A Tractable Model of Reciprocity and Fairness

James C. Cox
University of Arizona
jcox@bpa.arizona.edu

Daniel Friedman
University of California
Santa Cruz
dan@cats.ucsc.edu

September 2002

A Tractable Model of Reciprocity and Fairness[1]

By James C. Cox and Daniel Friedman

Abstract.
We introduce a parametric model of other-regarding preferences. The income distribution, other status considerations, and the kindness or unkindness of others' choices ("intentions") systematically affect a person's emotional state. The emotional state then determines the marginal rate of substitution between own and others' payoffs, and thus the person's subsequent choices.

Model applications are illustrated with two sets of laboratory data: simple binary choice mini-ultimatum games and more complex choice moonlighting games. The results support the usefulness of modeling other-regarding preferences as being conditional on others' intentions as well as the income distribution.

1. Introduction

Everyone knows that people care about other people. Even economists know this,[2] but only recently have we begun to recognize the need for explicit models. Under what circumstances will I bear a personal cost to help or harm you? What is the marginal rate of substitution between my own payoff and yours? In this paper we propose a model that addresses such questions and, using some existing laboratory data, illustrate its application.

Many things may affect how I care about you, but two general motives stand out. First is status, or relative position: are you a member of my family, or my boss, my employee, a wealthy or poor neighbor? In the laboratory data we examine, the most obvious such variable is the distribution of income: what is your current payoff relative to my current payoff?

A second motive is reciprocity: how I respond to your intentions towards me. If I think you have helped me in the past or want to help me in the future, I am more likely to value your welfare. Of course, economists are familiar with folk theorem arguments that I help you so that you will help me in the future and thereby increase the net present value of my payoff stream.

[1] We are grateful to the National Science Foundation for support under grant SES-9818561. We are also grateful to Gary Charness, Steffen Huck, Lori Kletzer, Lisa Rutstrum, and Daniel Zizzo for helpful comments.

Reciprocity here refers to something quite different, although complementary: if you are my friend, I find it pleasurable to increase your material payoff, whether or not it affects the present value of my own material payoff. Negative reciprocity is also included: if you are my foe (e.g., I think you have harmed me or my friends, or will do so when you have the opportunity), I enjoy decreasing your material payoff.

The basic modeling idea is that status and reciprocity affect my emotional state, summarized in a scalar variable θ, and my emotional state affects my choices. We retain the conventional assumption that I choose an available alternative that maximizes my utility function, and we follow recent contributions in allowing the utility function to depend on your material payoff y as well as my own material payoff m. The simplest example is $u(m, y) = m + \theta y$. Our innovation is to model the emotional state θ as systematically affected by the reciprocity motive r as well as by the status motive s.

Section 3 below proposes specifications of the model elements r, s, and θ, and proposes a somewhat more general utility function that allows non-linear indifference curves. Section 2 sets the stage by summarizing recent related literature. Section 4 applies the model to laboratory data from mini-ultimatum games, simple extensive form games where both players have binary choices. Section 5 applies the model to laboratory data from moonlighting games, more complicated two-player extensive form games where both players have a range of choices. Section 6 suggests further applications, and Section 7 offers a concluding discussion. Technical details from Sections 4 and 5 appear in an Appendix.

2. Recent Approaches

Economic models traditionally assume that decision-makers are exclusively motivated by material self-interest. Maximization of own material payoff predicts behavior quite well in many

[2] It is the central theme of Adam Smith (1759). Fehr and Falk (2002) summarize recent evidence on the economic impact of motives beyond self-interest.

contexts. Important examples include: competitive markets, even when gains from trade go almost entirely to sellers or almost entirely to buyers (Smith and Williams, 1990); one-sided auctions with independent private values (Cox and Oaxaca, 1996); procurement contracting (Cox, et al., 1996); and search (Cason and Friedman, forthcoming; Cox and Oaxaca, 1989, 2000; Harrison and Morgan, 1990).

Maximization of own material payoff predicts poorly in a variety of other contexts. Examples include ultimatum games (Güth, Schmittberger, and Schwarze, 1982; Slonim and Roth, 1997), voluntary contribution of public goods games (especially such games that allow costly opportunities for punishing free riders, e.g., Fehr and Schmidt, 1999), and experimental labor markets (e.g., Fehr, Gächter, and Kirchsteiger, 1997).

Such laboratory data, together with suggestive field data, has encouraged the development of models with other-regarding preferences. This literature falls into two broad classes (see Fehr and Schmidt, 2001, for a more complete survey of models and evidence). First there are the distributional models of Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Charness and Rabin (2001) and Cox, Sadiraj and Sadiraj (2002a). To facilitate comparison with our specifications, we write out two-player versions of these models.

The Fehr-Schmidt model has piecewise linear indifference curves over my income m and your income y, with two marginal rate of substitution parameters $0 \leq \beta \leq \alpha \leq 1$ for the cases that my income is less than or greater than yours. The utility function is

(1) $u(m, y) = m - \alpha(y - m)$, if $m < y$

$= m - \beta(m - y)$, if $m \geq y$.

That is, I like own income and dislike income inequality, especially when I have the short end. For two players, the Charness and Rabin distributional model looks the same except that the MRS parameters have fewer restrictions, and so can include competitive preferences ($\beta < 0 < \alpha$), inequality- or difference-averse preferences ($\beta, \alpha > 0$), and quasi-maximin preferences ($1 > \beta >$ -

$\alpha > 0$). The Bolton-Ockenfels (2000) model also assumes that I like own income and dislike income inequality, but the utility function takes the non-linear form

(2) $\quad v = v(m, m/(m+y))$.

The function *v* is assumed to be globally non-decreasing and concave in the first argument; to be strictly concave in the second argument, relative income $m/(m+y)$; and to satisfy $v_2(m,½) = 0$, for all m. The Cox, Sadiraj and Sadiraj (2002a) model includes nonlinear indifference curves for egocentric other-regarding preferences. The utility function has the form,

(3) $\quad u(m,y) = [m^a + \theta_-(y^\alpha - m^\alpha)]^{1/\alpha}$, if $m < y$,

$\quad = [m^a + \theta_+(y^\alpha - m^\alpha)]^{1/\alpha}$, if $m \geq y$,

with parameter restrictions, $0 < \alpha < 1$; $\quad 0 \leq \theta_- \leq \theta_+ < 1$; $\quad$ and $\theta_- < 1 - \theta_+$. Thus I am not averse to income inequality; I like own income, and your income, but my marginal rate of substitution depends on whose income is higher, and in comparing payoff pairs, $(m,y) = (c,d)$ and $(m,y) = (d,c)$, I prefer (c,d) to (d,c) when $c > d$.

