MEASURING PLAYERS' LOSSES IN EXPERIMENTAL GAMES[*]

Drew Fudenberg and David K. Levine

October 7, 1996

Abstract: In some experiments rational players who understand the structure of the game could improve their payoff. We bound the size of the observed losses in several such experiments. To do this, we suppose that observed play resembles an equilibrium because players learn about their opponents' play. Consequently, in an extensive form game, some actions that are not optimal given the true distribution of opponents' play could be optimal given available information. We find that average losses are small: $0.03 to $0.64 per player with stakes between $2 and $30. In one of the three experiments we examine this also implies a narrow range of outcome.

[*] We thank Eddie Dekel, Jack Hirshleifer, David Kreps, Tom Palfrey, Alvin Roth, Jeroen Swinkels, and Steve Tadelis for helpful comments and conversations. NSF Grants SBR-9223320 and SBR-9223175 provided financial support.

I. Introduction

Some observations in experimental games clearly involve "losses", in that they are not consistent with the hypotheses that players understand the structure of the game and act to maximize the payoff function specified by the experimental design. For example, some players refuse positive offers in the ultimatum game even though this means the game ends and they get nothing. However, there are other cases where whether an action is a "mistake" from the viewpoint of maximizing dollar payoff depends on what information the players are assumed to have when making their decisions.

This paper develops a theoretical tool for analyzing and reporting the extent of monetary losses that tries to reflect the information available when decisions are made. Our approach combines two theoretical ideas. The first is the relaxation of exact optimization to optimization with small losses, which leads us to study ε-equilibrium, a concept introduced by Radner [1980]. Second, rather than treating Nash equilibrium or one of its refinements as an implication of the hypothesis that players are rational, we suppose that the reason observed play resembles an equilibrium is that players learn about their opponents' play through repeated observations. As noted by Fudenberg and Kreps [1988], a player need not learn how an opponent would respond to an action that has never been taken. Consequently, from the viewpoint of learning theory, the appropriate solution concept is not Nash equilibrium, but rather the self-confirming equilibrium we introduced and characterized in Fudenberg and Levine [1993a].

We will argue that some observations that might seem to involve monetary losses are in fact consistent with players maximizing their expected monetary payoffs under beliefs that incorporate the sort of off-path prediction errors permitted by self-confirming equilibrium, and that self-confirming equilibrium is more appropriate and more useful than Nash equilibrium for analyzing game theory experiments. Of course, actions that lead to lower monetary payoffs regardless of opponents' play, such as refusing positive offers in the ultimatum game, cannot be rationalized by prediction error. Such observations can only be explained as a result of the players being "irrational" in the sense of not maximizing the monetary payoffs specified in the experimental design. Thus we know from the outset that even the properly measured monetary losses are not always 0; our interest is in measuring the average losses in various experiments.

More formally, we try to compute the minimum loss required to explain the experimental observations, where the minimum is over all beliefs that are consistent with the players' information and all mixed strategies consistent with observed behavior strategies. These minimizations arise because in the experiments we examine, the experimenters observe neither the subjects' beliefs nor their full contingent strategies.[1] Our analysis is based on the aggregate distribution of subject's play in each period, as opposed to the play of individual subjects, so that we identify all individuals who play the same actions in a given round of the experiment. We compare this approach to the study of the round-by-round play of individual players in section III .

Our approach is to look at an *ex ante* loss averaged over all contingencies. It is important to emphasize that a measure of the largest contingent loss would yield a very different picture. For example, in the centipede experiments we study, in the final move some subjects choose to give up a certain gain of $1.60. Since this happens in a relatively small fraction of the games that are played, it makes a small contribution to the average loss as we measure it.

Using our approach we measure the average losses in a number of experiments in the literature. We look for regularities in the losses: are they roughly constant, or do they vary in a systematic way? We also ask whether the theoretical concept of ε-self-confirming equilibrium is a useful tool for analyzing and predicting experimental play. More specifically, in games where the play resembles a stronger equilibrium concept, is this because the same size distribution of losses leads to a smaller set of ε-self-confirming outcomes?

In the experiments that we have examined, the average loss of a player is small in absolute terms: $0.03 to $0.64 per player in games involving stakes between $2 and $30, and where the maximum possible loss ranged from $0.80 to $5.00. As the stakes in the game are increased, the losses tend to increase at roughly the same rate, indicating that the types of mistakes made do not change as more money (up to four times as much in one case) is involved. As a benchmark, we also estimate the losses computed according to the Nash theory where players are supposed to have correct beliefs, even about play at information sets that they have never seen played. As a matter of definition, these Nash losses cannot be smaller than the self-confirming losses described above. Moreover, with one exception, these losses were four or more times as large as the self-confirming losses, showing that off-path errors can explain most of them.[2]

How does our approach differ from previous analyses of experimental data? In the case of simultaneous move games, where the issue of off-path prediction errors does not arise, Harrison [1989] argued that the cost of player errors is a useful metric for measuring departures from the theory. In a series of experiments with sealed-bid auctions[3] Harrison showed that for stakes on the order of $5, losses per player game were on the order of several cents. These stakes and losses are similar to the types of losses we find in the extensive-form game experiments we analyze. Notice also that it is consistent with our theory: in the case of simultaneous move games the theory of self-confirming equilibrium predicts the same outcomes (and same losses) as Nash equilibrium.

We should, however, distinguish our program from the argument that the observed losses are small enough to be ignored. This latter view, expressed most forcefully in Harrison [1992], says that observed departures from rational play are not surprising given the small stakes used in most experiments, and suggests that observed play would be closer to the predictions of standard theory if the stakes were substantially increased. While it may be that losses, properly measured, will shrink in relative size as the payoff scale grows, our concern is with the prior question of measurement. Moreover, we think it is interesting to develop tools for analyzing the outcomes of experiments with the stakes that are commonly used, even if these stakes give greater prominence to non-monetary considerations.

There is also a substantial methodological difference between our work and previous work on extensive-form games. Attempts to reconcile experimental data with game-theoretic predictions, such as the “home-made priors” (that an opposing player’s payoffs are different than those specified in the experimental design) used by Camerer and Weigelt [1988] and McKelvey and Palfrey [1992], proceeded on a case-by-case basis that seems difficult to generalize to other games, or to formalize in a standard way. Two different researchers might propose different forms of homemade priors, and then estimate different proportions of irrational types.[4] In contrast we propose an algorithm for computing the distribution of losses by the players that can be applied to any game.

II. The Environment

We study games with I players; the game tree X, with nodes $x \in X$ is finite. Terminal nodes are $z \in Z$. For notational convenience, we represent nature by player 0. Information sets, denoted by $h \in H$ are a partition of $X \setminus Z$. The information sets where player i has the move are denoted by $H_i \subset H$; information sets belonging to nature $h \in H_0$ are singletons. The feasible actions at information set $h \in H$ are denoted $A(h)$. We generally use $-i$ for all players except player i, so that for example H_{-i} are information sets for all players other than i.

A pure strategy for player i, s_i, is a map from information sets in H_i to actions satisfying $s_i(h_i) \in A(h_i)$; S_i is the set of all such strategies. Mixed strategies are $\sigma_i \in \Sigma_i$, the mixed strategy σ_0 represents any random moves by “Nature.” We generally omit subscripts to represent Cartesian products, so that for example $\Sigma \equiv \times_{i \in I} \Sigma_i$. Each player except nature receives a payoff $r_i(z)$. that depends on the terminal node.

In addition to mixed strategies, we define behavior strategies $\pi_i \in \Pi_i$. These are probability distributions over actions at each information set for player *i*. From Kuhn's theorem, there is an equivalent behavior strategy for any given mixed strategy σ_i; denote this by $\hat{\pi}_i(\cdot|\sigma_i)$. For any given profile of behavior strategies π it is also useful to define the induced distribution over terminal nodes $\hat{\rho}(\pi)$. We will also use the shorthand notation $\hat{\rho}(\sigma) \equiv \hat{\rho}(\hat{\pi}(\sigma))$.

Since we assume that all players know the structure of the extensive-form, their own payoff function, and the probability distribution over nature's moves, the only uncertainty each player faces concerns the strategies opponents will use. To model this "strategic uncertainty" we let μ_i be a probability measure over Π_{-i}, the set of other players' behavior strategies. For any such beliefs, we may, in the obvious way, compute the expected utility $u_i(s_i, \mu_i)$.

For any mixed profile σ, we let $\overline{H}(\sigma) \subset H$ be the information sets that are reached with positive probability when σ is played. Note that this set is entirely determined by the distribution over terminal nodes ρ, so we may equally well write $\overline{H}(\rho) = \overline{H}(\hat{\rho}(\sigma))$. For any subset $J \subset H$ and any profile σ we may define the subset of behavior strategies consistent with players other than *i* playing σ_{-i} at the information sets in J by $\Pi_{-i}(\sigma_{-i}|J) \equiv \{\pi_{-i}|\pi_j(h_j) = \hat{\pi}_j(h_j|\sigma_j), \forall j \neq i, h_j \in H_{-i} \cap J\}$.

Nash equilibrium is usually defined as a strategy profile such that each player's strategy is a best response to his or her opponents. For our purposes, though, it is instructive to give an equivalent definition that parallels the way in which we will define self-confirming equilibrium.

Definition 1: A *Nash equilibrium* is a mixed profile σ such that for each $s_i \in \text{supp}(\sigma_i)$ there exist beliefs μ_i such that

- $u_i(s_i|\mu_i) \geq u_i(s_i'|\mu_i)$ for all $s_i' \in S_i$, and
- $\mu_i(\Pi_{-i}(\sigma_{-i}|H)) = 1$.

