the performance of the maximum likelihood estimator for two samples of equal size: a sample with 5 pairs of players playing matching pennies for 40 rounds and a sample with 1 pair of player playing matching pennies for 200 rounds. The estimation of the parameters is more accurate for the sample in which we observe more cross sectional variation: the standard deviation of the empirical distribution of the three parameters is smaller 16 out of 24 times; the parameter ω is clearly better estimated, since both bias and standard deviation are smaller. The standard deviation of the distribution of $\hat{\phi}$ is smaller 7 out of 8 times, the standard deviation of the distribution of $\hat{\lambda}$ is smaller only 2 out of 8 times, but the maximal difference between the standard deviations is very small.

Summary 3 *Exploiting cross sectional variation in the data helps to accurately estimate the model by maximum likelihood estimation.*

4.2 Finite Sample Results of the MSD estimator

We repeat the Monte Carlo experiment and evaluate the performance of the MSD estimator. We specialize the set of true values to sets 1–4 as reported in table 2. Once again, in the first part of the experiment we vary the number of rounds holding the number of players constant (sample (a)). In the second part of the experiment, we vary the number of pairs of players holding the number of rounds constant. The means, standard deviations and medians of

the empirical distributions of the estimators minimizing the objective function MSD are displayed in tables 5 and 6 in the appendix. For the true values $\omega_0 = 0.00$, $\phi_0 = 0.20$ and $\lambda_0 = 0.20$, the kernel density estimates of the sampling distribution of $\hat{\omega}$, $\hat{\phi}$ and $\hat{\lambda}$ are shown in figures 3 and 4 in the appendix. The panel to the top reports results obtained from sample (a) and the panel to the bottom displays results obtained from sample (b). For both samples, the MSD estimator of the model performs well. Estimation of the parameter λ is most accurate with little bias in the estimated means and small standard deviation even in small samples. The larger the true value of the parameter ϕ, the more accurate the estimation of both ϕ and ω. In sample (a) with little cross sectional variation, there is a large bias in the estimated means of ω. The standard deviation is large, too. Observing cross sectional variation, as we do in sample (b), greatly helps to accurately estimate ω.

Summary 4 *On sample (a) and (b), the performance of the MSD estimator of λ and ϕ is good. On sample (a), the MSD estimation of ω is not satisfactory, whereas it is much better on sample (b).*

The performance of the MSD estimator is similar to the performance of the MLE. We calculate the ratios of the standard deviation of the empirical distribution of the MSD estimator to the standard deviation of the empirical distribution of the MLE for both sample (a) and (b) and for each of the 4 sets of true values. We report on the results in table 13 in the appendix. Recall the bias in the estimated means of the parameter ω on sample (a) when estimated by both MSD estimation and ML estimation. Therefore, we specialize to the parameters ϕ and λ for sample (a). On this subset, the MLE is more accurate than MSD 22 out of 32 times. The average ratio across 16 values for parameter ϕ is 1.000 and the average ratio across 16 values for parameter λ is 1.024.

For sample (b), the MLE estimator is more accurate than the MSD estimator 37 out of 48 times. The average ratio across 16 values for parameters ω, ϕ and λ is $1.036, 1.009$ and 1.027 respectively.

Summary 5 *On finite samples, the maximum likelihood estimator is more efficient than the MSD estimator.*

4.3 Finite Sample Results of the Deviation Estimators

In this subsection, we report on the performance of estimators minimizing the prediction error the entire path of play. A Monte Carlo study was carried out as described above. We report the summary statistics of the sampling distributions of the three deviation estimators, MSD_{RA}, log score_{RA} and MSD_{RA_ave} in tables 7 and 8, 9 and 10, 11 and 12 in the appendix, respectively. The kernel density estimator of the sampling distributions of the three estimators are shown in figures 5-10 in the appendix, both for sample (a) and sample (b). First, note that the sampling distributions of the estimators minimizing the distance of the predictions to the behavioral strategies, MSD_{RA_ave} and log score_{RA} are very

similar. They are so for both sample (a), as depicted in figures 5 and 7 in the appendix, and sample (b), as depicted in figures 6 and 8 in the appendix. The salient features of the performance of the MSD_{RA} and log score_{RA} estimators are as follows. Estimation of the parameter ω is very accurate. The parameter λ is estimated well, whereas parameter ϕ is estimated satisfactory. The larger the true value of ϕ, the better the estimation of all three parameters, ω, ϕ and λ. There is some bias, at times large, in the estimated means of the estimators of ϕ and λ. In addition the bias does not always get smaller as sample size increases.

Summary 6 *The deviation estimators MSD_{RA} and log $score_{RA}$ perform satisfactory on both sample (a) and (b). The estimation is most accurate for ω and satisfactory for ϕ and λ.*

Next, we compare the performance of the two deviation estimators with the performance of the MLE. Recall the large bias in the estimated means of ω of the MLE on sample (a). We specialize our comparison to parameters ϕ and λ for this subset. We note the bias in the estimated means of those two parameters for the deviation estimators. Nevertheless, we calculated the ratios of the standard deviation of the empirical distribution of the deviation estimator to the standard deviation of the empirical distribution of the MLE for sample (a). We report on the results in table 14 in the appendix. The average ratio across 16 values for parameter ϕ is 1.96 and 1.53 for the MSD_{RA} and log score_{RA} estimator, respectively. The average ratio across 16 values for parameter λ is 3.26 and 3.10 for the MSD_{RA} and log score_{RA} estimator, respectively. For sample (b), the average ratio across 16 values for parameter ω is 0.03 and 0.06, for parameter ϕ 2.58 and 3.00 and for λ 2.57 and 3.00 for the MSD_{RA} and log score_{RA} estimator, respectively.

Summary 7 *The maximum likelihood estimator is more efficient than the deviation estimators for the parameter ϕ and λ, but less efficient for parameter ω.*

Next, we compare the performance of the two deviation estimators for two samples of equal size: a sample with 5 pairs of players playing matching pennies for 40 rounds and a sample with 1 pair of player playing matching pennies for 200 rounds. Observing more cross sectional variation appears to help a tiny bit with accurate estimation of ω, whereas observing more time variation helps a bit with estimation of λ and ϕ.

Summary 8 *The different kinds of variations in the data hardly can be exploited for accurate estimation by the deviation estimators.*

Next, we report on the performance of the estimator minimizing the distance of the predictions to the aggregate behavior, MSD_{RA_ave}.

The salient features of the performance of the estimator are as follows. The sampling distributions of $\hat{\phi}$ and $\hat{\lambda}$ are at times not correctly centered and the

bias does at times increase while increasing the sample size. Even for some of the largest sample size, the bias is substantial both for $\hat{\phi}$ and $\hat{\lambda}$. The sampling distributions of $\hat{\omega}$ are correctly centered and display a tiny variation.

The standard deviations of the empirical distributions of the three estimators decrease by a factor of 2 while increasing the sample size by a factor of 4 23 out of 48 times. The standard deviations of the empirical distributions of the estimator $\hat{\omega}$ decrease by a factor of 2 while increasing the sample size by a factor of 4 10 out of 16 times. The standard deviations decrease 4 out of 16 times for the estimator $\hat{\phi}$. The standard deviations decrease 9 out of 16 times for the estimator $\hat{\lambda}$.

Summary 9 *MSD_{RA_ave} estimator does not perform satisfactory.*

5 Concluding Remarks and Future Research

The paper makes three basic contributions. First, we discuss identification of the EWA model and its special cases, reinforcement and belief learning. We note that on Matching Pennies Camerer and Ho's model specification does not allow identification of learning models when players follow equilibrium play. This implies that estimation of learning model parameters leads to larger standard errors when players choose strategies closer to equilibrium play. Another implication is that numerical optimization using their parameterization becomes hard to carry out. This hinders Monte Carlo simulation as it requires repeated estimation of the parameters. We derive an explicit solution to the difference equation that defines the learning model and show that a certain reparameterization overcomes this difficulty.

Second, we have investigated, via Monte Carlo simulation, five estimators of the reparameterized reinforcement learning model. The estimators fall in two broad classes: Estimators minimizing the error of the period to period transitions and estimators minimizing the error of the entire simulated path of play and observed choices. We have addressed questions about both the way variation in the data helps to estimate parameters and the performance of the estimators on the sample size used in practise. We have shown that the MLE of period to period transitions performs well (on the sample sizes used in practise). The payoff sensitivity parameter λ and the discounting parameter ϕ are accurately estimated. Observing cross sectional variation is crucial in obtaining more precise estimates of the parameter ω which determines the initial conditions of the model. When we have 40 pairs of players the sampling distributions are correctly centered. The standard deviations of the distributions do always decrease substantially as sample size increases. We have compared the maximum likelihood-like MSD estimator to the MLE. We have found that the MSD estimator performs very similar and almost as well as the MLE. Estimators of the entire path of play, MSD_{RA} and $\log \text{score}_{RA}$, do not perform as well as the estimators of period to period transitions. While those estimators provide more accurate estimation of ω, even if there is little cross sectional variation, the two

other parameters are not as well estimated. The sampling distribution of those two parameters slightly tend not to be correctly centered. The standard deviations of the distributions are at times large and they do not always decrease as sample size increases. The estimator of the entire path of play, MSD_{RA_ave}, which averages both predictions and observations over time into blocks, does not perform satisfactory as at times both the bias and the standard deviations of the sampling distributions are large.

We propose to estimate the reinforcement learning model by maximum likelihood estimation as this technique performs well on the sample size used in practice. To accurately estimate the parameters of the model, it is important to observe cross sectional variation. This implies that one should collect experimental data with shorter time series and a larger number of players. This is the main contribution of the paper.

In future research, we intend to expand on our analysis by both investigating identification of the EWA model on alternative data sets and evaluating the performance of the estimators of the general and belief learning models.

References

Brier, G. (1956): "Verification of Forecasts Expressed in Terms of Probability," *Monthly Weather Review*, 78, 1–3.

Camerer, C., and T. Ho (1996): "Experience-Weighted Attraction Learning in Games: A Unifying Approach," Mimeo, Caltech.

——— (1998): "EWA Learning in Games: Heterogeneity and Time Variation," *Journal of Mathematical Psychology*, 42, 305–326.

——— (1999a): "Experience-Weighted Attraction Learning in Games: Estimates from Weak-Link Games," in *Games and Human Behavior, Essays in Honor of Amnon Rapoport*, ed. by D. Budescu, I. Erev, and R. Zwick. Kluwer Academic, Dordrecht/Norwell, M.A.

——— (1999b): "Experience-Weighted Attraction Learning in Normal Form Games," *Econometrica*, 67, 827–874.

Chen, Y., and F.-F. Tang (1998): "Learning and Incentive Compatible Mechanisms For Public Good Provision," *Journal of Political Economy*.

Erev, I., and A. Roth (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. Belief-Based Learning Models in Experimental Asymmetric-Information Games," *Econometrica*, 68, 605–642.

