

# *Friend or Foe?*

## 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*

March 2008

### **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 from two seasons of a television game show, we provide evidence about how individuals implement conditionally cooperative preferences. We show that (1) contestants forgo large sums of money to be cooperative, (2) players cooperate at heightened levels when their opponents are predictably cooperative, and (3) players whose observable characteristics predict less cooperation fare worse (monetarily) over time, as opponents avoid cooperating with them.

**JEL:** H41, K42, A13, C93

We thank the editor and two anonymous referees for insightful suggestions that substantially improved the paper. 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. Max Bazerman, Gary Bolton, Rachel Croson, Daniel Kessler, Michael Wheeler, Peter Zemsky, and seminar participants at the Kennedy School of Government, Pennsylvania State University and Wharton provided useful and thought-provoking comments.

## I. Introduction

In recent years, economists have endeavored to explain why individuals frequently cooperate in competitive situations. Early work emphasizes reputation and reciprocity in the context of repeated games (Kreps and Wilson, 1982). More recently, theorists and experimentalists have turned their attention to norms and preferences as explanations of cooperative behavior. Rabin (1993) posits that individuals have conditionally cooperative preferences—they like to cooperate with those who are cooperative but punish those who are not. In theory and experiments, 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 without repeated interaction, however, conditionally cooperative preferences require individuals to be able to discern their opponents' preferences prior to play. Missing from the literature is evidence about how individuals implement such preferences in one-shot interactions. This paper attempts to fill that gap, using data from a high-stakes game show.

In June 2002, the Game Show Network began airing a cable television show in which contestants play a game called *Friend or Foe*. On the show, individuals play the one-shot, simultaneous-move game in Figure 1.

		<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,  $x > 0$ .

This game is a variant of the classic Prisoner's Dilemma and is similar in structure to games analyzed previously in experimental studies.<sup>1</sup> Here playing foe is a weakly-dominant strategy for each player. The stakes in *Friend or Foe* are high: The payoff  $x$  ranges from \$200 to over \$16,000, with an average of \$3,300 at stake per game.

This paper presents a model of conditional cooperation that indicates how players might im-

---

<sup>1</sup> For surveys, see Ledyard (1995) and Laury and Holt (forthcoming). Analysis of data from television game shows 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 (2006a) uses data from the first 40 episodes of *Friend or Foe* to draw inferences about discrimination.

plement conditionally cooperative strategies, and its consequences for equilibrium payoffs to cooperative and non-cooperative players. We use this model to interpret the evolution of play over two seasons of *Friend or Foe*. 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.<sup>2</sup>

Player behavior changes markedly across the two seasons of *Friend or Foe*. In the initial season, players’ choices are statistically independent of their opponents’, but players’ choices vary systematically with their observable characteristics. Season 2 players therefore have some basis for predicting how preferences for cooperation vary. Consistent with learning and conditional cooperation, players appear to change their strategies to condition on opponents’ observable attributes—cooperating with opponents who are predictably cooperative, while punishing opponents who are not.

Our second finding concerns the payoffs to conditional cooperation. By conditioning strategies on opponents’ observable characteristics, players who are expected to cooperate should fare better (monetarily) relative to players who appear less friendly. In the data, changes in take-home winnings over time are highly consistent with this prediction. 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 model of preferences and learning to organize our analysis. Section 3 describes the setting and the data. Section 4 presents and interprets the empirical results. A brief summary concludes.

## **II. Theoretical Background**

### **A. Fairness and the Augmented Game**

The high rate of friendly play in the show—45% of contestants cooperate—leads us to con-

---

<sup>2</sup> This structure, and our analysis of players’ behavior based on learning unknown type distributions in the population, are similar to the recurring games in Jackson and Kalai (1997).

sider player motivations beyond the payoffs in Figure 1. 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. One implication 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).

While the current literature does not agree on the underlying mechanism that produces fair outcomes in equilibrium (see the conflicting accounts by Falk and Fischbacher, 1998; Levine, 1998; Fehr and Schmidt, 1999; Bolton and Ockenfels 2000; Charness and Rabin, 2002; Rotemberg 2004), some forces are common to many models. This section presents a simple model of player preferences that captures some of these forces. This ‘augmented game’ provides a coherent framework for interpreting the evolution of play that we observe in the data.<sup>3</sup>

We assume that a player’s preferences reflect non-monetary considerations that depend on the friendliness of the opponent’s play as shown in Figure 2. 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. The non-monetary terms in this augmented game,  $S_i$  and  $G_i$ , thus reflect unobservable heterogeneity in preferences.

		<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 2: The augmented game.

