A Tractable Model of Reciprocity and Fairness

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

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

Steven Gjerstad
University of Arizona
gjerstad@econlab.arizona.edu

May 26, 2004

Abstract We introduce a parametric model of other-regarding preferences. The income distribution and the kindness or unkindness of others' choices ("intentions") systematically affect a person's emotional state. The emotional state systematically affects the marginal rate of substitution between own and others' payoffs, and thus the person's subsequent choices. The model is applied to two sets of laboratory data: simple binary choice mini-ultimatum games, and Stackelberg duopoly games with a range of choices. The results confirm that other-regarding preferences respond to others' intentions as well as to the income distribution.

We are grateful to the National Science Foundation for research support (grant numbers SES-9818561 and DUE-0226344). We are also grateful to Ken Binmore, Gary Charness, Steffen Huck, Lori Kletzer, Lisa Rutstrum, and Daniel Zizzo for helpful comments.

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)

1 Introduction

Everyone knows that people care about other people. Economists have known it at least since Adam Smith, but only recently have 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? The goal of this paper is to propose a model that addresses such questions and, using some existing laboratory data, to 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 or employee, or a wealthy or poor neighbor? In the laboratory data, the most prominent such variable is the distribution of income: what is your current payoff relative to my current payoff?

A second motive is reciprocity: how do 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 now so that you will help me later and thereby increase the net present value of my payoff stream. 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. Smith (1759)

refers to these emotions as the "moral sentiments" of gratitude and resentment, and suggests three necessary conditions for their proper expression.

Our model formalizes the idea. In the model, status and reciprocity affect my emotional state, summarized in a scalar variable θ, and my emotional state affects my choices. Smith's resentment corresponds to negative θ and gratitude corresponds to positive θ. The model retains the conventional assumption that I choose an available alternative that maximizes my utility function, and follows 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 $u(m, y) = m + \theta\, y$. The key 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 2 sets the stage by summarizing recent related literature. Section 3 below proposes specifications of the model elements r, s, and θ, and proposes a more general utility function that allows non-linear indifference curves. 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 Stackelberg duopoly 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 3 and 5 appear in the appendices.

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 contexts. 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, Isaac, Cech, and Conn, 1996); and search (Cason and Friedman, 2003; 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, 1998), 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). Fehr and Gächter (2000) summarize recent evidence on the economic impact of motives beyond self-interest.

The laboratory data, together with suggestive field data, have encouraged the development of models of other-regarding preferences. This literature falls into two broad classes. First there are the relative payoff (or distributional) models of Fehr and Schmidt (1999), Charness and Rabin (2002), Bolton and Ockenfels (2000), 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

$$u(m,\, y) = \begin{cases} m - \alpha\,(y - m), & \text{if } m < y, \\ m - \beta\,(m - y), & \text{if } m \geq y. \end{cases}$$

That is, I like own income and dislike income inequality, especially when I have the short end. For two players, the Charness-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 ($\alpha > 0$, $\beta > 0$), and quasi-maximin preferences ($1 > \beta > -\alpha > 0$). The Bolton-Ockenfels model also assumes that I like own income and dislike income inequality, but the utility function takes the non-linear form

$$u(m,\, y) = v\left(m,\, \frac{m}{m+y}\right).$$

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 $\frac{m}{m+y}$), and to satisfy $v_2(m, \frac{1}{2}) = 0$ 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

$$u(m,\, y) = \begin{cases} (m^{\,\alpha} + \theta_{-}\, y^{\,\alpha})^{1/\alpha}, & \text{if } m < y, \\ (m^{\,\alpha} + \theta_{+}\, y^{\,\alpha})^{1/\alpha}, & \text{if } m \geq y, \end{cases}$$

with parameter restrictions $0 \leq \alpha \leq 1$, $0 \leq \theta_{-} \leq \theta_{+} \leq 1$, 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 two 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

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

where $\pi_i(a_i,\, b_j)$ is the monetary payoff to player i, $\tilde{f}_j(b_j,\, c_i)$ is player 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 ($\tilde{f}_j < 0$ and $f_i < -1$) as well as positive reciprocity increases utility. The model looks for equilibria in actions and beliefs about intended kindness; typically there are many such equilibria.