In this definition, the first condition requires that each player's strategy be optimal given his beliefs about the opponents' strategies. The second requires that each player's beliefs are correct at every information set.[5] If, however, player *i* continually plays σ_i, he will only observe opponents play at information sets in $\overline{H}(\sigma)$, and will not learn about his opponents play at other information sets. For learning to yield a Nash equilibrium, players must not merely learn passively, but must learn actively by experimentation, that is, play actions that do not maximize their current expected payoff in order to gain information that may be useful in the future. Unless they are very patient and will have many opportunities to play the same game, they will have no incentive to do this. This suggests the following weaker equilibrium concept:

Definition 2: A *unitary self-confirming equilibrium* is a mixed profile σ such that for each $s_i \in \text{supp}(\sigma_i)$ there exist beliefs μ_i such that

- $u_i(s_i|\mu_i) \geq u_i(s_i'|\mu_i)$ for all $s_i' \in S_i$, and
- $\mu_i(\Pi_{-i}(\sigma_{-i}|\overline{H}(\sigma))) = 1$.

Here is assumed only that player *i* is correct in his beliefs at information sets that are actually observed. Fudenberg and Levine [1993a] showed that unitary self-confirming equilibrium has the same outcomes as Nash equilibrium in two-player games, and that the two concepts are also equivalent in multistage games with more than two players, provided that beliefs satisfy an additional independence condition.[6]

The experiments we examine use a matching design in which there is a population of subjects in each role ("player 1", "player 2", and so forth). Individual subjects are matched each period against different individuals in the other role, and each subject observes the outcomes of play in his or her own matches, but does not observe the hypothetical off-path play of the opponents nor the outcomes of play in other matches.[7] In such a setting, there is no reason that two subjects assigned the same player role should have the same prior beliefs. If subjects draw from a large common pool of observations, we might expect them to have the same posterior beliefs; and indeed, we might expect that subjects who have repeatedly played the same pure strategy will have learned the consequences of doing so. However, given that subjects only observe the outcomes in their own matches, if two subjects have always played different pure strategies, their beliefs may remain different.[8] This motivates the following weaker notion of self-confirming equilibrium:

Definition 3: A *heterogeneous self-confirming equilibrium* is a mixed profile σ such that

for each $s_i \in \operatorname{supp}(\sigma_i)$ there exist beliefs μ_i such that

- $u_i(s_i|\mu_i) \geq u_i(s_i'|\mu_i)$ for all $s_i' \in S_i$, and
- $\mu_i(\Pi_{-i}(\sigma_{-i}|\overline{H}(s_i,\sigma_{-i}))) = 1.$

This definition allows different beliefs μ_i to be used to rationalize each pure strategy s_i in the support of σ_i, and allows the beliefs that rationalize a given s_i to be mistaken at information sets that are not reached when s_i is played, but are reached under a different s_i' also in the support of σ_i. Figure I gives simple example from our [1993] paper showing how this allows outcomes that cannot arise with unitary beliefs. Since this is a two player multi-stage game, Nash equilibrium and unitary self-confirming equilibrium yield the same outcomes. The game has two types of Nash equilibria: the subgame perfect *RU* and the equilibria in which player 1 plays *L* and player 2 plays *D* at least 50 percent of the time. However, there is no Nash equilibrium in which player 1 randomizes between *L* and *D*. There is however a heterogeneous self-confirming equilibrium in which player 1 does randomize: player 2 plays *U*, and while those player 1's that play *R* know this, those that play *L* incorrectly believe that player 2 would play *D*.[9]

III. Measurement of Losses

The main purpose of this paper is to propose a method for reporting the distribution of losses in experimental games. To avoid potential confusion, we should make it clear at the outset that we will not propose and test a particular econometric model. Rather, we propose an accounting convention that has some partially arbitrary features. Our hope is that this way of looking at experimental data will prove useful in identifying empirical regularities.

Our analysis takes as data the frequency with which particular terminal nodes are reached, which is a commonly used method of summarizing observed play in experimental studies of extensive-form games. We will follow the common practice of concentrating attention on data from the "last few" rounds of the experiment, so that subjects will have had some chance to learn their opponents' strategies, and the play is more likely to have converged.[10] Moreover, our analysis implicitly presumes that play has converged, so that each subject is repeatedly using the same strategy. However, the strategies of the individual subjects need not be revealed by the aggregate distribution of play: for example, the distribution (1/2 L, 1/2 R) results if each subject mixes with equal probability on L and R, and also if half the subjects always play L while the other half always plays R.[11]

Under different assumptions about how much subjects know about the true distribution over terminal nodes we compare the amount of money that players actually made with the amount of money that they could have made. (Roughly speaking, we are measuring the size of ε in an ε-equilibrium.[12]) We focus on the monetary payoffs because they, unlike the players' "true" utility functions, are clearly specified in the experimental design. Our goal is not to test the obviously false null hypothesis that all subjects act to maximize monetary payoffs, as in some cases players clearly "give away" nontrivial amounts of money . Rather we will try to measure the of their losses, in an effort to uncover empirical regularities, and ideally to develop predictions about play in future experiments.[13]

We should emphasize that we do not try to explain the patterns in such departures from maximizing monetary payoffs. There have been a number of interesting attempts to develop "behavioral" theories that explain these departures, based on, for example, ideas of fairness, altruism, and spite. Our concern here is on what we see as the logically prior question of measuring the frequency of such "irrational" (non-money-maximizing) play. In our view, observations that *can* be explained as the result of players trying to maximize their dollar payoffs *should* in general be explained in that way, so that the appropriate goal of the behavioral theories is to explain the "epsilons" that this paper measures.

To avoid confusion, we should also emphasize that, although the measured losses are small in the experiments we analyze here, our method is valid in any game, including those where measured losses seem likely to be large, such as the voluntary-contribution experiments of e.g. Andreoni [1988] and Isaac and Walker [1988].

We should also point out that experiments contain (and some experimenters report) more detailed information than the distribution over terminal nodes, namely the period-by-period play of each individual subject. A number of studies have examined this data.[14] The general conclusion seems to be that theories of learning do much better at predicting aggregate play than individual play. In particular, the play of individual subjects can follow suboptimal rules-of-thumb quite rigidly, even when the aggregate distribution resembles a Nash equilibrium. Our goal in this paper is to examine the extent to which the theory fails in predicting aggregate play, in instances where aggregate play fails to resemble a Nash equilibrium.[15] This is not to suggest that understanding the period-by-period play of individual subjects is unimportant, although from the point of view of applying the theory outside of the laboratory, the most easily used prediction of the theory is that of the aggregate play.

We note that our approach of focusing on the distribution over terminal nodes both overstates and understates losses. The heterogeneous calculation overstates losses in that typically a subject will have played some strategies other than the one currently being played. The unitary version understates losses in that a subject will typically not have played some strategies that have been tried by other subjects of the same player type. Moreover, both calculations ignore the fact that individuals may have too small a sample from the distribution over terminal nodes to be confident that they have learned their opponent's response, even if the subject has chosen the same action in every round of the experiment. (This problem is particularly acute if the opponent's strategy is mixed, for then the observations may have a large variance.)

Let us denote by ρ the probability distribution over terminal nodes that corresponds to the empirical frequency in a particular experiment. Our goal is to define, for each of the three observation functions J corresponding to heterogeneous self-confirming, unitary self-confirming, and Nash equilibrium, the expected loss $\varepsilon_i(J(\cdot),\rho)$.

For any given pure strategy and beliefs, there is a clearly defined loss relative to those beliefs that we denote by $\varepsilon_i(s_i,\mu_i)=\max_{s'_i} u_i(s'_i|\mu_i)-u_i(s_i|\mu_i)$. However, the experiments we examine did not collect data on either the subjects' beliefs or their strategies.[16] Our approach is to be as charitable as possible, in the sense of looking for the smallest departure from utility maximization that is required to explain the observations. Thus, if the observed distribution of play can be generated by a unitary self-confirming equilibrium, we will set the "unitary loss" to be zero. Likewise, if the observed distribution corresponds to a heterogeneous self-confirming equilibrium, we set the heterogeneous loss equal to zero.

More generally, for a given distribution ρ and information function J , we look for the mixed strategy profile σ and beliefs for the players μ that minimize the resulting average loss over all strategies and beliefs consistent with ρ and *J*. In the unitary case, for a given mixed strategy profile σ, this requires finding for each player *i* the beliefs μ_i that minimize *i's* loss over all beliefs that are correct on $\overline{H}(\sigma)$. In the heterogeneous case, when a player *i* is observed to play s_i, we require only that player has correct beliefs about opponents' play at all information sets in $\overline{H}(s_i,\sigma_{-i})$, so that the loss-minimizing beliefs $\mu_i(s_i)$ may depend on *i's* strategy s_i. This leads to the following definition of the average loss for the information functions $J(\cdot)$ corresponding to heterogeneous and unitary beliefs:

$$\begin{aligned}&\varepsilon_i(J(\cdot),\rho)\equiv\min_{\mu_i(s_i),\sigma}\left\{\sum\nolimits_{s_i}\varepsilon_i(s_i,\mu_i(s_i))\sigma_i(s_i)\right\}\\&s.t.\ \mu_i(\Pi_{-i}(\sigma_{-i}|J(s_i,\sigma))=1,\quad\hat{\rho}(s)=\rho\end{aligned}$$

In the heterogeneous case this minimization implies that each subject is playing a pure strategy, as this minimizes the amount of information that each subject has. Thus the mixture over strategies is attributed entirely to different subjects of the same type playing in different ways.[17]