McFadden, D. (1974): "Conditional Logit Analyses of Qualitative Choice Behavior," in *Frontiers in Econometrics*, ed. by P. Zarembka. Academic Press, N.Y.

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

Roth, A. E., I. Erev, and R. Slonim (1998): "Learning and Equilibrium as Useful Approximations: Accracy of Prediction on Randomly Selected Constant Sum Games," Mimeo, University of Pittsburgh.

Selten, R. (1998): "Axiomatic Characterization of the Quadratic Scoring Rule," *Experimental Economics*, 1, 43–62.

Tang, F.-F. (1996): "Anticipatory Learning in Two-Person Games: An Experimental Study," Dissertation Paper B-393, University of Bonn.

Yellot, J. (1977): "The Relationship Between Luce's Choice Axioms, Thurstone's Theory of Comparative Judgement, and the Double Exponential Distribution," *Journal of Mathematical Psychology*, 15, 109–144.

Table 1: Results of the Monte Carlo simulations for 8 sets of true values of 3 parameters of the reinforcement learning model; maximum likelihood estimation; sample (a): 50, 125, 200 and 500 rounds of play of 1 pair of players

ML Estimators 50, 125, 200 and 500 rounds								
Set 1	ω_0 = 0.0				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-73.644	-41.194	-25.702	-7.8865	0.2372	0.2092	0.2166	0.1995
Stdev	225.48	170.45	129.70	76.001	0.4252	0.2235	0.1833	0.1231
Med	-0.0034	-0.0015	-0.0004	-0.0002	0.2376	0.2005	0.2038	0.1927
Set 2	ω_0 = 0.0				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-43.557	-65.471	-49.819	-21.641	0.2047	0.1616	0.1625	0.1673
Stdev	170.87	210.95	183.97	128.05	0.4254	0.2883	0.2130	0.1265
Med	0.0020	0.0010	0.0009	0.0004	0.1101	0.1272	0.1606	0.1805
Set 3	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0398	0.0021	0.0045	0.0447	0.7856	0.7895	0.7914	0.7946
Stdev	1.2880	1.1046	0.7490	0.7413	0.0777	0.0368	0.0278	0.0183
Med	0.0002	0.0002	0.0002	0.0001	0.8002	0.7957	0.7929	0.7943
Set 4	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-0.3007	-0.2248	-0.0581	-0.1370	0.7899	0.7962	0.8029	0.8021
Stdev	2.9156	1.8914	0.8974	0.7046	0.2336	0.1155	0.0700	0.0401
Med	-0.0002	0.0001	0.0001	0.0001	0.8421	0.8142	0.8153	0.8057
Set 5	ω_0 = 0.80				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-38.801	-16.687	-9.4718	-3.2638	0.2210	0.2087	0.2152	0.1989
Stdev	180.29	129.43	95.561	75.965	0.3662	0.2302	0.1783	0.1211
Med	0.8368	0.5098	0.4092	0.3813	0.1818	0.1927	0.1867	0.1927
Set 6	ω_0 = -0.80				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-84.705	-74.996	-109.90	-33.953	0.1885	0.1610	0.1628	0.1660
Stdev	224.284	214.10	258.71	146.46	0.4032	0.2818	0.2120	0.1275
Med	-0.6133	-0.4407	-0.3422	-0.4201	0.1020	0.1215	0.1469	0.1780
Set 7	ω_0 = 0.80				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	1.1495	0.9782	0.9676	0.9706	0.7734	0.7865	0.7912	0.7946
Stdev	1.6770	1.2202	1.2259	1.1650	0.0897	0.0405	0.0280	0.0181
Med	0.9335	1.0342	1.0459	1.0663	0.7891	0.7924	0.7945	0.7957
Set 8	ω_0 = -0.80				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-10.304	-0.9667	-0.8810	-0.8470	0.8046	0.7949	0.8023	0.8027
Stdev	75.744	1.2020	1.0747	1.0152	0.1838	0.1052	0.0709	0.0395
Med	-0.7832	-0.8571	-0.8292	-0.7826	0.8400	0.8137	0.8125	0.8084

Table 2: Results of the Monte Carlo simulations for 8 sets of true values of 3 parameters of the reinforcement learning model; maximum likelihood estimation; sample (a): 50, 125, 200 and 500 rounds of play of 1 pair of players

ML Estimator 50, 125, 200 and 500 rds				
Set 1	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.1916	0.1966	0.1966	0.2001
Stdev	0.0960	0.0484	0.0405	0.0258
Med	0.1878	0.1945	0.1923	0.2028
Set 2	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	-0.1989	-0.1972	-0.1982	-0.1980
Stdev	0.0799	0.0505	0.0389	0.0245
Med	-0.1893	-0.2008	-0.1975	-0.1974
Set 3	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.2115	0.2099	0.2069	0.2051
Stdev	0.0611	0.0365	0.0267	0.0177
Med	0.2111	0.2028	0.2070	0.2059
Set 4	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	-0.2302	-0.2070	-0.2030	-0.1965
Stdev	0.0682	0.0461	0.0364	0.0193
Med	-0.2202	-0.2026	-0.1971	-0.1950
Set 5	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.1954	0.1956	0.1968	0.2001
Stdev	0.0921	0.0500	0.0403	0.0258
Med	0.1937	0.1935	0.1922	0.2031
Set 6	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	-0.1989	-0.1977	-0.1977	-0.1977
Stdev	0.0798	0.0510	0.0388	0.0245
Med	-0.1918	-0.2008	-0.1984	-0.1975
Set 7	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.2119	0.2093	0.2059	0.2047
Stdev	0.0617	0.0361	0.0267	0.0177
Med	0.2071	0.2022	0.2046	0.2050
Set 8	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	-0.2323	-0.2093	-0.2042	-0.1971
Stdev	0.0738	0.0464	0.0363	0.0194
Med	-0.2261	-0.2073	-0.2010	-0.1964

Table 3: Results of the Monte Carlo simulations for 8 sets of true values of the 3 parameters of the reinforcement learning model; maximum likelihood estimation; sample (b): 40 rounds played by 5,10,20 and 40 pairs of players

ML Estimators 5, 10, 20 and 40 pairs								
Set 1	ω_0 = 0.0				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0289	0.0404	0.0292	-0.0017	0.2037	0.1977	0.2024	0.1973
Stdev	0.7208	0.4377	0.3245	0.2071	0.1930	0.1187	0.0748	0.0597
Med	-0.0125	0.0008	0.0041	-0.0004	0.2030	0.2054	0.2028	0.2035
Set 2	ω_0 = 0.0				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0278	0.0487	0.0327	0.0003	0.2188	0.2117	0.2014	0.2064
Stdev	0.7152	0.4484	0.3306	0.2076	0.1683	0.1370	0.0887	0.0618
Med	0.0015	0.0074	0.0004	0.0000	0.2094	0.2217	0.2047	0.2117
Set 3	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0082	0.0273	0.0226	0.0098	0.7982	0.8008	0.7996	0.7988
Stdev	0.3989	0.2519	0.1968	0.1055	0.0349	0.0222	0.0186	0.0119
Med	0.0002	0.0002	0.0002	0.0001	0.8056	0.7984	0.7998	0.8009
Set 4	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0460	0.0384	0.0453	0.0084	0.7881	0.8022	0.8011	0.8040
Stdev	0.4750	0.3115	0.2242	0.1312	0.0760	0.0509	0.0350	0.0265
Med	0.0002	0.0003	0.0004	0.0001	0.8019	0.8038	0.8015	0.8056
Set 5	ω_0 = 0.80				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	1.0060	0.8972	0.8455	0.8199	0.1978	0.1953	0.2014	0.1975
Stdev	0.9208	0.5068	0.3484	0.2467	0.1723	0.1101	0.0708	0.0598
Med	0.8616	0.7954	0.8108	0.8231	0.2038	0.1989	0.2053	0.1999
Set 6	ω_0 = $-$0.80				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-8.4570	-0.8168	-0.7960	-0.8101	0.2020	0.2083	0.2014	0.2055
Stdev	74.4259	0.5332	0.3176	0.2213	0.1690	0.1345	0.0881	0.0604
Med	-0.9240	-0.7422	-0.7598	-0.8306	0.2014	0.2256	0.1982	0.2086
Set 7	ω_0 = 0.80				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.8780	0.8661	0.8288	0.8242	0.7968	0.7987	0.7997	0.7988
Stdev	0.4237	0.3167	0.2181	0.1488	0.0349	0.0242	0.0170	0.0103
Med	0.8553	0.8435	0.8149	0.8205	0.7983	0.7989	0.8003	0.7987
Set 8	ω_0 = $-$0.80				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-0.8115	-0.7522	-0.7588	-0.7821	0.7817	0.7987	0.7996	0.8025
Stdev	0.4357	0.2978	0.2190	0.1302	0.0813	0.0498	0.0340	0.0257
Med	-0.7887	-0.7295	-0.7511	-0.7858	0.8001	0.8000	0.8026	0.8025

Table 4: Results of the Monte Carlo simulations for 8 sets of true values of the 3 parameters of the reinforcement learning model; maximum likelihood estimation; sample (b): 40 rounds played by 5,10,20 and 40 pairs of players

ML Estimator	5, 10, 20 and 40 pairs			
Set 1	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.1986	0.2003	0.1983	0.1981
Stdev	0.0442	0.0320	0.0210	0.0139
Med	0.1995	0.2023	0.1979	0.1984
Set 2	λ_0 = $-$0.20			
	$\hat{\lambda}$			
Mean	-0.2065	-0.1990	-0.1977	-0.1977
Stdev	0.0388	0.0267	0.0185	0.0145
Med	-0.2020	-0.1970	-0.1974	-0.1985
Set 3	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.1998	0.1989	0.2004	0.2003
Stdev	0.0252	0.0213	0.0164	0.0096
Med	0.1997	0.1973	0.2020	0.2006
Set 4	λ_0 = $-$0.20			
	$\hat{\lambda}$			
Mean	-0.2103	-0.1999	-0.2006	-0.1989
Stdev	0.0361	0.0240	0.0175	0.0123
Med	-0.2086	-0.1997	-0.2029	-0.1992
Set 5	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.1993	0.2004	0.1985	0.1979
Stdev	0.0433	0.0318	0.0204	0.0133
Med	0.1979	0.2002	0.1977	0.1988
Set 6	λ_0 = $-$0.20			
	$\hat{\lambda}$			
Mean	-0.2056	-0.1985	-0.1976	-0.1976
Stdev	0.0394	0.0273	0.0190	0.0145
Med	-0.2020	-0.1956	-0.1979	-0.1981
Set 7	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.2000	0.2005	0.2013	0.2007
Stdev	0.0282	0.0212	0.0140	0.0086
Med	0.1976	0.1998	0.2016	0.1998
Set 8	λ_0 = $-$0.20			
	$\hat{\lambda}$			
Mean	-0.2089	-0.2009	-0.2009	-0.1991
Stdev	0.0343	0.0231	0.0177	0.0124
Med	-0.2114	-0.1993	-0.2004	-0.2010