Dufwenberg and Kirchsteiger (2004) propose an extension to extensive form games with N players, and Falk and Fischbacher (2001) propose a different extension that also covers incomplete information but uses a distributional preference utility function. Charness and Rabin (2002), 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 m and in others' monetary payoffs y_j. For two player games, utilities are of the form

$$u(m, y) = m + \frac{a_m + \lambda\, a_y}{1 + \lambda}\, y,$$

where $a_m \in (-1, 1)$ is my type or "coefficient of altruism," $a_y \in (-1, 1)$ is my current estimate of your type, 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 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 (2001) 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 4 below), Falk, Fehr and Fischbacher (2003) 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 (2002) has similar results: intentional helpful (hurtful) actions were rewarded (punished) more frequently than identical but randomly generated actions. See also Ahlert, Crüger, and Güth (2001), Charness (2002), Güth and Kovács (2001), Gibbons and Van Boven (2001), and Kagel, Kim, and Moser (1996).

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 (1998) and Bolton, Brandts and Ockenfels (1998). Cason, Saijo, and Yamato (2002) 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 (2002, 2004) 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 by 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.

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 impact of the various motives.

3 Model Specifications

This section presents a new model of preferences that incorporates objectively defined variables r and s capturing reciprocity and status motives. For pedagogical and comparative purposes, the presentation here considers only two player extensive form games of complete information with first mover F receiving material payoff y, and second mover S receiving material payoff m. The model shows how the emotional state of S defines the marginal rate of substitution (MRS) between own payoff m and other's payoff y, and how the emotional state responds to the values of r and s that arise from F's prior choice.

Due to its importance in existing literature, the distribution or relative payoff is separated from other aspects of the status motive, and is captured in the shape of indifference curves in (m, y) space. 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. If preferences are convex, the MRS increases as one moves along any indifference curve in the direction of

increasing y/m ratio; see figure 1 (a). 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 $u(m, y) = (m^{\alpha} + \theta\, y^{\alpha})^{1/\alpha}$; see also Cox, Sadiraj, and Sadiraj (2002a).

We modify this function slightly. The exponent $1/\alpha$ is problematic when it applies to a negative expression, which will arise when θ is sufficiently negative. Of course, the outside exponent doesn't affect the shape of the indifference curves, but its sign affects their ordering. The ordering is preserved and the negativity issue is finessed by using $1/\alpha$ as a coefficient. Hence the preference model is defined for convexity parameter $\alpha \in (-\infty, 1]$ by

$$u(m, y) = \begin{cases} \frac{1}{\alpha}\,(m^{\alpha} + \theta\, y^{\alpha}), & \alpha \neq 0; \\ m\, y^{\theta}, & \alpha = 0. \end{cases} \tag{1}$$

With these preferences we have $MRS = \frac{\partial u/\partial m}{\partial u/\partial y} = \theta^{-1}\,(\frac{y}{m})^{1-\alpha}$. Hence the emotional state θ is the willingness to pay ($WTP = 1/MRS$) at an allocation on the equal payoff line $m = y$. Preferences are linear (and *MRS* is constant) if $\alpha = 1$, and preferences are strictly convex (and *MRS* strictly increases in relative payoff y/m along indifference curves) if and only if $\alpha < 1$. Appendix A.1 shows that indifference curves for $\alpha \neq 0$ converge pointwise to indifference curves of the Cobb-Douglas preferences $u(m, y) = m\, y^{\theta}$ as $\alpha \to 0$. A standard textbook argument shows that as $\alpha \to -\infty$, the indifference curves converge to Leontief indifference curves with corners on the ray $y/m = \theta^{-1}$.

