Department of Economics

On the Behavior of Proposers in Ultimatum Games

Thomas Brenner and Nicolaas J. Vriend

Working Paper No. 502 October 2003 ISSN 1473-0278

On the Behavior of Proposers in Ultimatum Games [*]

Thomas Brenner, *Max Planck Institute*, Jena, Germany
Nicolaas J. Vriend, *Queen Mary, University of London*, UK

September 2003

Abstract:

We demonstrate that one should *not* expect convergence of the proposals to the subgame perfect Nash equilibrium offer in standard ultimatum games. First, imposing strict experimental control of the behavior of the receiving players and focusing on the behavior of the proposers, we show experimentally that proposers do not learn to make the expected-payoff-maximizing offer. Second, considering a range of learning theories (from optimal to boundedly rational), we explain that this is an inherent feature of the learning task faced by the proposers, and we provide some insights into the actual learning behavior of the experimental subjects. This explanation for the lack of convergence to the subgame perfect Nash equilibrium in ultimatum games complements most alternative explanations.

J.E.L. classification codes: C72, C91, D81, D83

Keywords: Ultimatum game, Non-equilibrium behavior, Laboratory experiment, Multi-armed bandit, Optimal learning, Gittins index, Bounded rationality

---

corresponding author: N.J. Vriend, Queen Mary, University of London, Department of Economics, Mile End Road, London E1 4NS, UK, <n.vriend@qmul.ac.uk>, www.qmul.ac.uk/~ugte173/.

[*] We thank Antonio Cabrales, Ido Erev, Steffen Huck, Jon Leland, Paolo Patelli, Robin Pope, Giulio Spelanzon, and Massimo Warglien for helpful comments and discussions. The usual disclaimer applies.

1. Introduction

One of the games most extensively studied in the literature in recent years is the ultimatum game. The reason that this game is so intriguing seems to be that the game-theoretic analysis is straightforward and simple, while the overwhelming experimental evidence is equally straightforward but at odds with the game-theoretic analysis (see, e.g., Güth et al. [1982], Güth & Tietz [1990], or Thaler [1988]).

In the basic ultimatum game there are two players and a pie. Player A proposes how to split the pie between herself and player B. Upon receiving player A's proposal, player B has two options. First, to accept the proposal, which will then be carried out. Second, to reject it, after which both get nothing. Many variants of this basic setup have been considered in the literature. There are many Nash equilibria in this game. Every strategy for player A combined with any strategy for player B that accepts that offer but rejects all lower offers is one. But there is a unique subgame perfect equilibrium: player A offers the minimal piece, and player B accepts that.[1]

Empirical evidence shows time and again that this is not what happens in the laboratory. Players A usually offer somewhat less than half the pie to players B, and players B usually reject small offers. Concerning player A's behavior, there are two main explanations for this anomaly offered in the literature. First, some argued that fairness and reciprocity considerations are the force driving players A to offer more than the standard game-theoretic analysis would suggest (see, e.g., Forsythe et al. [1994]). An alternative explanation found in the literature is that players A are basically following an adaptive, best-reply seeking approach to the behavior of players B. In a multi-period setup where players played the game repeatedly but each time against different players, some papers showed how it can happen that players A `unlearn' to play the subgame perfect equilibrium strategy as players B have not learned yet that they should play their perfect equilibrium strategy. Once players A do not play that strategy anymore, players B will never learn to play theirs. Such learning dynamics are shown in Roth & Erev [1995] who follow a reinforcement learning approach, and Gale et al. [1995] who use replicator dynamics.

Both explanations are somehow based on the assumption that players A learn to play best-replies to the behavior of players B. The deviation from the predictions of subgame

[1] Strictly speaking, in case it is a discrete choice problem including zero, there are two subgame perfect equilibria, with player A offering either zero or the smallest possible strictly positive piece to player B.

perfect Nash equilibrium is, therefore, mainly explained by deviating behavior of players B. Players A only react on this deviation which is caused by a slow learning process or fairness considerations. In this paper, in contrast, we will show that the adaptive behavior of players A as such may also cause deviations from the subgame perfect Nash equilibrium, keeping players A away from the optimal minimal offer independent from the adaptive behavior of players B.[2]

In order to focus on the behavior of players A, we design an ultimatum game experiment in which the behavior of a large population of players B is fixed by some computer algorithm. This was known to the players. Our experimental design has two advantages. First, as there are no payoffs to other people influenced by the behavior of players A, fairness considerations cannot play a role. Second, learning in ultimatum games is essentially a coevolutionary process. Players learn about the behavior of other players who learn about the behavior of other players who learn .... Our experimental design allows to focus on the learning behavior of players A, abstracting from the complications and peculiarities related to coevolutionary processes.[3]

Basically, the problem faced by a player A is a multi-armed bandit problem. There are three treatments that differ in the general level of acceptance rates. The experimental parameters in each treatment are such that two monotonicity properties are satisfied: higher offers are more likely to be accepted, whereas lower offers are giving higher expected payoffs to the proposer.

Two stylized facts stand out in the experimental data. First, although the experiment comprises 100 periods, there is only little tendency for the average offer to come down to the minimal offer, the one that maximizes a proposer's expected payoff. Second, although the incentive structure of the proposers is the same in each treatment, higher general acceptance rates lead to significantly lower offers.

We consider a range of learning theories, from optimal learning (based on the Gittins index) to more boundedly rational learning methods. We show that both stylized facts can be

---

[2] Vriend [1997] presented some theoretical considerations why paying more attention to the behavior of the proposing players as such could be worthwhile.