As a practical matter, the minimization in the definition of $\varepsilon_i(J(\cdot),\rho)$ is most easily accomplished in two stages. First for each pure strategy s_i we find the beliefs that yield the smallest loss

$$\varepsilon_i(s_i,J(\cdot),\rho,\sigma)\equiv\min\left(\varepsilon_i(s_i,\mu_i)\middle|\quad\sigma,\mu_i\quad s.t.\quad\hat{\rho}(\sigma)=\rho,\quad\mu_i(\Pi_{-i}(\sigma_{-i}|J(s_i,\sigma))=1\right).$$

Although this definition involves a minimization over σ, that minimization is moot: the beliefs that opposing players will coordinate to minmax player *i* off of $J(s_i,\sigma)$ will obviously minimize the loss from playing s_i, and the set $J(s_i,\sigma)$ and hence the loss-minimizing beliefs are the same for every σ such that $\hat{\rho}(\sigma)=\rho$. Thus we can refer instead to the loss as $\varepsilon_i(s_i,J(\cdot),\rho)$. Averaging over the pure strategies with the frequencies given by σ_i then yields

$$\varepsilon_i(J(\cdot),\rho) \equiv \min_{\sigma|\hat{\rho}(\sigma)=\rho} \sum_{s_i} \varepsilon_i(s_i, J(\cdot),\rho)\sigma_i(s_i).$$

The practicality of computing average losses using this two-step procedure depends on the number of pure strategies available to players. In games with several stages, the number of pure strategies can quickly become overwhelming. For this reason, it is useful to note that if there is a player who does not have a move prior to a subgame the computation of losses can be simplified. We separately compute the loss in the subgame and in the game in which the subgame is replaced with a zero utility for that player. We then average these losses together with the probability that subgame is (or is not) reached. In particular, in a game in which player one moves, player two moves, then the game ends, we may compute player two losses by computing the difference between his actual and optimal strategy for each player one move, then averaging over player one moves, weighted by the probability that player one assigns to those moves.[18]

IV. The Centipede Game

The first experiment we analyze is the Centipede game experiment conducted by McKelvey and Palfrey [1992]. There were several versions played. The base case extensive-form is the perfect information game shown in Figure II. This game has a unique self-confirming equilibrium; in it player 1 with probability 1 plays T_1 (drops out). Naturally this is also the unique subgame-perfect equilibrium. The uniqueness of the self-confirming equilibrium may be proven recursively.[19]

We will now compute the unitary and heterogeneous losses implied by the observed outcomes specified by the square brackets in the figure. Since there are a small number of pure strategies in this game, the computations are fairly straightforward.

In the unitary case, we observe that every information set is reached a positive fraction of the time. Consequently, the unitary loss must computed assuming that players know their opponent's play at every information set, and so is measured relative to the optimized payoff against the true distribution. For player 1 this is to play P_3 , for an expected payoff of $1.02[20]; for player 2 this is to play T_4 which also, by coincidence, has an expected payoff of $1.02. So to compute the unitary losses, for each pure strategy we subtract the expected utility of that strategy against the empirical distribution of opponents play from $1.02. This is reported in the *Unitary* column of Table I . The empirical frequencies of the pure strategies are noted in the *Frequency* column, and the overall loss is computed by averaging the loss to each pure strategy over pure strategies. This leads us to compute the average unitary losses to be ($0.12,$0.17)

In the heterogeneous case, the strategies T_1 and T_2 that "drop out" early have 0 loss, because a player who drops out early can believe that the opposing player would take (play *T)* in the next round. The only loss is the loss to the strategy P_4, which loses $1.60 irrespective of beliefs about the opponent's play.[21] The average heterogeneous losses are then calculated to be ($0.00,$0.03).

So far we have analyzed data from the last 5 rounds of play only. In fact, each player played the game 10 times against different opponents. (Each time the game is played by every player is a round of play.) The first two rows of Table II give the unitary and heterogeneous losses computed above for the last 5 rounds, base-case experiment. Table II also gives the losses corresponding to the entire 10 rounds of play of the base case, and for the entire 10 rounds of an alternative treatment which involved the same game tree but payoffs that are four times as large.[22] In the interests of brevity, we have omitted the calculations of these losses; the calculations are much the same as those above.

The row in Table II labeled "WC" is a theoretical calculation of the "worst-case" losses; it is not based on the result of the experiment. This case gives the losses for the distribution over outcomes that gives the highest expected loss per player in the game under heterogeneous beliefs. When this number is small, it means that reported heterogeneous losses will necessarily be small regardless of the realized play. As we will see, though, the realized losses are much smaller than this worst case.

The row labeled “Random” is also a theoretical calculation, intended to measure what the heterogeneous loss would be under “completely random” play, which we take to be r the distribution over outcomes generated when players play each pure strategy with equal probability of 1/3. That is, when player 1 has a 1/3 chance of taking in period 1, a 1/3 chance of taking in period 3 and a 1/3 chance of passing in period 3, for example.[23] Like the worst-case loss, this calculation can also serve both as a benchmark and as a test of whether the method for measuring losses has any force: As a benchmark, we would expect that play would converge to a setting with lower losses than either of the theoretical calculations, while a test, we would be disappointed if the theoretical values for non-equilibrium play were typically zero or even small. In that light we should point out that the losses under “random” play will be zero if random play is an equilibrium, as it is for example in matching pennies. We should note that as the data suggest that heterogeneous self-confirming equilibrium is a much better description of the data than unitary, we compute only the heterogeneous losses for random play.

The first column of Table II indicates how many games were played in each round. (Since this is a two player game, the number of players playing is twice the number of trials/round.) The second column indicates which rounds were included in the particular sample. We feel that the most interesting case is when only the latter rounds (6-10) are included, as this eliminates the learning taking place during the early rounds, and gives players a chance to settle into equilibrium.

The third column indicates the payoffs as a multiple of the extensive-form above. These are as in the above game tree in the cases labeled “1x;” the entry “4x” describes one series of experiments carried out with the same extensive-form, but payoffs four times as large as those shown above. The fourth column indicates the basis of the loss computation: there are two cases, the unitary case (U), the heterogeneous case (H). The next three columns contain statistics about the losses. The first two columns contain the average expected loss $\varepsilon_i(J(\cdot),\rho)$ for players $i=1,2$; the column labeled "Both" simply averages the losses for the two players together to get an overall summary statistic of expected loss per player per game. The penultimate column labeled "Max Gain" is the greatest per player payoff possible in the game, and is used to summarize the magnitude of payoffs in the game. The final column reports the ratio of the loss per player per game to the greatest per player payoff possible.

The salient features are:

- The heterogeneous loss per player is very small. Player 1's heterogeneous loss is 0, because player 2 gives money away sufficiently frequently in the final stage that it is optimal for player 1 to stay in to the end, while the player 1's that drop out early have no way of knowing that player 2 is giving away money in this way. Similarly, the best response for player 2 to the empirical distribution of play is to drop out in the final stage, so the only mistake is to give away money at this stage. The worst-case outcome is thus probability 1 of player 2's last node being reached, and player 2 then choosing to give away money, which would result in a heterogeneous loss of $0.80 per player. In the experiments, , money is given away sufficiently infrequently that the average loss with 1x stakes is only $0.02, and even in the quadruple stakes case (where the loss to playing P_4 is $6.40), the expected loss is only $0.14. Thus the prediction that losses will be small compared to the worst case has substantial predictive power, even though it allows a wide variety of approximate equilibria. [24] This is reinforced by the fact that the observed heterogeneous losses were substantially smaller than would be generated by random play. On the other hand, in this particular case, actual play is relatively close to random play, so the losses from random play are comparable to those from actual play. However, while random play does a good job of explaining what happened in this experiment, it does relatively poorly in the other experiments we examine.

- The unitary losses while still only $0.15 per player per game in the ordinary stakes last 5 rounds, are still 7 times as large as in the heterogeneous case. Indeed, even player 2 loses quite a bit more from dropping out too early in round 2 (which is not irrational if player 2 does not learn how player 1 would play at the next node) than by giving money away at the end of the game.
- Quadrupling the stakes very nearly causes ε to quadruple indicating that increasing the amount of money involved does not seem to significantly change the way that players play.
- As indicated on the game tree, 18 percent of player 2's chose to pass in the last 5 rounds conditional on actually reaching the final stage. This means that the losses conditional on reaching the final stage are quite large, something that is inconsistent with subgame perfection. To reflect this problem, McKelvey and Palfrey proposed (and estimated) an incomplete information model where some "types" of player 2 liked to pass in the final stage. This accounts for the heterogeneous losses, but still faces the problem that many players dropped out early, as the sequential equilibrium concept they use requires that all players correctly predict the average distribution of play at all information sets. Hence their estimated model fits fairly poorly.[25]

V. The Best Shot Game

The second experiment we analyze is the "best shot" game introduced and first studied by Harrison and Hirshleifer [1989]. In fact we report the results from Prasnikar and Roth [1992] who used a larger sample, and provided a broader variety of experimental conditions. (We will also indicate how their results differ from Harrison and Hirshleifer.)

The best shot game is a sequential public goods contribution game in which the provision of public good is determined by the larger of the two contributions.[26] This extensive-form is shown in Figure III. Here x_i is player *i*'s contribution, *W* is the utility of the public good, and *C* is the cost of private contribution. Players could contribute any integer amount between 0 and 8, and the functions *W* and *C* are given in Table III..

