

***Friend or Foe?* Coordination, Cooperation, and Learning in High-Stakes Games**

Felix Oberholzer-Gee
Harvard Business School

Joel Waldfogel
The Wharton School and NBER

Matthew W. White
The Wharton School and NBER

July 7, 2004~~July 6, 2004~~

Abstract

Why do people frequently cooperate in defiance of their immediate incentives? One recent explanation is that individuals are *conditionally cooperative*: They prefer to cooperate with cooperative persons but would rather punish those who are not. As an explanation of behavior in one-shot settings, such preferences require individuals to be able to discern their opponents' preferences prior to play. Using data on contestant play from two seasons of a television game show, we provide evidence about how individuals implement conditionally cooperative preferences. We show (1) that contestants forgo large sums of money to be cooperative, and (2) that players with some historical basis for predicting their opponents' type cooperate at heightened levels only when both they and their opponent are predictably cooperative.

JEL: H41, K42, A13, C93

We are grateful to Mary Benner and Hannah Waldfogel for introducing us to *Friend or Foe*, and to Melanie Haw and Sarah Waldfogel for spirited research assistance. Gary Bolton, Rachel Croson, Daniel Kessler, Peter Zemsky, and seminar participants at Pennsylvania State University and Wharton provided useful and thought-provoking comments.

I. Introduction

Why is there order? Why do people frequently cooperate in defiance of their immediate incentives? One class of explanations involves legal sanctions (Becker, 1968). Alternatively, “order without law” can come about when individuals have incentives to invest in reputation in the context of repeated games (Kreps and Wilson, 1982; Tirole, 1988).¹ More recently, theorists and experimentalists have turned their attention to norms as explanations of cooperative behavior. Rabin (1993) posits that individuals have conditionally cooperative preferences – they would like to cooperate with those who are cooperative but would like to punish those who are not. In theory, models with conditionally cooperative players explain behavior in a fairly wide range of games (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002). As an explanation of order outside repeated games, conditionally cooperative preferences require individuals not only to prefer “fair” divisions but also to be able to discern their opponents’ types prior to play. What’s missing from the literature is evidence about how individuals implement such preferences in one-shot games. This paper attempts to fill that gap using data from a television game show.

In June 2002, the Game Show Network began airing a cable television show in which contestants play a high-stakes, one-shot game called *Friend or Foe*. In this game, each of two players simultaneously choose whether to play ‘friend’ or ‘foe’. Each player’s payoff depends on the action chosen by the other contestant in the following way:

		<i>Player 2</i>	
		Friend	Foe
<i>Player 1</i>	Friend	$x/2, x/2$	$0, x$
	Foe	$x, 0$	$0, 0$

Figure 1: The *Friend or Foe* game, $x > 0$.

¹ Ellickson’s (1991) work coins the term “order without law.”

This game is a variant of the classic Prisoner’s Dilemma, and is similar in structure to games analyzed previously in experimental studies.² Here playing foe is a weakly-dominant strategy for each player. In contrast to the prior literature, however, the stakes in *Friend or Foe* are astonishingly high.³ The payoff x ranges from \$200 to over \$16,000, with an average of \$3,300 at stake per game.

We are able to use a feature of the game’s production history to both demonstrate that players have conditionally cooperative preferences as well as to illuminate how players implement those preferences. The show was filmed in two “seasons,” with the first 40 episodes produced before the show’s on-air debut. The remaining 65 episodes were taped after the airing of the first season. Players on the first season therefore had little show-related basis for forming beliefs about opponent play, while players in the second season could observe the play in 120 prior games. In this respect, *Friend or Foe* can be viewed as a recurring game with a sequence of non-overlapping generations of players.⁴

We document a number of interesting findings. First, in this game, individuals choose friend at a remarkably high rate, even at very high stakes. In 315 games, 55 percent of players choose friend when the stakes are under \$3,000, and 54 percent do so when the stakes exceed \$3,000. It remains 55 percent when the stakes exceed \$5,000. The evidence from *Friend or Foe* is that players’ cooperative tendencies are surprisingly *stakes invariant*. Frequent cooperative play despite payoffs of this size strongly reinforces the experimental finding that many people simply prefer to cooperate.

Second, during the game’s first season a player’s choice is statistically independent of his or her opponent’s, but this choice varies systematically with the player’s observable characteristics. Subsequent players therefore have some basis for predicting how preferences for

² For surveys, see Ledyard (1995) and Laury and Holt (forthcoming). Analyses of data from television game shows also has numerous precedents. Gertner (1993) examines attitudes toward risk on *Card Sharks*, Metrick (1995) studies betting behavior on *Jeopardy!*, and Berk, Hughston, and Vandevande (1996) examine learning and bounded rationality on *The Price is Right*. List (2003) uses data from the first 40 episodes of *Friend or Foe* to draw inferences about discrimination.

³ For reasons of cost, most high-stakes experiments have been undertaken in low-income countries. The evidence in this paper is (to our knowledge) the first to study subjects from an advanced Western economy play such a high-stakes one-shot game.

⁴ This structure, and our analysis of players’ behavior based on learning unknown type distributions in the population (see Section IV), are similar to the recurring games in Jackson and Kalai (1997).

cooperation vary. If some players have conditionally cooperative preferences, then in the second season friendly play should occur at elevated levels only when both player in a game have observable characteristics associated with cooperativeness. We document this using the change in play for “friendly” players when paired with “friendly” vs. “unfriendly” opponents. For example, we find a large increase in foe rates between Season 1 and 2 for (presumably cooperative) women paired with (presumably uncooperative) men, yet no change in foe rates for women paired with other women. We find similar patterns by age and race. As a result, the proportion of games with split outcomes – when one chooses friend and the other foe – is substantially lower in the second season. Players learn to “coordinate” on outcomes on the main diagonal of the game in Figure 1, despite the simultaneous-move nature of play.

If second-season players learn to condition their strategies on their opponents’ observable characteristics, player types who are expected to cooperate will come to fare better (monetarily) over time, relative to players without observable characteristics initially associated with cooperation. The changes in take-home winnings over time in the data are highly consistent with this implication. In essence, the players ‘stereotyped’ by observable characteristics associated with uncooperative play in early generations are shunned by later opponents (who play foe against them). We document that such stereotyped groups fare progressively worse monetarily.