[3] Our approach is similar in spirit to a dictator game (see, e.g., Bolton et al. [1998]), in the sense that dictator games were also invented to cut out players B. But there are two advantages of our setup. First, in a dictator game players A know the behavior of players B (accepting anything), whereas this is not the case in our setup. Second, in a dictator game fairness considerations still play a role.

explained by these learning theories. In other words, having imposed strict experimental control on the behavior of the receiving players, we show that one should *not* expect convergence to the subgame perfect Nash equilibrium in standard ultimatum games. One of the main contributions of our paper, then, is that this offers an explanation for the lack of convergence to the subgame perfect Nash equilibrium in ultimatum games, an explanation that complements most existing explanations.

The rest of this paper is organized as follows. In section 2 we present the experimental design, and in section 3 the experimental results. Various learning theories to explain the experimental data are discussed in section 4, and their predicted dynamics are examined in section 5. Section 6 concludes.

2. Experimental design

The underlying idea for our design is the following. First, we wanted to set up a stylized ultimatum game in which the optimal strategy for players A would coincide with the subgame perfect equilibrium strategy of the standard ultimatum game of offering only a minimal slice. Second, as we wanted to focus on players A's learning behavior, we wanted to be in a position to exclude as much as possible other well-known explanations for players A staying away from the optimal action. In particular, this implies that we needed to be in a position to abstract from the learning behavior of the receiving players B.

We play an ultimatum game in which the pie has size 9, and we allow only integers from 1 to 9 to be chosen as offers (see also Roth & Erev [1995]). The experimental subjects are players A, who play against players B who form a large population of artificial agents, making their decisions using some computer algorithm. Every period a given player A is randomly matched with a player B that he has not met yet. Player A enters his offer, and then the reply of player B and the corresponding payoff for player A are shown. There are 100 periods to be played. Notice that this is more periods than in experimental ultimatum games typically reported in the literature. Figure 1 shows a player's screen during the experiment. A player could at any moment during the experiment scroll through his complete history. The identity of players B is listed on the screen to make clear that every period the opponent is a different player. Each period, after the choice of player A, it takes 5 to 15 seconds (uniform randomly chosen) before the reply of player B is listed (saying "please wait for reply player B"). This suggests players B make serious choices, and it avoids players A getting rushed

too much by the speed of players B.

ROUND 2

Your opponent is player B with id.: 649,021

Please choose your offer to player B: . . . . and press Enter

HISTORY

period	id. player B	your offer	reply player B	your payoff
1	231,896	2	accept	7

Figure 1. Sample interface during experiment

Using artificial players B allows us to control the environment for players A. Players A is told that players B are artificial players. Given that the players B are artificial players, altruism considerations are irrelevant. We told players A: *"Each of those players B's behavior is systematic in the following sense: If a specific player B has accepted an offer x then that player B would have accepted as well any offer greater than x. And if offer x had been rejected by that player B, so would have been all offers smaller than x. Of course, different players B might have different opinions about which offers are acceptable or not. The players B do not change their behavior over time"* (see instructions in the appendix). As a result, the population of players B can be characterized by a probability density function that a given offer will be accepted. The probability that a given offer is accepted is monotonically increasing in the size of the offer.[4] We organize three treatments, that differ in the general

[4] Hence, in our experiment there is a heterogeneous population of players B who play pure strategies characterized by a reservation value property. That is, every player B has a reservation value, but different players may have different values. An alternative possible interpretation concerning the population of players B is that the population consists of identical players who use a mixed strategy characterized by probabilities of accepting offers that are monotonically increasing in the size of the offer.

level of acceptance probabilities. The probabilities that a randomly chosen player B will accept a given offer in each treatment is listed in Table 1.

	offers								
treatment	1	2	3	4	5	6	7	8	9
ult3	0.30	0.31	0.32	0.35	0.39	0.47	0.64	1.00	1.00
ult5	0.50	0.51	0.54	0.58	0.66	0.79	1.00	1.00	1.00
ult7	0.70	0.72	0.76	0.82	0.92	1.00	1.00	1.00	1.00

Table 1. Acceptance probabilities

These probabilities are based on the following considerations. First, the expected payoff maximizing offer is 1, which coincides with the subgame perfect equilibrium strategy of a standard ultimatum game. Second, the acceptance probabilities increase monotonically with the size of the offer, thus maintaining realistic assumptions concerning the behavior of players B. Third, given the two considerations already mentioned, we wanted to make the learning task as easy as possible. Therefore, the minimal offer of 1 gives an expected payoff that is clearly higher than for any other possible offer. Moreover, there are no local optima in the strategy space of players A. This avoids that many adaptive modes of behavior can be locked in too easily at sub-optimal peaks in the range of possible offers. Starting from the minimal offer of 1, in each treatment each next higher offer gives an expected payoff that is at least 10% lower. Figure 2 shows these resulting expected payoffs for each treatment.