Table 5: Results of the Monte Carlo simulations for 4 sets of true values of 3 parameters of the reparameterized reinforcement learning model; MSD estimator; panel to the top: sample (a): 50, 125, 200 and 500 rounds of play of 1 pair of players; panel to the bottom: sample (b): 40 rounds of play of 5,10,20 and 40 pairs of players

MSD Estimators 50, 125, 200 and 500 rounds								
Set 1	ω_0 = 0.0				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-4.5e+24	-7.1e+10	-6.9e+21	-3.1e+10	0.2232	0.2075	0.2203	0.2012
Stdev	3.9e+25	7.1e+11	6.9e+22	3.1e+11	0.4167	0.2321	0.1830	0.1249
Med	-0.0028	-0.0010	-0.0006	-0.0001	0.2431	0.2140	0.2192	0.1963
Set 2	ω_0 = 0.0				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-2.7e+25	-5.8e+17	-8.5e+17	9.055	0.2014	0.1660	0.1722	0.1705
Stdev	2.7e+26	5.8e+18	8.5e+18	92.57	0.4031	0.2751	0.2058	0.1272
Med	0.0065	0.0009	0.0008	0.0003	0.1308	0.1520	0.1632	0.1805
Set 3	ω_0 = 0.0				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.2216	0.0334	0.0957	0.1346	0.7824	0.7881	0.7906	0.7945
Stdev	2.495	1.975	2.213	3.323	0.0810	0.0412	0.0310	0.0182
Med	0.0000	0.0002	0.0001	0.0001	0.7965	0.7936	0.7925	0.7932
Set 4	ω_0 = 0.0				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	6.942	-0.1672	-0.0528	-0.059	0.7842	0.7926	0.8012	0.8024
Stdev	71.01	1.536	1.096	0.9479	0.2267	0.1137	0.0725	0.0395
Med	-0.0015	0.0001	0.0001	0.0001	0.8358	0.8079	0.8114	0.8082
MSD Estimators 5, 10, 20 and 40 pairs								
Set 1	ω_0 = 0.0				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0226	0.0414	0.0078	-0.0016	-2.3342	-1.7960	-0.8532	0.1983
Stdev	0.7043	0.4383	0.2768	0.2077	10.824	6.7380	4.1121	0.0593
Med	0.0000	-0.0000	0.0000	-0.0001	0.2008	0.2008	0.2000	0.2034
Set 2	ω_0 = 0.0				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.3426	0.0477	0.0232	0.0002	0.2143	0.2103	0.2074	0.2050
Stdev	3.3630	0.4526	0.3241	0.2080	0.1737	0.1377	0.1442	0.0622
Med	0.0021	0.0064	0.0373	-0.0002	0.2085	0.2178	0.2197	0.2135
Set 3	ω_0 = 0.0				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0234	0.0330	-0.0017	0.0098	0.7975	0.8003	0.8013	0.7986
Stdev	0.4042	0.2786	0.1792	0.1199	0.0344	0.0229	0.0306	0.0124
Med	0.0003	0.0002	0.0001	0.0001	0.8024	0.7980	0.7993	0.7999
Set 4	ω_0 = 0.0				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0446	0.0341	0.0169	0.0130	0.7903	0.8027	0.7936	0.8042
Stdev	0.4731	0.3141	0.2186	0.1317	0.0751	0.0508	0.0496	0.0264
Med	0.0004	0.0005	0.0222	0.0002	0.8024	0.8060	0.7992	0.8062

Table 6: Results of the Monte Carlo simulations for 4 sets of true values of 3 parameters of the reparameterized reinforcement learning model; MSD estimator; panel to the top: sample (a): 50, 125, 200 and 500 rounds of play of 1 pair of players; panel to the bottom: sample (b): 40 rounds of play of 5,10,20 and 40 pairs of players

MSD Estimator 50, 125, 200, 500 rds				
Set 1	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.1970	0.1978	0.1968	0.2000
Stdev	0.0948	0.0498	0.0410	0.0260
Med	0.1970	0.1939	0.1917	0.2034
Set 2	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	-0.2053	-0.1995	-0.1995	-0.1982
Stdev	0.0814	0.0498	0.0393	0.0248
Med	-0.1971	-0.2021	-0.1988	-0.1968
Set 3	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.2259	0.2134	0.2091	0.2047
Stdev	0.0817	0.0453	0.0318	0.0178
Med	0.2190	0.2009	0.2086	0.2051
Set 4	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	-0.2337	-0.2087	-0.2047	-0.1972
Stdev	0.0723	0.0475	0.0376	0.0204
Med	-0.2268	-0.2009	-0.2009	-0.1954
MSD Estimator 5, 10, 20 and 40 pairs				
Set 1	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	-3.3529	-2.4569	-1.2900	0.1981
Stdev	11.562	7.2036	4.3320	0.0140
Med	0.2008	0.2009	0.2000	0.1985
Set 2	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	-0.2061	-0.1988	-0.1971	-0.1975
Stdev	0.0388	0.0269	0.0292	0.0143
Med	-0.2027	-0.1962	-0.1965	-0.1982
Set 3	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.2028	0.1999	0.2046	0.2004
Stdev	0.0295	0.0221	0.0258	0.0104
Med	0.1987	0.1990	0.2067	0.2010
Set 4	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	-0.2107	-0.1998	-0.1973	-0.1988
Stdev	0.0374	0.0244	0.0266	0.0127
Med	-0.2125	-0.1991	-0.1962	-0.1996

Table 7: Results of the Monte Carlo simulations for 4 sets of true values of 2 parameters of the reinforcement learning model; MSD RA estimator; panel to the top: sample (a): 50, 125, 200 and 500 rounds of play of 1 pair of players; panel to the bottom: sample (b): 40 rounds of play of 5,10,20 and 40 pairs of players

MSD_{RA} Estimators 50, 125, 200, 500 rounds								
Set 1	ω_0 = 0.00				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-0.0016	-0.0054	-0.0029	-0.0031	0.2417	0.1164	0.1494	0.1922
Stdev	0.1319	0.1152	0.09438	0.0485	0.5715	0.4425	0.3812	0.2108
Med	0.0003	0.0005	0.0004	0.0004	0.2814	0.1925	0.2000	0.1989
Set 2	ω_0 = 0.00				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0067	-0.0009	0.0002	-0.0015	0.2469	0.2331	0.2201	0.1997
Stdev	0.0808	0.0541	0.0324	0.0177	0.3770	0.2452	0.3027	0.1980
Med	-0.0002	0.0001	0.0007	0.0002	0.2044	0.2075	0.2046	0.2000
Set 3	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0004	0.0011	0.0004	-0.0004	0.7158	0.7902	0.8009	0.8008
Stdev	0.0820	0.0433	0.0196	0.0100	0.3637	0.0782	0.0518	0.0536
Med	-0.0001	0.0001	0.0000	0.0001	0.8258	0.7903	0.8017	0.8035
Set 4	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-0.0170	–0.0001	0.0001	0.0000	0.7436	0.8115	0.8053	0.8136
Stdev	0.0822	0.0006	0.0008	0.0002	0.2390	0.0672	0.0663	0.0455
Med	-0.0002	-0.0000	0.0000	0.0000	0.8026	0.8225	0.8229	0.8189
MSD_{RA} Estimators 5, 10, 20 and 40 pairs								
Set 1	ω_0 = 0.00				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-0.0135	-0.0094	-0.0103	0.0070	0.0915	0.1590	0.1426	0.2895
Stdev	0.0675	0.0640	0.0437	0.0379	0.3755	0.2953	0.2076	0.1203
Med	-0.0001	-0.0003	0.0000	-0.0003	0.2186	0.2026	0.1817	0.2633
Set 2	ω_0 = 0.00				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-0.0121	0.0020	-0.0042	-0.0077	0.2290	0.1689	0.3011	0.2800
Stdev	0.0513	0.0242	0.0273	0.0372	0.3825	0.2821	0.2413	0.1499
Med	-0.0009	-0.0001	-0.0001	-0.0007	0.2366	0.1896	0.2526	0.2791
Set 3	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0003	0.0019	0.0006	0.0001	0.7702	0.7621	0.7527	0.8107
Stdev	0.0027	0.0244	0.0024	0.0001	0.0818	0.0708	0.0753	0.0287
Med	0.0001	0.0001	0.0003	0.0000	0.7651	0.7676	0.7618	0.8108
Set 4	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0001	0.0004	0.0003	-0.0002	0.8114	0.8304	0.7919	0.8190
Stdev	0.0005	0.0006	0.0036	0.0002	0.0931	0.0845	0.0741	0.0138
Med	0.0000	0.0002	0.0001	-0.0002	0.8071	0.8365	0.8034	0.8227

Table 8: Results of the Monte Carlo simulations for 4 sets of true values of 3 parameters of the reparameterized reinforcement learning model; MSD RA estimator; panel to the top: sample (a): 50, 125, 200 and 500 rounds of play of 1 pair of players; panel to the bottom: sample (b): 40 rounds of play of 5,10,20 and 40 pairs of players

MSD_{RA} Estimator 50, 125, 200, 500 rds				
Set 1	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	-0.3465	0.1085	0.1511	0.1487
Stdev	3.852	0.7154	0.2855	0.2284
Med	0.2014	0.2084	0.2021	0.2013
Set 2	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	-0.1624	-0.1577	-0.1831	-0.1702
Stdev	0.2497	0.2145	0.0952	0.1306
Med	-0.1932	-0.2092	-0.2048	-0.2094
Set 3	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.2619	0.2126	0.2081	0.2032
Stdev	0.1234	0.0281	0.0190	0.0103
Med	0.2529	0.2129	0.2048	0.2027
Set 4	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	-2.817	-0.2186	-0.2125	-0.2041
Stdev	14.00	0.0330	0.0281	0.0122
Med	-0.2343	-0.2122	-0.2035	-0.2015
MSD_{RA} Estimator 5, 10, 20, 40 pairs				
Set 1	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.1786	0.1929	0.2319	0.1562
Stdev	0.1937	0.1558	0.0628	0.0623
Med	0.2111	0.1995	0.2188	0.1600
Set 2	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	-0.1671	-0.1693	-0.1661	-0.1578
Stdev	0.1959	0.1044	0.1166	0.0781
Med	-0.1923	-0.1578	-0.1894	-0.2037
Set 3	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.2121	0.2135	0.1919	0.2024
Stdev	0.0209	0.0202	0.0223	0.0041
Med	0.2091	0.2106	0.1950	0.2005
Set 4	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	-0.2036	-0.1775	-0.2025	-0.2102
Stdev	0.03561	0.03461	0.01929	0.0060
Med	-0.2103	-0.1788	-0.2025	-0.2105