The paper proceeds in three sections. The next section presents a simple model of preferences and learning to organize our analysis of the game. Section 3 describes the game context and the data. Section 4 then presents the major empirical results with respect to stakes, learning, and coordination. A brief summary concludes.

II. Theoretical Background

A. Fairness and the Augmented Game

Given the high overall rate of friendly play and stakes invariance reported in the introduction, we are led to consider player motivations beyond the payoffs in Figure 1 alone. Rabin (1993) argues, in essence, that people want to be nice to those who treat them fairly and want to punish those who hurt them. The fairness of actions depends on the players’ inten-

tions, which can be inferred from the distribution of payoffs that these actions induce. One implication of Rabin’s theory is that contestants who expect their opponents to choose foe might prefer to punish their partner by destroying the endowment. Many experiments show that people are willing to punish unfriendly play, even if punishment is costly and does not affect future play (Güth, Schmittberger and Schwarze, 1982; Fehr and Gächter, 2000).

Following the recent literature (Fehr and Schmidt, 1999; Bolton and Ockenfels 2000; Charness and Rabin, 2002), this section presents a simple model of player preferences and learning in our intergenerational setting. This ‘augmented game’ provides a coherent framework for understanding the evolution of play that we observe in the data. We assume that a player’s preferences reflect non-monetary considerations that depend on the friendliness of the opponent’s play in the following manner:

		<i>Player j</i>	
		Friend	Foe
<i>Player i</i>	Friend	$x/2, x/2$	$-S_i, x - G_j$
	Foe	$x - G_i, -S_j$	$0, 0$

Figure 3: The augmented game.

The term $S_i > 0$, for *sucker’s dismay*, motivates a player to destroy the entire endowment if he or she believes the opponent will try to grab the entire pie. The term G_i , for *guilt*, captures feelings of guilt or shame for having played foe when the partner played friend.⁵ The non-monetary terms in this augmented game, S_i and G_i , thus reflect unobservable heterogeneity in preferences.

Strategies in this augmented game are as follows: If p_i denotes player i ’s belief that his or her opponent will play foe, then player i prefers to play friend if and only if

⁵ On the show, this type of embarrassment is frequently observed. Many apologize for having chosen foe when the other contestant was friendly. Some explain they really needed the money, while others say they chose foe only because they thought—incorrectly—that the other player would choose foe.

$$G_i - S_i \cdot \lambda_i > x / 2$$

where $\lambda_i = p_i / (1-p_i)$ is the foe/friend odds ratio. That is, a player chooses friend if her guilt from taking the entire stakes exceeds $x/2$ by a multiple of the dismay if her opponent does so; the potential dismay weighs more heavily in the decision as the prior on an opponent playing foe becomes larger.

In this setting, we distinguish between two ‘types’ of players. Given a game at stakes x , players with $G_i > x / 2$ are *conditional cooperators*. For such types, there exists a set of beliefs about the likelihood of an opponent choosing foe for which it is optimal to also play foe. For a sufficiently low foe-prior p_i , however, i will prefer to play friend. That is, a conditional cooperator prefers to play friend against an opponent she believes is (sufficiently) likely to also play friend, but prefers to meet foe with foe. The second type of player that is relevant to understanding the augmented game has $G_i < x / 2$, which is a lower level of guilt from taking the entire pie than a conditional cooperator has. Such players have a dominant strategy of playing foe in the augmented game in Figure 2, assuming that all players have $S_i > 0$ (that is, no one is truly indifferent to being played the sucker in this environment).

Before proceeding further, it is worth noting what conditionally cooperative types have at stake empirically in *Friend or Foe*. Although G_i and S_i are not observable directly, information on the former can be inferred for a sizeable share of the population based on observed play. In the data, 45% of players choose friend; for these players, the median stakes x is approximately \$2,700. Thus for nearly half of the 630 players, the money-metric “cost” of playing foe against a possibly friend-playing partner—a cost we interpret as guilt, or shame—*must be upwards of \$1,350*.

This strikes us as a remarkably large sum, especially given the truly end-game nature of players’ *Friend or Foe* dilemma and their quite brief pre-game interactions as shown contestants. Yet there is no way around the facts of how people play, or the stakes they faced. Rabin (1993, p. 1283) speculated that anecdotal evidence suggests “people sacrifice substantial amounts of money to reward or punish kind or unkind behavior.” This indeed appears to be the case. Since the magnitude of the revealed-preference value for G_i (for roughly half the

players) does not depend on the players' (unknown) prior beliefs, we infer that for much of the population such 'fairness' considerations must be quite substantial—even in one-shot social interactions.

B. Learning and Coordination in the Augmented Game

An appealing feature of the augmented game is that it helps us understand how conditionally cooperative strategies might emerge in practice. To see this, imagine two contexts in which players know nothing and everything, respectively, about whether or not the opponents are conditionally-cooperative types. In each case, suppose it is common knowledge that 50 percent of individuals are conditional cooperators that prefer to play friend if $p_i \leq 0.5$, while the other 50 percent are "money players" whose dominant strategy is to play foe (we assume $S_i > 0$ for all i). In the case without further information, we would expect the individual foe rate to be 50 percent, half of the games to end in asymmetric (split) outcomes, and a quarter each to end with friend-friend and foe-foe outcomes.

How do outcomes change if players can *perfectly* predict an opponent's (G_j, S_j) values? Then three-quarters of games should end foe-foe, with the remaining quarter friend-friend. Half of the time, conditional cooperators are paired with money players and both now play foe (since common knowledge implies that $p_i = p_j = 1$). Greater player knowledge thus has two effects. First, it raises the individual foe rate, in this case (of perfect information) from 50 to 75 percent. Second, it also increases the degree of coordination, in the sense of reducing asymmetric game outcomes.

