

## **Appendix 1**

### **ADDITIONAL MODEL EVALUATION**

#### *Macro-level statistics*

Using the estimated parameter values from the individual level estimation we can generate the macro level statistics that describe the distributions of the model parameters. Since the parameter estimation is based on individual observations, ignoring the interactions of the agents, we wonder whether the estimated models generate the same macro level metrics as observed in the data. In fact we generate simulated data with the model we first calibrated on the empirical experimental data, and explore whether it generates the macro-level statistics as we have observed in the data. We will do this by using the estimates of the various types of models as discussed in the main paper (representative agent, different types of agents, individual estimation).

Using the estimated parameters of a representative agent estimation, Model SP+L+S in Tables 3 and 4, we simulate the experiments and compare the average contribution levels and standard deviations between the actual experimental data and the simulated data. We do this for the 10 round experiments (Figure A1), and for the 40 and 60 round experiments (Figures A2). We will show the results from single simulation (Figures A1a and A2a) and for the average of 100 simulations (Figures A1b and A2b). The results are quite close to the real data for the 10 round experiment, but for the 40/60 round experiment the contribution levels decline much faster than in the real data. If we look at the standard deviation, we see a larger standard deviation than most empirical observations (Figure A3). Thus the simulated agents differ more in the token investments within a group, than the real players. This might be caused by the probabilistic nature of the model.

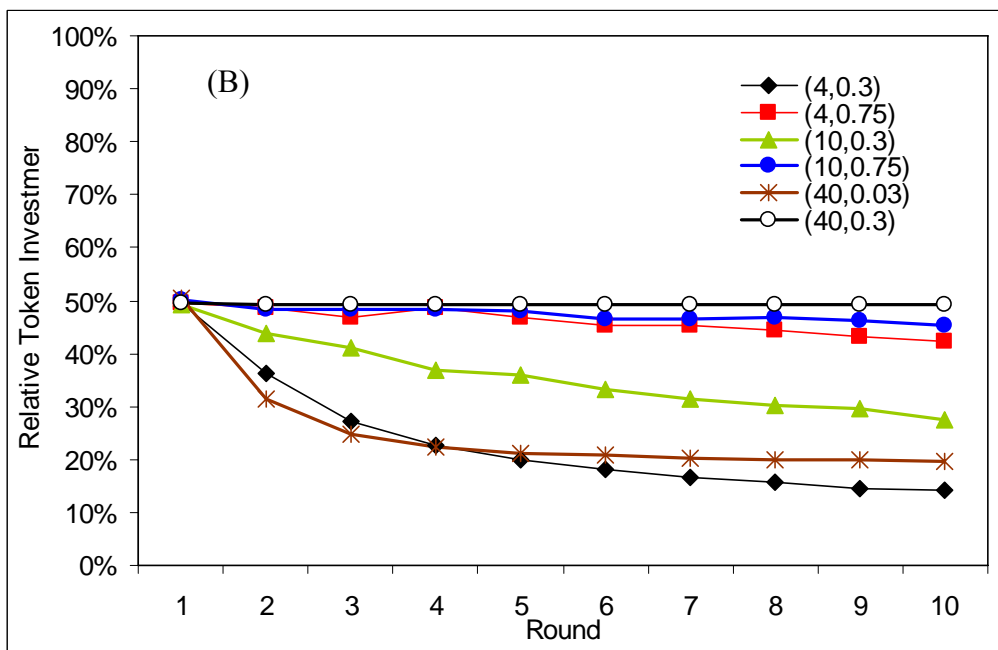
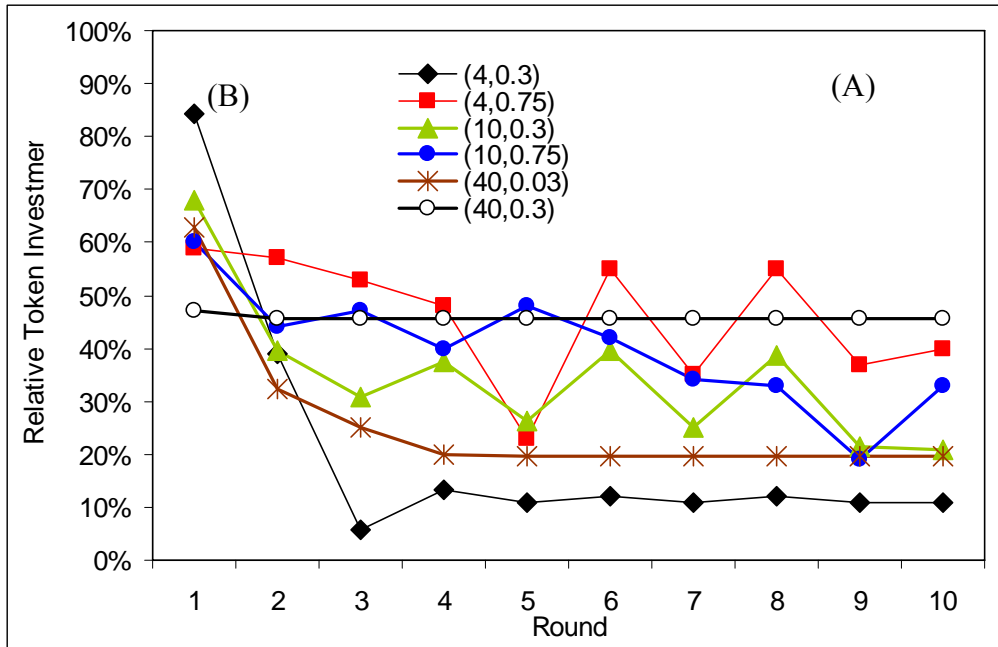