Equating the expected payoff from playing friend versus playing foe in the augmented game yields the following strategy: If  $\pi_i$  denotes player  $i$ ’s belief that his or her opponent will play foe, then player  $i$  prefers to play friend if  $\pi_i < \tau_i(x) < 1$  where  $\tau_i(x) = (G_i - x/2) / (G_i - x/2 + S_i)$ . This

---

<sup>3</sup>Alternatively, one might interpret the changes in play we observe between seasons as Kandori (1992)-type contagion. The mechanisms that sustain cooperation in these models differ from the setting here, however. For instance, in a contagious equilibrium players cooperate so as not to destroy their future payoffs, and the only benefit of cooperating is to slow down (or prevent) the process of contagion. This benefit does not exist in a recurring game with non-overlapping generations of players, such a *Friend or Foe*.

implies a player chooses friend if her guilt from taking the entire stakes ( $G_i$ ) exceeds  $x/2$  by a multiple of  $S_i$ ; the potential dismay weighs more heavily in the decision as the prior on an opponent playing foe increases. We refer to  $\tau_i(x)$  as an agent's *type*.<sup>4</sup>

In this setting, we distinguish between two sets of players. Given a game at stakes  $x$ , players with types between 0 and 1 (equivalently,  $G_i > x/2$ ) are *conditional cooperators*. For such players, there exists a set of beliefs about the likelihood of an opponent choosing foe for which it is optimal also to play foe. For a sufficiently low foe-prior  $\pi_i$ , however, a conditional cooperator will prefer to play friend. The other set of players have types outside the unit interval, which imply a lower level of guilt than conditional cooperators have. Such players have a dominant strategy of playing foe in the augmented game in Figure 2, assuming  $S_i > 0$  (that is, no one is truly indifferent to being played the sucker in this environment).

## B. Learning and Conditional Cooperation in the Augmented Game

An appealing feature of the augmented game is that it helps us to understand how conditionally cooperative strategies might emerge in real-world settings. To see this, suppose each player has an observable attribute  $A_i \in \{0,1\}$  – for example, a player's sex. We assume this attribute provides imperfect information about the player's preferences, in the sense that

$$0 < \Pr(0 < \tau_i(x) < 1 \mid A_i = 0) < \Pr(0 < \tau_i(x) < 1 \mid A_i = 1) < 1 \quad (1)$$

A player with  $A_i = 1$  is more likely to be a conditional cooperator than a player with  $A_i = 0$ , although neither attribute perfectly predicts a player's type. Given strategies in the augmented game, the relative frequency with which players' having attribute  $A_i$  will choose foe, as a function of beliefs and the stakes, is

$$f_{A_i}(\pi_i, x) \equiv 1 - \Pr(\pi_i < \tau_i(x) < 1 \mid A_i).$$

With experienced players,  $i$ 's prior  $\pi_i(A_j)$  facing an opponent with attribute  $A_j$  should reflect the relative frequency of foe-playing behavior by observationally similar opponents in the past:

$$\pi_i(A_j) = f_{A_j}(\pi_j(A_i), x). \quad (2)$$

If we combine this restriction on beliefs with the (Bayes-Nash) equilibrium condition that

---

<sup>4</sup> It is natural to expect a player's guilt and dismay to vary with the stakes, and this variation is consistent with our data. We suppress this dependence for notational simplicity, as it is encompassed by the type index  $\tau_i(x)$ .

players choose best replies given their beliefs, we obtain a useful characterization of how outcomes should vary with players' attributes. The relative frequencies of foe-playing behavior one should observe (in a pure-strategy equilibrium) among games with stakes  $x$  and player attributes  $(A_i, A_j)$  are the solutions to (2) and its symmetric complement,  $\pi_j(A_i) = f_{A_i}(\pi_i(A_j), x)$ .

Although there can be multiple equilibria in this augmented game, the case of empirical interest is shown in Figure 3.<sup>5</sup> The solid lines indicate the foe rate  $f_{A_i}$  as a function of the prior  $\pi_i$  for players with attributes  $A_i = 0, 1$ . The vertical intercept indicates the probability that an  $A_i$  player is a conditional cooperator; each function increases with  $\pi_i$  up to the boundary condition  $f(1, x) = 1$ . When the game's players are observationally similar ( $A_i = A_j$ ), the equilibrium fraction of players choosing foe is given by a point where  $f_A$  crosses the 45-degree line. When the players are observationally dissimilar ( $A_i \neq A_j$ ), the relative frequency of foe differs by player attribute. Although the exact shapes of  $f_0$  and  $f_1$  depend on the distribution of preferences  $G_i$  and  $S_i$  in the population, the relative frequency of foe-playing behavior will differ for the four attribute matchups  $(A_i, A_j)$ , as indicated in Figure 3.

The assumption of experienced players whose priors reflect actual frequencies of play might be strong—perhaps too strong to characterize play in Season 1 when there was little or no history of this game for players to observe. This raises the central question of interest: How might conditionally cooperative play emerge over time? To examine this, consider a simpler information context: Suppose an observable attribute  $A_i$  provides the same information about a player's preferences as in (1), but players are initially unaware of this association. We call this situation *uninformed play*.<sup>6</sup> Figure 4 indicates one possible outcome with uninformed play. In this situation the relative frequency of foe by attribute  $A_i$  players is still determined by a point along  $f_{A_i}(\pi_i, x)$ . However, an uninformed player's foe-prior might not be particularly well-calibrated to the (as yet unknown) relative frequency of game outcomes. As Figure 4 indicates, for a prior of  $\hat{\pi}_i$  the frequency of foe among players with  $A_i = 1$  is  $p_1$  and among players with  $A_i = 0$  is  $p_0 > p_1$ . Although foe rates are higher for  $A_i = 0$  players than  $A_i = 1$  players (due to preferences), with uninformed play an individual's actions are statistically independent of an *opponent's* attributes.