With the payoffs as specified, this game has the striking property that if the other player makes any contribution at all, it is optimal to contribute nothing. There is a unique subgame perfect equilibrium: player 1 contributes nothing and player 2 contributes 4. There is another Nash equilibrium, for player 1 to contribute 4 and player 2 to contribute nothing regardless of player 1's play. There are no mixed strategy Nash equilibria. Moreover, since all of the players are in the same population and do not have access to a public randomizing device, it is not consistent with Nash equilibrium for some player 1's to play 0 and others 4.[27] However, this and any other probability distribution over the two Nash equilibria are heterogeneous self-confirming equilibria: those player 1's who play 0 correctly perceive that 2 will respond with 4, while those choosing 4 fallaciously believe that if they contribute nothing, their opponent will not contribute.

The computation of losses is quite easy in this game despite the fact that player 2 has 64 pure strategies: as we noted above, when a player's only information set on any path is at the start of a proper subgame, so that the player in question cannot influence whether this information set is reached, the losses for that player may be computed conditional on the previous moves of the opponents, and then averaged over the observed distribution of opponents' moves. In this game things are even simpler, because player 2's information set ends the game, and so the loss to any action of player 2's is independent of 2's beliefs about1's (nonexistent) future play. To calculate the benchmark losses from completely random play, we assume that players simply choose each contribution level with equal probability of 1/9.

Table IV provides loss statistics. The columns are generally similar to those in the centipede game, except that there is only one set of stakes, and two different information conditions labeled full and partial. The full information experiment is conducted under the “standard” conditions, with players informed of the monetary payoffs that would be given to their opponents. In the partial information case, players were not informed of their opponent’s payoffs This corresponds to the only case analyzed by Harrison and Hirshleifer. However, in Harrison and Hirshleifer, after the first 4 of 10 rounds only the subgame perfect equilibrium was ever observed, so losses of all sorts are equal to zero. This is in contrast to Prasnikar and Roth, where the partial information losses are not only positive, but significantly higher than in the full information case. However there is an important difference in the way the two experiments were conducted[28]: in Harrison and Hirshleifer players alternated between moving first and second, while they did not in Prasnikar and Roth.

The salient features of best-shot losses:

- In the full information case and partial information heterogeneous case, losses are modest, $0.12-$0.15. This is almost entirely due to player 2 contributing less than 4 when player 1 has contributed nothing. In this context it is worth noting that the player who contributes nothing gets a far larger profit than the contributing player $3.70 against $0.42.
- Since player 2 only moves at the end of the game, the player 2 losses are all independent of player 2’s beliefs about player 1’s play. These losses correspond almost entirely to player 2 not contributing as much as is optimal when player 1 has failed to contribute, although in one case a player 2 wasted money by contributing when player had already contributed. (It is hard to find much of a rationale for this, since neither player benefited by 2’s action.)
- The losses are several times larger than in the centipede game despite the fact that the overall stakes are lower.

- In the full information case heterogeneous losses are as large as the unitary losses. This is because player 1 never contributed anything, and so never had a loss with either type of information, while all losses by player 2 are independent of 2's beliefs about 1's play.
- In the partial information case, heterogeneous losses are quite a bit smaller than the unitary ones, with per player per game losses 1/3 as large. The reason for this is that in the partial information case frequently player 1 contributed nothing with player 2 contributing 4, but there were also a number of cases in which player 1 contributed 4 and player 2 contributed nothing. What is observed is therefore very much like a public randomization between the two Nash equilibria. This is inconsistent with Nash equilibrium (or its unitary equivalent), but (because the game is sequential move) is consistent with self-confirming equilibrium.

One of the most striking features about the best shot game is that subgame perfection does quite well in the full information case. Even in the partial information case it is rare for both players to make positive contributions. This is shown in Figure IV, which plots the data from that case. It turns out that there is a theoretical reason to expect this regularity, for in this game ε-self-confirming equilibrium (with heterogeneous beliefs) makes quite strong predictions, even for the moderately large[29] estimate of ε implied by the data. This can partially be seen in the worst case column of Table IV , in which worst case losses are significantly worse than observed in the experiment.

A better way to see this, however, is to look at the size of the set of approximate equilibria. In the partial information case heterogeneous losses per player game are \$0.08. In Figure V and Table V we characterize which probability distributions over terminal nodes are consistent with a loss per player game this small.[30] (Figure V simply graphs the numbers in Table V .) Take a subset of the set of pairs of contributions, for example $\{(3,2),(2,2),(2,3)\}$. How much probability can this subset have if the per player expected loss is no more than $0.08? Since the smallest loss to any strategy in this set is .80, the probability of the set of strategies must be under .1 in order for the average loss to be less than .08. A similar calculation shows that the combined probability of all outcomes in which player 1 has contributed 1 or more and player 2 has contributed 2 or more is no more than 0.10. (This upper bound is loose; for strategies that loose more than .8, the probability must be even smaller.) In general, the table is calculated so that if we choose any subset of profiles, the combined probability of that subset can be no greater than the largest entry in the table for the members of the subset.

Generally speaking, we should not expect to see both players contributing at the same time (at most 31 percent of the time). On the other hand, if the other player is contributing zero, we should not be that surprised if the other player fails to contribute 4, as the loss from failing to do so is not great. This, of course, is exactly what is observed: one player contributes nothing, the other usually contributes 4, but occasionally something else.

VI. The Ultimatum Game

In the ultimatum game the first player proposes to divide a given amount of money. The second player may accept or reject this offer. If accepted, the money is divided as proposed; if rejected, neither player gets anything. This is illustrated in the extensive form in Figure VI , where the offer *x* must be in pennies.

In every subgame perfect equilibrium of this game, the first player's strategy is some mixture, possibly degenerate, over demanding the whole pie and demanding one penny less; the second player accepts any positive offer, and may mix or reject the offer of 0. Nash equilibrium, by contrast, permits player 1 to make any offer with probability 1. It also allows a variety of mixed equilibria. As usual in games of perfect information, heterogeneous self-confirming equilibrium adds the public randomizations between the various Nash equilibria.

These ultimatum games have been studied by a wide variety of authors especially Guth and his co-authors [Guth and Tietz 1988, 1990; Guth et al. 1982, 1990]. The results are generally similar: most proposals are for the first player get more than 50 percent of the money, but much less than 100 percent, and ungenerous offers tend to be rejected. The specific experimental results we analyze here are taken from Roth, Prasnikar, Okuna-Fujiwara and Zamir [1991], who systematically study ultimatum games in a number of experimental settings. We report loss statistics below in the usual format. Here we report the results of the final round. The variation in experimental treatment is the country in which the experiment was conducted, Israel, Japan, the US and Yugoslavia. In addition, in the US, an experiment was conducted with stakes 3 times those indicated above. Outside the US payments were in local currency, calibrated to a total of $10 adjusted for purchasing power parity.

The computation of losses is quite easy in this game despite the fact that player 2 has 1,000,000 pure strategies: as in the best-shot game, the only move by player two is a subgame, and so, as in best-shot, the losses for player 2 may be computed conditional on the particular first move by player 1, then averaged over player 1's moves.[31] The losses are reported in Table VI.

The salient features of the experimental results:

- Because every offer by player 1 is a best response to beliefs that all other offers will be rejected, player 1's heterogeneous losses are always zero.
- Player 1's have substantial losses in the unitary case. This should not be surprising: given the large number of possible offers, no player has much chance of learning very much about the responses to all offers in ten rounds, and so, unless the players have extremely accurate prior information, they are not likely to actually hit upon the best response to the true distribution. Indeed, even with data on all games played, it is not

that easy for us as observers to have much confidence that we have identified the distribution of responses, and so we do not know whether our computed optimal offer is indeed the optimum.[32]. Note the contrast to Roth et al [1991] who argue that mean (or modal) offers are nearly a best response to the acceptance rate of offers. From our perspective this ignores the fact that there is a substantial variance in the offers made, and a substantial fraction of the offers involve losses that are considerably greater than those suffered in the second period from the rejection of offers.

- The player 2 losses all stem from rejected offers. The magnitudes of these losses are an indication that subgame perfection does quite badly in this setting. Note that the losses if both players were to play completely at random are considerably larger than those observed.
- As is the case in centipede, tripling the stakes increases the size of losses a bit less than proportionally (losses roughly double).
- Although the expected losses are larger than in centipede or best shot, they are not large in absolute terms: They range in the ordinary stake games from $0.38 in Israel to $0.99 in Yugoslavia, out of the $10 on the table. These losses do, however, serve to refute the naive hypothesis that the extent of observed losses properly measured will be roughly constant across games. Rather, because the losses reflect the players choosing to consider other factors than their monetary payoffs, we should expect the distribution of losses to be larger in games where other features such as fairness are particularly salient. In particular our project should not be viewed as a substitute for studies and models of such psychological factors. Rather, our methods provide a better way of measuring the prevalence and magnitude of such factors.

In Table VII we report raw data for the US $10 games: surprisingly, the reason for the heterogeneous (player 2) losses is the fact that offers even very close to $5.00 are rejected a non-negligible fraction of the time.

VII. Concluding Remarks

The purpose of this paper has been to develop the experimental implications of the idea that even rational subjects may have incorrect beliefs about the off-path play of their opponents. This idea, when coupled with the recognition that some subjects take actions that do not maximize their expected dollar payoffs under any beliefs, leads to the idea that if the play in an experiment converges, the limit should be one of the ε-self-confirming equilibria of the game. The crude analysis in this paper suggests that the associated ε's are typically small compared to the stakes of the game. Moreover, we found that the size of the set of ε-self-confirming equilibria for typical ε's varies quite a bit from game to game.