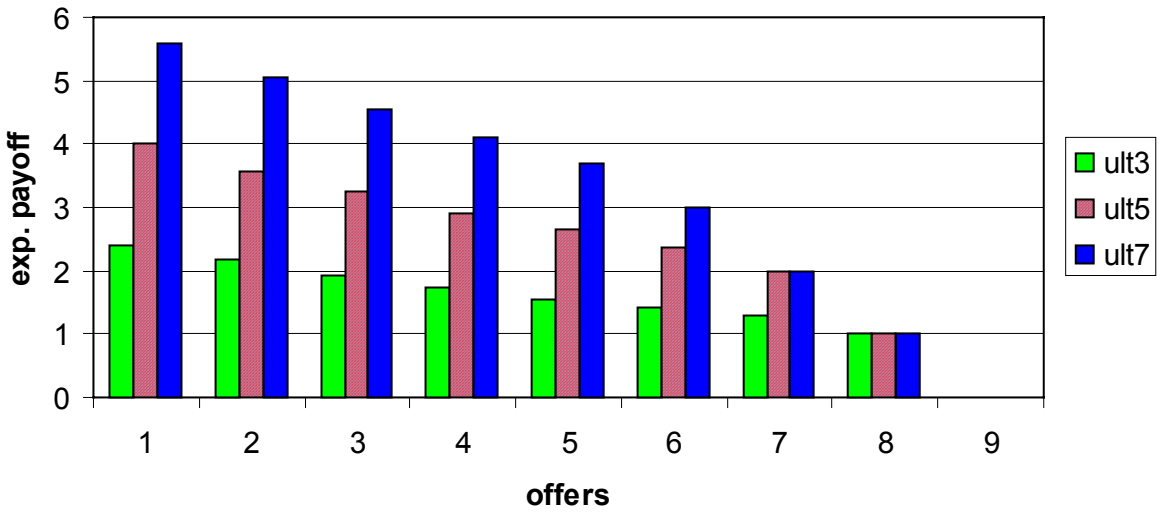

Figure 2. Expected payoffs across treatments

The experiments were conducted in the computerized experimental laboratory CEEL of the University of Trento in November 1997. With only a few exceptions, all players for a given treatment were simultaneously in the laboratory. The players, 19 for the ult3 treatment and 20 for the other treatments, went through the experiment in about an hour. The exchange rates (Italian Lires per point) were 83.3 (for ult3), 50.0 (ult5), and 35.7 (ult7). These exchange rates were chosen such that the monetary incentives were essentially identical in each treatment. The only differences are due to rounding, and to the fact that the acceptance probabilities cannot exceed 1, which is of relevance only for some of the highest offers in some treatments. The average monetary reward was just over Lit. 15,000 per player (about US $ 8.80 at the time).

3. Experimental data

Figure 3 presents the time-series of the offers averaged over the players in each of the three treatments.

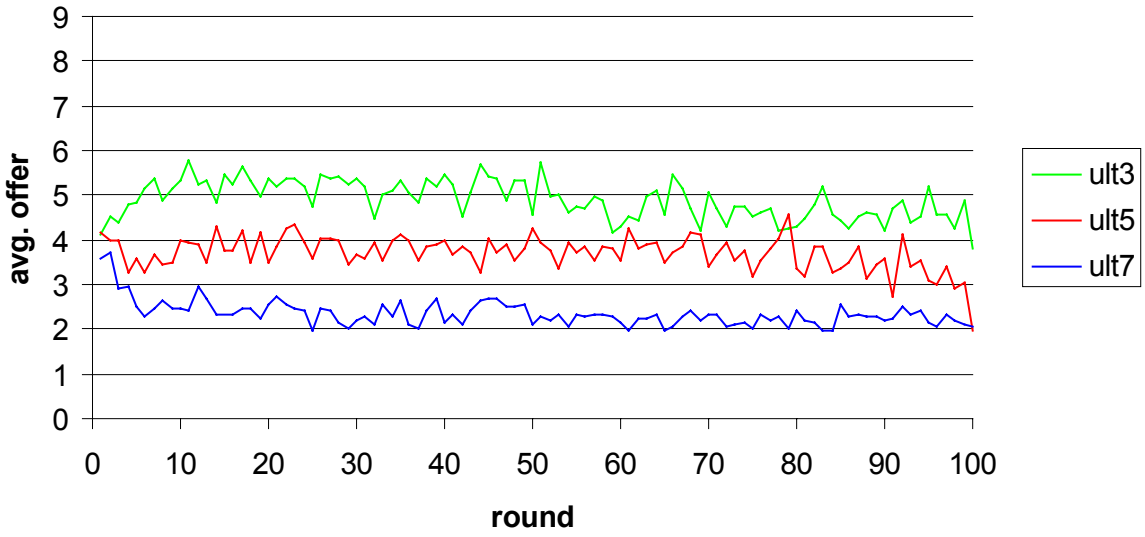

Figure 3. Average offers in ultimatum game experiment

We observe the following. First, in each treatment the average initial offer is about 4, slightly below the 50% offer. Second, from period 10 to 90, the average offers in each treatment decline significantly but slowly,[5] and in none of the treatments do the players learn to make the optimal offer of 1 (although there is a downward end effect, in particular in the treatments ult3 and ult5).[6] Third, although the monetary incentives were the same in each treatment, there is a systematic difference across treatments, with higher acceptance probabilities leading to lower offers.[7] These differences emerge in particular during the first 10 periods.

[5] A simple regression against time gives coefficients of -0.012, -0.004 and -0.004 for the ult3, ult5 and ult7 treatments respectively (all significant at 0.005; 1-sided). Extrapolation of the observed rates implies that it would take hundreds of periods more in each treatment for the average offer to reach 1.

[6] Such end effects are relatively common in experiments.

[7] A Wilcoxon-Mann-Whitney test, based on the average offer for each individual player over the 100 periods of the experiment, shows that the players offer significantly more in the ult3 than the players in the ult5 treatment, who in turn offer significantly more than those in the ult7 treatment (at 0.005 significance level; 1-sided).

4. Some learning theories

In this section we present a number of learning theories that might help to explain the experimental observations.

4.1 Modeling rational learning

As explained above, in our experimental design we simplified the standard ultimatum game into a one-person decision problem. This decision problem in our experiment resembles a multi-armed bandit. For many specific kinds of repeated multi-armed bandit situations optimal behaviors are given in the literature (see, e.g., Gittins [1989], Bergemann & Välimäki [2001] and Brezzi & Lai [2002]). One standard multi-armed bandit situation is characterized by people knowing the value of the payoffs, π_i, that they might receive with each of the arms *i*. They also know in this situation that they receive each payoff with a fixed but unknown probability p_i. If, in addition, the probabilities p_i are independent of each other, the *Gittins index* can be used to determine the optimal behavior.