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 $r(x) = m(x) - m_{_0}$, 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.[1] Often it is convenient to normalize $r(x)$ so that it lies in the range $[-1, 1]$. Let $m_g = \max_x m(x)$ and $m_b = \min_x m(x)$. The normalized version is $r(x) = (m(x) - m_0)/(m_g - m_b)$, 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_F and s_S 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 $s = s_F - s_S$. For example, under some social norms the first mover's status and hence s would increase if she had to earn the right to be the first mover.

In estimating the model, we maintain the following four assumptions.

A.1 Individuals choose so as to maximize a utility function of the form in equation (1).

A.2 The emotional state function $\theta = \theta(r, s)$ is identical across individuals except for a mean zero idiosyncratic term.

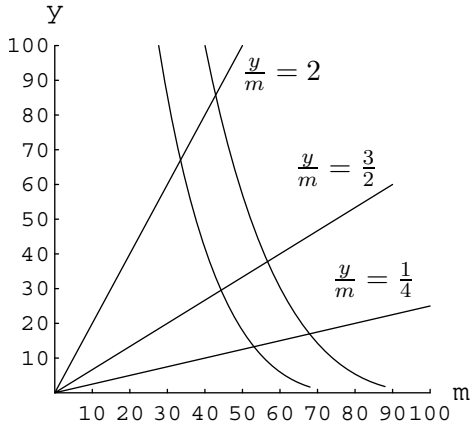
A.3 $\theta(r, s)$ is weakly increasing in r and s.

A.4 $\theta(0, 0)$ is non-negative but $\theta(r, s)$ is negative when its arguments r and s are sufficiently negative.

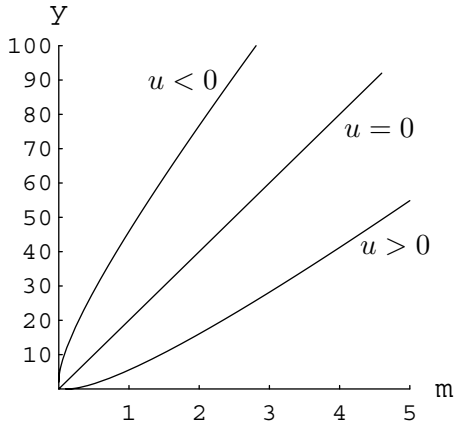
The case of negative θ deserves a brief comment before presenting sample applications. A person with negative θ is willing to pay to reduce other's payoff. That is, y is a "bad"

[1] 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.) Konow (2003) surveys relevant moral philosophy and evidence. Gächter and Riedl (2003) offer a general discussion and demonstrate the impact of m_0 (which they call moral property rights or entitlements) in new laboratory data.

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 slopes of other indifference curves converge towards the slope of this ray as in figure 1 (b).



(a) Indifference curves for $\theta = 0.3$.



(b) Indifference curves for $\theta = -0.3$.

Figure 1: Indifference curves for the utility function $u(m, y) = 2.5\,(m^{0.4} + \theta\, y^{0.4})$.

4 Evidence from Mini-Ultimatum Games

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

With standard self-interested preferences, S always accepts a positive payoff because refusing gives him zero payoff. Ultimatum games are interesting because S often rejects

[2] Binmore condemns the term mini-ultimatum game or MUG, which we perpetuate, and favors ultimatum mini-game. As a compromise, we urge readers to parse MUG as mini-[ultimatum game].

positive offers, and the mini-ultimatum game is especially interesting because, contrary to the distributional models reviewed earlier, the rejection rate of the offer (2, 8) varies systematically across games with different x = "Share" alternatives. We show that our model accounts for this effect via the impact of the reciprocity variable $r(x)$ on the WTP parameter θ.