The main alternatives so far to these distributional preference models are equilibrium models that try to capture the reciprocity motive in terms of beliefs regarding intentions. Building on the psychological games literature (e.g., Geanakoplos, Pearse and Stacchetti, 1989), Rabin (1993) develops a theory of fairness equilibria (for 2 player games in normal form) based on the following representation of agents' utilities. Define a_i, b_j, and c_i, respectively, as the strategy chosen by player *i*, the belief of player *i* about the strategy chosen by player *j*, and the belief by player *i* about the belief by player *j* about the strategy chosen by player *i*. Rabin (1993, pp. 1286-7) writes the expected utility function for player *i* as

(4) $\quad U_i(a_i, b_j, c_i) = \pi_i(a_i, b_j) + \tilde{f}_j(b_j, c_i) \cdot [1 + f_i(a_i, b_j)]$,

where $\pi_i(a_i, b_j)$ is the monetary payoff to player i, $\widetilde{f}_j(b_j, c_i)$ is players i's belief about how kind player j is being to him, and $f_i(a_i, b_j)$ is how kind player i is being to player j (relative to a benchmark taken to be the average of the highest and lowest possible payoffs). Thus negative reciprocity ($\widetilde{f}_j$ and f_i both negative) as well as positive reciprocity increase utility. The model looks for equilibria in actions and beliefs about intended kindness; typically there are many such equilibria.

Dufwenberg and Kirchsteiger (1999) propose an extension to extensive form games with N players, and Falk and Fischbacher (1999) propose a different extension that also covers incomplete information but uses a distributional preference utility function. Charness and Rabin (forthcoming), in addition to their distributional model, also propose an equilibrium model involving distributional preferences and beliefs about other players' intentions. All the models are complex and have many equilibria, and so seem intractable in most applications. Such problems seem unavoidable for models that assume equilibrium in higher order beliefs.

Levine (1998) improves tractability by replacing beliefs about others' intentions by estimates of others' types. In his model, players' utilities are linear in their own monetary payoff y_i and in others' monetary payoffs y_j, $j \neq i$. For two player games, utilities are of the form

$$(5) \qquad v_i(y_i, y_j) = y_i + \frac{a_i + \lambda a_j}{1 + \lambda} y_j,$$

where $a_k \in (-1,1)$ is agent k's type or "coefficient of altruism" (actual for $k = i$ or estimated for $k = j$), and $\lambda \in [0,1]$ is a weight parameter. Levine demonstrates that his model is consistent with data from some ultimatum game and market experiments, and it clearly is more tractable than the previous equilibrium models.

We shall propose a more drastic simplification. Instead of beliefs or type estimates we use emotional states based on actual experience: my attitude towards your payoffs depends on my

state of mind, e.g., kind or vengeful, and your actual behavior systematically alters my emotional state. Our model is consistent with the axiomatic approaches of Sobel (2000) and Guttman (2000) but is more explicit. It is simply a preference model, not an equilibrium model, and therefore sidesteps many of the complications involving higher order beliefs. But unlike the distributional preference models discussed above, in our model an agent's distributional preferences are conditional on the revealed intentions of others.

Recent experiments compare the explanatory power of earlier models. Evidence contrary to the (unconditional) distributional preference models includes the following. Kagel and Wolfe (2001) find that rejection rates in the ultimatum game are essentially unaffected by unequal (high or low) contingent payments to a third (strategic dummy) player. In four separate public goods experiments, Croson (1999) finds positive relations between own contribution and (a) own beliefs about others' contributions and (b) actual contributions of others, especially with the median of others' contributions. In mini-ultimatum games (discussed further in section IV below), Falk, Fehr & Fischbacher (2001) find that the rejection rate for a [2 of 10] offer declined as the alternative offer (not chosen by the proposer) became less favorable to the respondent. They also find that people punish even when the punishment does not reduce payoff inequality. Brandts and Charness (2000) find that deception in the prior cheap talk stage significantly increases the punishment rate, and some subjects reward favorable sender behavior. Blount (1995) finds that responders in her ultimatum games accepted lower offers when they were randomly generated than when they were chosen by human subjects. Offerman (1999) has similar results: intentional helpful (hurtful) actions were rewarded (punished) more frequently than identical but randomly generated actions. See also Ahlert et al (1999), Charness (1996), Guth and Kavacs (2001), Gibbons and Van Boven (1999), and Kagel, Kim and Moser (1995).

On the other hand, there are some empirical studies that seem more favorable to unconditional distributional preferences than to reciprocal preferences, including Bolton, Katok and Zwick (1997) and Bolton, Brandts and Ockenfels (1997). Cason, Saijo, and Yamoto (1999)

look at voluntary contributions public good games with a prior participation decision. They conclude that "spite" is more prevalent in Japan than in US subject pools, but eventually outcomes are more efficient in Japan.

Cox (2002a, 2002b) uses a triadic experimental design to discriminate between actions motivated by unconditional distributional preferences and actions motivated by reciprocity considerations, in the context of the Berg, Dickhaut, and McCabe (1995) investment game. Using dictator game treatments, as controls, the experiments support the conclusion that behavior is significantly motivated by altruism as well as trust and positive reciprocity. Cox, Sadiraj, and Sadiraj (2002b) use a triadic design in the context of the moonlighting game introduced to the literature by Abbink, Irlenbusch, and Renner (2000). Cox, et al. report that altruism and positive reciprocity (but not negative reciprocity) are significant motives for behavior in the moonlighting game. The data are re-examined in section 5 below.

Cox and Deck (2002) report data from eleven experimental treatments involving 692 subjects that provide a systematic exploration of the existence and nature of motives for reciprocal behavior in two-person games. The triadic experimental design supports discrimination between motivations of reciprocity and (non-reciprocal) altruism. They find significant positive reciprocity in the trust (or mini-investment) game when it is run with a single-blind protocol but not when it is run with a double-blind protocol. They do *not* find significant negative reciprocity in the "punishment" game (i.e., the (5, 5) mini-ultimatum game) when it is run with a double-blind protocol in a triadic design.

In summary, the laboratory evidence confirms that people do care about others' payoffs as well as their own. The marginal rate of substitution (between my payoff and yours) is not constant, however, and may be affected by reciprocity as well as distributional and other status motives. There is room for a tractable model that can assess empirically the significance of the various motives.

3. Model Specifications

We now propose a new model of preferences that incorporates objectively defined variables *r* and *s* capturing reciprocity and status motives. Due to its importance in existing literature, the distribution or relative payoff is separated from other aspects of the status motive. For pedagogical and comparative purposes, we restrict attention here to two player extensive form games of complete information. The model shows how the second mover's emotional state defines his marginal rate of substitution (MRS) between final own payoff *m* and final first player payoff *y*, and how the emotional state responds to the values of *r* and *s* that arise from the first mover's choice.

First note that the shape of indifference curves in (*m*, *y*) space carries a particular reaction to relative payoff. To see this clearly, suppose for the moment that both payoffs are positive and that the second mover has kind preferences (i.e., increasing in both own and other's payoff). The indifference curves then have the usual negative slope = -MRS. If preferences are convex, the MRS increases as one moves along any indifference curve in the direction of increasing *y/m* ratio; see panel A of Figure 1. But *y/m* is a natural way to specify relative payoff. The MRS is independent of *y/m* when indifference curves are linear, and greater sensitivity to *y/m* takes the form of more convex preferences.