<sup>5</sup> For example, there is always an equilibrium in which all players choose foe with probability 1. We ignore this outcome here, as it is inconsistent with the data.

<sup>6</sup> A good example of uninformed play is the finding by Fershtmann and Gneezy (2001) that Ashkenazi Jews expect Eastern Jews to be less trustworthy. Although this belief is inconsistent with experimental evidence, it informs the groups' play in trust games.

The history of play by others allows subsequent generations to update their beliefs. This makes uninformed play disappear and allows conditional cooperation to emerge. Players with conditionally cooperative preferences who would be approximately indifferent between choosing friend or foe in uninformed play will adopt a conditional strategy: cooperate with opponents whose attributes predict similarly cooperative behavior, and play foe against opponents whose attributes do not.

This implies that, over time, we should expect a *tâtonnement* to a separating equilibrium where play depends on observable proxies for an opponent's preferences. An example is indicated by the four arrows in Figure 4. Here (and in general) changes in the overall foe rate over time depend on players' priors during uninformed play, which are not directly observable. The model's predictions are that later generations of players will choose to cooperate in ways consistent with the initially-observed association of attributes and actions among earlier players.

More precisely, if we let  $p_{ij}$  denote the fraction of players with attribute  $A_i$  that chooses foe against opponents with  $A_j$ , then this model predicts a separating equilibrium with  $p_{00} > p_{01}$  and  $p_{11} < p_{10}$  (see Figure 3). This means that, among players whose attributes initially predict higher relative rates of friendly play, we expect a higher rate of friendly play in later generations against observationally similar opponents, and a lower rate against opponents whose attributes predict relatively uncooperative play (that is,  $p_{11} < p_{10}$ ). Analogously, among players whose attributes initially predict higher relative foe rates we should anticipate a higher relative frequency of foe against observationally similar opponents than against dissimilar opponents (or  $p_{00} > p_{01}$ ).

This characterization suggests a set of empirical tasks. First, we document whether players' tendencies to act cooperatively (i.e., choose friend) varies systematically with their observable attributes in Season 1. Second, we examine whether the second-season data reflects conditionally cooperative behavior. This proceeds by asking (a) whether players' second-season strategies are now conditional on opponents' observable attributes, and (b) whether these conditional strategies are consistent with the predicted variation in foe rates above.

### 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 answers to trivia ques-

tions, and a *distribution phase*, in which contestants play the game in Figure 1 to determine how the pie they have created will be divided between them. The number of trivia questions a pair correctly answers determines the game's stakes, which can range from a minimum of \$200 to a maximum of \$20,000 (although the highest observed is \$16,400). There are three pairs of contestants on each episode.

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. The six contestants on each show were matched into pairs via their stated preferences after the show's producer provided the contestants with brief background information about each possible partner. The show aired this partner matching process during the first season, but did not air it during the second season.<sup>7</sup> The non-random nature of player-pair selection requires some important caveats in interpreting our empirical results, which we address below.

## **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 recurring game with exceptionally high stakes. Balanced against this advantage are a few features that distinguish our context from standard laboratory experiments. First, players interact in person on the show. 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 assess whether appearances, show banter, and the like influence observed decisions.

Second, our contestants are on a televised show where play is not anonymous. As List (2006a) observes, one might expect that the more an individual's actions are widely communicated (e.g., by being televised), the more fairness concerns are likely to receive increased emphasis. For example, decreases in anonymity frequently increase cooperative behavior in public

---

<sup>7</sup> A more complete description of this matching process is in Oberholzer-Gee, et al. (2003); see also List (2006a).

goods contribution experiments (Ledyard 1995, Masclet et al. 2003, Rege and Telle, 2004). Moreover, it is plausible that both  $G_i$  and  $S_i$  (guilt and dismay) are larger in non-anonymous play. That said, we do not view the absence of anonymity here as clearly bad. In real life, only rarefied examples of one-shot interactions have no chance of being observed by third parties. Still, the absence of anonymity in the environment examined here should be noted as a caveat against generalization to social-dilemma settings in which anonymity prevails.

Third, as noted above the players appearing on each episode are not randomly paired. The matching-via-stated-preferences process that is used to form pairs among the six *ex ante* strangers on each episode might produce different levels of cooperation than would occur with true randomization (List, 2006a). For purposes of drawing inferences about how players' strategies change over time, however, the main issue is whether the matching process used by the show changed between Season 1 and Season 2. Our communications with the show's producer on this point provide no evidence to suggest changes in the matching process between seasons.<sup>8</sup>

### C. Data

A total of 105 *Friend or Foe* episodes were produced, with 630 players in 315 games. Our data include each player's gender, age, race, occupation, stakes, the number of correct and incorrect answers each player offered to the trivia questions during 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, occupations, and home towns were self-reported on-air at the start of every show. 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.<sup>9</sup> Player name generally allows one to corroborate gender, which is available for 627 observations. The distribution of players' demographic characteristics is similar to the U.S. adult population, except that contestants tend to be younger (the median contestant

---

<sup>8</sup> Melissa Rudman, e-mail communication with authors, April 8, 2003.