Table 9: Results of the Monte Carlo simulations for 4 sets of true values of 3 parameters of the reparameterized reinforcement learning model; log score RA estimator; panel to the top: sample (a): 50, 125, 200 and 500 rounds of play of 1 pair of players; panel to the bottom: sample (b): 40 rounds of play of 5,10,20 and 40 pairs of players

log $score_{RA}$ Estimators 50, 125, 200, 500 rounds								
Set 1	ω_0 = 0.00				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0046	-0.0021	-0.0004	0.0001	0.2402	0.1285	0.1550	0.1974
Stdev	0.1472	0.1284	0.1067	0.0585	0.5722	0.4577	0.3937	0.2165
Med	0.0003	0.0005	0.0007	0.0009	0.2772	0.1887	0.1999	0.1914
Set 2	ω_0 = 0.00				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0053	0.0008	-0.0006	0.0024	0.2484	0.2329	0.2190	0.1946
Stdev	0.0844	0.0683	0.0446	0.0303	0.3766	0.2574	0.3327	0.2016
Med	-0.0004	-0.0002	0.0006	0.0006	0.2057	0.1879	0.2124	0.1997
Set 3	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-0.0153	-0.0020	0.0004	-0.0004	0.7156	0.7904	0.8012	0.8010
Stdev	0.1175	0.0553	0.0303	0.0117	0.3639	0.0777	0.0511	0.0533
Med	-0.0003	-0.0001	0.0001	0.0001	0.8258	0.7915	0.8014	0.8040
Set 4	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0006	-0.0010	-0.0030	-0.0012	0.7561	0.7828	0.8084	0.7870
Stdev	0.0608	0.0365	0.0310	0.0153	0.2112	0.1196	0.0631	0.0991
Med	0.0001	0.0001	-0.0001	0.0001	0.8039	0.8121	0.8169	0.8033
log $score_{RA}$ Estimators 5, 10, 20 and 40 pairs								
Set 1	ω_0 = 0.00				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-0.0135	-0.0096	-0.0105	0.0070	0.0914	0.1569	0.1427	0.2888
Stdev	0.0674	0.0644	0.0442	0.0379	0.3735	0.2934	0.2074	0.1211
Med	0.0001	-0.0003	0.0002	-0.0003	0.2186	0.2028	0.1817	0.2633
Set 2	ω_0 = 0.00				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-0.0124	0.0020	-0.0041	-0.0076	0.2315	0.1724	0.3012	0.2802
Stdev	0.0517	0.0243	0.0273	0.0373	0.3823	0.2842	0.2416	0.1498
Med	-0.0009	–0.0001	-0.0001	-0.0007	0.2366	0.1904	0.2526	0.2791
Set 3	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0003	0.0022	0.0006	0.0001	0.7702	0.7621	0.7524	0.8107
Stdev	0.0027	0.0239	0.0024	0.0002	0.0818	0.0708	0.0751	0.0286
Med	0.0001	0.0001	0.0003	0.0001	0.7651	0.7676	0.7618	0.8108
Set 4	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0001	0.0004	0.0003	-0.0002	0.8114	0.8304	0.7919	0.8190
Stdev	0.0005	0.0006	0.0036	0.0002	0.0931	0.0845	0.0740	0.0138
Med	0.0000	0.0002	0.0001	-0.0001	0.8071	0.8365	0.8034	0.8227

Table 10: Results of the Monte Carlo simulations for 4 sets of true values of the 3 parameters of the reparameterized reinforcement learning model; log score RA estimator; panel to the top: sample (a): 50, 125, 200 and 500 rounds of play of 1 pair of players; panel to the bottom: sample (b): 40 rounds of play of 5, 10, 20 and 40 pairs of players

log score_{RA} Estimator 50, 125, 200, 500 rds				
Set 1	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.0141	0.1144	0.1496	0.1488
Stdev	0.8989	0.7158	0.2859	0.2290
Med	0.2027	0.2084	0.2013	0.2012
Set 2	λ_0 = $-$0.20			
	$\hat{\lambda}$			
Mean	-0.1663	-0.1576	-0.1958	-0.1691
Stdev	0.2410	0.2164	0.1095	0.1332
Med	-0.2001	-0.2105	-0.2091	-0.2075
Set 3	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.2494	0.2125	0.2077	0.2034
Stdev	0.1608	0.0284	0.0190	0.0103
Med	0.2528	0.2125	0.2048	0.2033
Set 4	λ_0 = $-$0.20			
	$\hat{\lambda}$			
Mean	-0.2014	-0.2107	-0.2023	-0.2053
Stdev	0.06341	0.02968	0.0247	0.0250
Med	-0.2030	-0.2084	-0.2030	-0.2045
log score_{RA} Estimator 5, 10, 20, 40 pairs				
Set 1	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.1792	0.1925	0.2320	0.1561
Stdev	0.1561	0.1992	0.1562	0.0629
Med	0.2093	0.1995	0.2188	0.1600
Set 2	λ_0 = $-$0.20			
	$\hat{\lambda}$			
Mean	-0.1673	-0.1697	-0.1664	-0.1577
Stdev	0.1956	0.1044	0.1164	0.0780
Med	-0.1923	-0.1578	-0.1894	-0.2037
Set 3	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.2121	0.2135	0.1920	0.2024
Stdev	0.0209	0.0202	0.0224	0.0041
Med	0.2091	0.2106	0.1950	0.2005
Set 4	λ_0 = $-$0.20			
	$\hat{\lambda}$			
Mean	-0.2036	-0.1775	-0.2025	-0.2102
Stdev	0.0356	0.0346	0.0193	0.0060
Med	-0.2103	-0.1788	-0.2025	-0.2105

Table 11: Results of the Monte Carlo simulations for 4 sets of true values of 3 parameters of the reparameterized reinforcement learning model; MSD RA AVE estimator; panel to the top: sample (a): 50, 125, 200 and 500 rounds of play of 1 pair of players; panel to the bottom: sample (b): 40 rounds of play of 5,10,20 and 40 pairs of players

MSD_{RA_ave} Estimators 50, 125, 200, 500 rounds								
Set 1	ω_0 = 0.00				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0338	0.0194	0.0030	0.0032	0.3613	0.0274	0.1216	0.2318
Stdev	0.1846	0.1298	0.0771	0.0564	0.9192	0.6745	0.5464	0.4887
Med	-0.0002	0.0005	0.0007	0.0003	0.5188	0.1621	0.1907	0.2072
Set 2	ω_0 = 0.00				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-0.0079	0.0017	0.0023	0.0008	0.1946	0.2751	0.1850	0.2363
Stdev	0.1245	0.0745	0.0517	0.0367	0.7637	0.7242	0.7104	0.6317
Med	-0.0005	-0.0050	0.0010	-0.0001	0.2100	0.4112	0.2359	0.2527
Set 3	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-0.0328	-0.0099	0.0019	0.0009	0.8748	0.7627	0.8266	0.8035
Stdev	0.1499	0.0791	0.0457	0.0190	1.721	0.2900	0.0429	0.0648
Med	-0.0004	0.0000	0.0001	0.0001	0.8402	0.8181	0.8248	0.8140
Set 4	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-0.0102	-0.0053	-0.0017	-0.0007	0.6256	0.744	0.7889	0.7901
Stdev	0.0589	0.0363	0.0219	0.0096	0.5552	0.3825	0.1505	0.1928
Med	0.0001	0.0000	0.0000	0.0001	0.81	0.8282	0.8108	0.8049
MSD_{RA_ave} Estimators 5, 10, 20 and 40 pairs								
Set 1	ω_0 = 0.00				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-0.0084	-0.0064	-0.0767	0.0611	-0.0631	0.2289	0.01236	0.2352
Stdev	0.1204	0.1978	0.2668	0.4447	0.5635	0.4181	0.3883	0.1418
Med	0.0002	-0.0000	-0.0007	-0.0001	0.2103	0.2176	0.1567	0.2225
Set 2	ω_0 = 0.00				ϕ_0 = 0.20			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-0.0120	-0.0006	0.0045	-0.0023	0.4473	0.2088	0.1442	0.1820
Stdev	0.0823	0.0921	0.1210	0.0293	0.4841	0.3054	0.4810	0.2088
Med	-0.0050	-0.0010	0.0002	0.0000	0.4725	0.2270	0.2260	0.2217
Set 3	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	-0.0002	0.0002	0.0002	0.0001	0.7878	0.7656	0.7832	0.7979
Stdev	0.0034	0.0003	0.0004	0.0001	0.07799	0.0974	0.1043	0.0442
Med	0.0001	0.0001	0.0002	0.0000	0.8008	0.7895	0.7980	0.8055
Set 4	ω_0 = 0.00				ϕ_0 = 0.80			
	$\hat{\omega}$				$\hat{\phi}$			
Mean	0.0000	-0.0021	0.0031	-0.0001	0.8240	0.7789	0.8012	0.8037
Stdev	0.0004	0.0263	0.0263	0.0008	0.0608	0.2827	0.0834	0.0526
Med	0.0000	0.0002	0.0000	0.0000	0.8061	0.8334	0.8095	0.8141

Table 12: Results of the Monte Carlo simulations for 4 sets of true values of 3 parameters of the reinforcement learning model; MSD RA AVE estimator; panel to the top: sample (a): 50, 125, 200 and 500 rounds of play of 1 pair of players; panel to the bottom: sample (b): 40 rounds of play of 5,10,20 and 40 pairs of players

MSD_{RA_ave} Estimator 50, 125, 200, 500 rds				
Set 1	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	-0.8325	-0.4192	-0.1698	0.1592
Stdev	4.096	2.475	2.270	0.6987
Med	0.1742	0.2006	0.2086	0.2029
Set 2	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	0.2632	-0.1210	-0.1403	-0.1469
Stdev	3.201	1.277	0.7934	0.4896
Med	-0.0239	-0.2076	-0.2025	-0.2118
Set 3	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.0038	0.2096	0.1985	0.2000
Stdev	1.328	0.1445	0.0185	0.0185
Med	0.2102	0.2050	0.2013	0.2007
Set 4	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	-0.0947	-0.174	-0.199	-0.204
Stdev	0.361	0.1774	0.1321	0.0270
Med	-0.1858	-0.195	-0.2007	-0.2016
MSD_{RA_ave} Estimators 5, 10, 20, 40 pairs				
Set 1	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.1891	0.1706	0.2845	0.1737
Stdev	0.2122	0.4262	0.1267	0.0532
Med	0.1953	0.1813	0.2393	0.206
Set 2	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	-0.1557	-0.2051	-0.1899	-0.1853
Stdev	0.2231	0.1255	0.1341	0.0538
Med	-0.1721	-0.1921	-0.1956	-0.2028
Set 3	λ_0 = 0.20			
	$\hat{\lambda}$			
Mean	0.2093	0.2111	0.1952	0.2059
Stdev	0.0253	0.0158	0.0226	0.0081
Med	0.2072	0.2081	0.1993	0.2035
Set 4	$\lambda_0 = -0.20$			
	$\hat{\lambda}$			
Mean	-0.2055	-0.1801	-0.2080	-0.2065
Stdev	0.02083	0.05516	0.0231	0.0122
Med	-0.2106	-0.1990	-0.2045	-0.2088