With homothetic preferences, all indifference curves have the same slope where they cross any given ray, *y/m* = constant; in this case relative payoff dependence is well defined. Fortunately the convenient and well-known constant elasticity of substitution (CES) utility function represents homothetic preferences. Written in general form, the CES utility function is $u(m, y) = (m^{\alpha} + \theta y^{\alpha} + c)^{1/\alpha}$; see also Cox, Sadiraj, and Sadiraj (2002a). For convenience we use the ordinally equivalent expression

(6) $\quad u(m, y) = m^{\alpha} + \theta y^{\alpha}$.

With CES preferences we have MRS = $(\partial u/\partial m)/(\partial u/\partial y) = \theta^{-1}(y/m)^{1-\alpha}$. Hence the emotional state θ is the willingness to pay (WTP=1/MRS) at an allocation on the 45° line $m = y$. Preferences are linear (and MRS is constant) if $\alpha = 1$, and preferences are strictly convex and MRS is strictly increasing in y/m along indifference curves iff $\alpha < 1$. It is well known (and one can verify from the MRS expressions) that preferences are Cobb-Douglas if $\alpha = 0$. Preferences approach Leonieff as $\alpha \to -\infty$.

The emotional state θ is a function of the reciprocity motive r and the (residual) status motive s. A natural specification for the reciprocity variable is $r = m(x) - m_0$, where $m(x)$ is the maximum payoff the second mover can guarantee himself given the first mover's choice x, and m_0 is $m(x)$ when x is neutral in an appropriate sense. The idea is that the second mover regards additional payoff as kindness to be reciprocated, and shortfalls from m_0 as violations of his property rights, to be negatively reciprocated.[3] Often it is convenient to normalize r so that it lies in [-1, 1]. Let $m_g = \max_x\{m(x)\}$ and $m_b = \min_x\{m(x)\}$. The normalized version is

$r = (m(x) - m_0)/(m_g - m_b)$, when $m_g > m_b$, and $r = 0$ otherwise.

The variable s represents relative status (other than relative payoff, which is already accounted for.) Assume that social norms assign real (possibly integer) status values s_1 and s_2 to the first and second movers in the context of the game currently played; these may depend on the roles played as well as on observable personal characteristics such as gender, age, job title, etc. Then a natural specification is $s = s_1 - s_2$. For example, under some social norms the first mover's status and hence s would increase if her gender were female and if she had to earn the right to play the role.

[3] Konow (2001) elaborates an objective theory of m_0 as a function of the agent's relative actual effort levels ("accountability"), the efficient effort levels, the agents' basic material needs, and the context. Konow (2000) extends (part of) this theory to allow for self-serving subjective distortions of the objective m_0, and confronts evidence from dictator games. (In our framework, this game entails a strategic dummy first mover.) Gächter and Riedl (2002) offer a general discussion and demonstrate the impact of m_0 (which they call moral property rights or entitlements) in new laboratory data.

In estimating the model, we maintain

Assumptions A.

1. Individuals choose so as to maximize a utility function of the form (6).
2. The emotional state function $\theta = \theta(r, s)$ is identical across individuals except for a mean zero idiosyncratic term.
3. $\theta(r, s)$ is weakly increasing in r and s.
4. $\theta(0, 0)$ is non-negative but θ is negative when its arguments are sufficiently negative.

The case of negative θ deserves a brief comment before presenting sample applications. A person with negative θ is willing to pay to *reduce* other's payoff. That is, y is a "bad" rather than a "good," and the indifference curves slope upward. CES preferences then have one straight line indifference curve, the ray $y/m = |\theta|^{-1/\alpha}$ corresponding to $u = 0$, and the other indifference curves fan in slightly towards this ray as in Panel B of Figure 1.

4. Evidence from Mini-Ultimatum Games

Mini-ultimatum games (Bolton and Zwick, 1995; Gale et al, 1995) have an especially simple structure that is amenable to our approach. As illustrated in Figure 2, the first mover (the "proposer") offers one of two possible positive payoff vectors, and the second mover (the "responder") either accepts the offer, which then becomes the actual payoff vector, or else refuses, in which case the payoff is $0 = m = y$. In the 5/5 game, for example, if the proposer chooses left (x = "Take") then the responder chooses between payoff vectors $(y(x), m(x)) = (8, 2)$ and $(0, 0)$, and if the proposer choses right (x = "Share") then the responder chooses between $(5, 5)$ and $(0, 0)$.

We want to explain responder choice, coded

(7) $Z = 0$, if (0,0) is chosen,

= 1, otherwise.

It is natural to use probit estimation, with explanatory variables derived as follows. A reasonable candidate for the responder's property right m_o is his feasible payoff that is closest to equal split but not higher than the proposer's. Thus we set $m_o = \min\{5, m_g\}$. In the 8/2 game in Figure 2, the reciprocity variable is $r = 0$ because the proposer has no real choice and $m_g = m_b$. In the other three games $m_g = \max_x\{m(x)\} > \min_x\{m(x)\} = m_b$ and the normalized reciprocity variable is $r = (m(x) - m_o)/(m_g - m_b)$.

The mini-ultimatum game data reported Falk, Fehr, and Fischbacher (2001, henceforth denoted FFF01) contain no variation in the status variable (other than relative payoff), so s is constant. By Assumption A and a first order Taylor series approximation, responder i has trade-off parameter $\theta_i = A + Br + \sigma e_i$, where (for the constant value of s) A is the population average value of θ at $r = 0$, and B is the non-negative responsiveness to r. Slightly strengthening A.2, we assume here that idiosyncratic individual differences are Normally distributed with variance $\sigma^2 > 0$.

By utility maximization, $Z = 1$ iff $0 = u(0, 0) < u(m(x), y(x)) = m(x)^{\alpha} + \theta_i y(x)^{\alpha}$. Dropping the (x) arguments for notational simplicity, the condition is easily seen to be $0 < [m/y]^{\alpha} + \theta_i = [m/y]^{\alpha} + A + Br + \sigma e_i$, or $-e_i < \sigma^{-1}([m/y]^{\alpha} + A + Br)$. Hence the probability that $Z = 1$ is the standard cumulative Normal distribution evaluated at $\sigma^{-1}([m/y]^{\alpha} + A + Br)$, and probit estimation will recover the structural parameters.