Figure A1: Relative token investments when we use the parameters of estimation 3 (Table 3) and simulate the token investments for one experiment of 10 rounds (A), or 100 experiments of 10 rounds (B).

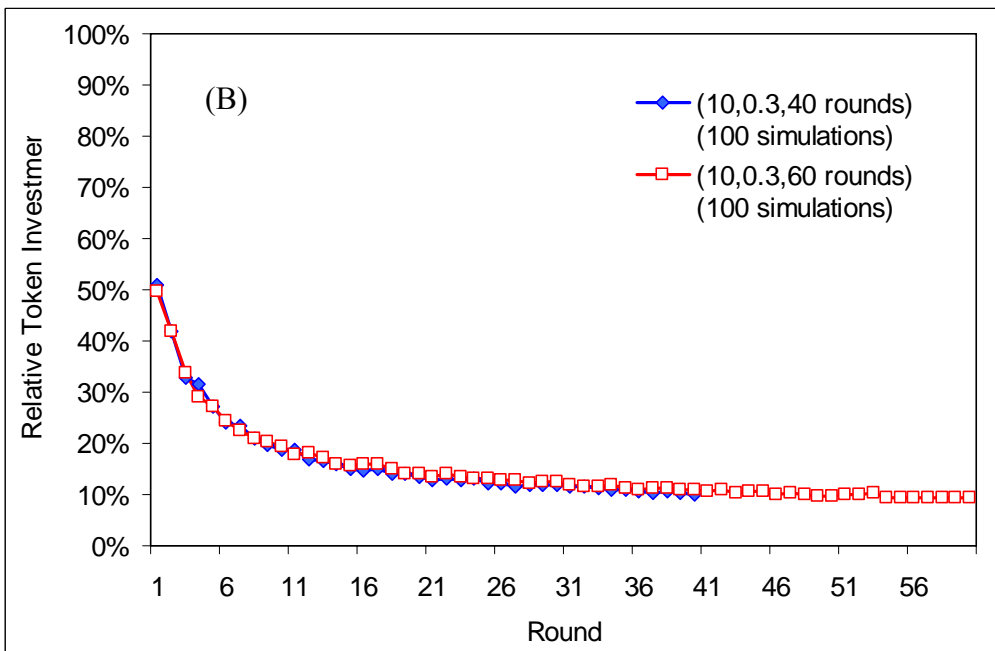
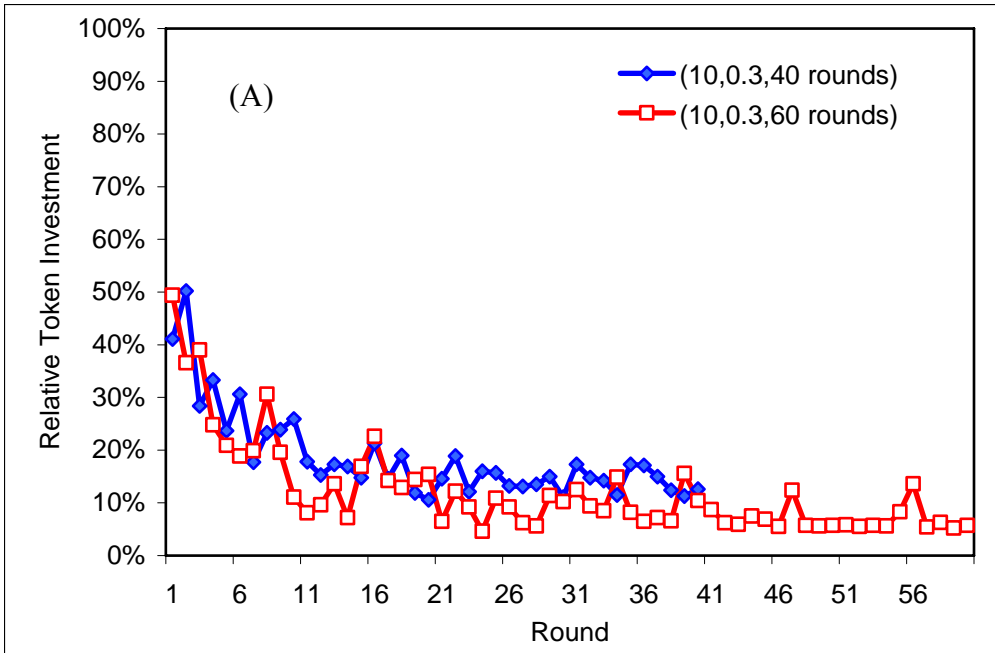


Figure A2: Relative token investments when we use the parameters of estimation 3 (Table 3) and simulate the token investments for one experiment of 40 and 60 rounds (A) or 100 experiments of 40 and 60 rounds (B).