Table 13: Results of the Monte Carlo study for 4 sets of true values of the 3 parameters of the reinforcement learning model; Ratio of the standard deviation of the empirical distribution of the MSD estimator to the standard deviation of the MLE; panel to the top: sample (a): 50, 125, 200, 500 rounds of play of 1 pair of players; panel to the bottom: sample (b): 40 rounds of play of 5, 10, 20, 40 pairs of players

Ratio of standard deviations: MSD estimator to MLE 50, 125, 200, 500 rds												
Set 1	ω_0 = 0.0				ϕ_0 = 0.20				λ_0 = 0.20			
MSD	2E+23	4E+09	5E+20	4E+09	0.98	1.04	1.00	1.015	0.99	1.03	1.01	1.008
Set 2	ω_0 = 0.0				ϕ_0 = 0.20				λ_0 = -0.20			
MSD	2E+24	3E+16	5E+16	0.723	0.95	0.95	0.97	1.006	1.02	0.99	1.01	1.012
Set 3	ω_0 = 0.0				ϕ_0 = 0.80				λ_0 = 0.20			
MSD	1.94	1.79	2.96	4.483	1.04	1.12	1.12	0.995	1.34	1.24	1.19	1.017
Set 4	ω_0 = 0.0				ϕ_0 = 0.80				λ_0 = -0.20			
MSD	24.4	0.81	1.22	1.345	0.97	0.98	1.04	0.985	1.06	1.03	1.03	1.057
Ratio of standard deviations: MSD estimator to MLE 5, 10, 20, 40 pairs												
Set 1	ω_0 = 0.0				ϕ_0 = 0.20				λ_0 = 0.20			
MSD	0.98	1.00	0.85	1.003	56.1	56.8	55	0.993	262	225	206	1.007
Set 2	ω_0 = 0.0				ϕ_0 = 0.20				λ_0 = -0.20			
MSD	4.7	1.01	0.98	1.002	1.03	1.01	1.63	1.006	1.00	1.01	1.58	0.986
Set 3	ω_0 = 0.0				ϕ_0 = 0.80				λ_0 = 0.20			
MSD	1.01	1.11	0.91	1.136	0.99	1.03	1.65	1.042	1.17	1.04	1.57	1.083
Set 4	ω_0 = 0.0				ϕ_0 = 0.80				λ_0 = -0.20			
MSD	1.00	1.01	0.98	1.004	0.99	1.00	1.42	0.996	1.04	1.02	1.52	1.033

Table 14: Results of the Monte Carlo study for 4 sets of true values of the 3 parameters of the reinforcement learning model; Ratio of the standard deviation of the empirical distribution of the MSD RA, log score RA and MSD RA ave estimator to the standard deviation of the MLE; panel to the top: sample (a): 50, 125, 200 and 500 rounds of play of 1 pair of players; panel to the bottom: sample (b): 40 rounds of play of 5, 10, 20 and 40 pairs of players

Ratio of standard deviations: Deviation estimators to MLE 50, 125, 200, 500 rds												
Set 1	ω_0 = 0.0				ϕ_0 = 0.20				λ_0 = 0.20			
MSD_{RA}	0.183	0.263	0.291	0.234	2.961	3.728	5.096	3.531	87.15	22.36	13.60	16.43
log $score_{RA}$	0.204	0.293	0.329	0.282	2.965	3.856	5.263	3.626	20.34	22.37	13.61	16.47
MSD_{RA_ave}	0.256	0.297	0.238	0.272	4.763	5.682	7.305	8.186	92.67	77.34	108.1	50.27
Set 2	ω_0 = 0.0				ϕ_0 = 0.20				λ_0 = $-$0.20			
MSD_{RA}	0.113	0.121	0.098	0.085	2.240	1.790	3.413	3.204	6.436	8.034	5.146	9.007
log $score_{RA}$	0.118	0.152	0.135	0.146	2.238	1.879	3.751	3.262	6.211	8.105	5.919	9.186
MSD_{RA_ave}	0.166	0.174	0.156	0.177	4.538	5.286	8.009	10.22	82.50	47.83	42.89	33.77
Set 3	ω_0 = 0.0				ϕ_0 = 0.80				λ_0 = 0.20			
MSD_{RA}	0.206	0.172	0.100	0.095	10.42	3.523	2.780	4.504	4.897	1.319	1.159	1.073
log $score_{RA}$	0.295	0.220	0.154	0.111	10.43	3.500	2.753	4.479	6.381	1.333	1.159	1.073
MSD_{RA_ave}	0.376	0.314	0.233	0.180	49.32	13.06	2.306	5.445	52.70	6.784	1.128	1.917
Set 4	ω_0 = 0.0				ϕ_0 = 0.80				λ_0 = $-$0.20			
MSD_{RA}	0.173	0.002	0.004	0.002	3.145	1.32	0.189	1.717	387.8	1.375	1.606	0.992
log$score_{RA}$	0.128	0.117	0.138	0.117	2.779	2.35	1.800	3.740	1.756	1.238	1.411	2.033
MSD_{RA_ave}	0.124	0.117	0.098	0.073	7.305	7.515	4.300	7.275	10	7.392	7.549	2.195
Ratio of standard deviations: Deviation estimators to MLE 5, 10, 20, 40 pairs												
Set 1	ω_0 = 0.0				ϕ_0 = 0.20				λ_0 = 0.20			
MSD_{RA}	0.094	0.146	0.134	0.183	1.946	2.488	2.775	2.015	4.382	4.869	2.990	4.489
log $score_{RA}$	0.094	0.147	0.136	0.183	1.935	2.472	2.773	2.028	4.507	4.881	2.995	4.489
MSD_{RA_ave}	0.256	0.297	0.238	0.272	4.763	5.682	7.305	8.186	92.67	77.34	108.1	50.27
Set 2	ω_0 = 0.0				ϕ_0 = 0.20				λ_0 = $-$0.20			
MSD_{RA}	0.072	0.054	0.083	0.180	2.273	2.059	2.720	2.426	5.049	3.910	6.303	5.386
log $score_{RA}$	0.072	0.054	0.083	0.180	2.272	2.074	2.724	2.424	5.041	3.910	6.292	5.379
MSD_{RA_ave}	0.174	0.166	0.156	0.177	4.538	5.286	8.009	10.22	82.50	47.83	42.89	33.77
Set 3	ω_0 = 0.0				ϕ_0 = 0.80				λ_0 = 0.20			
MSD_{RA}	0.007	0.097	0.012	0.002	2.344	3.189	4.048	2.412	0.829	0.948	1.360	0.427
log $score_{RA}$	0.007	0.095	0.012	0.002	2.344	3.189	4.043	2.403	0.829	0.948	1.366	0.427
MSD_{RA_ave}	0.376	0.314	0.233	0.180	49.32	13.06	2.306	5.445	52.7	6.784	1.128	1.917
Set 4	ω_0 = 0.0				ϕ_0 = 0.80				λ_0 = $-$0.20			
MSD_{RA}	0.001	0.002	0.016	0.002	1.225	1.660	2.117	0.521	0.986	1.442	1.103	0.488
log $score_{RA}$	0.001	0.002	0.016	0.002	1.225	1.660	2.117	0.521	0.986	1.442	1.103	0.488
MSD_{RA_ave}	0.124	0.117	0.098	0.073	7.305	7.515	4.300	7.275	10.00	7.392	7.549	2.195

Figure 1: Empirical distributions of the MLE of the 3 parameters of the model; 50, 125, 200 and 500 rounds; ω_0 = 0.00, ϕ_0 = 0.20, λ_0 = 0.20

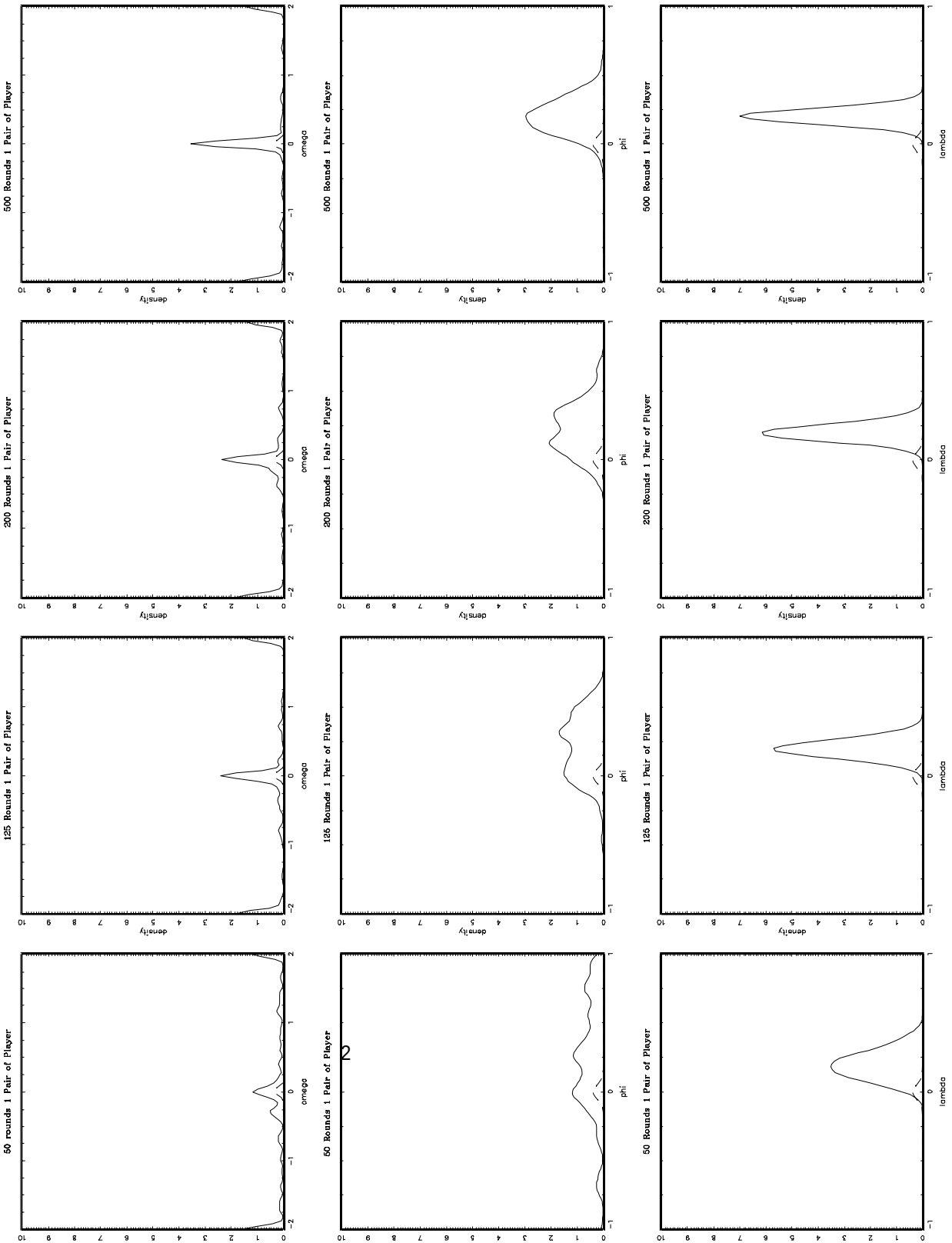

Figure 2: Empirical distributions of the MLE of the 3 parameters of the model; 5, 10, 20 and 20 pairs; ω_0 = 0.00, ϕ_0 = 0.20, λ_0 = 0.20