Our method of estimating the losses was to identify the empirical distribution of play in the "last few rounds" with the theoretical distribution of outcomes in a steady state, and then use this distribution to compute the expected payoff to the actions the players actually used. Since many experiments are only run for ten periods, this identification of the empirical and theoretical distributions is often unjustified, particularly in games, like the ultimatum game, with a large number of choices for the first mover. One way of refining our analysis would be to use more sophisticated methods to obtain either a point estimate, or a distribution, over the distribution of play at on-path information sets.[33] Another potential refinement would be to track the period-by-period play of each subject and estimate the loss-minimizing beliefs for each subject in light of the observations the subject has received. This approach does run in to the problem of increased sampling error we mentioned in section 3, but that problem need not be insurmountable, particularly in an experiment that was run for more than the usual ten rounds.

Finally, our approach suggests some new experimental designs that could be used to further clarify the role of incorrect off-path beliefs in determining experimental outcomes. One design would involve two treatments that are identical except that one has the standard observation structure where players observe only the outcomes in their own matches, while in the other each player is informed of the aggregate distribution of play in all matches. We would expect the unitary losses to be much smaller in the second treatment. Another possibility would be to ask players their beliefs about the opponents' actions at the end of each round, and then test whether the player's beliefs are consistent with their information and a "reasonable" prior, and also whether the players seem to be maximizing the money payoff given their beliefs. Of course, asking for beliefs to be reported might well lead to different behavior than in the "standard" treatment, but that seems unavoidable if one wants period-by-period information on beliefs. Yet another experimental issue is to explore extensive-form games with more than two players. The theory shows that in such games there is an additional way that self-confirming equilibria can fail to be Nash, namely that two players can have differing beliefs about the off-path play of a third. It would be interesting to see how important this theoretical possibility turns out to be in the lab.

Department of Economics, Harvard University

Department of Economics, University of California, Los Angeles

References

Andreoni, James, "Why Free Ride?: Strategies and Learning in Public Goods Experiments," *Journal of Public Economics,* XXXVII (1988), 291-304.

Brown, J.N., and Robert W. Rosenthal, "Testing the Minmax Hypothesis: A Reexamination of O'Neill's Game Experiment," *Econometrica*, LVIII (1990), 1065-1082.

Camerer, Colin, and K. Weigelt, "Experimental Tests of a Sequential Equilibrium Reputation Model," *Econometrica*, LVI (1988), 1-36.

Cox, James C., Bruce Roberson, and Vernon L. Smith, "Theory and Behavior of Single Object Auctions," *Research in Experimental Economics* (Greenwich, CT: JAI Press, 1982).

Davis, Douglas D., and Charles A. Holt, *Experimental Economics* (Princeton, NJ: Princeton University Press, 1993).

Erev, I., and Alvin R. Roth, "Learning in Extensive-Form Games: Experimental Data and Simple Dynamic Model in the Intermediate Term", *Games and Economic Behavior* VIII (1995), 164-212.

Fudenberg, Drew, and David Kreps, "A Theory of Learning, Experimentation and Equilibrium in Games," mimeo, MIT, 1988.

Fudenberg, Drew, and David K. Levine, "Self-Confirming Equilibrium," *Econometrica*, LXI (1993a), 547-573.

Fudenberg, Drew, and David K. Levine, "How Large Are Players' Losses In Extensive Form Games?," mimeo, UCLA, 1995.

Fudenberg, Drew, and Jean Tirole, *Game Theory* (Cambridge, MA: MIT Press, 1991).

Guth, W., and R. Tietz, "Ultimatum Bargaining for a Shrinking Cake: An Experimental Analysis," in R. Tietz, W. Albers and R. Selten, eds., *Bounded Rational Behavior in Experimental Games and Markets (*Berlin, Germany: Springer, 1988).

Guth, W., and R. Tietz, "Ultimatum Bargaining Behavior: A Survey and Comparison of Results," *Journal of Economic Psychology*, +XI (1990), 417- 449.

Guth, W., P. Ockenfels, and R. Tietz, "Distributive Justice Versus Bargaining Power: Some Experimental Results," *Frankfurter Arbeiten zur Experimentellen Wirtschaftsforschung,* (1990).

Guth, W., R. Schmittberger, and B. Schwartz, "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization*, III (1982), 367-388.

Harrison, Glen W., "Theory and Misbehavior in First-Price Auctions," *American Economic Review*, LXXIX (1989), 749-762.

Harrison, Glen W., "Rational Expectations and Experimental Methods," in Gross, B.A. (ed.), *Rational Expectations and Efficiency in Futures Markets* (London: Routledge, 1991).

Harrison, Glen W., "First-Price Auctions: Reply," *American Economic Review*, LXXXII (1992), 1426-1443.

Harrison, Glen W., and Jack Hirshleifer, "An Experimental Evaluation of Weakest Link/Best Shot Models of Public Goods," *Journal of Political Economy,* XCVII (1989), 201-225.

Harrison, Glen W., and K. McCabe, "Testing Noncooperative Bargaining Theory in Experiments," in Isaac, R.M. (ed.), *Research in Experimental Economics* (Greenwich, CT: JAI Press, 1992).

Hey, J.D., and Chris Orme, "Investigating Generalizations of Expected Utility Theory Using Experimental Data," *Econometrica*, LXII (1994), 1291-1396.

Isaac, R. Mark, and James M. Walker, "Group Size Effects in Public Goods Provision: The Voluntary Contributions Mechanism," *Quarterly Journal of Economics*, CIII (1988), 179-200.

Kreps, David, and Bob Wilson, "Reputation and Imperfect Information," *Journal of Economic Theory,* L (1982), 253-79.

McKelvey, Richard, and Thomas Palfrey, "An Experimental Study of the Centipede Game," *Econometrica*, LX (1992), 803-836.

Majure, R., "Fitting Learning and Evolution Models to Experimental Data," mimeo, in *Equilibrium Game Theory*, unpublished Ph.D. dissertation, MIT, 1994.

Milgrom, Paul, and Johyn Roberts, "Predation, Reputation and Entry Deterrence," *Econometrica*, L (1982), 443-60.

Mookerjhee, D., and B. Sopher, "Learning Behavior in an Experimental Matching Pennies Game," *Games and Economic Behavior*, VII (1994), 62-91.

Ochs, J., "Games with Unique Mixed Strategy Equilibria: An Experimental Study," *Games and Economic Behavior*, X (1994), 202-217.

Ochs, J., and Alvin E. Roth, "An Experimental Study of Sequential Bargaining," *American Economic Review*, LXXIX (1989), 355-384.

O'Neill, B., "Nonmetric Test of the Minimax Theory of Two-Person Zerosum Game," *Proceedings of the National Academy of Sciences*, U.S.A., LXXXIV (1987), 2106-2109.

Prasnikar, V., and Alvin E. Roth, "Considerations of Fairness and Strategy: Experimental Data from Sequential Games," *Quarterly Journal of Economics*, CVII (1992), 865-888.

Radner, Roy, "Collusive Behavior in Non-cooperative Epsilon Equilibria of Oligopolies with Long but Finite Lives," *Journal of Economic Theory*, XXII (1980), 136-154.

Rapoport, Amnon and M. A. Fuller “Bidding Strategies in a Bilateral Monopoly with Two-Sided Incomplete Information,” *Journal of Mathematical Psychology, XXXIX* (1995), 179-176.

Roth, Alvin E., and Michael W.K. Malouf, "Game-Theoretic Models and the Role of Information in Bargaining," *Psychological Review*, LXXXVI (1979), 803- 836.

Roth, Alvin E., V. Prasnikar, M. Okuno-Figiware, and S. Zamir, "Bargaining and Market Behavior in Jerusalem, Liubljana, Pittsburgh, and Tokyo: An Experimental Study," *American Economic Review*, LXXXI (1991), 1068-1095.

Roth, Alvin E. and Francis Schoumaker (1983) " Expectations and Reputations in Bargaining: An Experimental Study", American Economic Review, LXXIII, 1983, 362-372.

Selten, Reinhardt, "Die Strategiemethode zur Erforschung des Eingeschränkten Rationalen Verhaltens im Rahmen eines Oligopolexperiments," in *Beiträge zur Experimentellen Wirtschaftsforschung*, H. Sauermann, ed. (Tübingen, Germany: JCB Mohr, 1967, 136-168).

Selten, Reinhardt, "Spieltheoretische Behandlung eines Oligopmodells mit Nachfrageträgheit," *Z. Ges. Staatswiss.*, CXXI (1965), 301-324.

Notes

1. Some experiments have required subjects to prespecify complete contingent strategies, as for example, Selten [1967]. This experimental design is not widely used, perhaps because games rarely present themselves this way in practice. Also, some experiments have asked players to report their beliefs about the opponents' play, either at the time of play or *ex post;* see Harrison [1991] for a review.
2. The exception was the full-information treatment of the best-shot game, where the two losses were almost identical because play closely resembled that of a Nash equilibrium. The best-shot game has the interesting property that set of approximate self-confirming equilibria is quite small However, this fact on its own does not imply that the Nash and self-confirming losses are similar, for in the partial information treatment of the game the Nash losses were again about four times as large as the self-confirming ones.
3. Based on the earlier work of Cox, Smith and Walker [1985].
4. See however Harrison and McCabe [1992] and Roth and Schoumaker [1983] for experiments designed to control the homemade priors. Harrison and McCabe's design showed that giving players in a three-stage bargaining game experience playing the subgame corresponding to the last two stages resulted in outcomes more like the subgame-perfect equilibria. This can be interpreted as showing that the divergence of the outcomes from subgame-perfection when players are not given this experience is due to their having incorrect (but self-confirming) beliefs about off-path play.
5. Note that the fact that beliefs are correct forces all players to share the same (correct) beliefs, even though the notation allows each player to have different beliefs.
6. Note that the independence condition is moot in two-player games.