If players from some demographic groups have higher initial rates of friendly play than others in Season 1, then in the context of the augmented game, an opponent's demographic attributes in Season 2 provide a signal—*i.e.*, imperfect information—about the opponent's (G_j, S_j) values. Even imperfect information is valuable to a conditionally-cooperative player, however, who may switch from playing friend to foe on the basis of it.

To see how coordination affects winnings, consider again the two extreme cases in which players have no information or perfect information, respectively, about opponents' preferences in the augmented game. If half of players are conditionally-cooperative types but

have uninformed beliefs about opponents, then players will arrive at foe-foe—no winnings for either player—only $\frac{1}{4}$ of the time. With perfect information, by contrast, there are no winnings for either player $\frac{3}{4}$ of the time. Information only helps those who are conditionally cooperative, and then only when they are paired with another conditional cooperator. If we denote the probability that a randomly-selected player chooses foe by p , and we normalize the stakes x to 1, then expected winnings per player under independence (*i.e.*, no information) are $1-p^2$. By contrast, with perfect information average winnings fall to $(1-p)^2$. Information reduces average winnings regardless of p , with the decline larger as the prior probability of playing foe increases. [MWW1]

This characterization suggests a set of empirical tasks. First, we examine the data to determine whether players' tendencies to choose friend (*i.e.*, act cooperatively) varies systematically with their observable characteristics in Season 1. Second, if this is the case, we can then use the second-season data to test for conditionally cooperative behavior. This proceeds by asking whether players whose characteristics predict relatively higher rates of friendly play in Season 1 move to play foe against opponents they should expect to play foe, and play friend against opponents more likely to act friendly as well.

III. The Quasi-Experimental Context

A. The Game

Friend or Foe aired on the Game Show Network beginning in June 2002. The game has two components: A *production phase*, in which player pairs jointly contribute to answering trivia questions, and a *distribution phase*, in which contestants play the game in Figure 1 to determine how the pie they have produced will be divided between them. The Game Show Network (2003) provides the following description of the game:⁶

The show consists of six strangers who pair up at the start of the show to form three teams. Each team is separated into isolation chambers where all trivia rounds [are] played. The [two members of each team] answer trivia questions in order to build a bank account. At the

⁶ Descriptions provided by the Game Show Network here and below have been edited for clarity.

end of each round (there are 3 rounds total) the lowest-scoring team is eliminated. Before each team is dismissed, they enter a “Trust Box” where their account is divided [by playing the game in Figure 1].

The first round has four trivia questions, worth \$500 each. The second round has four questions, worth \$1000 each. In the third round, the remaining team answers up to ten \$500 questions. If all ten are answered correctly, the entire score is doubled. Given an initial endowment of \$200 per team, the winnings to be divided can therefore range from \$200 to \$22,200.⁷

The show aired in two seasons. The first season consisted of 40 episodes taped prior to the show’s premiere on June 3, 2002. These episodes aired twice daily on weekdays, and were re-run on weekends. A second season of 65 new episodes was taped in late summer 2002. These were aired beginning October 1, 2002. Contestants on the show during the second season therefore can have seen the play from the first season, but contestants during the first season could not.

Partner assignment within a given show is not completely random; instead, the selection worked as follows (according to the show producers):

Prior to the taping of each episode the six game players will be gathered together backstage. There, three contestants and three potential [partners] will be introduced to one another via a producer. The producer will first expose the three contestants’ [self-reported] positive and negative attributes. The producer will then disclose all three potential partners’ positive and negative attributes. The producer will then ask the three contestants to select one person they would like to partner with. These choices, made in security, will be written down and then displayed one by one. If all three have selected different partners, the producer will identify each team. If two or even three contestants have selected the same partner, then the choice falls to this selected partner. After the selected partner has chosen, the remaining contestants select their second choice partner. This process will continue until three teams of two have been formed.⁸

The show aired this partner selection process during the first season, but did not air it during the second season. To our knowledge there was no contact between players prior to

⁷ In the second season, the show started teams with zero but gave teams answering no questions correctly \$200 to split in the dilemma game. Thus the top possible score in the second season is \$22,000.

⁸ Producer Melissa Rudman, e-mail communication with authors, April 8, 2003.

this selection process, and no systematic way in which the producer divided each show's six participants into contestants (the initial choosers) versus potential partners.

B. TV Shows and Laboratory Experiments

Our context has some advantages and disadvantages. The first and foremost advantage is that *Friend or Foe* allows us to observe decisions in a social dilemma situation with very high stakes. Balanced against this advantage are a few features that distinguish our context from standard laboratory experiments.

The game show differs from many experiments in that players interact in person. While face-to-face one-shot interaction is not less realistic than double-blind exchanges—many business and social situations constitute one-shot games where people countenance their opponents—personal interactions reduce the degree of control in the experiment because it is difficult to empirically assess whether appearances, show banter, and player attributes disclosed during the assignment process influence observed decisions. Second, our contestants are on a televised show where play is not anonymous. While it is unlikely that acquaintances of the contestants would happen to see the show by accident—only 0.6 percent of cable television households watch *Friend or Foe* (Greco, 2002)—it is possible that some players informed friends or family that they would be on TV. While this visibility diverges from standard laboratory experiments, we do not view it as clearly bad. In real life, only very rarefied examples of one-shot interaction have no chance of being observed by third parties.

These issues aside, our context has several challenges. The first is that events in the production phase of the game may affect the distribution phase. For example, contestants contributing less in the production phase may feel gratitude toward their partner, altering their likelihood of playing foe in the distribution phase. Our strategy for dealing with this is to examine whether players' friend or foe decisions are related to their contributions during the production phase (more about this below). A related concern is that successful teams who play for higher stakes have a longer production-phase history and get to observe the decisions of the contestants exiting the show earlier in the episode. We address whether this matters by

examining variation in play both within and between rounds of the game. Finally, as with subjects in most controlled experiments, our subjects are not a random sample of the general population. Precisely how this might affect our results is difficult to assess.

C. Data

A total of 105 *Friend or Foe* episodes were produced, with 6 new contestants on each episode, for a total of 630 players and 315 games. The data include each player's gender, age, race, occupation, team score (the value of x in Figure 1, or the "stakes"), the number of positive and negative contributions to the team score (*i.e.* the correct and incorrect answers contributed in the production phase), and the amount each player ultimately takes home (his or her "winnings").