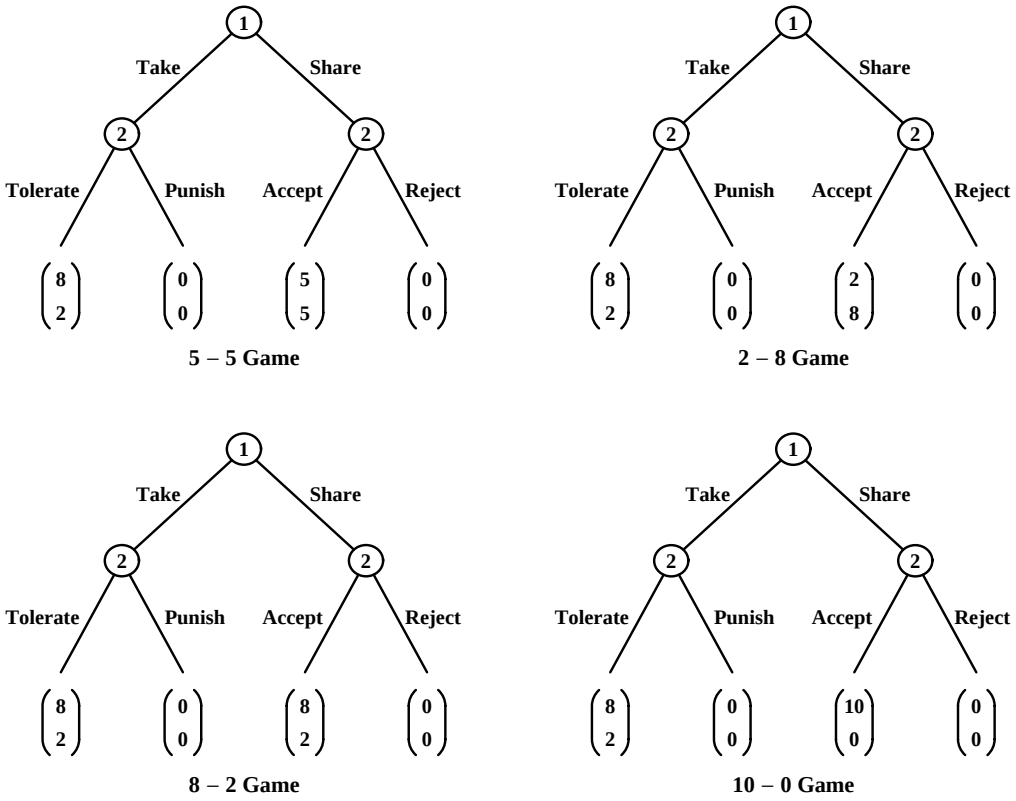


Figure 2: Extensive forms of mini-ultimatum games.

The empirical task is to explain responder choice, coded

$$Z = \begin{cases} 0, & \text{if } S \text{ chooses } (0,0), \\ 1, & \text{otherwise.} \end{cases}$$

It is natural to use probit estimation, with explanatory variables derived as follows. Let S's property right m_0 be his feasible payoff that is closest to equal split but not higher than the proposer's, so $m_0 = \min\{5, m_g\}$. 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 $r(x) = (m(x) - m_0)/(m_g - m_b)$ takes on a range of values.

The mini-ultimatum game data reported by Falk, Fehr, and Fischbacher (2003) contain no variation in the status variable (other than relative payoff), so s is constant. By Assumptions A.1 – A.4 and a first order Taylor series approximation, second mover i has WTP parameter $\theta_i = a + b\,r + \sigma\,\epsilon_i$, 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 $\sigma^2 > 0$.

For $\alpha < 0$, $u(0,\,0) = -\infty$ and $u(m(x), y(x))$ is finite, regardless of whether $x =$"Take" or $x =$"Share," so the predicted choice always would be $Z = 1$. In practice, this implies that for data sets that include rejections of the first-mover offer, the estimate of α will be positive. When $\alpha > 0$, we have $Z = 1$ if and only if $0 = u(0,0) < u(m(x), y(x)) = \frac{1}{\alpha}\,(m(x)^\alpha + \theta\,y(x)^\alpha)$ which is equivalent to $0 < (m/y)^\alpha + \theta_i = (m/y)^\alpha + a + b\,r + \sigma\,\epsilon_i$, or $-\epsilon_i < \sigma^{-1}\,((m/y)^\alpha + a + b\,r)$. Hence the probability that $Z = 1$ is the standard cumulative normal distribution evaluated at $\sigma^{-1}((m/y)^\alpha + a + b\,r)$, and probit estimation will recover the structural parameters.