The *Gittins index* has been introduced by Gittins [1979, 1989]. It assigns at each time to each arm the maximal average payoff that can be obtained by repeatedly choosing that arm for an event-dependent number of times. Each time an arm is chosen, it is determined randomly, according to the probability assigned to the arm, whether a payoff is obtained or not. After each choice it is decided, depending on the experience with this arm, whether the arm is chosen again or not. How this decision is made is called the *stopping rule.* Many different stopping rules can be imagined. If we consider the time span from the actual time *t* until the time at which the choice is changed because of the stopping rule, the average payoff that is received within this time span can be calculated. This average payoff depends on the stopping rule. The Gittins index is defined as the maximal average payoff that can be reached by any stopping rule. We denote the Gittins index for arm *i* at time *t* by $g_i(t)$. To calculate the average payoffs for each stopping rule it is necessary to calculate the probability for obtaining a payoff at each time. Since the real probabilities are not known, Bayesian updating is used for calculating these expected probabilities. Hence, the concept of Gittins indices is based on two basic features. First, Bayesian updating is used to calculate the expected probabilities of payoffs. Second, average expected payoffs are calculated for each arm separately according to the stopping rule approach. Gittins has

proved that choosing at each time, t, the arm, i, with the highest Gittins index, $g_i(t)$, is the optimal strategy for the situation described above.

Strictly speaking, the use of the Gittins index is not appropriate for the situation faced by the players in our experiment. The reason is that the arms are not independent. The information given to the players implied that the probability of acceptance was weakly increasing in the size of the offer. This requires two modifications of the standard Gittins index approach.

First, when updating the probability of acceptance for some arm i in the context of our ultimatum game, a player should also update the probabilities for all other arms, as he knows that $p_i \leq p_j$ for each $i < j$. As a consequence the hypotheses in Bayesian learning have to be formulated for all arms jointly. Hence, a hypothesis is characterised by nine probabilities: p_1, p_2, …, p_9, the probabilities assigned to each of the nine arms. The number of possible hypothesis is reduced by the condition $p_1 \leq p_2 \leq p_3 \leq p_4 \leq p_5 \leq p_6 \leq p_7 \leq p_8 \leq p_9$. Therefore, the set of all feasible hypothesis is given by

$$H = \{(p_1, p_2, ..., p_9) \mid p_1 \leq p_2 \leq p_3 \leq p_4 \leq p_5 \leq p_6 \leq p_7 \leq p_8 \leq p_9\}.$$

The probability $P(h,t)$ that is assigned to each hypothesis, $h \in H$, at each time, t, is updated according to Bayes' rule. The expected probability to obtain a payoff if choosing arm i is given by

$$E_i(t) = \sum_{h \in H} P(h,t) \cdot p_i(h)$$

or

$$E_i(t) = \int_0^1 \int_0^{p_9} \int_0^{p_8} \int_0^{p_7} \int_0^{p_6} \int_0^{p_5} \int_0^{p_4} \int_0^{p_3} \int_0^{p_2} P(p_1, p_2, ..., p_9, t)\, p_i\, dp_1 dp_2 dp_3 dp_4 dp_5 dp_6 dp_7 dp_8 dp_9 \,. \qquad (*)$$

The expected probabilities defined by Equation (*) are used in the calculation of the Gittins indices.[8] Hence, the Gittins indices are based on expected probabilities that are calculated using *all* the available information.

Second, when choosing an arm, a player should not simply try the arm with the highest Gittins index because he should take into account as well that the outcome with that arm will provide useful information about the other arms (as, according to Equation (*),

[8] This integral can be calculated in closed form for each possible history. However, for reasons of convenience we will do this numerically.

the expected payoffs of other arms change when using a given arm). No optimal strategy is known in the literature taking this second point into account.[9]

Therefore, we will compute adjusted Gittins indices (taking the first modification into account), and then let players simply choose the arm with the highest Gittins index. Hence, we obtain an approximation of optimal behavior. This approximation deviates from optimal behavior only by the fact that it assumes the expectation for the average payoffs of all other arms to remain constant while repeatedly choosing one arm. Since this deviation is similar for all arms, the ranking of the Gittins indices should be little influenced by this approximation.

The Gittins indices are calculated as described in the literature (see Gittins [1979, 1989]). At each time, *t*, for each arm, *i*, the stopping rule is calculated that leads to the highest average payoff for repeatedly choosing this arm. This is done through backward induction on all possible sequences of outcomes for repeatedly choosing arm *i*. Whenever continuing the choice of arm *i* decreases the average payoff, the choice is stopped. The expected probability $E_i(t)$ of arm *i* to lead to a positive payoff is calculated at each time, *t*, according to Equation (*). The probability $P(h,t)$ for each hypothesis, *h*, is updated according to Bayes' rule. This means that

$$P(p_1, p_2, ..., p_9, t+1) = \frac{p_1 \cdot P(p_1, p_2, ..., p_9, t)}{\sum_{h \in H} p_i \cdot P(p_1, p_2, ..., p_9, t)}$$

holds if arm *i* is chosen at time *t* and a positive outcome results and that

$$P(p_1, p_2, ..., p_9, t+1) = \frac{(1-p_1) \cdot P(p_1, p_2, ..., p_9, t)}{\sum_{h \in H} (1-p_i) \cdot P(p_1, p_2, ..., p_9, t)}$$

holds if arm *i* is chosen at time *t* and an outcome of zero results. The initial prior is that each hypothesis, *h*, is equally likely.

4.2 Modeling boundedly rational learning

To put the adjusted Gittins index approach further into perspective, we now describe two models of boundedly rational learning.