These data come from two sources. First, we obtained complete data for 300 games by taping 100 episodes and coding outcomes and player data from the tapes. Each player's gender, race, team score, contribution history, friend or foe decision, and final winnings is directly observable; players' ages and occupations were self-reported on-air at the start of every show, but are otherwise unverifiable. For corroborative purposes, we obtained supplementary data on all 105 episodes (specifically: airdates, player names, friend or foe decision, and each player's winnings) from a game show episode guide.⁹ Player name generally allows inference about gender, so the gender variable is available for 627 observations. The stakes were unavailable at this secondary source for cases where pairs played foe-foe (and therefore winnings were zero).

The distribution of players' demographic characteristics is similar to the U.S. population, except that contestants tend to be younger adults and disproportionately California residents.¹⁰ Approximately half of the contestants are male, and nearly a sixth of the contestants are black. The 25th and 75th percentile age players are 23 and 33, respectively; the median age is 27. About half of the players report a hometown in the West (as defined by Census Divisions), with 1/6th from each of the other three divisions. Table 1 shows how game out-

⁹ <http://gameshowfavorites.classictvfavorites.com/FriendorFoe/episodeguide.html> (accessed May 8, 2003).

¹⁰ The geographic distribution may reflect the fact that the show was produced and taped in Santa Monica, CA.

comes—scores, the tendency to play foe, and individuals’ winnings—vary with gender, age, and race. Table 2 shows how play varies with both own and opponent characteristics. We discuss these tables below.

IV. How Do Participants Play?

We first examine how the tendency to play foe varies with stakes. Second, we document how play varies with players’ observable characteristics during Season 1. Third, we test for conditional cooperation by examining whether players whose observable characteristics predict relatively higher rates of friendly play in Season 1 move to (a) play foe against opponents they should expect to play foe, and (b) play friend against opponents more likely to act friendly as well. Finally, we examine the evolution of winnings as players learn to conditionally cooperate.

A. Stakes and Play

Among the 630 players, the relative frequency of cooperative play (*i.e.*, choosing friend) is 45 percent. Figure 2 shows the relationship between stakes and the tendency to play foe. Each circle in the figure shows the relative frequency of foe at an observed stake level (shown in log-scale). Circle area reflects the number of players at each stakes. Other than a mildly depressed foe rate at the \$200 point, there is no discernable relationship between stakes and the individual tendency to play foe. The absence of any obvious stakes relationship also applies when play is examined within each round of show episodes (recall the game is repeated, with different players, three times per show). Nor is the stakes and individual tendency to play foe different between the two seasons of the show.¹¹

Although the data reveal no unconditional decline in the individual tendency to play friend at higher stakes, it is useful to explore this relationship conditional on commonly known player characteristics. Table 3 presents four groups of bivariate probit estimates of the likelihood players choose (foe, foe) as a function of the stakes, with and without explanatory

¹¹ The supporting figures are omitted here; for details, see our NBER working paper (2003).

covariates. Each group uses data from either Season 1 or Season 2. Stakes is not a significant predictor of how pairs play—statistically or practically—in any of the specifications.¹²

Except for the teams facing stakes of \$200—who are slightly more likely to play friend ($p = .08$)—there is no evidence that stakes affect cooperativeness, either within or across rounds of the game. What is remarkable about the data is that even with very large sums of money at stake, friendly play is quite stable with respect to stakes and represents about half of all the players’ decisions. If we are correct to conjecture that players understand the simple rules of this game (there being no evidence to the contrary), then the high frequency with which players choose friend must reflect non-monetary considerations that differ from the payoffs in Figure 1. Moreover, these non-monetary considerations must scale up in a roughly proportionate way with monetary stakes over quite a broad range—from \$200 to over \$16,000.

An important question in experimental economics is whether behavior observed in small-stakes environments can be generalized to situations with high risks and rewards. While subjects are more likely to approach Nash play with high stakes in some experiments (e.g., the centipede game studied by Rapoport et al., 2003), most studies on the role of stakes conclude that play is not greatly affected by the size of the incentives (Binswanger (1980), Kachelmeier and Shehata (1992), Fehr, Fischbacher and Tougareva (1995), Cameron (1995), Slonim and Roth (1998); for a survey see Camerer and Hogarth (1999)). One concern with these stakes experiments is that they are typically performed with subjects living in poorer countries. Given substantial uncontrolled differences in cultural norms (Roth, et al. 1991), it is not obvious if the high-stakes simulations in poor countries are in fact generalizable to Western economies. The stakes invariance documented here may allay some of these concerns about the generalizability.


¹² One might conjecture that players’ contributions during the production phase of the show may confound the relationship between stakes and play, inasmuch as a pair’s stakes are likely to be higher when both players are good at answering the trivia questions posed on the show. To examine this, we tabulated the number of questions that each player answers (correctly or incorrectly) in each round of the production phase. These counts, along with dollar-denominated analogs (constructed using each question’s value to the players, *cf.* Section II), are then used as an additional covariate in re-estimating the probit models. Their inclusion, in either count or dollar-value form, does not alter the stakes-invariance nature of play reported above.

B. Learning from Season One

We now consider how first-season play varies with contestants' observable characteristics. Overall, the tendency to play cooperatively varies markedly across demographic groups. Foe rates by group are listed in Table 1. Men play foe more often than women (53 percent of men in Season 1 play foe, vs. 46 percent of women). Players at or under the median age (27 years) choose foe much more often than players over the median age (65 vs. 39 percent), and blacks – almost always paired with whites – play foe more often than whites (58 vs. 48 percent). Column (3) of Table 3 reports bivariate probit model estimates of player pairs' tendencies to play foe as a function of each player's characteristics during Season 1. Here only a player's own characteristics are used to explain his or her decision. We resoundingly reject the hypothesis that a players' own characteristics are unrelated to his or her play ($p < 0.001$). A major explanatory factor here is the difference in play associated with age, as is evident in the probit coefficients.