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

age is 27) and disproportionately California residents (approximately 40% of players).<sup>10</sup>

#### IV. How Do Participants Play?

##### A. Stakes and Play

Among the 630 players, the relative frequency of cooperative play (*i.e.*, choosing friend) is 45 percent. Figure 5 shows the relationship between stakes and the tendency to play foe. These data reveal that even with the large sums used in *Friend or Foe*, friendly play is effectively *stakes invariant*. The absence of any obvious stakes relationship also applies when play is examined within each round of show episodes.<sup>11</sup> Nor is the stakes and individual tendency to play foe different between the two seasons of the show.

This stakes invariance finding is noteworthy for two reasons. First, it reinforces prior experimental results, suggesting that similar behavior observed in small-stakes laboratory experiments may hold generally for higher risks and rewards.<sup>12</sup> Second, the analysis in section II makes predictions about the fraction of players choosing foe conditional on the game's stakes,  $x$ . Figure 5 suggests that, in fact, this fraction does not vary systematically with  $x$ . Thus, when we analyze how play varies with individuals' observable attributes (below), we obtain similar results conditionally or unconditionally on the stakes.

##### B. How Large Might Non-Monetary Payoffs Be?

Before proceeding further, it is worth noting what conditionally cooperative types have at stake empirically in *Friend or Foe*. The high frequency with which players choose friend supports our assumption that play reflects non-monetary considerations beyond 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. 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-

---

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

<sup>11</sup> The supporting tables are omitted here; for details, see Oberholzer-Gee, et al. (2003).

<sup>12</sup> While subjects are more likely to approach Nash play with high stakes in some experiments (e.g., the centipede game of 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)).

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 *Friend or Foe*. Rabin (1993, p. 1283) speculated that “people sacrifice substantial amounts of money to reward or punish kind or unkind behavior.” This indeed appears to be the case. Of course, players’ values of  $G_i$  may loom larger in non-anonymous contexts, such as the (televised) arena of *Friend or Foe*. Nevertheless, 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.

### C. Learning from Season One

We now consider how first-season play varies with contestants’ observable characteristics. Foe rates by sex and race are listed in Table 1. On average, men play foe more often than women in Season 1 (54 percent vs. 47 percent), and black players—almost always paired with white players—choose foe more often than white players in Season 1 (57 vs. 49 percent). In both cases the relative risk of playing foe is approximately 1.15 (for men v. women and for blacks v. whites), so sex and race provide (imperfect) information about how contestants play. Column (1) of Table 2 reports bivariate probit estimates of player pairs’ tendencies to play foe as a function of each player’s characteristics during Season 1. We resoundingly reject the hypothesis that a player’s 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 estimates. 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 (3) of Table 2—we cannot reject the hypothesis that all opponent and player-pair interaction characteristics do not matter ( $p = 0.97$ ).

### D. 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. Such differentials might enable conditional cooperation to emerge in Season 2. Do later generations of players (*a*) learn to condition their strategies on opponents’ observable attributes, and if so, (*b*) do they play friend at a higher rate against opponents whose observables predict more frequent cooperation? These

questions correspond to the joint hypothesis that players' conditional foe rates satisfy  $p_{10} - p_{11} > 0$  and  $p_{00} - p_{01} > 0$ , as discussed in Section II.

Consider issue (a) first. We find that—in contrast to Season 1—players in Season 2 chose actions that vary with the *opponent's* observable characteristics. In the bivariate probit estimates of Season 2 play in column (4) of Table 2, we reject the hypothesis that an opponents' characteristics do not predict a player's decision ( $p=0.042$ ). This indicates that in Season 2, players condition on *some* observable information about their opponents.<sup>13</sup>

Next, consider issue (b): Is play conditional on opponents' characteristics in Season 2 consistent with the variation predicted by conditional cooperation? Consider first whether or not women implement conditional cooperation by gender. Table 1 indicates that in Season 2 women play foe in 66 percent of games versus men, but only 45 percent of games versus women, giving  $p_{10} - p_{11} = 0.21$  ( $p=0.01$ ). During Season 1, however, the analogous rates are statistically indistinguishable at 48 percent and 44 percent, respectively ( $p=0.36$ ). The evidence regarding men's conditional cooperation is less precise. The foe rate differential for men playing men versus men playing women is directionally consistent with conditionally-cooperative male players changing their behavior to cooperate at higher rates with women than with men ( $p_{00} - p_{01} = 0.03$ ). However, this difference is too small to be statistically meaningful ( $p=.37$ ). The small difference may reflect a smaller proportion of conditional cooperators among men than among women, or perhaps lower levels of guilt among male players who are conditional cooperators than among female conditional cooperators.

Do white players implement conditional cooperation by race? In season 2 white players choose foe in 75 percent of games versus black players, but only 53 percent of games versus white players, giving  $p_{10} - p_{11} = 0.22$  ( $p<0.01$ ). The corresponding rates for Season 1, however, are statistically indistinguishable 46 percent and 51 percent. The lack of games pairing two black opponents precludes examining whether black players' strategies differ by opponent race.