Using the FFF01 data and the LIMDEP probit procedure, we searched across various values of α, and found that likelihood was maximized in the vicinity of $\alpha = 1/4$ (with $\alpha = 1/8$ almost as good). The estimated equation is

$$\Pr(Z_t) = -0.49 + 0.69 r_t + 2.00(m/y)^{\alpha} + e_t. \quad (8)$$

The equation predicts correctly 302 of the 360 subjects' choices. The coefficient estimate for $(m/y)^{\alpha}$ implies that $\sigma^{-1} = 2.00$ and σ is 0.5, with a *p*-value of 0.0000. The coefficient estimate for *r*, with *p*-value of 0.001, implies that $B = \partial \theta / \partial r$ is about 0.69/2 or 0.35. That is, moving *r* from 0 to 1 (or from –1 to 0) would on average increase the probability that the second mover would accept the proposal by about 0.35 of a standard deviation. Likewise, other things equal, moving relative income *m*/*y* from 0.5 to 1 would increase the acceptance probability by about

$2.00(1 - 2^{-0.25}) \approx 0.32$ standard deviations.

The coefficient estimates are fairly robust to changes in α. For α = 1/8 the point estimates are within 10% of those given, and the coefficient on *r* doesn't change much even for α as low as –4. (With negative α, the portion of the data with $m = 0$ needs to be omitted or modified to avoid the zero divide problem.) The coefficient increases to 1.3 as α increases to its upper limit of 1, but the fit deteriorates substantially.

5. Evidence from Moonlighting Games

It is more challenging to apply our model to moonlighting game data. In the experiment of Cox, Sadiraj and Sadiraj (2002b, henceforth CSS02b), both first and second movers have an endowment of $10. Let *x* be the amount sent ($0 \le x \le 10$, a kind act) or taken ($-5 \le x < 0$, an unkind act) by the first mover. The experimenter triples any positive value of *x* before giving it to the second mover but does not transform negative *x* values. The second mover then chooses an amount *z* to increase or decrease the first mover's payoff. Positive values of *z* are simple transfers to the first mover, but negative values reduce both payoffs, the first mover's at three times the rate of the second mover's. That is, the final payoff vector has components

(9) $m = 10 + (1+2 I_x) x - |z|$ and

(10) $y = 10 - x + (1+2 I_{-z}) z$,

where the indicator function $I_w = 1$ if $w > 0$ and otherwise is 0. Constraints on z ensure that both m and y are non-negative.

Figure 3 illustrates the payoffs in terms of the second mover's choice set following the first mover's choice x. The interim allocation $m(x) = 10 + (1+2I_x)x$ and $y(x) = 10 - x$ defines a kink at $z = 0$. The choice segment for $z > 0$ has slope –1 as in the usual budget set with unit prices, but the choice segment for $z < 0$ has slope +3. The non-negativity constraints are $-(10-x)/3 \le z \le 10 + 3x$ when $x > 0$, and $-(10-x)/3 \le z \le 10 + x$ when $x \le 0$.

Under Assumption A.1 (utility maximization), choices on either segment's interior reveal precisely the second mover's MRS and trade-off parameters. Such choices are characterized by the first order condition –slope = MRS = $\theta^{-1}(y/m)^{1-\alpha}$, so interior choices $z > 0$ reveal $\theta_i = (y/m)^{1-\alpha} > 0$ and interior choices $z < 0$ reveal $\theta_i = -(1/3)(y/m)^{1-\alpha} < 0$, where m and y are given by equations (9) and (10) evaluated at the second mover i's choice of z, given x. Choices at the kink reveal intermediate values of θ_i and choices at corners reveal more extreme values.

The corners and kink complicate fitting the model to the CSS02b data. Of 30 observations, three are at the neutral kink $z = x = 0$, which are choices that are consistent with a very wide range of θ_i. Another twelve observations involve the extreme m_b case $x = -5$, and five other observations involve the extreme m_g case $x = 10$. The data do not include any variation in status (other than relative income). Hence it is impractical to estimate a smooth function $\theta_i(r, s)$ from these data.

Our empirical strategy, consistent with Assumptions A.2 and A.3, is to estimate a piecewise constant function with two values, one (call it a) representing second movers' average emotional state θ_i following a kind first move ($x \ge 0$), and the other (call it b) representing average θ_i following an unkind first move ($x < 0$). Note that the reciprocity variable r has the

same sign as x because $m(x)$ is increasing in first mover's choice x, and $m(0) = m_o = 10$. Hence by Assumption A.4, the model predicts that a is positive and that b is negative.

The estimates are derived as follows. At an interior solution $z > 0$, insert equations (9) and (10) into the tangency condition $(y/m)^{1-\alpha} = \theta_i = a + e_i$, evaluate at the mean error $e_i = 0$, and solve for z to obtain, for $\alpha < 1$,

$$(11) \qquad z_i = 10\frac{a^{\frac{1}{1-\alpha}} - 1}{a^{\frac{1}{1-\alpha}} + 1} + \frac{3a^{\frac{1}{1-\alpha}} + 1}{a^{\frac{1}{1-\alpha}} + 1}x_i.$$

At an interior choice $z < 0$, the tangency condition is $(y/m)^{1-\alpha} = -3\theta_i = -3(b + e_i)$, and the expression becomes

$$(12) \qquad z_i = 10\frac{(-3b)^{\frac{1}{1-\alpha}} - 1}{3 - (-3b)^{\frac{1}{1-\alpha}}} + \frac{(-3b)^{\frac{1}{1-\alpha}} + 1}{3 - (-3b)^{\frac{1}{1-\alpha}}}x_i.$$

Equations (11) and (12) suggest a linear regression on four explanatory variables—call them P $= I_z$, N $= I_{-z}$, XP = xI_z, and XN = xI_{-z} —representing the intercepts and slopes when z is positive and negative. To account for the non-negativity constraints (and the sign constraints that determine which expression applies), we use a tobit regression with lower limit $z \geq -(10 - x)I_{-x}/3$ and upper limit $z \leq (10 + 3x)I_x$.

The data, however, include only two interior points where z is negative, so we can't get meaningful estimates of the N and XN coefficients. Dropping these variables, the Tobit procedure with individual random effects in the STATA package yields the result

$$(13) \qquad z_t = -5.17P_t + 1.92XP_t + v_t.$$

The standard errors for P, XP, and v respectively are 4.06, 0.53 and 5.33. The P coefficient is not quite significant at even the 20% level, so we conclude from (11) that the parameter a is in the vicinity of 1.0. Substituting the XP coefficient estimate of 1.92 into equation (11), we obtain

$a^{\frac{1}{1-\alpha}} \approx 0.85$ or, for α = ¼, we have $a \approx 0.89$. That is, at an equal allocation following a kind move, the average second mover is willing to pay about 89 cents to increase the first mover's payoff by a dollar.

One can estimate WTP following an unkind first move ($x < 0$) from data at the corner and kink. Twelve of thirteen unkind first moves are at the $x = -5$ extreme; in this case the corner is at the origin, i.e., the choice set segment with slope +3 lies on the ray $y/m = 3$. Recall from the end of section 3 and Figure 1 that, when θ is negative, the indifference curves consist of the ray $y/m = |\theta|^{-1/\alpha}$ together with almost parallel curves that fan in slightly towards this ray. Hence in the extreme case in which $x = -5$, we should not expect interior solutions as assumed in equation (12) but rather corner solutions, either at $z = 0$ when $-\theta_i = -(b + e_i) < 3^{-\alpha}$, or at $z = -5$ when $-\theta_i > 3^{-\alpha}$. Using the estimate α = ¼ from the mini-ultimatum game data we have $3^{-\alpha} \approx 0.75$.