Using the Falk, Fehr, and Fischbacher 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

$$Pr[Z_i] = -0.49 + 0.69\,r_i + 2.00\,(m/y)^\alpha + \epsilon_i.$$

The equation predicts correctly 302 of the subjects' 360 choices. The coefficient estimate for $(m/y)^\alpha$ implies that $\sigma^{-1} = 2.00$ and $\sigma = 0.5$, 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 $2.00\,(1^{1/4} - (0.5)^{1/4}) \approx 0.32$ of a standard deviation.

The coefficient estimates are fairly robust to changes in α. For $\alpha = 1/8$ 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 Stackelberg duopoly

Huck, Müller, and Normann (2001, henceforth HMN01) present an experiment in which randomly matched pairs of subjects play a Stackelberg duopoly game. The first mover (F) chooses an output level $x \in \{3, 4, 5, \ldots, 15\}$. The second mover ($S$) observes x and chooses an output level $q \in \{3, 4, 5, \ldots, 15\}$. The price is $p = 30 - x - q$; both players have constant marginal cost 6 and no fixed cost, so the profit margin for each player is $M = 24 - x - q$. Payoffs therefore are $m = M\,q$ and $y = M\,x$.

Given F's choice x, the second mover's choice set is the locus in (m, y) space traced out by varying q from 3 to 15. As illustrated in figure 3, it is a parabolic arc that opens toward the y-axis whose vertex $(m, y) = \left(\frac{1}{4}(24 - x)^2, \frac{1}{2}(24 - x)\,x\right)$ corresponds to $q = \frac{1}{2}(24 - x)$. In figure 3, F's choice is $x = 4$; S's choice $q = 3$ produces payoff vector (51, 68) while $q = 10$ produces the vertex payoff vector (100, 40). With $x = 4$, choices $q < 10$ reduce m but increase y, while choices $q > 10$ reduce both m and y.

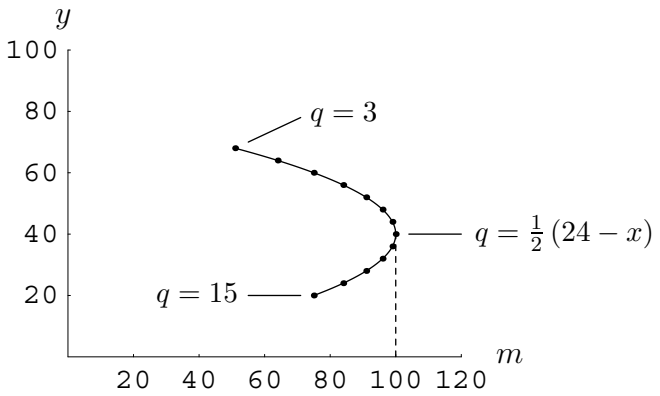


Figure 3: Feasible joint profits when first-mover output is $x = 4$.

The standard textbook analysis of this game is that S will always respond to F's choice x by choosing $q = \frac{1}{2}(24-x)$ to obtain the m-maximizing (vertex) payoff $m(x) = \frac{1}{4}(24-x)^2$, and that F therefore will choose $x = 12$ to maximize his component $y(x) = \frac{1}{2}(24-x)\,x$ of the vertex payoff. Hence at the classic Stackelberg equilibrium $x = 12$, $q = 6$, $p = 12$, and $M = 6$, yielding payoffs $m = 36$ and $y = 72$. In the symmetric, simultaneous move Cournot game, the classic equilibrium choices are $x = q = 8$ so that $p = 14$, $M = 8$, and $m = y = 64$.