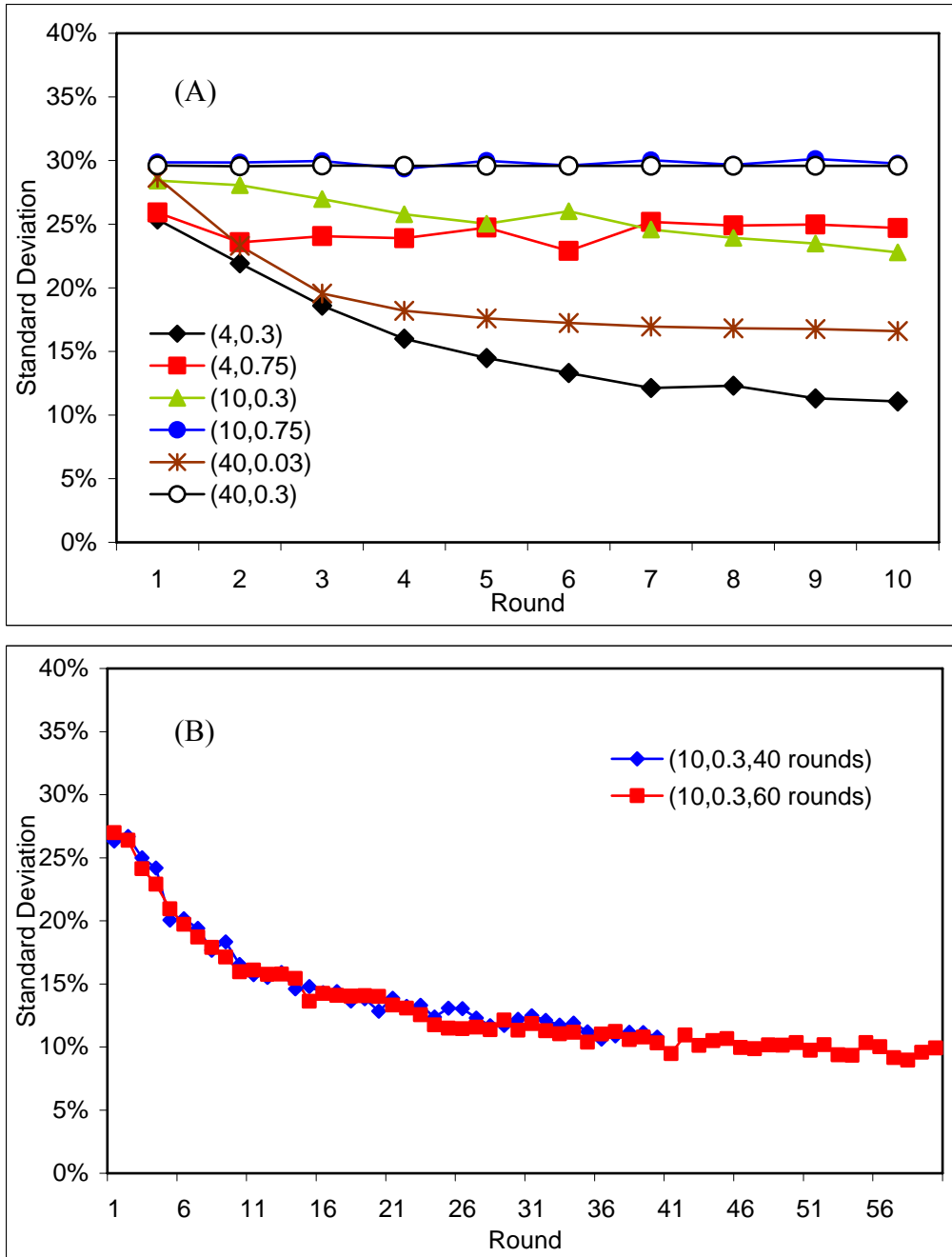


Figure A3: Relative standard deviation of token investments when we use the parameters of estimation 3 (Table 3) and simulate the token investments for 100 experiments of 10 rounds (A) and 40/60 rounds (B).

The next question is whether the models estimated for the 10 round experiments are able to generate the macro-level statistics of the 40 and 60 round experiments, and the other way around. The reason for doing this is to test the sensitivity of the data that is used. If a model is claimed to be general it should explain both data sets. Figure 10 show the results for the average of 100 experiments. It shows that the macro-level statistics are

comparable to the generated statistics for the original models. Thus the cross-validation test of the model gives satisfying results.