With reinforcement learning we assume that players A have no understanding of

[9] To calculate optimal behavior in such a situation requires to examine all possible sequences of actions, their potential outcomes, and the respective probabilities, which is computationally not feasible.

game theory, the structure of the ultimatum game, or the behavior of players B. Players A are boundedly rational agents who behave adaptively to their environment. They simply try actions, and are in the future more likely to choose those offers that had been more reinforced (through higher payoffs) in the past. One can imagine players A playing with a multi-armed bandit, where different arms might give different payoffs, and players A do not know at the start which is the best arm to pull. The basic reinforcement learning model is as follows (see also Roth & Erev [1995]): At time t=1 each player has an initial propensity to choose his ith arm given by some real number $q_i(1)$. We assume $q_j(1) = q(1)$ for each j, and $\Sigma q_j(1) = 10$ (following Roth & Erev [1995]). If a player plays arm i at time t, and receives a payoff of z, then the propensity to choose arm i is updated by setting $q_i(t+1)=q_i(t)+z$, while for all other arms j, $q_j(t+1)=q_j(t)$. The probability that the player selects his ith arm at time t is $p_i(t) = q_i(t)/\Sigma q_j(t)$, where the sum is over all the available arms j. Thus, given the reinforcements for all offers, a player chooses (with some experimentation) his most reinforced offer. Notice that in the reinforcement learning model players ignore the interdependence of the arms.

Players A who behave according to learning direction theory (see, e.g., Selten & Stoecker [1986]) look at the outcome of the most recent period, and reason in which direction a better offer could have been found. They, then, simply adjust their current offer into that direction. More specifically, if a player A found an offer i was rejected at time t, then at time t+1 he will offer i+1 to player B (unless offer i equaled the maximal possible offer). If, on the other hand, offer i was accepted at time t, then at time t+1 he will offer i-1 to player B (unless offer i equaled the minimal possible offer). Notice that a player learning according to learning direction theory can be seen as seeking myopically for a best-response against his latest opponent, and that implicitly he takes the interdependence of the arms into account.

5. Predicted dynamics for learning theories

We first consider the predicted choices for the model of optimal learning based on the adjusted Gittins indices, adjusted to take into account the interdependence of the arms in our ultimatum game.

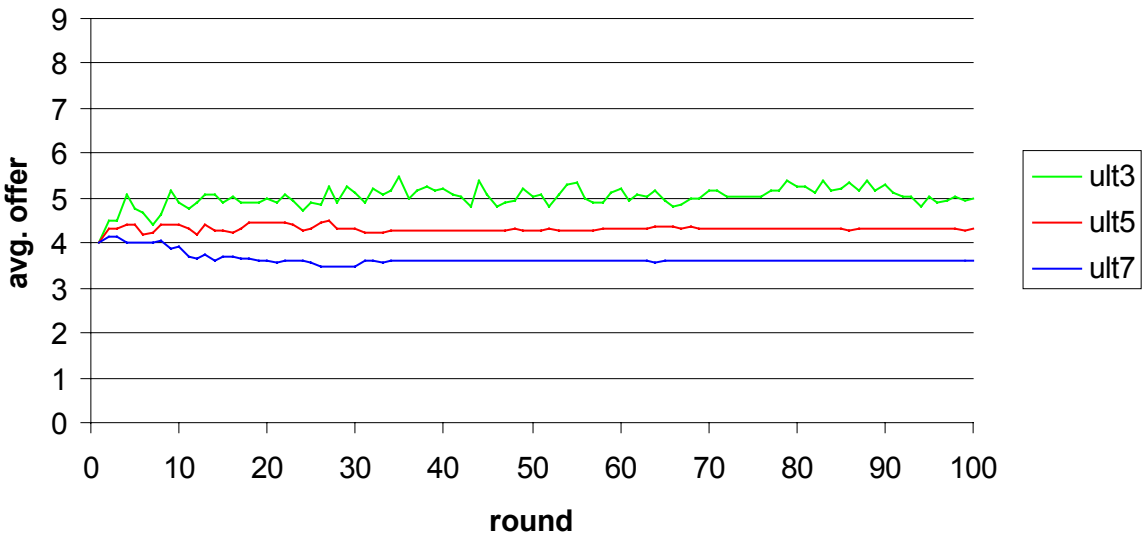

Figure 4. Theoretical prediction of the average behavior using adjusted Gittins indices

Figure 4 shows the average behavior of the model of optimal learning over 100 runs for the three different treatments. We make the following observations. First, the initial choice is 4. This is due to the fact that, according to the initial probabilities for the different hypotheses, the fourth arm has to the highest Gittins index in the first round. Second, differences between the treatments emerge early on, with the ult3 treatment showing an increase in average offers, the ult7 treatment more of a decline, and the ult5 treatment somewhere in between. Third, after the initial learning phase there is no further downward trend, and no convergence to the optimal minimal offer. The optimal choice (an offer of 1) is chosen with probability zero in all three treatments after 100 rounds, and the average offer only falls below 4 for treatment ult7.

We draw two conclusions from these observations. First, the situation faced by the players in our ultimatum game experiment is such that even with optimal learning no convergence to the optimal offer of 1 takes place within 100 rounds. Second, the predicted behavior of the model of optimal learning shows qualitative similarities with the actual experimental data. This concerns in particular, the initial choices, the early emergence of differences between the treatments, and the lack of convergence to the optimal offer.

There is, however, also some qualitative difference between the predicted behavior of the model of optimal learning and the experimental data. In the experimental data we

observe a weak downward trend, whereas the model of optimal learning does not show a trend at all. The reason for the lack of a downward trend towards the optimal offer with the model of optimal learning is a lack of experimentation by these players. Given the structure of the ultimatum game, each outcome provides not only information about the probability behind the chosen arm but also about the probabilities behind the other arms because the probabilities depend on each other. As a result, experimentation by actually trying other arms is relatively less attractive than in the case of independent arms, and the players tend to stick more to their choices. In the experimental data, we see that the players tend to experiment more than in the model of optimal learning, but not enough to learn to choose the optimal minimal offer.

To gain some further insights where we should expect learning to lead to in the ultimatum game experiment, we now consider the predicted behavior of a number of learning models that deviate from the optimal one.