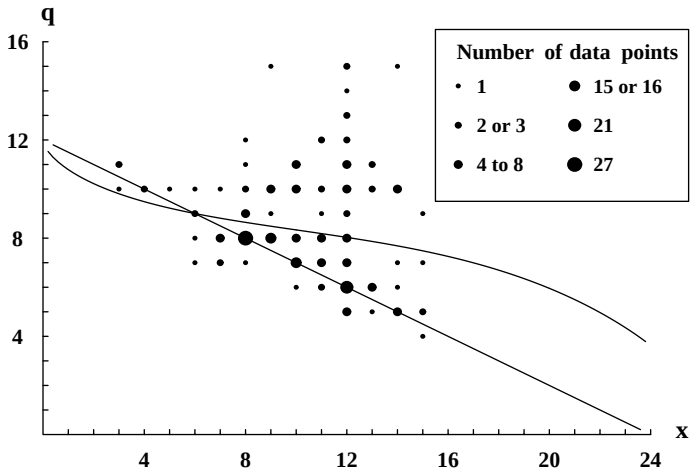
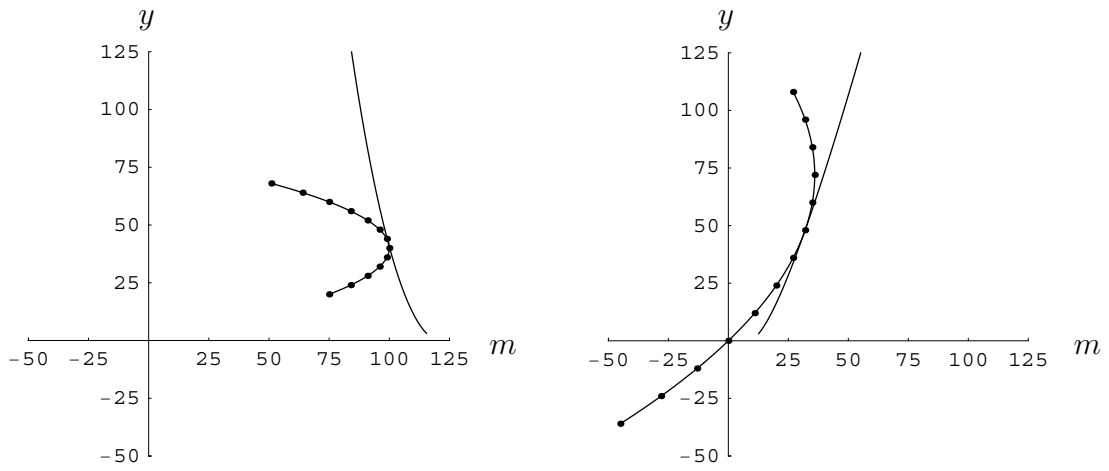


Figure 4: Actual choice pairs (x, q) and estimated best response function.

The HMN01 experiment produced a range of outcomes. Although the two most frequent outcomes are the Cournot equilibrium outcome and the Stackelberg outcome, as shown in figure 4, a large fraction (199 of 220) of second-mover outputs meet or exceed those from the standard, self-interested, best-response function, and this tendency becomes more pronounced as the first-mover output x increases. This second-mover choice pattern arises naturally from our emotional state-dependent utility function. The intuition is that F is being greedier when he chooses a larger x, and this pushes the reciprocity variable $r(x)$ towards more negative values. Hence S has a more negative emotional state θ, and therefore chooses a larger q to reduce F's payoff y. This intuition is confirmed in figure 4: for high values of the first-mover output x, observed choices q from the HMN01 data exceed the standard best-response, which is shown as the straight line. The estimated best-response

from our emotional state dependent utility model is shown in figure 4 as the curve.