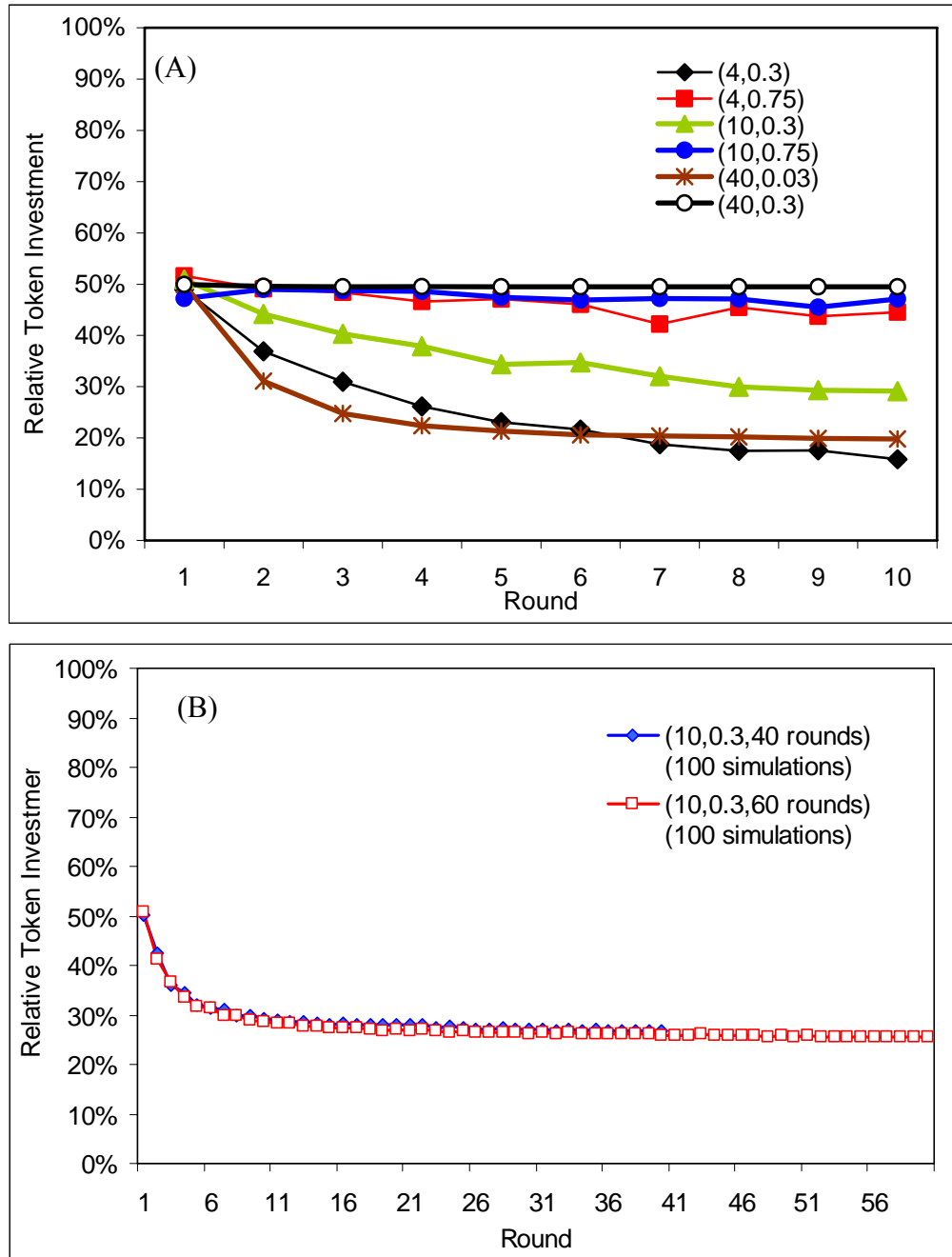


Figure A4: Relative token investments average over 100 runs for the 10 round experiments, using the estimated model for the 40/60 round experiments (A), and for the 40/60 round experiments, using the estimated model for the 10 round experiments (B).

We can compare the cross-validation results also on its scores of the maximum likelihood, the AIC and the BIC (Table A1). These scores confirm that the cross-validation is satisfying. We have done the same for the models that distinguish different types of agents. We can compare this with a simple one-parameter model, as we discuss

later, and the maximum likelihood of the cross-validated models performs much better than the null model.

Table A1: Comparison of different models. We compare the maximum likelihood statistic, the AIC and the BIC for different models on the data of the 10 round and 40/60 round experiments.

		Maximum Likelihood		AIC		BIC	
		10 rounds	40/60 rounds	10 rounds	40/60 rounds	10 rounds	40/60 rounds
Model	10 round ML	-5495.1	-4980.75	11006.2	9945.5	11052.7	9903.5
	40/60 round ML	-5626.1	-4730.4	11236.2	9476.8	11189.7	9518.7
	8 types (10)	-4780.5	-5315.5	9688.9	10551.1	10061.1	10341.3
	2 types (40/60)	-5499.2	-4621.0	10966.4	9274.0	10873.3	9358.0
	One-parameter	-7345.9	-6188.6	14693.8	12379.2	14699.0	12384.4

Another statistic to evaluate the performance of the models is the change of relative token investment between the rounds. The data provide some interesting distributions. Figure A4a shows that the estimated models perform well for the 10 round experiments. However, Figure A4b shows that the estimated models do not provide the increased frequency of big changes in the 40/60 round experiments, except the model that assumes parameter distributions (as given in Table 7).