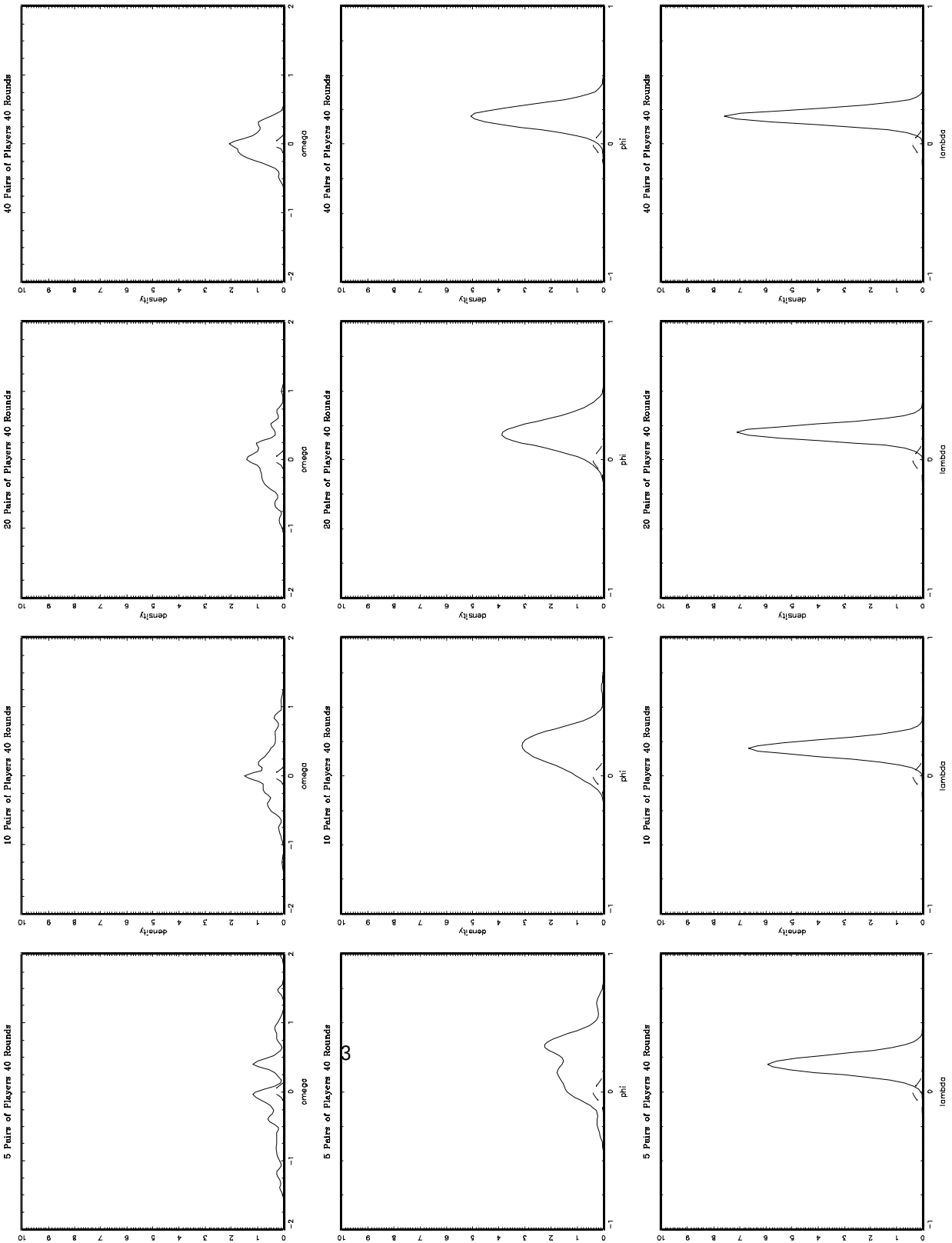

Figure 3: Empirical distributions of the MSD estimator of the 3 parameters of the model; 50, 125, 200 and 500 rounds; ω_0 = 0.00, ϕ_0 = 0.20, λ_0 = 0.20

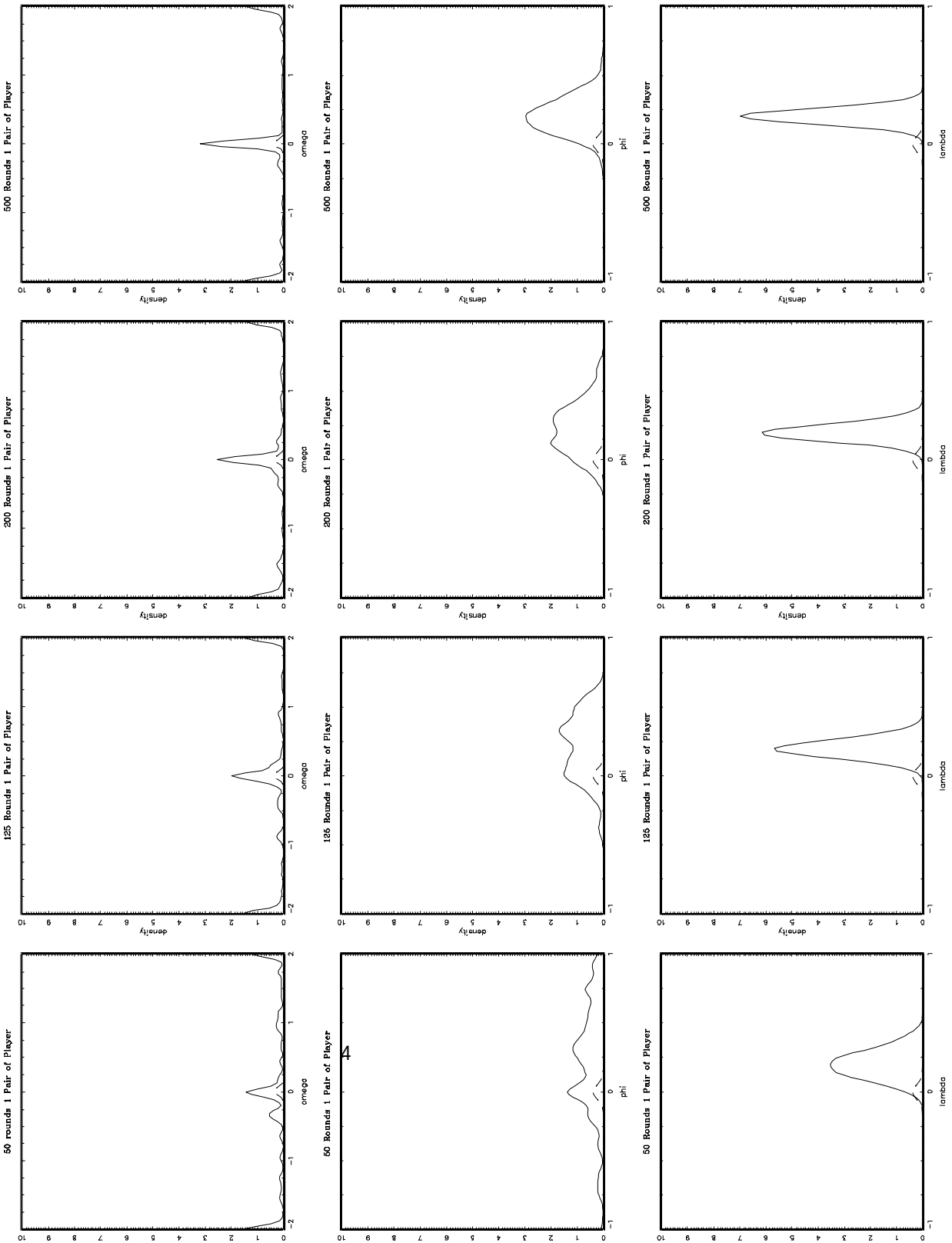

Figure 4: Empirical distributions of the MSD estimator of the 3 parameters of the model; 5, 10, 20 and 40 pairs; ω_0 = 0.00, ϕ_0 = 0.20, λ_0 = 0.20