The above analysis of the model of optimal learning has shown that players are unable to learn to make the optimal offer because exploration is not sufficiently attractive. This is caused by the interdependence of the probabilities of the arms. Furthermore, the experiment shows that people experiment more than predicted by the above modelling. Hence, it might be conjectured that perhaps people neglect the information about the relationship between the arms and actually treat them as *independent*. This leads to the standard Gittins index approach. That is, there is a separate set of hypotheses for each arm. The hypotheses for arm i are given by all possible probabilities p_i. The initial beliefs are that each hypothesis is equally likely. This leads to an initial prediction of 0.50 for each arm to lead to a positive outcome. The expected probability for each arm is determined only by the experience that has been made with this arm in the past. On the basis of the probabilities that result from Bayesian updating, the Gittins indices are calculated for each arm. Then, the arm which offers the highest Gittins index is chosen.

Figure 5 shows the average behavior for 100 runs of the standard Gittins index approach, neglecting the interdependence of the arms. We make the following observations. First, the initial offer made is 1. This is due to the unbiased priors, attaching a probability of acceptance of 0.50 to each arm. Second, the early learning effect leads to increased offers, with differences between the three treatments emerging. Third, this is followed by a weak downward trend, with only about half of the players making the optimal

minimal offer in the end (47%, 56% and 64% for the ult3, ult5 and ult7 treatments respectively), although they had all started there in the first round.[10]

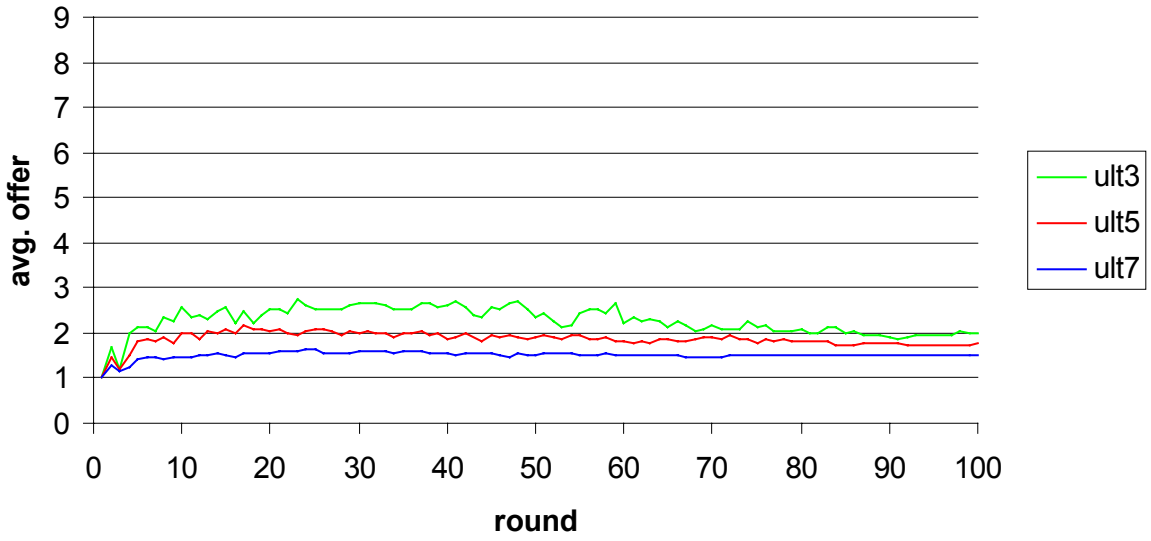

Figure 5. Theoretical prediction of the average behavior using Gittins indices

Hence, for this learning model that neglects the interdependence between the arms we see again that the players do not really learn to make the minimal offer, and, again, there are some qualitative similarities with the actual experimental data. The latter applies in particular to the weak downward trend.

It might be that the actual players in the experiment start out taking into account correctly the information about the situation (the interdependence of the arms, as predicted by the adjusted Gittins index approach), when they have no other information. However, as time progresses they increasingly neglect this information about the situation (being boundedly rational), and they start experimenting more and more with different arms (as predicted by the standard Gittins index approach).

This leads us to consider the predicted dynamics of the two boundedly rational models of learning presented in section 4; not so much to test which model fits the

[10] The fact that there is more of a downward trend towards the optimal minimal offer may seem counter-intuitive, as this model makes less use of the available information than the model of optimal learning analyzed above.

experimental data best, but to get a better idea as to where we should expect learning to lead to in the ultimatum game experiment.

Figure 6 shows the average offers for 1000 players using reinforcement learning as in Roth & Erev [1995]. We observe the following. First, the initial choices are about 5. Second, gradually differences between the treatments emerge. Third, there is a weak downward trend, and in none of the treatments is the optimal minimal offer approached.