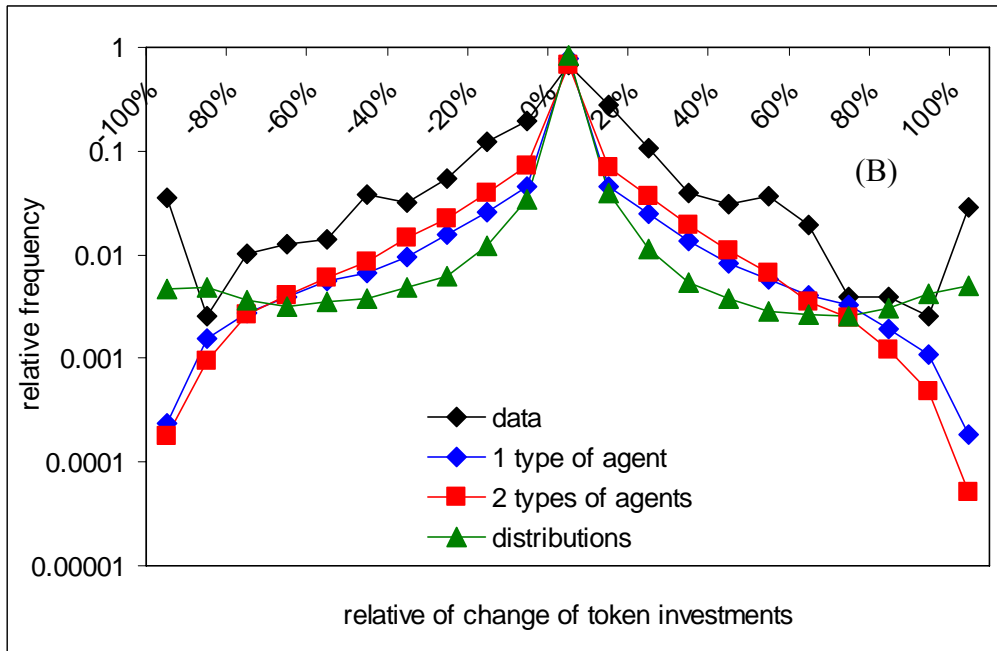
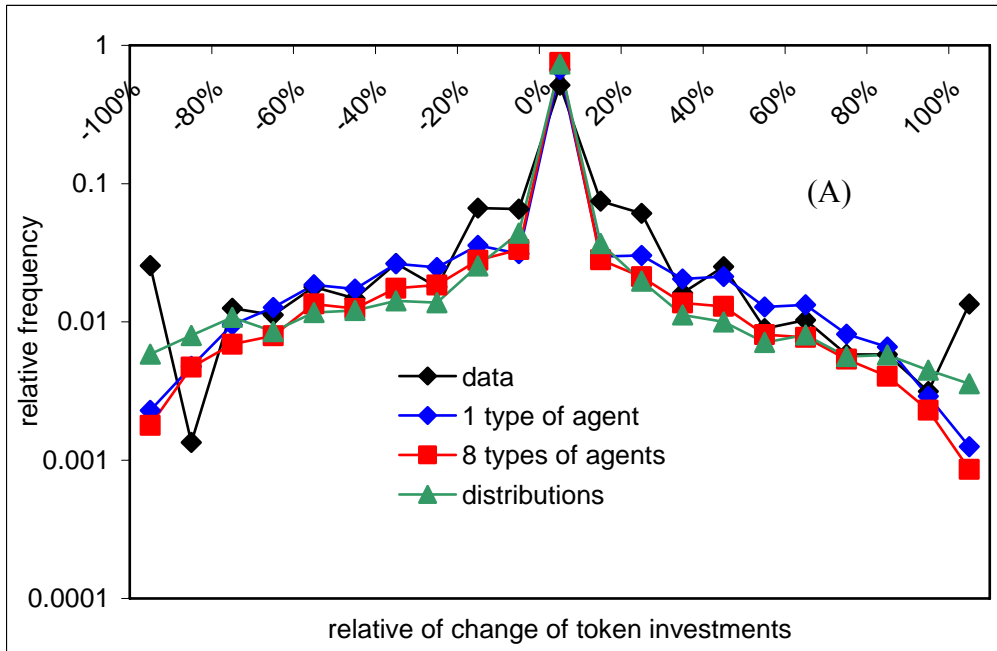


Figure A5: Relative token investments average over 100 runs for the 10 round experiments, using the estimated model for the 40/60 round experiments (A), and for the 40/60 round experiments, using the estimated model for the 10 round experiments (B). We show the best estimates of multiple type estimations. Distributions refer to model runs where the distributions from Table 7 are used.