Figure 5 further illustrates the intuition behind our model. Panel (a) redraws S's choice set from figure 3 given $x = 4$, and also includes a tangent indifference curve for positive θ. Here S chooses q slightly below the selfish best reply $q = \frac{1}{2}(24 - x) = 10$, reducing his payoff a bit below $m(x) = \frac{1}{4}(24 - x)^2 = 100$ while boosting F's payoff noticeably above $y(x) = \frac{1}{2}(24 - x)x = 40$. Panel (b) shows S's choice set given F's much less generous choice $x = 12$. The tangent indifference curve is for negative θ. Due again to the parabolic choice set, by increasing q above the selfish best response, S obtains a first-order decrease in F's payoff from $y(x) = 72$ while sacrificing only a second-order amount of his own payoff from $m(x) = 36$. The key insight is that the attitude of S toward F is a function of the action taken by F, i.e., the WTP parameter θ depends on x.



(a) Indifference curve for $x = 4$.

(b) Indifference curve for $x = 12$.

Figure 5: Indifference curves for utility function estimated from HMN01 data.

The second mover's utility function can be written in terms of the players' choices by substituting the payoff functions $m(x, q)$ and $y(x, q)$ into equation (1), while keeping $\theta(x)$ in general form. Simplifying slightly we get

$$U(x,q) = \begin{cases} \frac{1}{\alpha}(24 - x - q)^{\alpha}(q^{\alpha} + \theta(x)\, x^{\alpha}), & \alpha \neq 0, \\ \frac{1}{\alpha}(24 - x - q)^{1+\theta(x)}\, q\, x^{\theta(x)}, & \alpha = 0. \end{cases} \tag{2}$$

Equate to 0 the derivative of (2) with respect to q, and simplify to obtain the first order condition

$$0 = (24 - x - 2\,q)\,q^{\alpha-1} - \theta(x)\,x^{\alpha}. \tag{3}$$

Although (3) is valid for all $\alpha \leq 1$, it can be solved for $q = q^*(x\,;\,\theta,\,\alpha)$ in closed form only in special cases. Appendix A.2 demonstrates that a unique maximizer for equation (2) exists for every parameter vector $(\theta,\,\alpha) \in (-\infty,\,\infty) \times (-\infty,\,1]$, so the best response $q^*(x\,;\,\theta,\,\alpha)$ is well defined. Appendix A.3 describes the algorithm used to determine $q^*(x\,;\,\theta,\,\alpha)$.

The empirical task is to explain S's choice $q^*(x\,;\,\theta,\,\alpha)$ given F's choice x. The last model element that we need to specify is $\theta(x)$. Define F's neutral choice as the solution $x = 8$ to the equal payoff condition $m(x) = y(x)$. This condition also characterizes the Cournot equilibrium, and yields $m_{_0} = m(8) = 64$. In the normalized reciprocity expression $r(x)$, the denominator is $m_g - m_b = \max_x m(x) - \min_x m(x) = m(3) - m(15) = 90$, so $r(x) = \frac{1}{90}\,(m(x) - m_{_0}) = \frac{1}{360}\,(24 - x)^2 - \frac{32}{45}$. As in the previous application, the status variable s is constant and the first order Taylor series yields

$$\theta(x) = a + b\,r(x) = a + \frac{b}{90}\,(m(x) - m_{_0}). \tag{4}$$

The HMN01 data we analyze consist of all Stackelberg games with randomly matched players. These data include twenty-two first- and second-movers; each player participated in ten Stackelberg games. The estimation procedure finds the parameter vector that minimizes the sum of squared residuals $SSR = \sum_{i=1}^{220} (q^*(x_i\,;\,a,\,b,\,\alpha) - q_i)^2$ for these 220 choice pairs (x_i, q_i). Details of the estimation procedure appear in Appendix A.4. The resulting parameter estimates ($\pm$ standard errors) are $\hat{a} = -0.16 \pm 0.05$, $\hat{b} = 0.816 \pm 0.28$, and $\hat{\alpha} = 0.53 \pm 0.44$. The estimated best-response function is shown in figure 4; figure 5 shows the estimated utility function. The confidence region for these three parameter estimates is an ellipsoid. Five cross sections through the 95% confidence ellipsoid are depicted in figure 6.