7. The random-matching design avoids the "repeated game" effects that can arise if the same individuals face each other in subsequent rounds.

8. On the other hand, we would expect all players to eventually have the same beliefs if they observe the aggregate distribution of outcomes in the whole population. This observation condition has been used in some experiments, see Camerer and Weigelt [1988].

9. Notice in this example the heterogeneous self-confirming equilibrium is equivalent to a public randomization over Nash equilibria. This can be shown to be the case generally in games of perfect information. However, Fudenberg and Levine [1993a] give an example of a two-player two-period game in which an action is played with positive probability in a self-confirming equilibrium that is not played in any Nash or indeed even correlated equilibrium.

10. The prevalence of this practice among experimental economists suggests that they tend to subscribe to learning or some other adaptive process as the explanation for equilibrium, as opposed to explanations based on common knowledge of rationality.

11. In the sequel, our presumption will be that every player uses a pure strategy, and that the distribution of play arises because different individuals use different strategies. See Ochs [1994] for an attempt to test if subjects will use "mixed" (actually interior) strategies when asked to choose the proportion of time they will use each action over the next 10 rounds.

12. Note that ε-equilibria may look very different than exact equilibria, even for small ε: see for example Radner's (1980) work on finite repeated oligopoly and the work of the gang of four (Kreps and Wilson (1982), Milgrom and Roberts (1982)) on reputation.

13. This use of dollar losses as a metric is common in the literature on market experiments; see the discussion in Davis and Holt and the references cited there. Davis and Holt also discuss experimental designs intended to control for risk aversion

(such as Roth and Malouf [1979]) and designs intended to measure the preference for fairness as opposed to other concerns in certain bargaining games.

14. Rapoport and Fuller [1995], Hey and Orme [1994], Stahl and Wilson [1994,1995], Brown and Rosenthal [1990], O'Neill [1987], Mookerjhee and Sopher [1994], Crawford [1995], Majure [1994] and McKelvey and Palfrey [1992] are examples of such studies.
15. Our impression is that individual play exhibits some of the same inconsistencies with our theory that it does with more standard theory in cases in which the aggregate distribution does resemble a Nash equilibrium.
16. See footnote 1.
17. In the two-player case Nash and unitary self-confirming equilibria are observationally equivalent (Fudenberg and Levine [1993a]) so this results in exactly the same calculation as in the unitary case, and throughout this paper we consider only two player games. In games with three or more players there is a significant complication: Pairs of players are constrained to agree about the off path behavior of a third player, which can imply that the losses attributed to the various players are linked in a complicated way that we do not know how to handle. Fortunately, there is a large class of games called games with identified deviators (Fudenberg and Levine [1993a]), in which players cannot disagree in a meaningful way about the strategy followed by a third player.
18. A formal proof was given in an earlier draft of this paper.
19. If the final node is reached with positive probability player 2 drops out. This implies that if the next to last node is reached with positive probability and player 1 stays in he will find out that player 2 is dropping out. Hence, player 1 must drop out if the next to last node is reached with positive probability, implying the final node is not reached, and so forth.
20. The payoff to P_3 is . $0.49*\$0.20+0.51*0.82*\$0.80+0.51*0.18*\$6.40=\$1.02$.

21. Note that the loss reported of $0.37 is the expected loss using the strategy P_4; player 1’s play is such that there is only a 23 percent chance of reaching the final round, so the expected loss is 0.23x$1.60=$0.37.

22. Detailed information about the play of every player in every game can be found in the Appendix to McKelvey and Palfrey

23. Unlike the worst case there is not an unambiguous way to define “completely random” play. One alternative is the behavior strategy that, at each information set, assigns equal weight to all feasible actions. In the centipede game this corresponds to a ½ chance of dropping out at the start, and ¼ each for the other two pure strategies. In the two other experiments we consider each player has only one information set on any path of play so the two versions of “completely random” coincide. In centipede, a 50-50 randomization at each information set means that we will even more rarely see money given away at the end of the game, so the losses would be even smaller than reported here. Since the stakes rise so rapidly that it is always worth staying in for a period in exchange for a 50 percent chance of a gift next period, and is never a knowing mistake to drop out too early, if we extended the number of rounds of centipede, we could drive the loss from this type of random play to zero. This is just another way of saying that the approximate equilibrium set in centipede is large enough to include random play; the fact that the worst case losses are so much greater than the observed losses, indicates that there are other strategies that are not approximate equilibria.

24. This last fact - the large set of approximate self-confirming equilibria is due to the sensitivity of the equilibrium to the play of a small fraction of players at the final round.

25. In response to this McKelvey and Palfrey also estimated a model in which the prior beliefs of player 1 are random, and the two players’ beliefs are not consistent with a

common prior. Relaxing the common prior assumption is in some ways similar to allowing for heterogeneous beliefs.

26. Harrison and Hirshleifer ran experiments on both the sequential move game we discuss and its simultaneous-move analog. References in the literature to the "best-shot game" are to the sequential-move version of the game.

27. As an aside, let us emphasize a distribution of outcomes whose support consists entirely of Nash outcomes need not itself be consistent with Nash equilibrium. Thus the percentage of observed outcomes consistent with some Nash equilibrium, which is reported as a summary statistic in some analyses of game-theory experiments, cannot be grounded in theories that predict Nash equilibria.

28. This is confirmed by detailed information on the experimental results provided to us by Harrison and Hirshleifer

29. When compared to the centipede case.

30. Notice that strictly speaking this is not the same as a \$0.08-self-confirming equilibrium, although we loosely refer to it as such. In a \$0.08-self-confirming equilibrium neither player can have an expected loss of more than \$0.08. Here we allow one player to have a \$0.16 loss provided the other player has no loss.

31. Note, moreover, that apart from the number of choices available to player 2, the best-shot and ultimatum games have the same tree, and differ only in their payoffs. Moreover, as noted by Prasnikar and Roth, subgame-perfect equilibrium predicts very unequal (hence "unfair") payoffs in both games, which makes the dissimilar experimental results all the more interesting..

32. In this game our maintained assumption is that the empirical distribution of responses exactly equals the true one is particularly inappropriate. An alternative approach, suggested by David Kreps, would be to suppose that each player 2 is playing a cut-off strategy, and use the observed data to estimate the distribution of cutoffs in the population. We could then compute the payoff-maximizing offer against that

estimated distribution, and use the associated payoff as our benchmark for measuring the unitary losses.

[33] See footnote 21.

Player i	Pure Strategy s_i	Unitary $\varepsilon_i(s_i, J(\cdot), \rho)$	Heterogeneous $\varepsilon_i(s_i, J(\cdot), \rho)$	Frequency
1	T_1	$0.62	$0.00	.08
1	T_3	$0.11	$0.00	.69
1	P_3	$0.00	$0.00	.23
2	T_2	$0.28	$0.00	.49
2	T_4	$0.00	$0.00	.42
2	P_4	$0.37	$0.37	.09

Table I

Outcomes and Losses in the Centipede Game Base Case

Trials/	Rnds	Stake	Case	Expected Loss			Max	Ratio
Rnd				Pl 1	Pl 2	Both	Gain	
29*	6-10	1x	H	$0.00	$0.03	$0.02	$4.00	0.4%
29*	6-10	1x	U	$0.12	$0.17	$0.15	$4.00	4%
	WC	1x	H			$0.80	$4.00	20%
	Random	1x	H	$0.00	$0.05	$0.03	$4.00	0.6%
29	1-10	1x	H	$0.00	$0.08	$0.04	$4.00	1.0%
10	1-10	4x	H	$0.00	$0.28	$0.14	$16.00	0.9%

Rnds = Rounds, WC = Worst Case, H = Heterogeneous, U = Unitary

*The data from which this case is computed is reported above.

Table II

Summary of Losses in the Centipede Experiments

x	*W(x)*	*C(x)*
0	$0.00	$0.00
1	$1.00	$0.82
2	$1.95	$1.64
3	$2.85	$2.46
4	$3.70	$3.28
5	$4.50	$4.10
6	$5.25	$4.92
7	$5.95	$5.74
8	$6.60	$6.50

Table III

Payoffs in the Best Shot Game

Trials	Rnds	Info	Case	Expected Loss			Max	Ratio
				Pl 1	Pl 2	Both	Gain	
8	8-10	full	H	$0.00	$0.12	$0.06	$2.06	2.9%
8	8-10	full	U	$0.00	$0.12	$0.06	$2.06	2.9%
10	8-10	part	H	$0.01	$0.15	$0.08	$2.06	3.9%
10	8-10	part	U	$0.39	$0.15	$0.27	$2.06	13.0%
	WC		H			$3.41	$2.06	165%
	Random		H	$0.16	$2.10	$1.18	$2.06	57%

Rnds=Rounds, WC=Worst Case, H=Heterogeneous, U=Unitary

Table IV

Summary of Losses in the Best Shot Game

		Player 2 contribution								
		0	*1*	*2*	*3*	*4*	*5*	*6*	*7*	*8*
Player	*0*	0.38	0.67	1.00	1.00	1.00	1.00	1.00	0.84	0.50
1	*1*	0.67	0.31	0.10	0.10	0.10	0.10	0.10	0.10	0.10
contrib.	*2*	1.00	0.31	0.10	0.10	0.10	0.10	0.10	0.10	0.10
	3	1.00	0.31	0.10	0.10	0.10	0.10	0.10	0.10	0.10
	4	1.00	0.31	0.10	0.10	0.10	0.10	0.10	0.10	0.10
	5	1.00	0.31	0.10	0.10	0.10	0.10	0.10	0.10	0.10
	6	1.00	0.31	0.10	0.10	0.10	0.10	0.10	0.10	0.10
	7	1.00	0.31	0.10	0.10	0.10	0.10	0.10	0.10	0.10
	8	0.84	0.31	0.10	0.10	0.10	0.10	0.10	0.10	0.10