Do players implement conditional cooperation by age in Season 2? Here the evidence is

---

<sup>13</sup> In principle, it is possible that the changes in play in Season 2 are driven by both the observables in our data as well as factors that the players observe but are not in the data. In the bivariate probit estimates of players' foe decisions shown in Table 3, the estimated parameter  $\rho$  in columns (3) and (4) 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 .14 in Season 2, although it remains imprecisely estimated ( $SE = .13$ ). Thus the evidence is weak that players are systematically engaging Season 2 strategies contingent on information beyond what is observed in the data.

consistent with conditional cooperation but statistically imprecise. For instance, if we split players into younger and older groups based on the median age (27 years) as in Table 1, we find that in Season 2 older players choose foe 62 percent of the time against younger players but only 55 percent of the time against older players ( $p=0.25$ ). Younger players choose foe at the same rates against older and younger players in Season 2 (60 percent of games against each). Similar conclusions emerge using the probit results in column (4) of Table 3 and other age differentials than the (arbitrary) above-or-below median dichotomy. While these differences in season 2 behaviors with respect to age are consistent with some players adopting conditionally-cooperative strategies, we view the age evidence as weak support.<sup>14</sup>

Taken together, the univariate comparisons by gender, race, and age suggest two regularities. First, players with attributes that predict higher season 1 foe rates, such as men and younger players, show little evidence of playing conditional strategies: Their foe rates do not differ appreciably by opponent attributes in season 2. Within the context of the model, this implies the foe-rate curve  $f_0$  for these groups is basically flat between the two equilibrium points in Figure 3. In economic terms, this means few such players have excess guilt levels ( $G_i - x/2$ ) of the same magnitude as their dismay ( $S$ ) in the augmented game (Figure 2). This seems a plausible, albeit not immediately obvious, characteristic of preferences among groups who tend to play foe.

Second, we see large differences in conditional foe rates among two of the three groups whose attributes predicted cooperative play in Season 1 (women and white players). The changes by these players between season 1 and 2 are sufficiently large as to merit further discussion, which we address presently. Season 2 play by the other initially ‘friendly’ group, the older players, is more difficult to interpret. While not inconsistent with our model of behavior, it is not supportive either. We return to this observation in section V.

### **E. Simple Updating versus Equilibrium Play**

The changes in foe rates by women and white players are particularly striking. Although foe rates by sex in season 1 differed by seven percentage points, women’s conditional foe rates in season 2 differed by more than twenty percentage points. Similar magnitude changes occurred

---

<sup>14</sup> Similarly, players do not seem to condition on opponents from the West in season 2 (or other geographic areas, for that matter); see Table 2, column (4). This is plausible: players’ hometowns are mentioned at the start of taping for each show, but are unlikely to be as salient as an opponent’s sex, race, or age when the game in Figure 1 is played later.

with white players from season 1 to 2. Does it make sense for these players to exhibit larger differences in conditional foe rates in season 2 than the (unconditional) differences observed during season 1?

The sense in which one might view the season 1 differences as too ‘small’ to explain large changes in season 2 behavior is this: Simple Bayesian updating on season 1 foe-rate differences by sex or race would not move a player’s beliefs much, even with a diffuse prior. For example: If all players had a uniform prior over the foe rate, then Bayesian updating on season 1 play implies posterior beliefs of (approximately) .53 and .46 for men’s and women’s foe rates, respectively.<sup>15</sup> Clearly, simple Bayesian updating alone cannot explain the fact that in season 2 women chose to play foe against men 66% of the time, yet against women 45% of the time.

It is precisely this gap—the large difference in foe rates in season 2 (conditional on opponent) versus the smaller differences in season 1 (conditionally or not)—that our theoretical model of conditional cooperation attempts to bridge. The key observation is that simple Bayesian updating is not equilibrium play. To illustrate, consider Figure 4 and imagine players had a common (initial) prior of  $\hat{\pi} = .5$ . For concreteness’ sake, view the .53 posterior on men’s foe rate as a ‘small’ difference from the prior in the sense of .53 being a slight increment to the right of  $\hat{\pi}$  in Figure 4. Because the function  $f_0$  is flatter than (and above) the 45-degree line, men would actually play foe against women *more than 53%* of the time *if women believed men play foe 53%* of the time. This obviously cannot be equilibrium play—so women’s beliefs, and their foe rates against men, must be revised further upward. Beliefs and relative frequencies are in equilibrium when they coincide (that is, where  $f_1$  crosses the inverse of  $f_0$ )—which, as Figure 4 suggests, may be much larger than .53.

How much larger depends on the (conditional) distribution of  $G$  and  $S$  in the population, as they determine the shapes of the foe-rate curves  $f_0$  and  $f_1$ . Although  $G$  and  $S$  are unobservable, differences in equilibrium foe rates will exceed the ‘small’ differences in simple Bayes posteriors by larger amounts if there are many people in the population with high levels of  $G$  (guilt). This theoretical requirement is partly why we noted the remarkably high levels of  $G$  implied by the

---

<sup>15</sup> This is a binomial likelihoods calculation using the outcomes in Table 1. One might also consider the season 1 differences by sex and race to be ‘small’ in the sense that they are individually statistically insignificant in the bivariate probit results (Table 1). This is a peculiar basis for evaluating whether *players* viewed these differences as too small to act upon, however: it assumes contestants evaluate foe rates the way an empirical economist does (with classical statistics and arbitrary  $p$ -values), instead of updating subjective beliefs like players in games of incomplete information.

data in Section IV.B.