### *Minimum description length*

In our evaluation of the various versions of the model, we took into account the number of parameters and the number of observations. We did not include the complexity of the model in terms of functional forms. In recent years cognitive scientists have start including model complexity into model selection methodology (Pitt and Myung, 2002; Pitt et al., 2002). The model evaluation criterion minimum description length favors models that describe the same complexity with less complicated models. This methodology has been successfully applied to a number of decision problems (Pitt et al., 2002). The Minimum Description Length criterium is formulated as

$$MDL = -\ln L + \frac{k}{2} \ln\left(\frac{N}{2\pi}\right) + \ln \int dp \sqrt{\det[I(p)]}$$

Where  $p$  is the vector of parameter values,  $I(p)$  is the Fisher Information matrix, and  $\det[I(p)]$  the determinant of the Fisher Information matrix. The integral is over the space of eligible parameter values. When we calculate the MDL for the four cases of Table 3, then we have for model SP: 7057, for model SP+L: 5549, for model SP+L+S: 5553, and for model L: 5957. Thus according to the MDL criterion, model 2 is the most appropriate.

### *One parameter model*

Despite the fact that Minimum Description Length identifies a 6 parameter model as the most appropriate, we want to investigate the performance of a simple one parameter version of our original model. There are a specific of reason for this. When scholars want to apply empirically based agent-based models, they may want to include the simplest possible model that represents the data. We will show that a one-parameter model is able to reproduce a large part of the observations. We assume agents are selfish ( $\rho=0$  and  $\chi=0$ ), and do not signal ( $\theta=0$ ,  $\eta=0$ ). Further more we assume that the agents are belief learners ( $\lambda=1$ ,  $\delta=1$ ,  $\kappa=1$ ), which leads to the following equation of the attraction:

$$A_i^x(t) = A_i^x(t-1) + \pi_i$$

where  $\pi_i$  is the monetary payoff that player  $i$  obtained in round  $t-1$ .

The only parameter we use to calibrate the model is the response sensitivity  $\phi$  (0.029 for 10 rounds and 0.003 for 40/60 rounds dataset). In Table A1 we see that this simple model do much worse than the original model in terms of maximum likelihood. If we look at the macro-level statistics, we see that the most of the treatments provide a reasonable good aggregate fit with the data (Figure A6), in one case even better than the full model (40, 0.03).

With regard to the standard deviation, we see that the average standard deviation of the simple model performs better than the full model (Figure A7). Especially with the 40 and 60 round experiments, the standard deviation and the average investments stay longer at a higher level, as the observations also indicate.

The relative change of token investments between rounds shows an interesting difference between the data and full model, and the simple one parameter model (Figures A8). The one parameter model is not able to generate observed patterns. The one parameter model is more sensitive and agents change their token investments more frequently than observed.

This analysis shows the problem of model evaluation. Using maximum likelihood estimation we use individual level information, but do not test the resulting macro-level phenomena. Using micro-level parameterization to generate the macro-level statistics does not necessary better results that a very simple model that does worse on the micro-level. We need to develop model estimation methods that simultaneously evaluate micro-level differences as well as macro-level differences.