Table V

Frequency Bounds on Approximate Equilibrium for Best Shot

Trials	Rnd	Cntry	Case	Expected Loss			Max	Ratio
		Stake		Pl 1	Pl 2	Both	Gain	
27	10	US	H	$0.00	$0.67	$0.34	$10.00	3.4%
27	10	US	U	$1.30	$0.67	$0.99	$10.00	9.9%
10	10	USx3	H	$0.00	$1.28	$0.64	$30.00	2.1%
10	10	USx3	U	$6.45	$1.28	$3.86	$30.00	12.9%
30	10	Yugo	H	$0.00	$0.99	$0.50	$10?	5.0%
30	10	Yugo	U	$1.57	$0.99	$1.28	$10?	12.8%
29	10	Jpn	H	$0.00	$0.53	$0.27	$10?	2.7%
29	10	Jpn	U	$1.85	$0.53	$1.19	$10?	11.9%
30	10	Isrl	H	$0.00	$0.38	$0.19	$10?	1.9%
30	10	Isrl	U	$3.16	$0.38	$1.77	$10?	17.7%
	WC		H			$5.00	$10.00	50.0%
	Random		H	$0.00	$2.50	$1.25	$10.00	12.5%

Rnds=Rounds, WC=Worst Case, H=Heterogeneous, U=Unitary

Table VI

Summary of Losses in Ultimatum Bargaining

x	*Offers*	*Rejection Probability*
$2.00	1	100%
$3.25	2	50%
$4.00	7	14%
$4.25	1	0%
$4.50	2	100%
$4.75	1	0%
$5.00	13	0%
Total	27	100%

TABLE VII

Rejection Probabilities in US $10.00 Stake Games Round 10

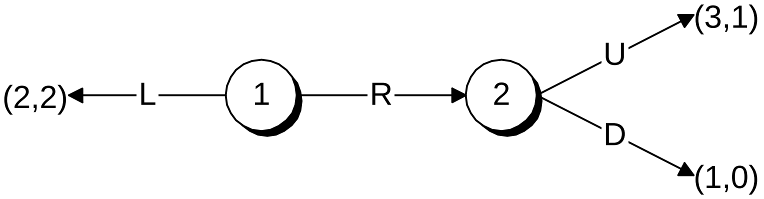


Figure I

Selten Game Used to Illustrate Self-Confirming Equilibrium

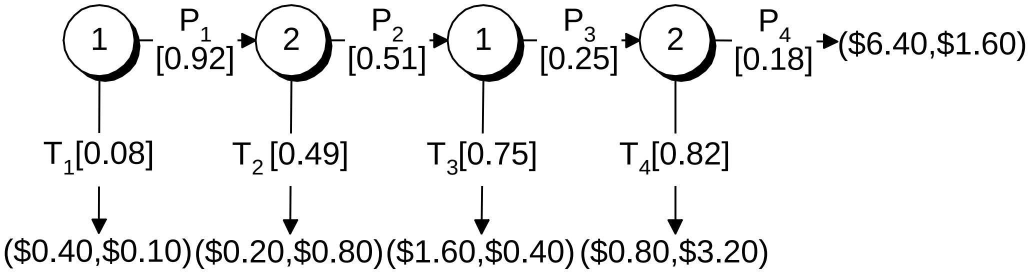


Figure II

Palfrey and McKelvey’s Centipede Game:

Numbers in square brackets correspond to the observed conditional probabilities of play at each information set in rounds 6-10, stakes 1x.

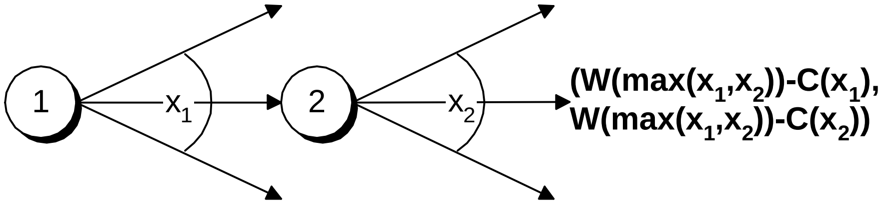


Figure III

Extensive Form for Best Shot

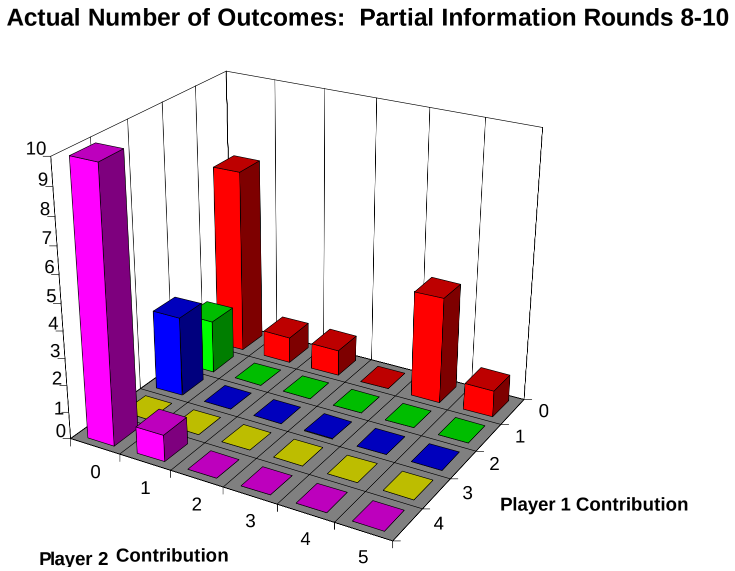


Figure IV

Observed Outcomes in Best Shot

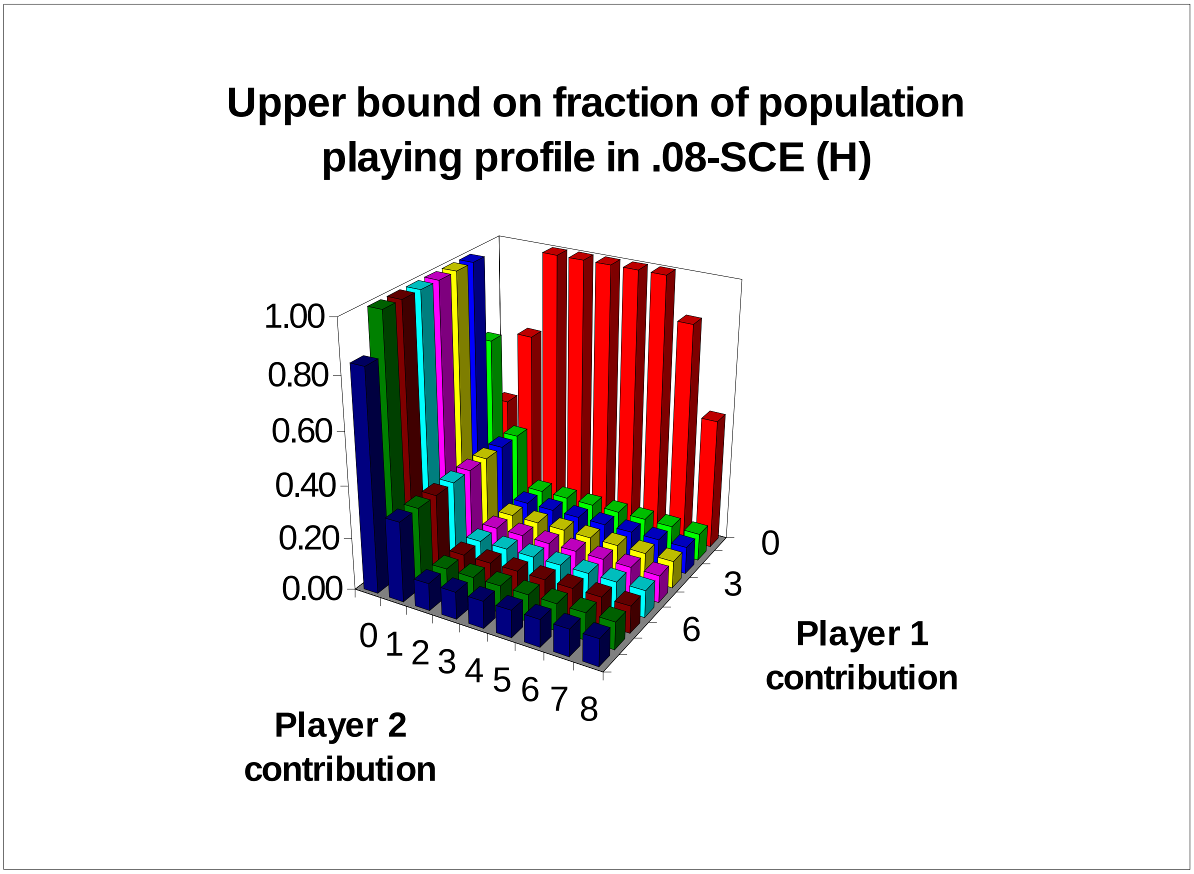


Figure V

Theoretical Probability Bounds in Best Shot

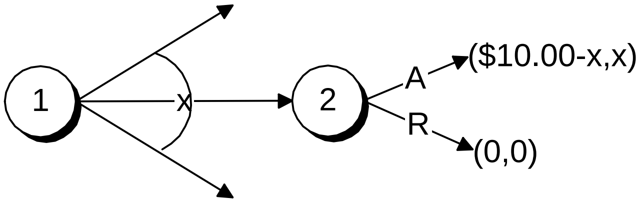
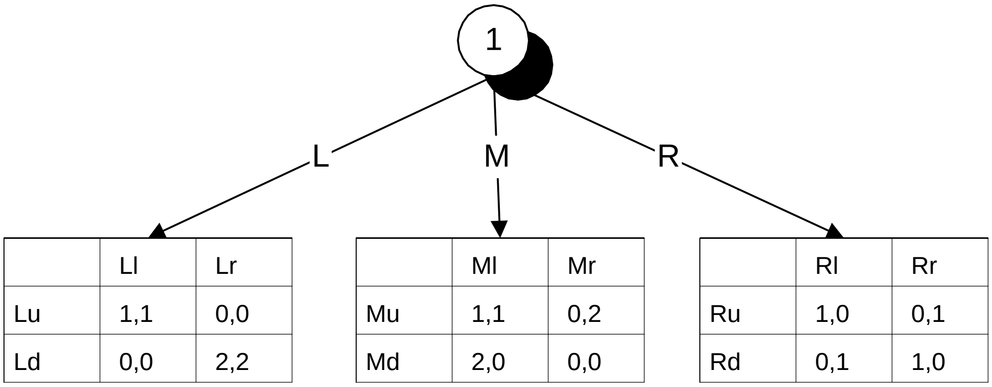