Importantly, during Season 1 play is completely unrelated to an *opponent's* observable characteristics. When we include both own and opponent characteristics in the bivariate probit estimates for Season 1 play—see column (5) of Table 3—we continue to reject the hypothesis that own characteristics do not matter ($p < 0.001$), but we cannot reject the hypothesis that opponent characteristics do not matter ($p = 0.96$).

C. Conditional Play in Season Two

Results on Season 1 play indicate that women, whites, and older players choose friend more frequently than men, blacks, and younger players, respectively. [MWW2]To the extent that these differentials are known by Season 2 players but were unknown prior to Season 1, they set up a test for conditional cooperation behavior in Season 2. Between Seasons 1 and 2, do individuals with characteristics that predict relatively higher rates of friendly play move to play foe against opponents they should expect to play foe, and continue to play friend against opponents like themselves? To explore this, we examine the inter-season change in play conditional on the players' observable characteristics (viz, gender, race, age). Note that the

theory predicts a differential change in friendly play only for conditional cooperators. As discussed in Section II.A, players who are not conditional cooperators have a dominant strategy of playing foe and should not change play at all.

Do women implement conditional cooperation by gender in Season 2? From Season 1 to 2, the rate at which women play foe against men rises from 48 to 66 percent ($p = .01$). By contrast, the rate at which women play foe against women does not change---it holds steady at 44-45 percent in both seasons. The difference between these changes is marginally significant (at the 7 percent level in a one-sided test), imprecisely estimated because of the double differencing. On the other hand, men exhibit no differential change in their tendency to play foe by opponent gender between Season 1 and Season 2.

Do older players implement conditional cooperation by age in Season 2? From Season 1 to 2 the rate as which older players chose foe against younger players jumped by 24 percentage points, from 38 percent of contests to 62 percent ($p < .001$). Older players' foe rates also increased against other older players, although by much less and not statistically distinguished from zero---from 40 to 55 percent of contests ($p = .1$). The difference in the changes is 9 percentage points, but is only significant at a 21 percent level in a one-sided test. In contrast, the differential change in younger players' tendencies to play foe by opponent age is only one percentage point.

Finally, do white players implement conditional cooperation by race? While whites played foe at indistinguishable rates against blacks (46 percent) and whites (51 percent) in Season 1, the Season 2 white foe rate against blacks jumps by 29 percentage points to 75 percent ($p < .01$) while there is essentially no change in white players' tendency to play foe against other whites (it remains 53 percent). The difference in these changes is significant at the 8 percent level in a one-sided test.¹³ The lack of games pairing two black opponents precludes our examining whether the change in the black foe rate between seasons differs by race.

¹³ While black players choose foe more often than white players, this should to be understood in light of the fact that black players essentially never face a black opponent in the data. Hence, black players' high average foe rates could simply reflect a high expectation that white players will typically play foe against them.

The bivariate probit estimates in Table 3 provide other evidence of conditional cooperation. Columns (7) and (8) include indicator variables for whether *both* players have observable characteristics initially associated with higher unilateral rates of friendly play. Specifically, they indicate whether players are both female, both old, or both white. The coefficient estimates predict that female, white, and older players are much less likely to play foe in Season 2 than in Season 1 when facing a demographically similar opponent.¹⁴

Results in this section show evidence that some players implement conditionally cooperative strategies in Season 2. Relative to the way that friendly players treat friendly opponents, they punish opponents they expect to be uncooperative. By contrast, players with characteristics associated with uncooperative play in Season 1 do not demonstrably condition their cooperativeness on opponent characteristics in Season 2.

1. Coordination of Play

Adopting the conditional strategies suggested by the results above implies an increase in coordinated play—that is, an increase in outcomes along the main diagonal of the normal-form game in Figure 1—among certain groups of players. Players whose characteristics predict relatively higher rates of friendly play in season 1 moved to play foe against opponents they should expect to play foe, and continued to play friend against opponents like themselves. This implies a fall in the overall rate of split (one plays foe, one plays friend) outcomes in match-ups between players from contrasting demographic groups.

Such changes can be seen directly in the data. Consider the pairwise outcomes reported in Table 2. In match-ups involving one male and one female player, the proportion of split-outcome games falls from Season 1 to Season 2 (from 45 to 40 percent). This reflects an

¹⁴ One might suspect that the changes in play in Season 2, if based on updating beliefs from Season 1, is driven by both the conditions in our data as well as factors that the players observe but are not in the data. Accordingly, we can ask how much of the increase in foe-playing behavior is attributable to our simple characterization of observables. In the bivariate probit estimates of players' foe decisions shown in Table 3, the estimated parameter ρ in columns (5) and (6) indicates the correlation of player pairs' tendencies not explained by observed own- and opponent-characteristics. It is effectively zero in Season 1 ($\hat{\rho} = -.02, SE = .15$). It increases to .15 in Season 2, although it remains imprecisely estimated ($SE = .12$). Thus the evidence is weak that players are systematically engaging Season 2 strategies contingent on information beyond what is observed in the data.

increase in the proportion of foe-foe outcomes (from 28 to 42 percent) and a *decrease* in the proportion of friend-friend games (from 26 to 19 percent). By contrast, match-ups involving two female players shift disproportionately to the symmetric friend-friend outcome. Among two-female player games, the proportion of split outcomes falls (from 58 to 40 percent) from Season 1 to Season 2, but the proportion of friend-friend games increases (from 27 to 35 percent).

A similar phenomenon occurs with respect to race and age. In match-ups with one black and one white player, the fraction of split-outcome games falls (from 64 to 43 percent) from Season 1 to Season 2, with a disproportionately large increase in the fraction of foe-foe outcome games (from 19 to 51 percent). In match-ups of one younger and one older player, the fraction of split-outcome games falls (from 61 to 40 percent) between the seasons, with a disproportionately large increase in the fraction of foe-foe games (from 20 to 40 percent).

Overall, these changes result in a decrease in the proportion of asymmetric-outcome games from Season 1 to Season 2. Such changes are equivalent to an increase in coordinated play, provided we interpret coordination as reflecting some players' preferences to meet friend with friend, and foe with foe. These results identify a natural mechanism by which later generations of players on *Friend or Foe* increased the relative frequency of symmetric game outcomes: They conditioned their individual strategies on opponents' observable characteristics, taking as given the association of cooperative and uncooperative play with these characteristics among the earlier generation of players.