The responses to $x = -5$ include three observations of $z = 0$ and six of $z = -5$, so we conclude that the parameter b is to the left of –0.75.[4] Using the error estimate $\sigma_v = 5.33$ from the regression (14) we obtain $b \approx$ –3.0. That is, at an equal allocation following an (extreme) unkind move, the average second mover is willing to pay about 3 dollars to decrease the first mover's payoff by a dollar. Of course, this estimate is much less precise than that for the parameter a; if the $x = -5$ responses had broken evenly between $z = 0$ and $z = -5$, then the estimate would be b = –0.75. See the Appendix for computation details and alternative estimates.

6. Further Applications

The preceding sections illustrate applications to existing data gathered for other purposes. The model, however, suggests new two player extensive form game experiments that elicit

[4] The remaining three x = -5 observations consist of two just off the corners (z = -1 and z = -4) that were noted earlier, plus one on the "wrong" segment (z=1). CSS02b interpret the last observation as a gesture of second mover's contempt rather than as the result of maximizing his preferences over feasible final allocations. There is only one other observation with $x < 0$, and the response was at the corner $z = 0$.

willingness to pay (WTP) own payoff for other's payoff while systematically varying relative income opportunities y/m, other aspects of status s, and reciprocity considerations r. With such data one could directly estimate the impact of each motive.

To illustrate simply, continue to hold s constant and take the linear Taylor series approximation of the systematic portion of the emotional state, $\theta = A + Br$, noting that the coefficients A and B depend on the particular value of s. Use a Taylor series expansion around the equal payoff position $y = m$ to obtain

$(m/y)^{1-\alpha} = 1 + (1-\alpha)(m-y)/y + \frac{\alpha^2 - \alpha}{2}(m-y)^2/y^2 + \ldots$. Use the reciprocity variable $r = m(x) - m_0$; this is observable given the first mover's choice $m(x)$ assuming that m_0 is unambiguous. Substitute these expressions into the basic CES relation WTP = $(m/y)^{1-\alpha}\theta$ from section 3, and use the Taylor series approximation of θ from section 3, to obtain

(14) WTP = $A + B(m(x) - m_o) + (1-\alpha)A(m-y)/y +$

$$(1-\alpha)B(m(x)-m_o)(m-y)/y + \frac{\alpha^2-\alpha}{2}A(m-y)^2/y^2 + \ldots$$

This equation suggests a simple OLS regression of the elicited WTP on variables formed from the observable interim allocation of my payoff, $m(x)$ and the final allocation of both payoffs, (m, y). From the coefficient estimates one recovers the desired structural parameters A, B, and α.

Future applications can explore the impact of other aspects of status. Possibly relevant treatments include the age, gender and observable socioeconomic characteristics, as well as the process that assigned the second and first mover roles. Available evidence suggests that the status variable s interacts strongly with the reciprocity variable r. For example, the CSS02b Treatment C data automate first movers, and the second movers' choices then are generally consistent with θ

= 0, suggesting a dominant interaction $r \times s$ with s = 0.[5] Zizzo and Oswald (2002) find that subjects with low status are particularly eager to "burn" the payoff of players with large unearned payoffs. Available theory, going back to Smith (1759), also suggests a positive interaction:

> Before any thing, therefore, can be the complete and proper object, either of gratitude or resentment, it must possess three different qualifications. First it must be the cause of pleasure in the one case, and of pain in the other. Secondly, it must be capable of feeling these sensations. And, thirdly, it must not only have produced these sensations, but it must have produced them from design, and from a design that is approved of in the one case and disapproved of in the other. –Adam Smith (1759, p. 181).

Future applications should also explore games with more than two players. The model extends directly. My utility function depends on every other player i's payoff y_i, via my emotional attitude towards him θ_i, and my utility function is simply

$$u(x, y_1, \ \dots, y_n) = x^\alpha + \theta_1 y_1^{\ \alpha} + \ \dots \ + \theta_n y_n^{\ \alpha}. \tag{15}$$

The way θ_i depends on the motives *r* and *s* is the same as in the two player case. Of course, in games where I can't separately identify the other players, there is only one θ. For games where I can observe individual player histories, the model could be enriched to include an indirect reciprocity motive as well as the direct motive captured in *r*.

7. Discussion

We hypothesize that a person's desire to help or harm others depends on emotional states that arise from universal motives such as reciprocity and status. In this paper we proposed a simple empirical model incorporating this hypothesis.

The first hurdle for an empirical model is tractability: can the model be estimated from available data? We obtained an affirmative answer by examining two existing data sets, laboratory studies of mini-ultimatum games (MUG) and moonlighting games (MLG). The MUG data consist of binary choices from the second mover following binary choices by a first mover.

[5] Alternatively, one could simply define m_o as the automated choice of the first mover and obtain $r = 0$ directly.

We derived and estimated a probit model that accounted for the data rather well and that produced parameter estimates consistent with a priori predictions (Assumption A.3-4). The MLG data consist of a range of choices by a second mover following a range of choices by a first mover, and the payoff is quite non-linear. Again we derived and estimated a model (this time using tobit regression) that accounts for the data and produces reasonable parameter estimates.

Of course, to be considered successful and important, an empirical model must jump further hurdles. Which variants work best? Can extensions deal with different sorts of data? How well do the best variants compare to alternative models? We close with a few thoughts on these matters.

Assumption A.2 states that individuals differ only in idiosyncratic additive components of the emotional state variable θ. It is possible that people also differ in their responsiveness to given reciprocity and status motives. Hence future work should compare variants of the model that allow multiplicative rather than additive individual random effects.

Specifying the property right m_o is crucial for the reciprocity motive. We chose the feasible allocation nearest to (but not exceeding) an equal split. This is quite reasonable, we think, in the games we consider. One can construct games in which it looks much less reasonable. But the same is true for any other specification of m_o, e.g., the midrange favored in the psychological games literature.[6] We propose that in empirical work one ask subjects (behind the veil of ignorance) to tell us what they think the second mover can reasonably expect. In games with a strong consensus on the property right we expect to see a relatively large coefficient on r.[7]

The definitions presented here extend directly to extensive form games in which some players have several moves. With suitable definitions of the interim endowment $m(x)$ and property right m_0 one could further extend the model to normal form games and some other

[6] A promising candidate in a variety of games is that m_o is my allocation in an efficient outcome.
[7] When there is little consensus one might expect a larger self-serving bias.

games of incomplete information. Future empirical work will show how successful such extensions are relative to available alternatives.

As we see it, our approach has several advantages that might survive beyond the current implementation. First of all, it is a model of preferences and choice, not equilibrium, and so is quite tractable and transparent. Second, it is more flexible than distributional preference models in that it takes other motives into account. Third, it is open to new findings in the psychology of emotions and so may facilitate interdisciplinary cross-fertilization.