### **F. Conditional Cooperation and 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,<sup>16</sup> and—to a large degree—because players were far more likely to walk away empty-handed. Table 3 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 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 players' efficiency.

This is not true across all demographic groups, however. Players with observable characteristics initially associated with low rates of friendly play fare worse (monetarily) over time, relative to players with characteristics associated with higher rates of cooperativeness. Table 3 shows that for women, white players, and older players, average winnings per player as a share of the stakes changed from Season 1 to 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 conditional strategies.

In contrast, the opposite is true for players whose observable characteristics are associated with less-friendly play. For male players, average winnings as a share of stakes fall a significant 17 percentage points between seasons. A similar decline of 15 percentage points occurs for younger players, when grouped by the median age. More dramatically still, black players experience a precipitous decline of 39 percentage points between seasons. In effect, the conditional strategies adopted by players in Season 2 yield a drastic fall in monetary gains for players tagged as having higher rates of unfriendly play.

### **V. Caveats**

Two aspects of our analysis give rise to caveats. One is the question of how players weigh

---

<sup>16</sup> The show's producers appear to have used more difficult trivia questions during Season 2, lowering the average stakes  $x$ .

multiple, potentially-conflicting opponent attributes when deciding how to play. If a player whose season 1 prior does not depend on opponents' attributes observes season 1 play, and then faces a season 2 opponent who has some attributes that predict cooperation and other attributes that do not, how does our player trade these off? It seems reasonable to suppose, as we do, that players can update on one dimension in the direction of equilibrium play. It seems more tenuous to assume they do in three. Thus, facing conflicting information there remains the interesting question of how players resolve the mixed signals. (This is the reason we focus our discussion on univariate comparisons in sections IV.C. and D.) Perhaps some attributes, such as sex, are simply more 'salient' to decision-making than others. For instance, older players choose foe against like opponents surprisingly often in season 2, given their relatively low foe rate in season 1, a result that is difficult to rationalize with our model of equilibrium play if individuals condition on age alone.

Second, it is conceivable that players' learning from season 1 influenced the pre-game player matching process and hence rates of cooperation.<sup>17</sup> We have little evidence that the frequencies of pairings (by observable attributes) changed between the two seasons. For example, women are no more frequently paired with other women in season 2 than season 1, despite the availability of information in season 2 that it was desirable for a female player to be paired with a female opponent. Still, because season 2 players were better able to judge the likelihood that their partner would choose foe, we cannot exclude the possibility that players' sense of having been matched with a less desirable partner might have influenced rates of cooperation in season 2.

## VI. Conclusion

Several interesting findings emerge from our study. First, the data suggest that contestants on *Friend or Foe* learned to evaluate players' cooperative tendencies from the history of past play by others. These conditional strategies identify a natural mechanism by which some players are able to implement conditionally cooperative preferences: They prefer to cooperate with opponents whom past play indicates are more likely to be similarly cooperative, and are less inclined to cooperate with players who are predictably less cooperative.

Second, the monetary winnings of players tagged as less likely to cooperate on the basis of their observable attributes declines substantially after conditional cooperation emerges. In con-

---

<sup>17</sup> We thank one of the referees for emphasizing this point.

trast, players who are predictably more cooperative seem to maintain an “island of cooperation” where winnings are stable.

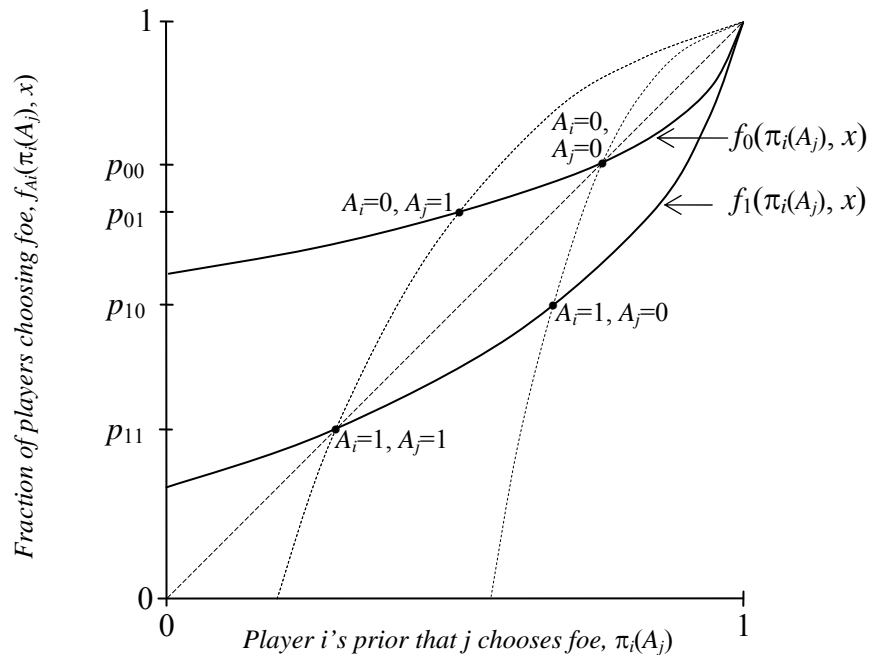
Third, we document that non-monetary payoffs—the value of playing fairly—must be remarkably large, and roughly proportional to the monetary stakes. It is one thing to forgo \$5 in order to be fair, as players typically do in laboratory games; the players on *Friend or Foe* commonly forgo over \$1,350. Preferences for fair divisions are apparently strong. Moreover, since this game is one-shot, it appears this preference for fairness is indeed a preference—or a norm—rather than an investment or signal whose value requires future interaction.