D. Winnings

What happens to winnings? Overall, we see a drastic decline in average winnings between Season 1 and Season 2. This occurs partly because the average stakes were lower in Season 2,¹⁵ and—to a large degree—because players were far more likely to walk away empty-handed. Table 1 indicates that individuals in Season 1 faced average stakes of \$3,718 and took home average winnings of 39 percent, or \$1,463. In Season 2 players faced average

¹⁵ The show's producers appear to have used more difficult trivia questions during Season 2, lowering the average stakes x somewhat.

stakes of \$3,063, and took home an average of 30 percent, or \$926. If we interpret inefficiency in this context as the foe-foe outcome in which both players destroy the contingent asset x they have produced, then the conditional strategies described above markedly lowered efficiency.

This is not true across all demographic groups, however. Players with observable characteristics initially associated with higher rates of friendly play come to fare better (monetarily) over time, relative to players with characteristics associated with lower rates of cooperativeness. Table 1 shows that for women, white players, and older players, average winnings per player as a share of the stakes changed between Season 1 and Season 2 by +1, -2, and -5 percentage points, respectively. None of these changes are statistically distinguishable from zero. Essentially, these players' ability to convert a game's stakes into take-home pay remained unchanged (on average) by adopting their conditional strategies.

By contrast, the opposite is true for players with characteristics associated with less-friendly play. For male players, average winnings as a share of stakes falls a significant 17 percentage points between seasons overall. Notably, it falls only 6 percentage points in games against another male player (comprising 66 of Season 2 games); in games against a female player, however, male players' winnings rate falls 27 percentage points (comprising 111 Season 2 games). Similar changes occur for younger players, who experience an overall decline of 15 percentage points, and black players, who experience a precipitous decline of 39 percentage points between seasons. In effect, the conditional strategies adopted by players in groups with higher rates of cooperative play stabilize their average winnings. However, these strategies yield a drastic decline in monetary gains for players tagged as having higher rates of unfriendly play based on past experience.

IV. Conclusion

Our study has a number of interesting findings. First, we document that contestants on *Friend or Foe* learn to evaluate players' cooperative tendencies from the history experienced by other players. In particular, they learn how players with different observable characteristics are likely to play, and using their predictions, many contestants choose to forgo large

sums of money in order to divide winnings “fairly” – even though the game is not repeated. In Season 2, cooperative individuals play friend at elevated rates when paired with other cooperative types, relative to their cooperative play against opponents with observable characteristics associated with unfriendly play. While better-informed players tend to bring about smaller dollar winnings, more cooperative types of players (e.g. women) can create an “island of cooperation” where winnings are fairly stable.

Second, we document that even very high stakes do not induce Nash play. This indicates that the non-monetary payoffs in *Friend or Foe* – the value of playing fairly – must be remarkably large, roughly proportional to the monetary stakes. It is one thing to forgo \$5 in order to be fair, as players typically do in laboratory contexts. Our players forgo over \$1000. Preferences for fair divisions are apparently very strong. Moreover, since our contestants play one-shot games, the apparent preference for fairness is indeed a preference – or norm – rather than an investment in expected future interaction.

References

- Becker, Gary S. (1968). Crime and Punishment: An Economic Approach. *Journal of Political Economy* 76(2): 169-217.
- Berk, Jonathan B., Eric Hughson and Kirk Vandezande (1996). The Price Is Right, but Are the Bids? An Investigation of Rational Decision Theory. *American Economic Review* 86(4): 954-970.
- Binswanger, Hans P. (1980) Attitudes toward Risk: Experimental Measurement in Rural India. *American Journal of Agricultural Economics* 62(3): 395-407.
- Bolton, Gary E. and Axel Ockenfels (2000). A Theory of Equity, Reciprocity, and Competition. *American Economic Review* 90(1): 166-93.
- Camerer, Colin F. and Robin M. Hogarth (1999). The Effects of Financial Incentives in Experiments: A Review and Capital-Labor-Production Framework. *Journal of Risk & Uncertainty* 19(1-3): 7-42.
- Cameron, Lisa A. (1995). Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia. *Economic Inquiry* 37(1): 47-59.
- Charness, Gary, Matthew Rabin (2002). Understanding Social Preferences with Simple Tests. *Quarterly Journal of Economics* 117(3): 817-69.
- Ellickson, Robert C. (1991). *Order without Law: How Neighbors Settle Disputes*. Cambridge, MA. Harvard University Press.
- Fehr, Ernst and Simon Gächter (2000). Cooperation and Punishment in Public Goods Experiments. *American Economic Review* 90(4): 980-94.
- Fehr, Ernst and Klaus M. Schmidt (1999). A Theory of Fairness, Competition, and Cooperation. *The Quarterly Journal of Economics* 114(3): 817-68.
- Fehr, Ernst, Urs Fischbacher and Elena Tougareva (2002). Do High Stakes and Competition Undermine Fairness? Evidence from Russia. Institute for Empirical Economics, University of Zurich, Working Paper No. 120.
- Game Show Network (2003). Friend or Foe. At <http://www.gameshownetwork.com/index.html>, accessed April 15, 2003.
- Gertner, Robert (1993). Game Shows and Economic Behavior: Risk-Taking on "Card Sharks." *The Quarterly Journal of Economics* 108(2): 507-21.
- Greco, Melissa (2002). Gamer Net Game for More Foe. *Daily Variety*, August 12: 15.
- Güth, Werner, Rolf Schmittberger, and Bernd Schwarze (1982). An Experimental Analysis of Ultimatum Bargaining. *Journal of Economic Behavior and Organization* 3(4): 367-388.
- Jackson, Matthew O., and Ehud Kalai (1997). Social Learning in Recurring Games. *Games and Economic Behavior* 21: 102-34.
- Kachelmeier, Steven J. and Mohamed Shehata (1994). Examining Risk Preferences under High Monetary Incentives: Reply. *American Economic Review* 84(4): 1105-06.
- Kreps, David M. and Robert Wilson (1982). Reputation and Imperfect Information. *Journal of Economic Theory* 27(2): 253-79.
- Laury, Susan K. and Charles A. Holt (forthcoming). Voluntary Provision of Public Goods: Experimental Results with Interior Nash Equilibria. In: C.R. Plott and V. Smith (eds.), *Handbook of Experimental Economic Results*. North Holland, Amsterdam.
- Ledyard, John O. (1995). Public Goods: A Survey of Experimental Research. In: Kagel, John H. and Roth, Alvin E. (eds.), *The Handbook of Experimental Economics*. Princeton: Princeton University Press, 111-194.