Appendix: Computations and Alternative Estimates

A.1 Error Propagation

The STATA package reports a standard error of $\sigma_v = 5.3 \pm 1.2$ for individual subject effects in the tobit regression (13). However, the package assumes that the error enters additively, while our model assumes that the error originates in the structural parameter $\theta_i = a + e_i$. To see that this makes no significant difference, replace a by $a + e_i$ in the XP coefficient of (11), and take first order Taylor series approximations. After some tedious algebra, one obtains

(11a) $$z_i = c + \beta_o x_i + \beta_1 e_i x_i ,$$

where c and β_o are exactly as in (11) but

(16) $$\beta_1 = \frac{3ka^{2k-1}}{(a^k+1)^2} ,$$

with $k = 1/(1-\alpha)$. It is easy to see that $\beta_1 > 0$ and that $\beta_1 \to 0$ as $a^k \to 0$ and also as $a^k \to \infty$. Hence it is maximized at some positive value of a^k. The expression for the maximum in general is messy, but it is not very sensitive to α. In the Cobb-Douglas case ($\alpha = 0$ and $k = 1$) the upper bound is ¾ and occurs conveniently at $a = 1$, the current estimate. From (11a) we see that at this point $\sigma_v = \beta_1 \bar{x} \sigma_e \approx (3/4)(1.33)\sigma_e \approx \sigma_e$, since in our data the mean first mover choice is $\bar{x} \approx$ 1.33.

Under the maintained assumption in section 5 that the individual effect error term e_i is independent and identically distributed, it creates no bias in the coefficient estimate in (13). However, if e_i is positively correlated with x_i (as it is when θ is a strictly increasing function of r), then the reported XP coefficient estimate is known to be biased upward by Cov (e,x)/ Var (x) = $\sigma_e \rho_{ex} / \sigma_x$ (e.g., Pindyck and Rubinfeld, 1976, pg. 127). Using the empirical values $\sigma_e = 5.3$ and $\sigma_x = 6.3$ from our data, and noting that the strong individual effects make it unlikely that ρ_{ex} exceeds

1/3, we conclude that the coefficient is between $1.92 - 5.3/(3 \times 6.3) \approx 1.64$ and 1.92, and hence the structural parameter a is between $0.85^{1-\alpha}$ and $0.47^{1-\alpha}$ or (using $\alpha = 1/4$) between 0.6 and 0.9.

Finally, recall that the parameter b was calculated given that the relevant opportunity set segment has slope corresponding to $\theta \approx$ -0.75, that twice as many subjects chose the corner $z = -5$ as chose the kink $z = 0$, and that the idiosyncratic component of θ is Normal with standard deviation 5.33. Using a table of the cumulative Normal distribution we find $(-0.75 - b)/5.33 = N(2/3) = 0.45$, whence $b \approx -0.75 - 0.45 \times 5.33 \approx -3.15$.

A.2 Alternative Estimate

Solving the tangency condition $y/m = a^{1/(1-\alpha)}$ for $z > 0$ in terms of the predetermined interim allocation $(m(x), y(x))$ yields

$$(11b) \qquad z_i = -\frac{1}{1 + a^{\frac{1}{1-\alpha}}} y(x) + \frac{a^{\frac{1}{1-\alpha}}}{1 + a^{\frac{1}{1-\alpha}}} m(x) + u_i.$$

For an interior choice $z<0$, the expression becomes

$$(12b) \qquad z_i = -\frac{1}{3 - (-3b)^{\frac{1}{1-\alpha}}} y(x) + \frac{(-3b)^{\frac{1}{1-\alpha}}}{3 - (-3b)^{\frac{1}{1-\alpha}}} m(x) + u_i.$$

The last two equations suggest a linear regression on four explanatory variables — call them YP = $y(x)I_z$, YN = $y(x)I_{-z}$, MP = $m(x)I_z$, and MN = $m(x)I_{-z}$ — representing the interim allocations when z is positive and negative. The unrestricted estimating equation is

$$(13b) \qquad z_t = \beta_1 YP_t + \beta_2 MP_t + \beta_3 YN_t + \beta_4 MN_t + v_t$$

We again use tobit estimation, to account for the constraints $z \geq -(10 - x)I_{-x}/3$ and $z \leq (10 + 3x)I_x$, and use random effects estimation to account for individual differences. But for efficiency sake we also should take into account the implied linear relations $\beta_2 - \beta_1 = 1$ and

$-3\beta_3 - \beta_4 = 1$ satisfied by (11b) and (12b). Imposition of these restrictions in equation (13b) yields the restricted estimating equation

(14b) $z_t - MP_t + MN_t = \beta_1(YP_t + MP_t) + \beta_3(YN_t - 3MN_t) + \xi_t$.

Due to the data limitations noted in the text, β_3 and β_4 cannot be estimated with these data. The estimate of β_1 is -0.584, with standard error 0.063. This implies an estimate of β_2 equal to 0.416, and a value of the structural parameter *a* equal to $(.416/.584)^{1-\alpha}$ or (for $\alpha = 1/4$), $a = .78$. This is consistent with the estimates reported in section 5 using the alternative specification reported there.

References

Abbink, Klaus, Bernd Irlenbusch, and Elke Renner, "The Moonlighting Game: An Empirical Study on Reciprocity and Retribution." *Journal of Economic Behavior and Organization*, 42, 2000, pp. 265-77.

Ahlert, Marlies, Arwed Cruger and Werner Guth, "How Paulus Becomes Saulus--An Experimental Study of Equal Punishment Games," Manuscript, Humboldt University Berlin, August 1999.

Andreoni, James and John H. Miller, "Giving According to GARP: An Experimental Test of the Rationality of Altruism," Discussion paper, University of Wisconsin, 2000.

Barkow, Jerome H., Leda Cosmides and John Tooby, *The Adapted Mind*, NY: Oxford University Press, 1992.

Berg, Joyce, John Dickhaut, and Kevin McCabe, "Trust, Reciprocity, and Social History." *Games and Economic Behavior*, July 1995, 10(1), pp. 122-42.

Blount, Sally, "When Social Outcomes Aren't Fair: The Effect of Causal Attributions on Preferences." *Organizational Behavior and Human Decision Processes*, August 1995, 63(2), pp. 131-44.

Bolton, Gary, Jordi Brandts, and Axel Ockenfels, "Measuring Motivations for the Reciprocal Responses Observed in a Simple Dilemma Game." *Experimental Economics*, 1998, 1(3), pp. 207-19.

Bolton, Gary E., Elena Katok, and Rami Zwick, "Dictator Game Giving: Rules of Fairness versus Acts of Kindness." *International Journal of Game Theory*, 1998, 27, pp. 269-99.

Bolton, Gary E. and Axel Ockenfels, "ERC: A Theory of Equity, Reciprocity and Competition." *American Economic Review*, March 2000, 90(1), pp. 166-93.