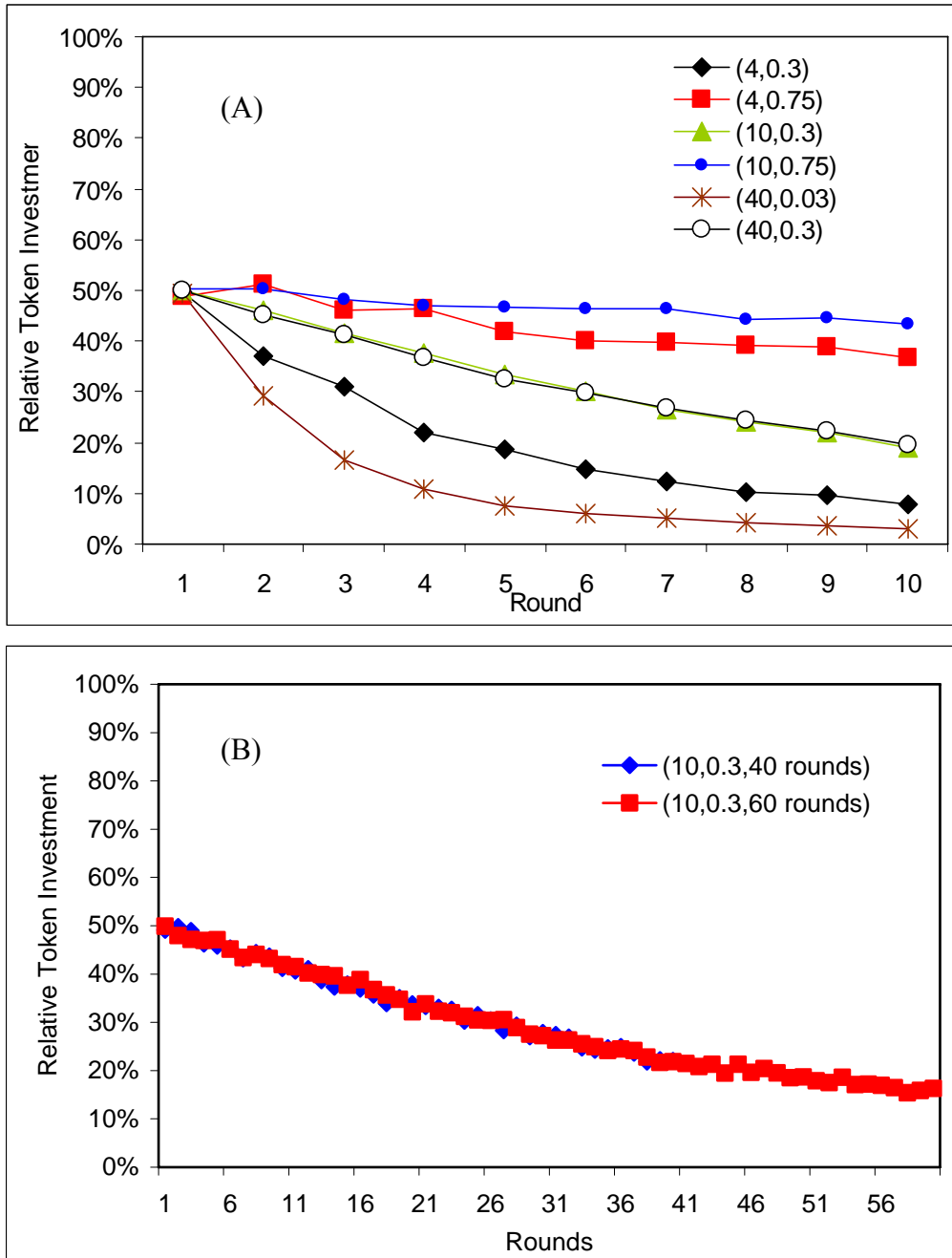


Figure A6: Relative token investments when we use the simple one-parameter model and simulate the token investments for 100 experiments of 10 rounds (A) or 40 and 60 rounds (B).