- List, John A. (2003) "Friend or Foe: A Natural Experiment of the Prisoner's Dilemma." Mimeo, University of Maryland.
- Metrick, Andrew (1995). A Natural Experiment in "Jeopardy!" *American Economic Review* 85(1): 240-53.
- Rabin, Matthew (1993). Incorporating Fairness into Game Theory and Economics. *American Economic Review* 83: 1281–1302.
- Rapoport, Amnon et al. (2003). Equilibrium Play and Adaptive Learning in a Three-Person Centipede Game. *Games & Economic Behavior* 43(2): 239-65.
- Roth, Alvin E., Vesna Prasnikar, Masahiro Okuno-Fujiwara, and Shmuel Zamir (1991). Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study. *American Economic Review* 81: 1068-1095.
- Slonim, Robert and Alvin E. Roth (1998). Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic. *Econometrica* 66(3): 569-96.
- Tirole, Jean (1988). *The Theory of Industrial Organization*, Cambridge, MA: The MIT Press.

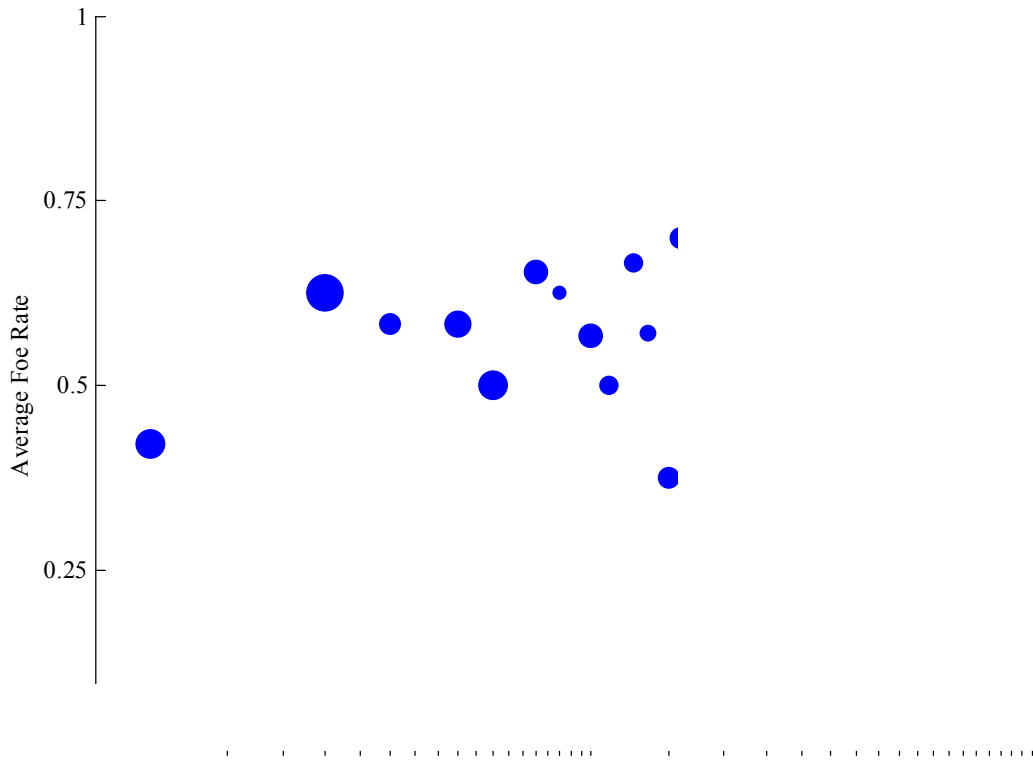


Figure 2. Average Foe Rates and Game Stakes, $n = 600$ players. The area of each circle is proportional to the number of players observed at that stakes level.

Table 1
Individual Outcomes by Group and Season

Group	Season	Number of Players	Foe Rate	Mean Stakes	Mean Winnings	Ratio of Winnings to Stakes
All	1	240	49%	\$ 3,718	\$ 1,463	39%
	2	390	58%	\$ 3,063	\$ 926	30%
	<i>Total</i>	630	55%	\$ 3,314	\$ 1,131	34%
Men	1	118	53%	\$ 4,331	\$ 1,848	43%
	2	195	59%	\$ 3,217	\$ 820	26%
	<i>Total</i>	313	57%	\$ 3,638	\$ 1,208	33%
Women	1	121	46%	\$ 3,101	\$ 1,074	35%
	2	193	58%	\$ 2,920	\$ 1,043	36%
	<i>Total</i>	314	53%	\$ 2,990	\$ 1,055	35%
White	1	183	48%	\$ 3,915	\$ 1,389	35%
	2	287	58%	\$ 3,237	\$ 1,082	33%
	<i>Total</i>	470	54%	\$ 3,501	\$ 1,201	34%
Black	1	40	58%	\$ 2,913	\$ 1,529	52%
	2	55	71%	\$ 2,787	\$ 368	13%
	<i>Total</i>	95	65%	\$ 2,840	\$ 857	30%
Young (Age ≤ 27)	1	96	65%	\$ 3,687	\$ 1,665	45%
	2	214	60%	\$ 3,336	\$ 999	30%
	<i>Total</i>	310	61%	\$ 3,445	\$ 1,205	35%
Old (Age > 27)	1	144	39%	\$ 3,740	\$ 1,328	36%
	2	176	56%	\$ 2,722	\$ 838	31%
	<i>Total</i>	320	48%	\$ 3,186	\$ 1,059	33%