Bolton, Gary E. and Rami Zwick, "Anonymity versus Punishment in Ultimatum Bargaining." *Games and Economic Behavior*, 1995, 10, pp. 95-121.

Brandts, Jordi and Gary Charness, "Retribution in a Cheap Talk Experiment." *UPF Barcelona manuscript,* September 2000.

Cason, Timothy and Daniel Friedman, "Buyer Search and Price Dispersion: A Laboratory Study," forthcoming, *Journal of Economic Theory*.

Cason Timothy, Tatsuyoshi Saijo and Takehika Yamoto, "Voluntary Participation and Spite in Public Good Provision Experiments: An International Comparison," Purdue University manuscript, August 1999; forthcoming in *Experimental Economics*.

Charness, Gary, "Attribution and Reciprocity in a Simulated Labor Market: An Experimental Investigation," Working paper, University of California at Berkeley, September 1996, revised April 2001.

Charness, Gary and Matthew Rabin, "Social Preferences: Some Simple Tests and a New Model," Discussion paper, University of California at Berkeley, 2000; forthcoming in *Quarterly Journal of Economics*.

Cox, James C., "Trust, Reciprocity, and Other-Regarding Preferences: Groups vs. Individuals and Males vs. Females," in Rami Zwick and Amnon Rapoport, (eds.), *Advances in Experimental Business Research*, Kluwer Academic Publishers, 2002a.

Cox, James C., "How to Identify Trust and Reciprocity," University of Arizona working paper, 2002b.

Cox, James C. and Cary A. Deck, "On the Nature of Reciprocal Motives," Discussion paper, University of Arizona, September 2000; revised 2002.

Cox, James C., R. Mark Isaac, Paula-Ann Cech, and David Conn, "Moral Hazard and Adverse Selection in Procurement Contracting," *Games and Economic Behavior*, 17, 1996, pp. 147-76..

Cox, James C. and Ronald L. Oaxaca (1989): "Laboratory Experiments with a Finite Horizon Job Search Model," *Journal of Risk and Uncertainty*, 2, 301-29.

Cox, James C. and Ronald L. Oaxaca (1996): "Is Bidding Behavior Consistent with Bidding Theory for Private Value Auctions?", in R. Mark Isaac (ed.), *Research in Experimental Economics*, vol.6. Greenwich: JAI Press.

Cox, James C. and Ronald L. Oaxaca (2000): "Good News and Bad News: Search from Unknown Wage Offer Distributions," *Experimental Economics*, 2, 197-225.

Cox, James C., Klarita Sadiraj, and Vjollca Sadiraj, "A Theory of Competition and Fairness for Egocentric Altruists," University of Arizona working paper, January 2001; revised 2002a.

Cox, James C., Klarita Sadiraj, and Vjollca Sadiraj, "Trust, Fear, Reciprocity, and Altruism," University of Arizona working paper, August 2001; revised 2002b.

Croson, Rachel T., "Theories of Altruism and Reciprocity: Evidence from Linear Public Goods Games." Wharton manuscript, July 1999.

Dufwenberg, Martin, Uri Gneezy, Werner Güth, Eric van Damme, "Direct versus Indirect Reciprocity: An Experiment." *Homo Oeconomicus*, 2001, 18, pp. 19-30.

Dufwenberg, Martin and Georg Kirchsteiger, "A Theory of Sequential Reciprocity." Discussion paper, CentER for Economic Research, Tilburg University, 1999.

Falk, Armin, Ernst Fehr, and Urs Fischbacher, "Testing Theories of Fairness – Intentions Matter." University of Zurich discussion paper, May 2001.

Falk, Armin and Urs Fischbacher, "A Theory of Reciprocity." Working Paper No. 6, Institute for Empirical Research in Economics, University of Zurich, 1999.

Fehr, Ernst and Simon Gächter, "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives*, Summer 2000b, 14(3), pp. 159-81.

Fehr, Ernst, Simon Gächter, and Georg Kirchsteiger, "Reciprocity as a Contract Enforcement Device: Experimental Evidence." *Econometrica*, July 1997, 65(4), pp. 833-60.

Fehr, Ernst and Klaus M. Schmidt, "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, August 1999, 114(3), pp. 817-68.

Fehr, Ernst and Klaus M. Schmidt, "A Theory of Fairness, Competition, and Cooperation." *Quarterly Journal of Economics*, August 1999, 114(3), pp. 817-68.

Gächter, Simon and Arno Riedl, "Moral Property Rights in Bargaining," CESifo Working Paper No. 697, Munich, 2002.

Gale, John, Kenneth G. Binmore and Larry Samuelson, "Learning to Be Imperfect: The Ultimatum Game," *Games and Economic Behavior*, 1995, 8, pp. 56-90.

Geanakopolos, John, David Pearce, and Ennio Stacchetti,``Psychological Games and Sequential Rationality." *Games and Economic Behavior*, 1989, 1, pp. 60-79.

Güth, Werner, and Judit Kovacs, "Why Do People Veto?" *Humboldt University manuscript*, 2001.

Güth, Werner, Rolf Schmittberger, and Bernd Schwarze, "An Experimental Analysis of Ultimatum Bargaining," *Journal of Economic Behavior and Organization*, 3, 1982, pp. 367-88.

Guttman, Joel M., "On the Evolutionary Stability of Preferences for Reciprocity," *European Journal of Political Economy* 16, 2000, pp. 31-50.

Harrison, Glenn W. and Peter Morgan, "Search Intensity in Experiments," *Economic Journal*, 100, 1990, pp. 478-86.

Hoffman, Elizabeth, Kevin McCabe, Keith Shachat, and Vernon L. Smith, "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior*, 1994, 7, pp. 346-80.

Kagel, John H., Chung Kim and Donald Moser, "Fairness in Ultimatum Games with Asymmetric Information and Asymmetric Payoffs," *Games and Economic Behavior* 13:1, March 1996, pp100-110.

Kagel, John H. and Katherine Wolfe, "Tests of Fairness Models Based on Equity Considerations in a Three Person Ultimatum Game," *Experimental Economics* 4:3, December 2001, pp. 203-220.

Konow, James, "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions," *American Economic Review* 90:4, September 2000, pp.1072-1091.

Konow, James, "Fair and Square: The Four Sides of Distributive Justics," *Journal of Economic Behavior and Organization* 46, 2001, pp.137-164.

Johanneson, M. and B. Persson, "Non-reciprocal Altruism in Dictator Games." *Economics Letters*, 2000, 69, pp. 137-42.

Levine, David K., “Modeling Altruism and Spitefulness in Experiments.” *Review of Economic Dynamics*, July 1998, 1(3), pp. 593-622.

Offerman, Theo, “Hurting Hurts More than Helping Helps: The Role of the Self-Serving Bias.” Working paper, University of Amsterdam, 1999.