Figure VI

Extensive Form for Ultimatum Bargaining



	Ll	Lr
Lu	1,1	0,0
Ld	0,0	2,2

	Ml	Mr
Mu	1,1	0,2
Md	2,0	0,0

	Rl	Rr
Ru	1,0	0,1
Rd	0,1	1,0

Figure VII

Game Used to Illustrate Use of Isolated Subgames to Compute Losses

Notes

[1] Some experiments have required subjects to prespecify complete contingent strategies, as for example, Selten [1967]. This experimental design is not widely used, perhaps because games rarely present themselves this way in practice. Also, some experiments have asked players to report their beliefs about the opponents' play, either at the time of play or *ex post;* see Harrison [1991] for a review.

[2] The exception was the full-information treatment of the best-shot game, where the two losses were almost identical because play closely resembled that of a Nash equilibrium. The best-shot game has the interesting property that set of approximate self-confirming equilibria is quite small However, this fact on its own does not imply that the Nash and self-confirming losses are similar, for in the partial information treatment of the game the Nash losses were again about four times as large as the self-confirming ones.

[3] Based on the earlier work of Cox, Smith and Walker [1985].

[4] See however Harrison and McCabe [1992] and Roth and Schoumaker [1983] for experiments designed to control the homemade priors. Harrison and McCabe's design showed that giving players in a three-stage bargaining game experience playing the subgame corresponding to the last two stages resulted in outcomes more like the subgame-perfect equilibria. This can be interpreted as showing that the divergence of the outcomes from subgame-perfection when players are not given this experience is due to their having incorrect (but self-confirming) beliefs about off-path play.

[5] Note that the fact that beliefs are correct forces all players to share the same (correct) beliefs, even though the notation allows each player to have different beliefs.

[6] Note that the independence condition is moot in two-player games.

[7] The random-matching design avoids the "repeated game" effects that can arise if the same individuals face each other in subsequent rounds.

[8] On the other hand, we would expect all players to eventually have the same beliefs if they observe the aggregate distribution of outcomes in the whole population. This observation condition has been used in some experiments, see Camerer and Weigelt [1988].

[9] Notice in this example the heterogeneous self-confirming equilibrium is equivalent to a public randomization over Nash equilibria. This can be shown to be the case generally in games of perfect information. However, Fudenberg and Levine [1993a] give an example of a two-player two-period game in which an action is played with positive probability in a self-confirming equilibrium that is not played in any Nash or indeed even correlated equilibrium.

[10] The prevalence of this practice among experimental economists suggests that they tend to subscribe to learning or some other adaptive process as the explanation for equilibrium, as opposed to explanations based on common knowledge of rationality.

[11] In the sequel, our presumption will be that every player uses a pure strategy, and that the distribution of play arises because different individuals use different strategies. See Ochs [1994] for an attempt to test if subjects will use "mixed" (actually interior) strategies when asked to choose the proportion of time they will use each action over the next 10 rounds.

[12]. Note that ε-equilibria may look very different than exact equilibria, even for small ε: see for example Radner's (1980) work on finite repeated oligopoly and the work of the gang of four (Kreps and Wilson (1982), Milgrom and Roberts (1982)) on reputation.

[13] This use of dollar losses as a metric is common in the literature on market experiments; see the discussion in Davis and Holt and the references cited there. Davis and Holt also discuss experimental designs intended to control for risk aversion (such as Roth and Malouf [1979]) and designs intended to measure the preference for fairness as opposed to other concerns in certain bargaining games.

[14] Rapoport and Fuller [1995], Hey and Orme [1994], Stahl and Wilson [1994,1995], Brown and Rosenthal [1990], O'Neill [1987], Mookerjhee and Sopher [1994], Crawford [1995], Majure [1994] and McKelvey and Palfrey [1992] are examples of such studies.

[15] Our impression is that individual play exhibits some of the same inconsistencies with our theory that it does with more standard theory in cases in which the aggregate distribution does resemble a Nash equilibrium.

[16] See footnote 1.

[17] In the two-player case Nash and unitary self-confirming equilibria are observationally equivalent (Fudenberg and Levine [1993a]) so this results in exactly the same calculation as in the unitary case, and throughout this paper we consider only two player games. In games with three or more players there is a significant complication: Pairs of players are constrained to agree about the off path behavior of a third player, which can imply that the losses attributed to the various players are linked in a complicated way that we do not know how to handle. Fortunately, there is a large class of games called games with identified deviators (Fudenberg and Levine [1993a]), in which players cannot disagree in a meaningful way about the strategy followed by a third player.

[18] A formal proof was given in an earlier draft of this paper.

[19]If the final node is reached with positive probability player 2 drops out. This implies that if the next to last node is reached with positive probability and player 1 stays in he will find out that player 2 is dropping out. Hence, player 1 must drop out if the next to last node is reached with positive probability, implying the final node is not reached, and so forth.

[20] The payoff to P_3 is . $0.49*\$0.20+0.51*0.82*\$0.80+0.51*0.18*\$6.40=\$1.02$.

[21] Note that the loss reported of \$0.37 is the expected loss using the strategy P_4; player 1's play is such that there is only a 23 percent chance of reaching the final round, so the expected loss is 0.23x\$1.60=\$0.37.

[22].Detailed information about the play of every player in every game can be found in the Appendix to McKelvey and Palfrey

[23] Unlike the worst case there is not an unambiguous way to define "completely random" play. One alternative is the behavior strategy that, at each information set, assigns equal weight to all feasible actions. In the centipede game this corresponds to a ½ chance of dropping out at the start, and ¼ each for the other two pure strategies. In the two other experiments we consider each player has only one information set on any path of play so the two versions of "completely random" coincide. In centipede, a 50-50 randomization at each information set means that we will even more rarely see money given away at the end of the game, so the losses would be even smaller than reported here. Since the stakes rise so rapidly that it is always worth staying in for a period in exchange for a 50 percent chance of a gift next period, and is never a knowing mistake to drop out too early, if we extended the number of rounds of centipede, we could drive the loss from this type of random play to zero. This is just another way of saying that the approximate equilibrium set in centipede is large enough to include random play; the fact that the worst case losses are so much greater than the observed losses, indicates that there are other strategies that are not approximate equilibria.

[24] This last fact - the large set of approximate self-confirming equilibria is due to the sensitivity of the equilibrium to the play of a small fraction of players at the final round.

[25] In response to this McKelvey and Palfrey also estimated a model in which the prior beliefs of player 1 are random, and the two players' beliefs are not consistent with a common prior. Relaxing the common prior assumption is in some ways similar to allowing for heterogeneous beliefs.

[26] Harrison and Hirshleifer ran experiments on both the sequential move game we discuss and its simultaneous-move analog. References in the literature to the "best-shot game" are to the sequential-move version of the game.

[27] As an aside, let us emphasize a distribution of outcomes whose support consists entirely of Nash outcomes need not itself be consistent with Nash equilibrium. Thus the percentage of observed outcomes consistent with some Nash equilibrium, which is reported as a summary statistic in some analyses of game-theory experiments, cannot be grounded in theories that predict Nash equilibria.

[28] This is confirmed by detailed information on the experimental results provided to us by Harrison and Hirshleifer

[29]When compared to the centipede case.

[30]Notice that strictly speaking this is not the same as a $0.08-self-confirming equilibrium, although we loosely refer to it as such. In a $0.08-self-confirming equilibrium neither player can have an expected loss of more than $0.08. Here we allow one player to have a $0.16 loss provided the other player has no loss.

[31] Note, moreover, that apart from the number of choices available to player 2, the best-shot and ultimatum games have the same tree, and differ only in their payoffs. Moreover, as noted by Prasnikar and Roth, subgame-perfect equilibrium predicts very unequal (hence "unfair") payoffs in both games, which makes the dissimilar experimental results all the more interesting..

[32] In this game our maintained assumption is that the empirical distribution of responses exactly equals the true one is particularly inappropriate. An alternative approach, suggested by David Kreps, would be to suppose that each player 2 is playing a cut-off strategy, and use the observed data to estimate the distribution of cutoffs in the population. We could then compute the payoff-maximizing offer against that estimated distribution, and use the associated payoff as our benchmark for measuring the unitary losses.

[33] See footnote 21.