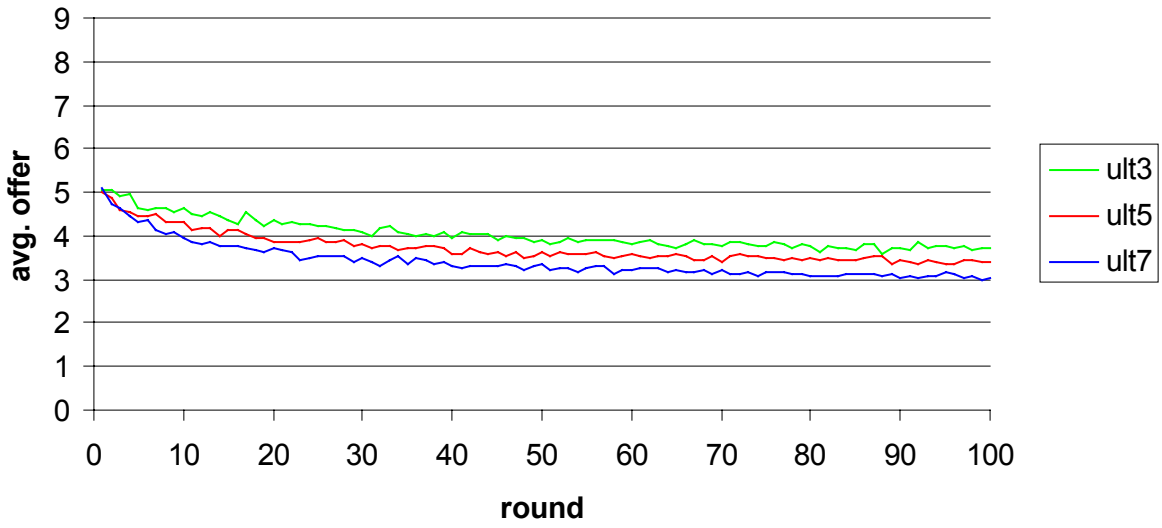

Figure 6. Theoretical prediction of the average behavior using reinforcement learning

Figure 7 shows the average offers for 1000 learning direction theory players. We observe the following. First, the initial offers are again around 5. Second, after a steep initial learning effect, the average offers are constant in each treatment.[11] Third, during this initial learning phase, marked differences between the treatments emerge.

Hence, we can conclude that the situation faced by the players in an ultimatum game is such that also players learning according to these models of boundedly rational behavior would not learn to make the optimal minimal offer. Moreover, these models of boundedly rational learning display some qualitative similarities with the actual experimental data. This concerns in particular the differences between the treatments and the lack of a clear trend towards the optimal minimal offer. Interestingly, similar to what we observed with the

[11] This is not surprising given that this is a discrete Markov process. Notice also that the stationary

adjusted and standard Gittins indices, the reinforcement learning model, which does not take any interdependence of the offers into account, predicts more of a downward trend than learning direction theory, which is based on an assumed interdependence of the offers.[12]

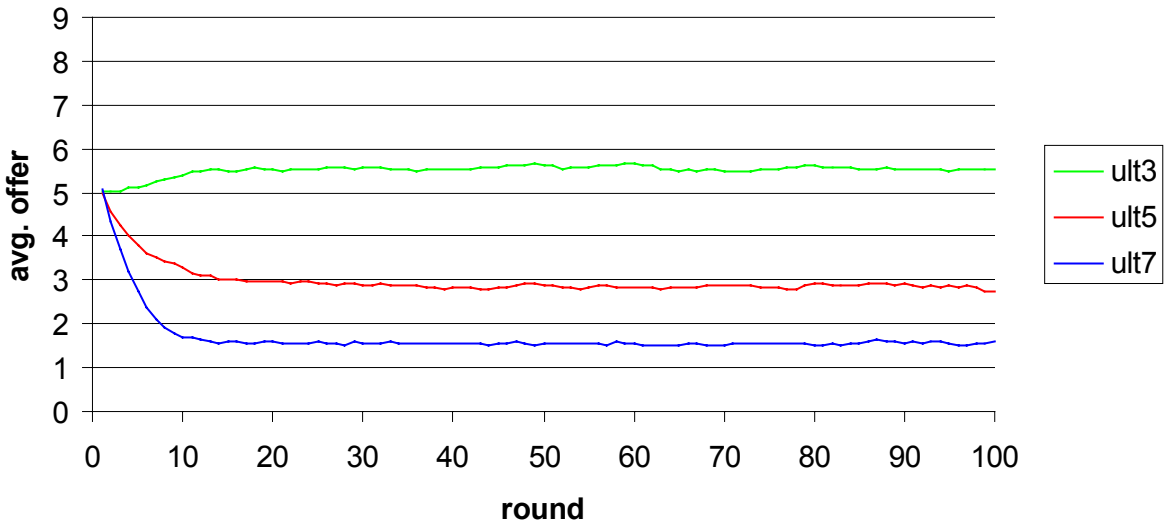

Figure 7. Theoretical prediction of the average behavior using learning direction theory

6. Concluding remarks

This paper makes three contributions to the literature. First, we designed a laboratory experiment of a stylized ultimatum game, in which we abstract from the coevolutionary aspects of adaptive behavior in a standard ultimatum game experiment, and in which there is also no scope for altruism or fairness considerations. That is, we organized an experiment that matches the situation studied in multi-armed bandit problems. The behavior of the receiving players B is fixed from the outset in such a way that making the minimal offer of 1 is optimal. We show that *even* if the learning task for the proposing players A is made as easy as possible, while maintaining realistic assumptions concerning the behavior of players

(..continued)

distributions of offers are independent from the initial guesses.

[12] We also considered a 2-stage model combining reinforcement learning and learning direction. This model has two different levels of learning. At the base level a player learns which offer to make. He can do this using either reinforcement learning or using learning direction theory. At the higher level a player learns which of these two modes of learning to use. This model gives a rather good fit for the average offers in each of the three treatments.

B, players A do not really learn this, notwithstanding a learning opportunity of 100 periods. Average offers are coming down, but very slowly. Hence, the lack of convergence to the subgame perfect equilibrium offer in ultimatum game experiments is not necessarily related to coevolutionary aspects of learning or to fairness considerations.

Second, we show that the lack of convergence to the optimal minimal offer is inherent in the learning task faced by players A, independent from the learning of players B. Analyzing a range of learning theories (from optimal learning based on the Gittins index to boundedly rational learning) in a setup that matches exactly the experimental design, we show why one should not expect convergence to the optimal offer, not even if the learning task is made as easy as possible.[13] That is, we offer a theoretical explanation for the experimentally observed lack of convergence to the optimal offer.