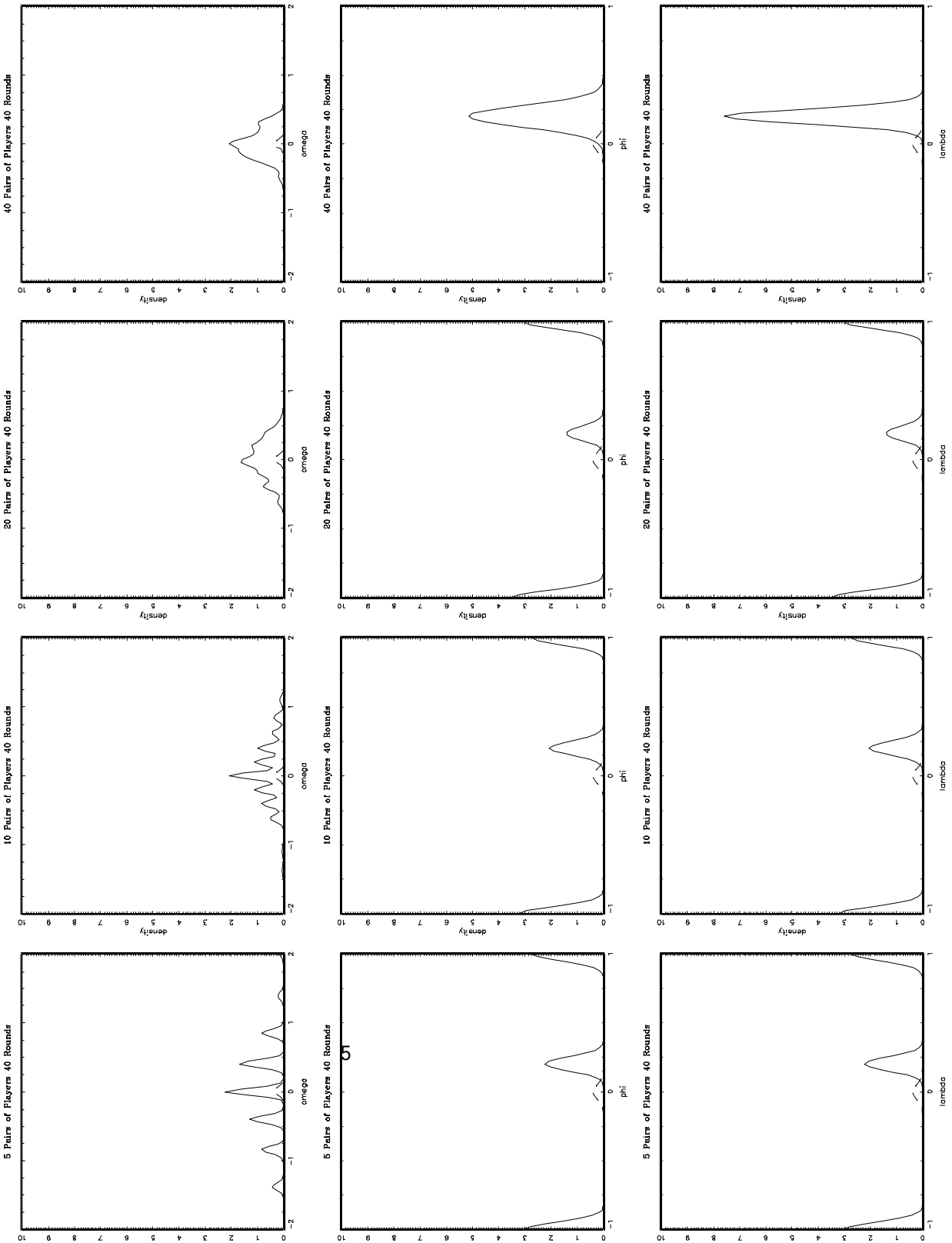

Figure 5: Empirical distributions of the MSD_{RA} estimator of the 3 parameters of the model; 50, 125, 200 and 500 rounds; ω_0 = 0.00, ϕ_0 = 0.20, λ_0 = 0.20

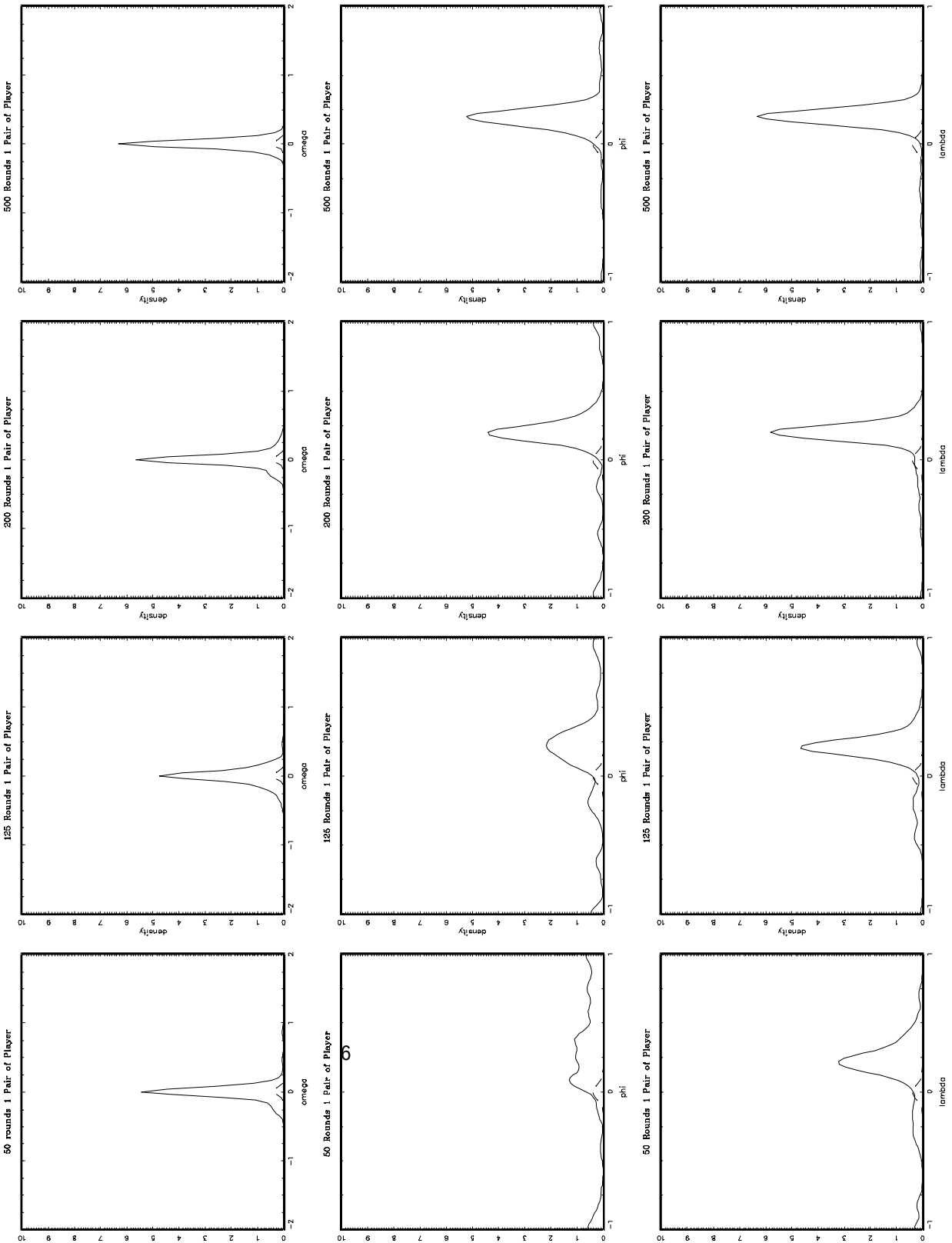

Figure 6: Empirical distributions of the MSD_{RA} estimator of the 3 parameters of the model; 5, 10, 20 and 40 pairs; ω_0 = 0.00, ϕ_0 = 0.20, λ_0 = 0.20

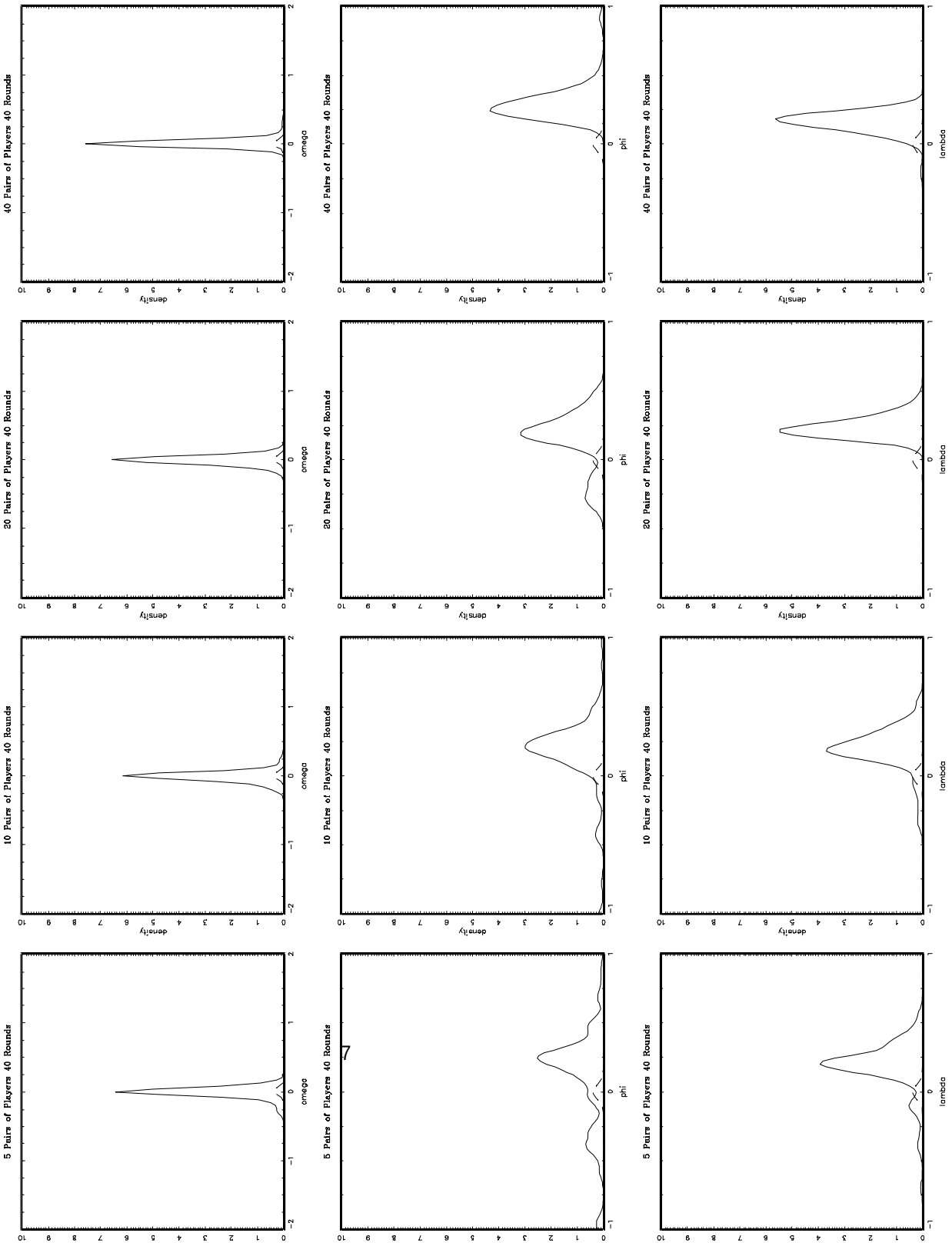

Figure 7: Empirical distributions of the log score of the 3 parameters of the model; 50, 125, 200 and 500 rounds; ω_0 = 0.00, ϕ_0 = 0.20, λ_0 = 0.20