Pindyck, Robert S. and Daniel L. Rubinfeld, *Econometric Models and Economic Forecasts*, NY: McGraw Hill, 1976

Rabin, Matthew, “Incorporating Fairness into Game Theory and Economics.” *American Economic Review*, 1993, 83, pp. 1281-1302.

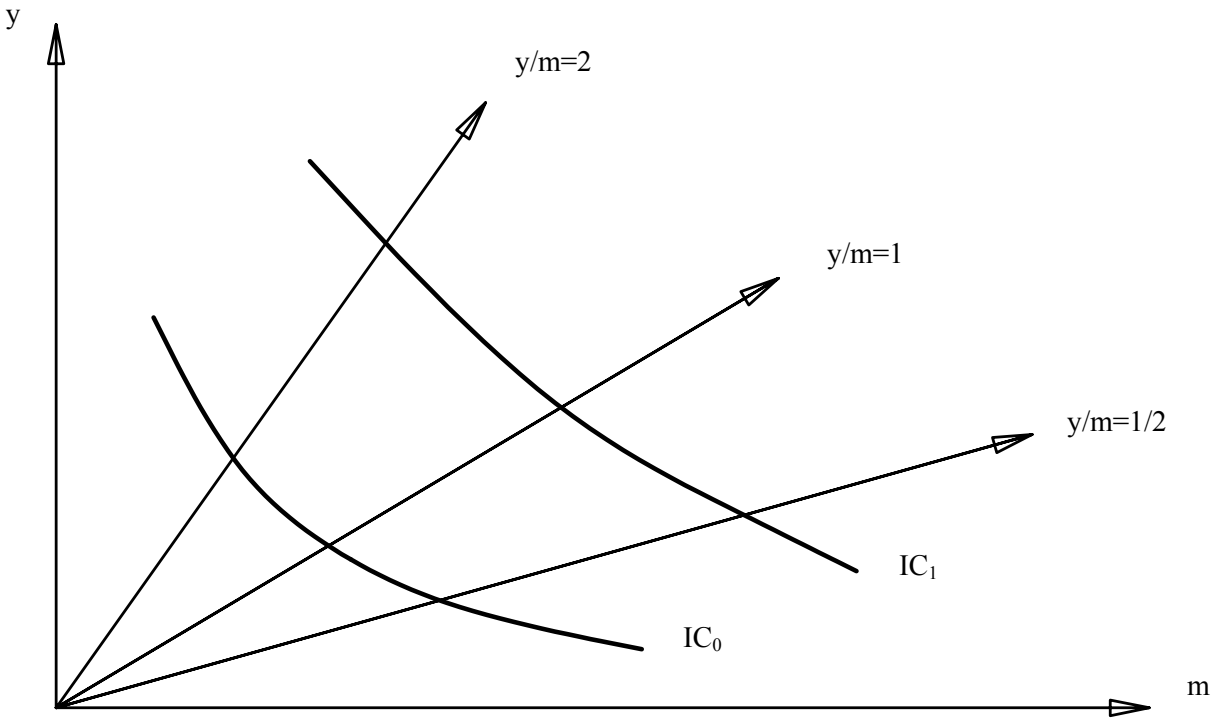
Smith, Vernon L. and Arlington W. Williams, “The Boundaries of Competitive Price Theory: Convergence Expectations and Transaction Costs,” in L. Green and J.H. Kagel, eds., *Advances in Behavioral Economics*, vol. 2, Norwood, NJ: Ablex Publishing Corp., 1990.

Smith, Adam, *The Theory of Moral Sentiments*, 1759; reprinted by Indianapolis: Liberty Classics, 1976.

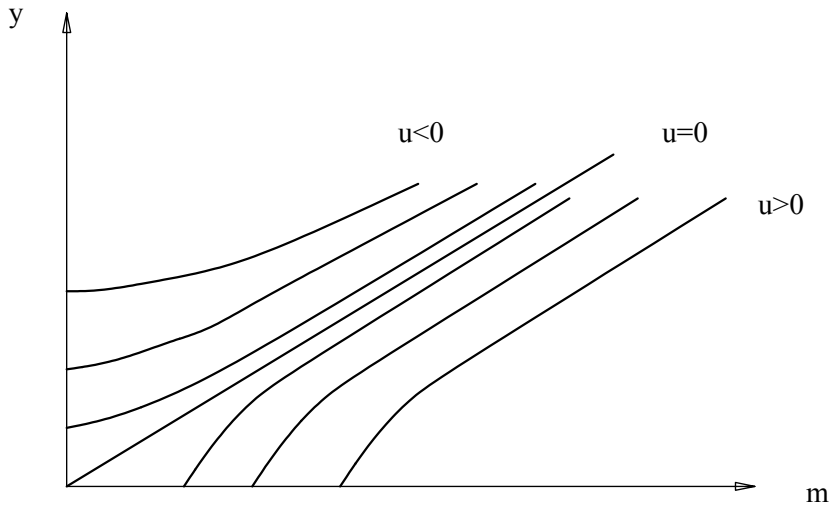
Sobel, Joel, "Social Preferences and Reciprocity," UC San Diego manuscript, December 2000.

Zizzo, Daniel J. and Andrew Oswald, "Are People Willing to Pay to Reduce Others' Income?" Oxford University Manuscript, July 2001.

Figure 1: Indifference Curves

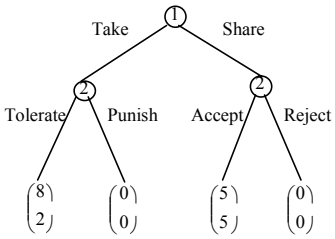


Panel A: Positive θ.
The MRS=-slope of IC increases as y/m increases. It increases more rapidly on the more convex IC_0 than on IC_1. With CES preferences, IC_1 would come from a larger $\mathbb{E}$ than IC_0.

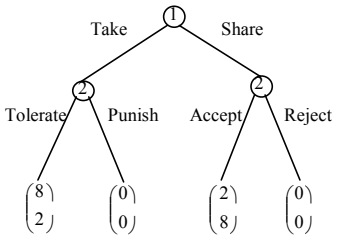


Panel B: Negative θ.
The slope of the u=0 indifference curve everywhere is (-1/$\mathbb{E}$)
The slope of other ICs converge to this value as m, y→∞.

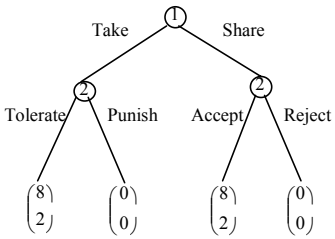
Figure 2. Mini-Ultimatum Games



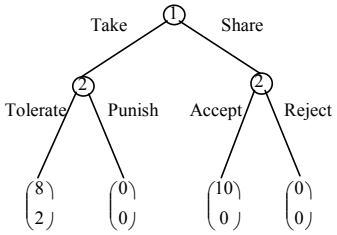
5/5 - game



2/8 - game



8/2 - game



10/0 - game

Figure 3. Second Mover's Choice Set