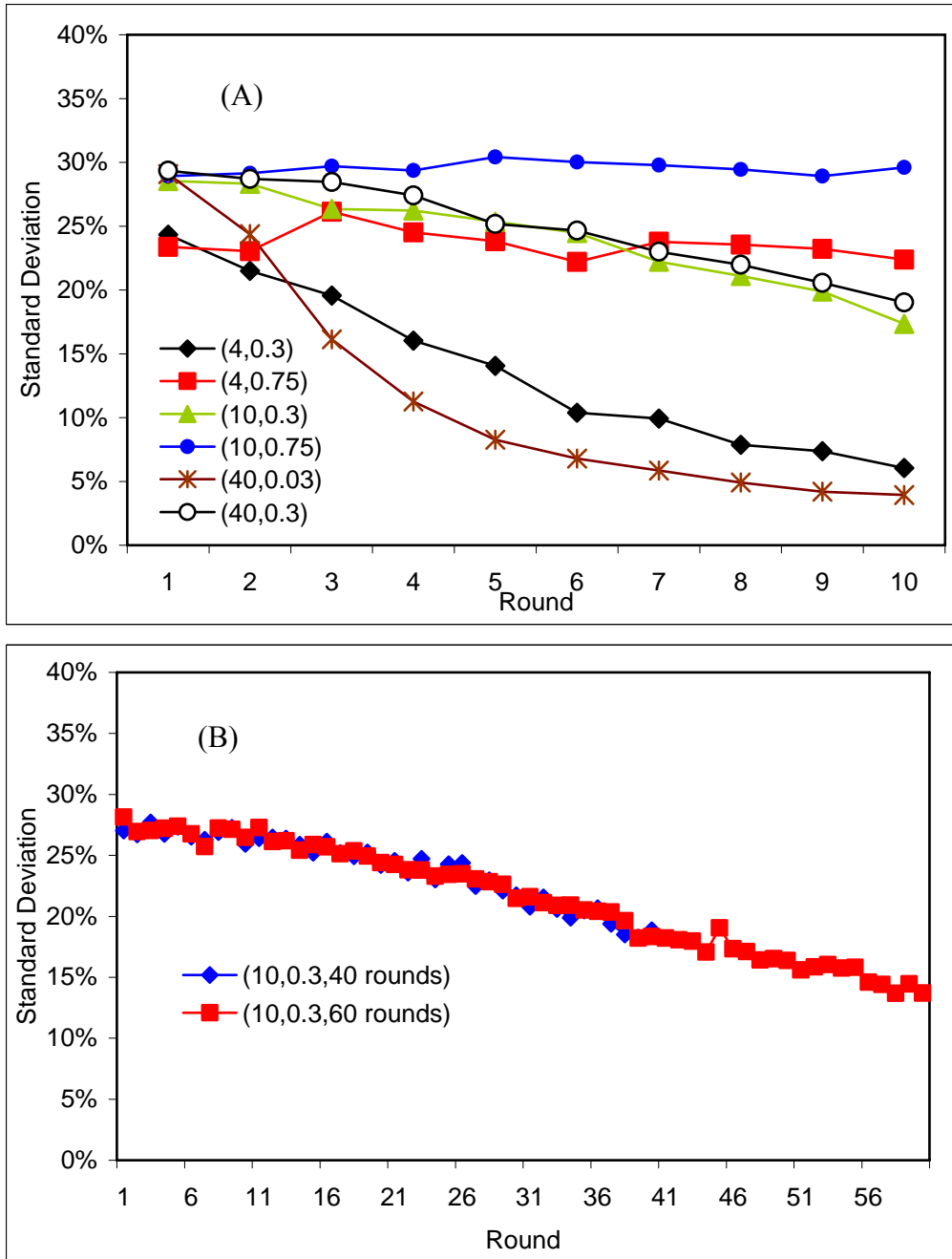


Figure A7: Relative standard deviation of token investments when we use the simple one parameter model and simulate the token investments for 100 experiments of 10 rounds (A) and 40 to 60 rounds (B).

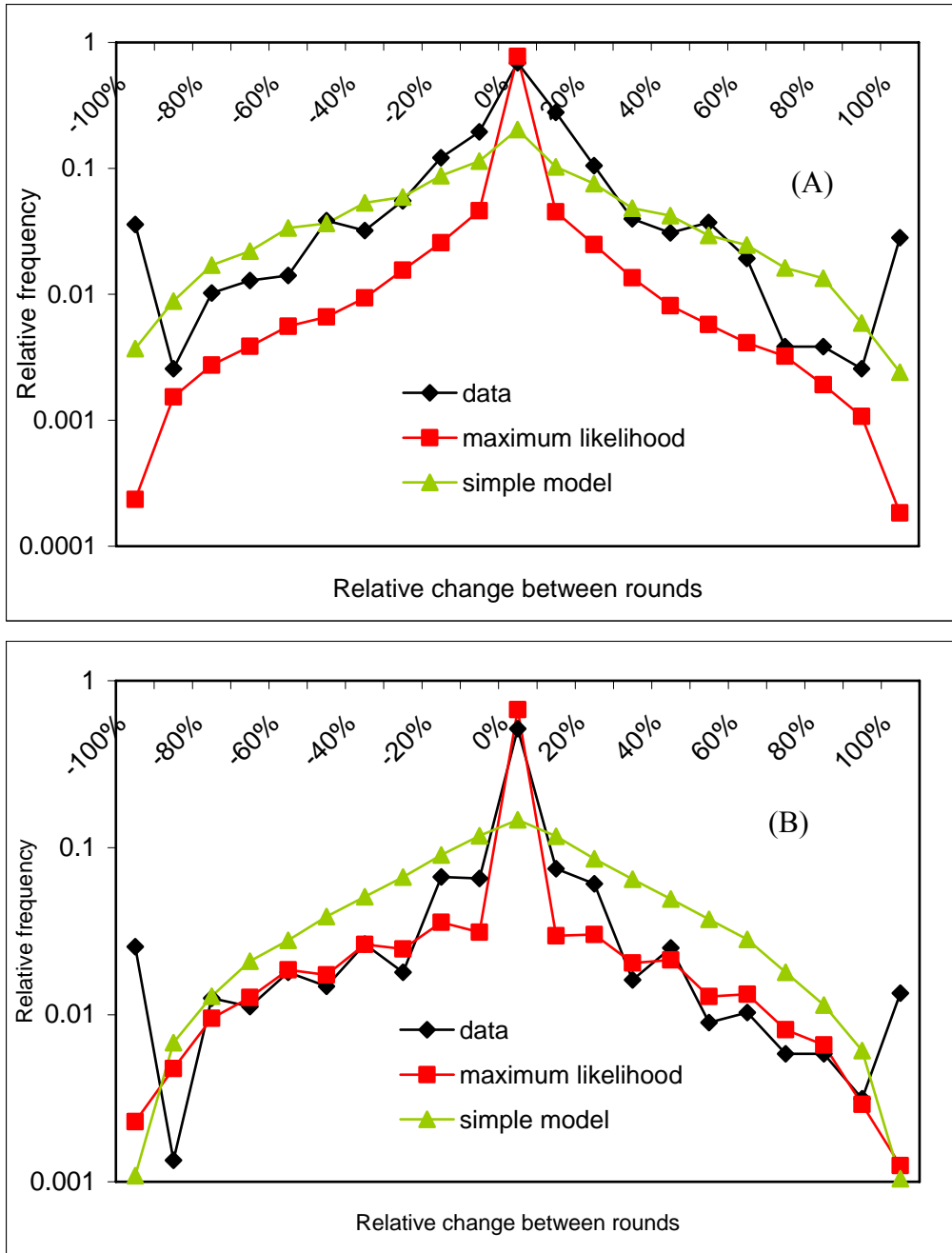


Figure A8: Frequency of change of token investments between rounds for 10 round experiments (A) and for 40/60 round experiments (B). The model simulations include 100 simulated experiments per treatment.