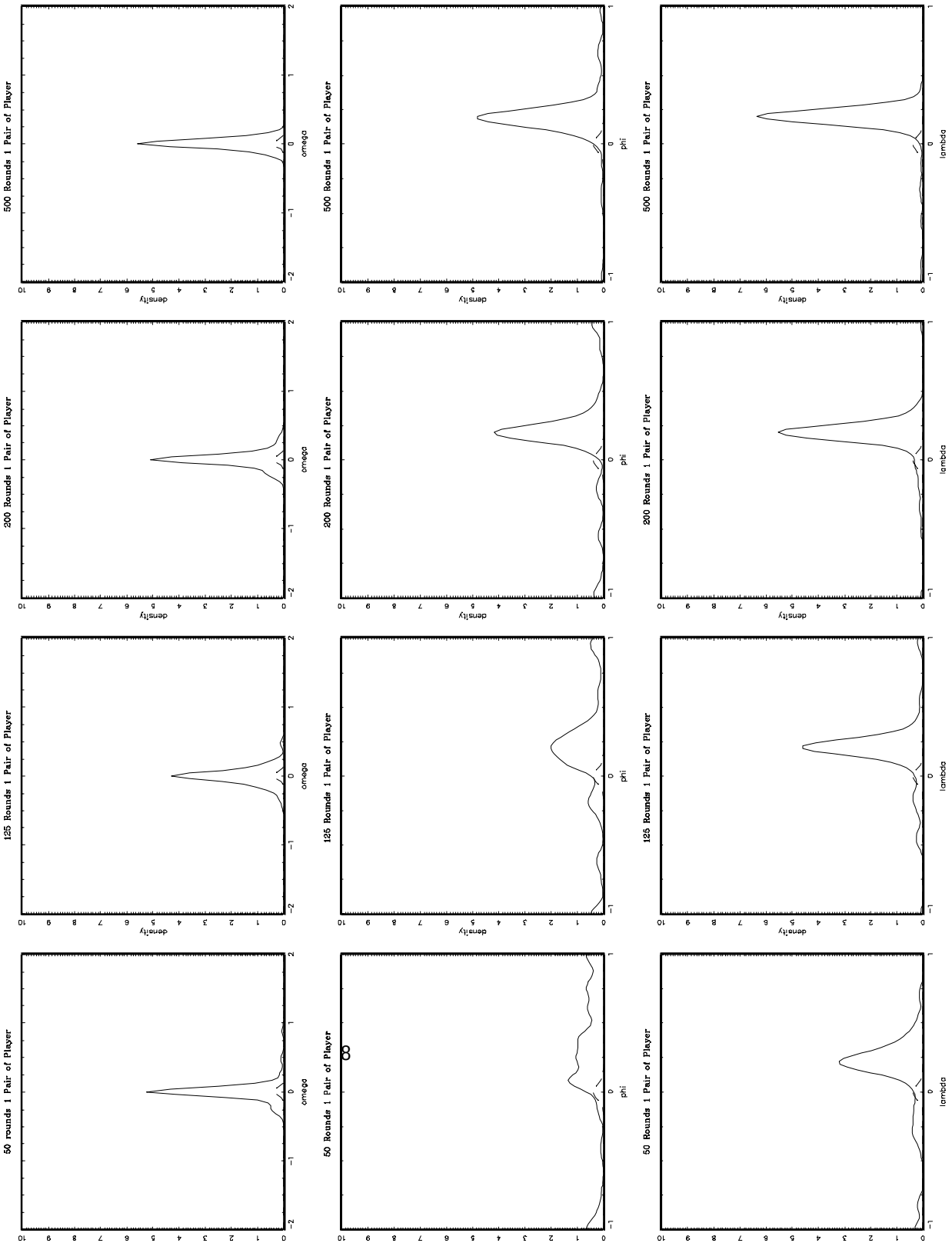

Figure 8: Empirical distributions of the log score of the 3 parameters of the model; 5, 10, 20 and 40 pairs; ω_0 = 0.00, ϕ_0 = 0.20, λ_0 = 0.20

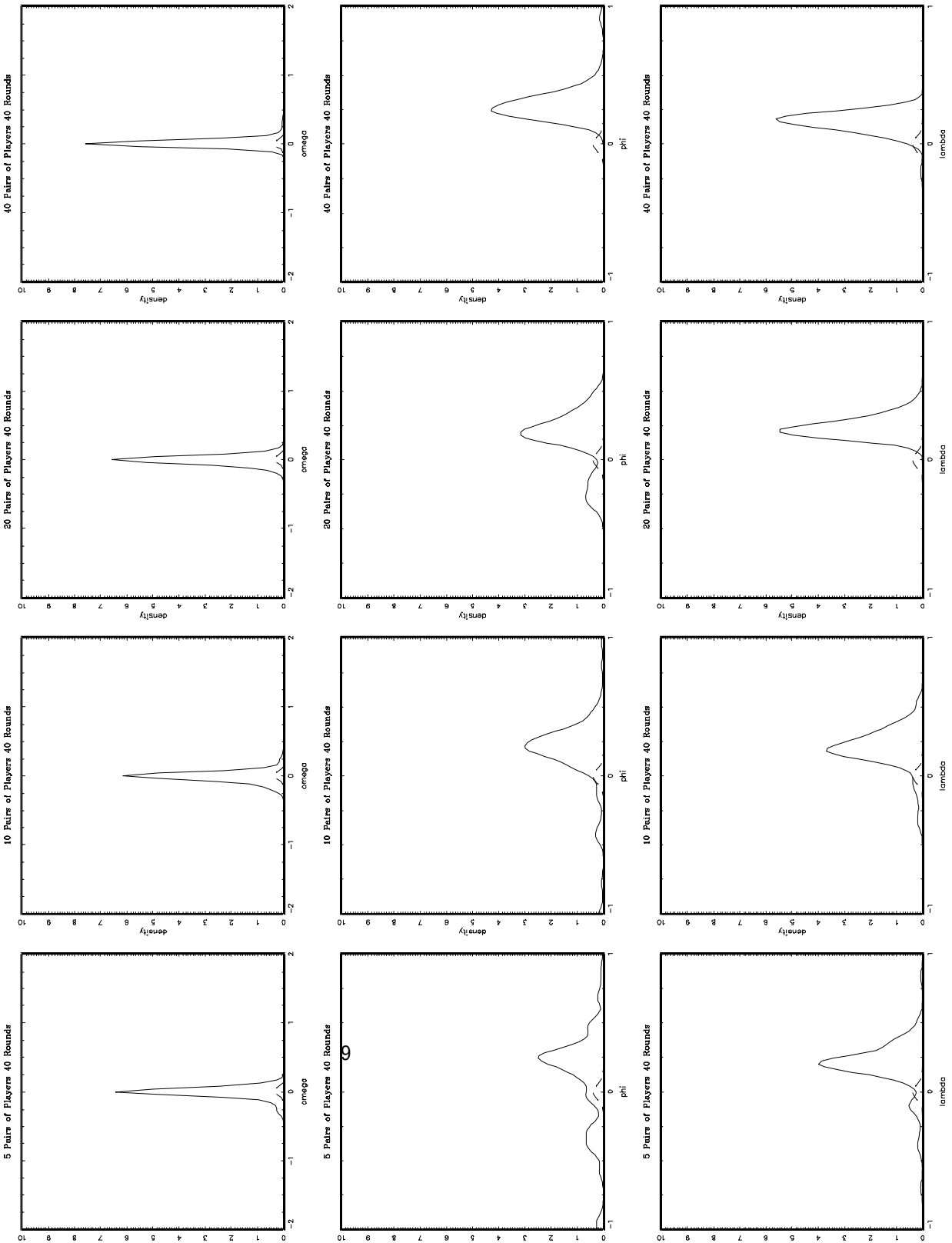

Figure 9: Empirical distributions of the MSD_{RA_ave} of the 3 parameters of the model; 50, 125, 200 and 500 rounds; ω_0 = 0.00, ϕ_0 = 0.20, λ_0 = 0.20

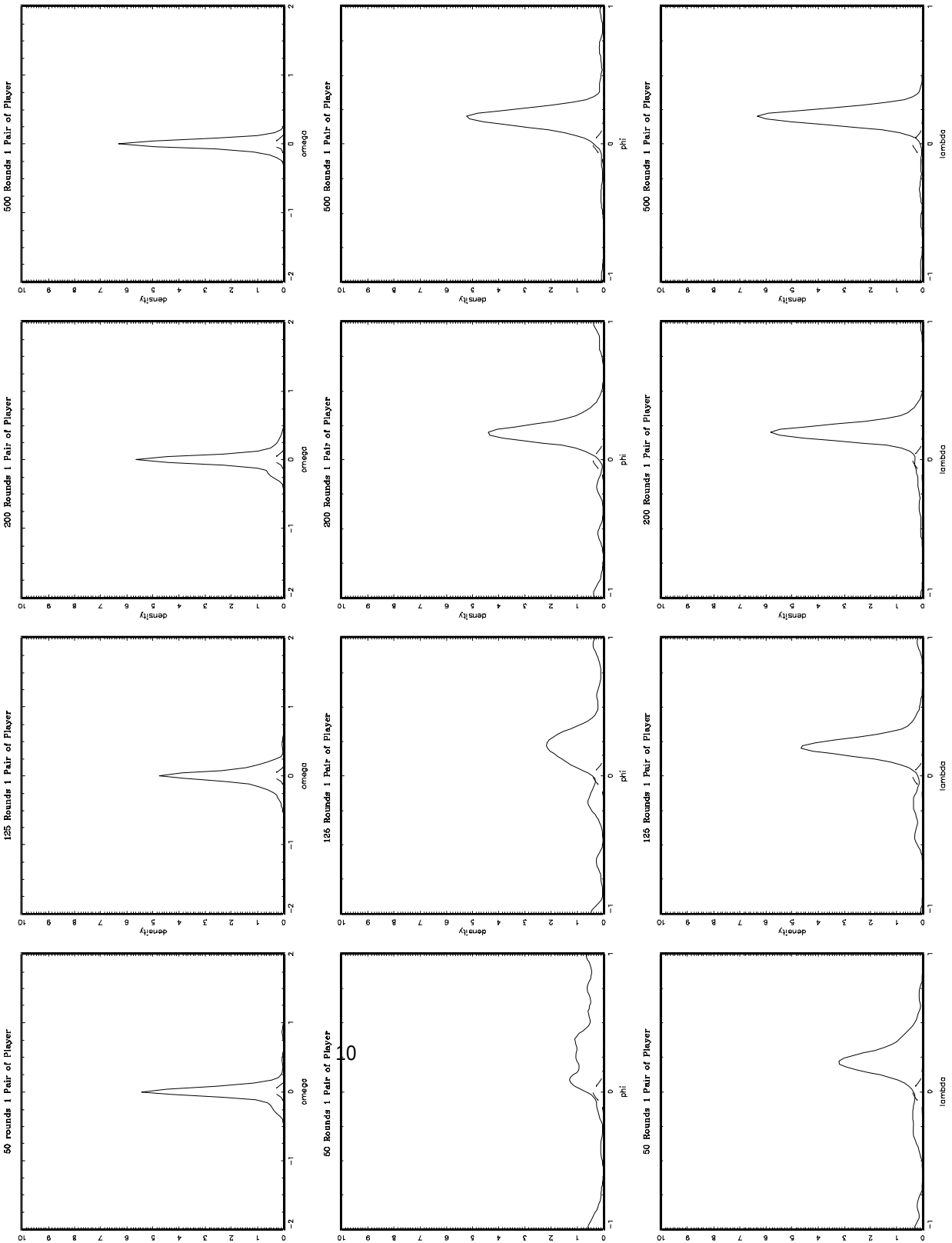

Figure 10: Empirical distributions of the MSD_{RA_ave} of the 3 parameters of the model; 5, 10, 20 and 40 pairs; ω_0 = 0.00, ϕ_0 = 0.20, λ_0 = 0.20