Third, although the objective of this paper is not to find the learning model that fits the experimental data best, our analysis does yield some insights into the actual learning behavior of the experimental subjects. Their initial choices suggest that they are reasonably good at taking the structure of the choice situation they face into account (in particular the interdependence of the offers), as their initial choices are consistent with the unbiased guesses of the optimal learning model. But these initial guesses happen to be far away from the real probabilities that actually determine the optimal minimal offer. The weak downward trend in the data suggests that the experimental subjects increasingly forget the initial information about this interdependence between the offers, and continue to experiment, as predicted by those learning models that assume independent arms. Although this implies more experimentation than predicted by the model of optimal learning, and this helps to overcome misleading initial guesses, convergence towards the optimal minimal offer takes place only very partly.[14]

References

Bergemann, D., & Välimäki, J. (2001). Stationary Multi-choice Bandit Problems. *Journal of Economic Dynamics & Control*, *25*, 1585-1594.

[13] Leloup [2000] shows that one should not expect convergence to the optimal offer in a situation in which there is almost no difference in expected payoffs for offers between 10% and 50% of the possible offer range.

[14] Interestingly, notice that the models of boundedly rational learning predict more convergence towards the optimal minimal offer than the model of optimal learning.

Bolton, G.E., Katok, E., & Zwick, R. (1998). Dictator Game Giving: Rules of Fairness versus Acts of Kindness. *International Journal of Game Theory*, *27 (2)*, 269-299.

Brezzi, M., & Lai, T.L. (2002). Optimal Learning and Experimentation in Bandit Problems. *Journal of Economic Dynamics & Control*, *27*, 87-108.

Forsythe, R.J., Horowitz, J.L., Savin, N.E., & Sefton, M. (1994). Fairness in Simple Bargaining Experiments. *Games and Economic Behavior*, *6*, 347-369.

Gale, J., Binmore, K., & Samuelson, L. (1995). Learning to Be Imperfect: The Ultimatum Game. *Games and Economic Behavior*, *8*, 56-90.

Gittins, J.C. (1979). Bandit Processes and Dynamic Allocation Indices. *Journal of the Royal Statistical Society, Series B*, *41*, 148-177.

Gittins, J.C. (1989). *Multi-armed Bandit Allocation Indices*. John Wiley & Sons.

Güth, W., & Tietz, R. (1990). Ultimatum Bargaining Behavior: A Survey and Comparison of Experimental Results. *Journal of Economic Psychology*, *11*, 417-449.

Leloup, B. (2000). *May Learning Explain the Ultimatum Game Paradox?* (GRID Working Paper No. 00-03) Ecole Normale Supérieure de Cachan.

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

Selten, R., & Stoecker, R. (1986). End Behavior in Sequences of Finite Prisoner's Dilemma Supergames. A Learning Theory Approach. *Journal of Economic Behavior and Organization*, *7*, 47-70.

Thaler, R.H. (1988). The Ultimatum Game. *Journal of Economic Perspectives*, *2*, 195-206.

Vriend, N.J. (1997). Will Reasoning Improve Learning? *Economics Letters*, *55, No. 1*, 9-18.

Appendix: Instructions to the players

Introduction

- This is a decision experiment. The instructions are simple, and if you pay attention, you can gain a reasonable amount of money. From now on till the end of the experiment you are not allowed to communicate with each other. If you have a question, please raise your hand. You are not allowed to use paper, pen, calculator, or any other material not provided by the organizers of the experiment.
- Each of you will play repeatedly the same basic game. Before explaining how often, and with whom you will play this game, we will first explain the basic game as such.

The Basic Game

- There are two players: player A, and player B. Player A has a pie cut in 9 equal slices. Player A makes a proposal to player B concerning the distribution of the pie. Player A can offer to player B from 1 up to 9 slices. Only whole slices are allowed. Player B can do 2 things. First, player B can accept the proposal of player A, which will then be carried out, player B getting the number of slices proposed by player A, and player A keeping the rest of the pie. Second, player B can reject the proposal of player A, in which case the pie perishes immediately, and both players will get nothing.
- Example: if player A proposes to give player B 1 slice, and player B accepts, then player B's payoff will be 1 slice, and player A will keep 8 slices. If, however, player B rejects the proposal,

then the payoff for both players will be 0 slices.

The Experiment

- You will play the same basic game for 100 rounds. In each round you will play the role of the proposer: player A. Each round you will be matched `at random' with some player B. You will never play more than once against the same player B.
- The players B with whom you will be matched are drawn from a large population of Artificially Intelligent agents, making their decisions using some computer algorithm.
- Each of those players B's behavior is systematic in the following sense: If a specific player B has accepted an offer x then that player B would have accepted as well any offer greater than x. And if offer x had been rejected by that player B, so would have been all offers smaller than x. Of course, different players B might have different opinions about which offers are acceptable or not. The players B do not change their behavior over time.
- During the experiment, your computer screen will be divided into 2 windows. The upper window will give you general messages, ask for input, etc. The lower window will display the history of your experiment. This window will be scrollable (using the arrows ↑ and ↓), such that you have always access to the complete history. The history will list all rounds, the identity of the specific player B you were matched with, the offer you made, player B's response, and the resulting payoff for you.
- To make your offer, please enter a number. Remember that only integer values from 1 to 9 can be chosen. Please, before pressing Enter, always make sure that you did not make a typing-error.
- There is no time limit for your decisions.

Payment

- You will be paid according to the total payoffs you realized. For each slice of a pie gained you will get 83.3 Lire. At the end of the experiment, we will add up your payoffs, and calculate your monetary rewards. This will be done in a separate room, so you will not see what other players earned.

This working paper has been produced by
the Department of Economics at
Queen Mary, University of London

Department of Economics
Queen Mary, University of London
Mile End Road
London E1 4NS
Tel: +44 (0)20 7882 5096 or Fax: +44 (0)20 8983 3580
Email: j.conner@qmul.ac.uk
Website: www.econ.qmul.ac.uk/papers/wp.htm