Finally, our study provides mixed news about the prospects for voluntarily cooperative behavior in simultaneous-move, one-shot contexts. Although many people are willing to forgo a great deal to implement equal divisions, interactions do not always unfold cooperatively because others may not share this preference. Successful conditional cooperation requires information about the relationship between players’ observable characteristics and preferences. As this information became available in the limited way that *Friend or Foe* allows, cooperation became less common overall. We conclude that while preferences for fair play are evidently very strong among many individuals, they alone are not sufficient to bring about more frequent cooperative outcomes.

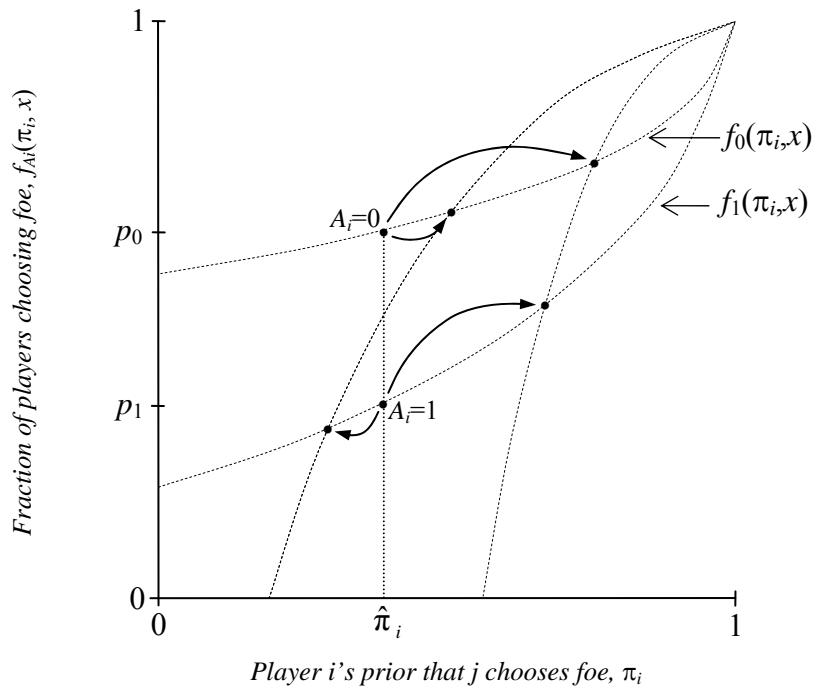
## References

- 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, and Matthew Rabin (2002). Understanding Social Preferences with Simple Tests. *Quarterly Journal of Economics* 117(3): 817-69.
- Falk, Armin, and Urs Fischbacher (1998). A Theory of Reciprocity. Working Paper No. 6, University of Zurich.
- 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.
- Fershtman, Chaim and Uri Gneezy (2001). Discrimination in a Segmented Society: An Experimental Approach. *Quarterly Journal of Economics*, Vol. 116(1): 351-377.
- Gertner, Robert (1993). Game Shows and Economic Behavior: Risk-Taking on "Card Sharks." *The Quarterly Journal of Economics* 108(2): 507-21.
- 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.
- Kandori, Michihiro (1992). Social Norms and Community Enforcement. *Review of Economic Studies*, 59: 63-80.
- 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.

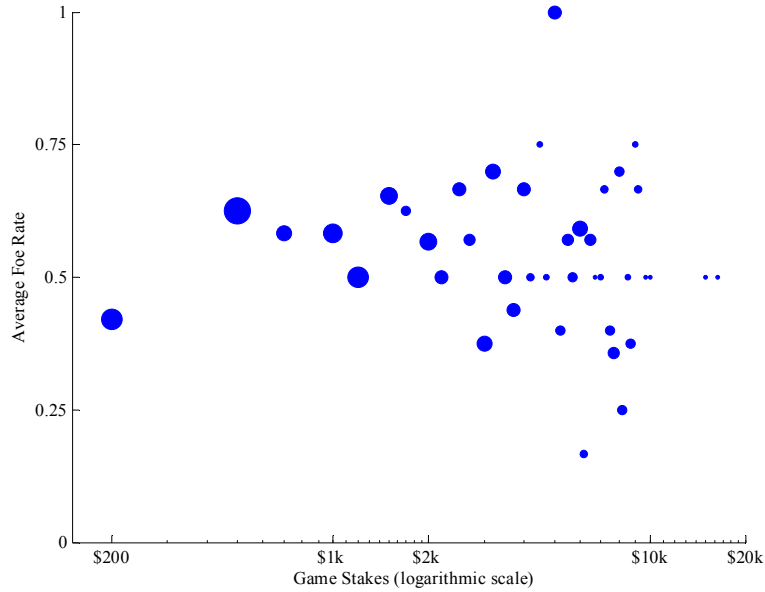
- Levine, David K. (1998) Modelling Altruism and Spitefulness in Experiments. *Review of Economic Dynamics* 1: 593-622.
- List, John A. (2006a) Friend or Foe: A Natural Experiment of the Prisoner's Dilemma. *Review of Economics and Statistics* (forthcoming).
- List, John A. (2006b) The Behavioralist Meets the Market: Measuring Social Preferences and Reputation Effects in Actual Transactions. *Journal of Political Economy* 114(1): 1-37.
- Metrick, Andrew (1995). A Natural Experiment in "Jeopardy!" *American Economic Review* 85(1): 240-53.
- Masclet, D., C. Noussair, S. Tucker, and M.C. Villeval (2003) Monetary and Nonmonetary Punishment in the Voluntary Contribution Mechanism. *American Economic Review* 93(1): 366-80.
- Oberholzer-Gee, Felix, Joel Waldfogel, and Matthew White (2003) Social Learning and Coordination in High-Stakes Games: Evidence From Friend or Foe. NBER Working Paper #9805.
- 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.
- Rege, M. and K. Telle (2004). The Impact of Social Approval and Framing on Cooperation in Public Good Situations. *Journal of Public Economics*, June, 1625-44.
- Rotemberg, Julio J. (2004). Minimally Acceptable Altruism and the Ultimatum Game. Working paper. Harvard Business School.
- 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.