Table 2
Game Outcomes by Player and Opponent Characteristics

Player	Opponent	Season	Number of Games	Player Foe Rate	<i>Game Outcome Frequencies</i>			Mean Stakes	Mean Player Winnings ^a
					Foe-Foe	Fr-Fr	Split		
Male	Male	1	24	52%	21%	17%	63%	\$ 5,542	\$ 2,292
		2	42	61%	33%	12%	55%	\$ 3,576	\$ 1,235
		<i>Total</i>	66	58%	29%	14%	58%	\$ 4,291	\$ 1,619
	Female	1	69	54%	28%	26%	46%	\$ 3,461	\$ 1,520
		2	111	58%	42%	19%	39%	\$ 2,942	\$ 507
		<i>Total</i>	180	56%	37%	22%	42%	\$ 3,142	\$ 895
Female	Male	1	69	48%	28%	26%	46%	\$ 3,461	\$ 1,023
		2	111	66%	42%	19%	39%	\$ 2,942	\$ 964
		<i>Total</i>	180	59%	37%	22%	42%	\$ 3,142	\$ 986
	Female	1	26	44%	15%	27%	58%	\$ 2,623	\$ 1,142
		2	40	45%	25%	35%	40%	\$ 2,920	\$ 1,173
		<i>Total</i>	66	45%	21%	32%	47%	\$ 2,803	\$ 1,161
White	White	1	68	51%	29%	26%	45%	\$ 4,225	\$ 1,543
		2	113	53%	33%	27%	44%	\$ 3,314	\$ 1,115
		<i>Total</i>	181	52%	31%	27%	42%	\$ 3,656	\$ 1,276
	Black	1	37	46%	19%	16%	65%	\$ 2,930	\$ 834
		2	47	75%	51%	6%	43%	\$ 2,860	\$ 786
		<i>Total</i>	84	62%	37%	11%	52%	\$ 2,890	\$ 807
Black	White	1	37	57%	19%	16%	65%	\$ 2,930	\$ 1,653
		2	47	70%	51%	6%	43%	\$ 2,860	\$ 350
		<i>Total</i>	84	64%	37%	11%	52%	\$ 2,890	\$ 924
Young	Young	1	16	66%	50%	19%	31%	\$ 3,513	\$ 941
		2	60	60%	40%	20%	40%	\$ 3,672	\$ 1,163
		<i>Total</i>	76	61%	42%	20%	38%	\$ 3,638	\$ 1,116
	Old	1	64	64%	20%	19%	61%	\$ 3,773	\$ 2,027
		2	94	60%	40%	19%	40%	\$ 2,909	\$ 788
		<i>Total</i>	158	61%	32%	19%	49%	\$ 3,259	\$ 1,290
Old	Young	1	64	38%	20%	19%	61%	\$ 3,773	\$ 1,041
		2	94	62%	40%	19%	40%	\$ 2,909	\$ 863
		<i>Total</i>	158	52%	32%	19%	49%	\$ 3,259	\$ 935
	Old	1	37	40%	19%	38%	43%	\$ 3,741	\$ 1,547
		2	29	55%	28%	17%	55%	\$ 2,541	\$ 736
		<i>Total</i>	66	47%	23%	29%	49%	\$ 3,214	\$ 1,191

^a Winnings are first-player winnings when players are from different groups (e.g., male-female pairs), and average winnings of both players when players are from the same group (e.g., male-male pairs).

Table 3
Bivariate Probit Estimates of Pairwise Outcomes
 Dependent variable is a (Foe, Foe) pair. Coefficient estimates shown.
 (Standard errors in parentheses, except as noted)

<i>Model:</i>	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
<i>Season:</i>	First	Second	First	Second	First	Second	First	Second
Constant	-0.02 (0.08)	0.20 (0.07)	1.36 (0.78)	-0.45 (0.59)	1.49 (0.90)	-0.84 (0.69)	2.08 (1.00)	-0.16 (0.78)
Log Score			-0.05 (0.08)	0.06 (0.06)	-0.05 (0.09)	0.06 (0.07)	-0.07 (0.09)	0.06 (0.07)
Player Age			-0.04 (0.01)	-0.001 (0.01)	-0.04 (0.01)	-0.001 (0.01)	-0.05 (0.01)	0.001 (0.01)
Player Black			0.33 (0.23)	0.33 (0.19)	0.33 (0.23)	0.42 (0.20)		
Player Male			0.22 (0.17)	0.07 (0.14)	0.21 (0.18)	0.13 (0.14)	0.16 (0.25)	-0.16 (0.20)
Player West			0.23 (0.18)	0.45 (0.14)	0.23 (0.18)	0.45 (0.14)	0.19 (0.17)	0.46 (0.14)
Opponent Age					0.0001 (0.01)	0.006 (0.01)	-0.006 (0.01)	0.008 (0.01)
Opponent Black					-0.0001 (0.23)	0.40 (0.20)		
Opponent Male					-0.05 (0.18)	0.27 (0.14)	-0.11 (0.25)	-0.01 (0.20)
Opponent West					-0.13 (0.17)	-0.02 (0.14)	-0.12 (0.17)	-0.02 (0.14)
Both female							-0.11 (0.34)	-0.57 (0.29)
Both Older							0.21 (0.24)	-0.17 (0.28)
Both white							-0.13 (0.18)	-0.41 (0.16)
ρ	-0.05 (0.14)	0.19 (0.11)	-0.02 (0.15)	0.17 (0.12)	-0.02 (0.15)	0.15 (0.12)	-0.03 (0.15)	0.13 (0.12)
H_0 : Player Characteristic Effects All Zero (p-value)			20.2 (0.001)	14.1 (0.007)	20.2 (0.001)	15.5 (0.004)		
H_0 : Opponent Characteristic Effects All Zero (p-value)					0.64 (0.96)	8.69 (0.07)		
H_0 : Interaction Characteristic Effects All Zero (p-value)							1.37 (0.71)	11.1 (0.01)
Observations (pairs)	120	195	117	183	117	183	117	183