**FIGURE 3**



**FIGURE 4**



**FIGURE 5.** Average Foe Rates and Game Stakes,  $n = 600$  players. Circle area is proportional to the number of players at each stakes level.

**TABLE 1**  
**Foe Rates by Player and Opponent Characteristics**

Player	Opponent	<i>Season 1</i>		<i>Season 2</i>	
		Games	Player Foe Rate	Games	Player Foe Rate
Male	Male	24	52%	42	61%
	Female	69	54%	111	58%
Female	Male	69	48%	111	66%
	Female	26	44%	40	45%
White	White	68	51%	113	53%
	Black	37	46%	47	75%
Black	White	37	57%	47	70%
Younger (Age $\leq 27$ )	Younger	16	66%	60	60%
	Older	64	64%	94	60%
Older (Age $> 27$ )	Younger	64	38%	94	62%
	Older	37	40%	29	55%

**TABLE 2**  
**Bivariate Probit Estimates of Game Outcomes**  
 Dependent variable is a (Foe, Foe) pair. Table entries are marginal effects  
 on foe probability, evaluated at covariate means. Standard errors in parentheses.

	(1)		(2)		(3)		(4)	
<i>Season:</i>	First		Second		First		Second	
Log Score	-0.02	(0.03)	0.02	(0.03)	-0.03	(0.04)	0.02	(0.03)
Player Age	-0.02	(0.004)	-0.001	(0.003)	-0.03	(0.02)	0.002	(0.01)
Player Black	0.13	(0.09)	0.12	(0.07)	0.13	(0.09)	0.16	(0.07)
Player Male	0.09	(0.07)	0.03	(0.05)	0.10	(0.09)	0.17	(0.08)
Player West	0.09	(0.07)	0.17	(0.05)	0.09	(0.07)	0.18	(0.05)
Opponent Age					-0.01	(0.02)	0.004	(0.01)
Opponent Black					-0.01	(0.09)	0.15	(0.07)
Opponent Male					-0.01	(0.10)	0.22	(0.08)
Opponent West					-0.05	(0.07)	-0.006	(0.05)
Player Male × Op. Male					-0.04	(0.14)	-0.24	(0.11)
Player Age × Op. Age					0.001	(0.001)	-0.001	(0.001)
$\rho$	-0.02	(0.15)	0.17	(0.12)	-0.02	(0.15)	0.14	(0.13)
$H_0$ : <i>Player Characteristic Effects All Zero</i>	$\chi^2_4 = 20.2$		$\chi^2_4 = 14.1$					
	$p = 0.001$		$p = 0.007$					
$H_0$ : <i>Opponent and Interaction Effects All Zero</i>					$\chi^2_6 = 1.36$		$\chi^2_6 = 13.2$	
					$p = 0.97$		$p = 0.042$	
Observations (pairs)	117		183		117		183	

**TABLE 3**  
**Winnings by Group and Season**

Group	Season	Number of Players	Mean Stakes	Mean Winnings	Ratio of Winnings to Stakes
All	1	240	\$ 3,718	\$ 1,463	39%
	2	390	\$ 3,063	\$ 926	30%
Men	1	118	\$ 4,331	\$ 1,848	43%
	2	195	\$ 3,217	\$ 820	26%
Women	1	121	\$ 3,101	\$ 1,074	35%
	2	193	\$ 2,920	\$ 1,043	36%
White	1	183	\$ 3,915	\$ 1,389	35%
	2	287	\$ 3,237	\$ 1,082	33%
Black	1	40	\$ 2,913	\$ 1,529	52%
	2	55	\$ 2,787	\$ 368	13%
Younger (Age ≤ 27)	1	96	\$ 3,687	\$ 1,665	45%
	2	214	\$ 3,336	\$ 999	30%
Older (Age > 27)	1	144	\$ 3,740	\$ 1,328	36%
	2	176	\$ 2,722	\$ 838	31%