The parameter estimates allow us to test several hypotheses. Appendix A.5 details the calculations for the test statistics.

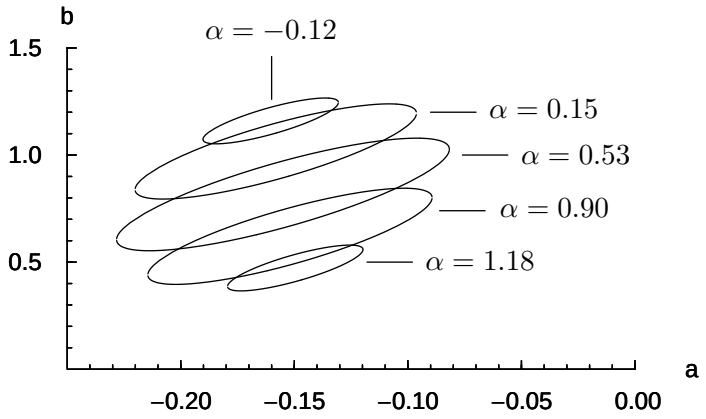


Figure 6: Cross sections through the 95% confidence region for a, b, and α.

H.1 *The parameter pair* $(a, b) = (0, 0)$.

This is the natural null hypothesis of selfish preferences, responsive neither to reciprocity nor to distributional concerns. The F test statistic for the data is $F(2, 217) = 70.9$, which implies rejection of the hypothesis at a p-value of less than 10^{-16}. The data firmly support other-regarding preferences.

H.2 *Under the maintained assumption that* $m_0 = 64$*, the parameter* a *is positive.*

In equation (4), the reference profit level m_0 and the parameter a are not separately identified, since $\theta(x) = A + \frac{b}{90}\, m(x)$ where $A = a - \frac{b}{90} m_0$. Consequently the choice $m_0 = 64$, even though we find it persuasive, has no impact on the estimates for the reciprocity parameter b or the distribution (or shape) parameter α. This null hypothesis can therefore be interpreted as stating that the typical second mover in the HMN01 experiment has a property right (or reference profit level) that is at or below the Cournot-Nash profit level $m_0 = 64$. The F statistic for the data is $F(1, 217) = 10.6$, which implies rejection at the p-value of 0.001. We conclude that the second-movers in this experiment typically maintain higher reference profit levels.

H.3 *The parameter* b *is zero or negative.*

This is the key null hypothesis. It states that, although they may respond to distributional concerns, second movers do not respond to reciprocity concerns (or else respond

perversely). The F statistic is $F(1, 217) = 11.7$, which implies rejection at the p-value of less than 0.001. The data firmly support reciprocal behavior by the typical subject.

6 Further Applications

In an earlier version of the present paper, Cox and Friedman (2002) fit the model in equation (1) to a fairly complex two player extensive form game called the Moonlighting Game (MLG). In the MLG, the first mover F can send money to or take money from the second mover S, and the amounts sent are tripled. Then S, at differing personal costs, can increase or reduce F's payoff. As in the Stackelberg game just analyzed, S's choice set depends on F's action, and contains a segment with positive slope as well as a segment with negative slope. Unlike the Stackelberg game, the MLG choice set has a sharp kink between two linear segments. Much of the data (from an experiment of Cox, Sadiraj and Sadiraj, 2002b) lies on or very close to the kinks and corners of the budget set. Therefore estimates of model parameters are not very precise. Qualitatively, the model explains the data quite well. It predicts correctly that very few interior solutions lie on the positively sloped segment; this follows from the almost linear indifference curves for negative θ shown in figure 1 (b). The model also captures the strong tendency of second movers to reward first mover generosity and the common tendency to punish first mover greed.

Future applications can explore the impact of other aspects of status. Possibly relevant treatments include 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 Cox, Sadiraj, and Sadiraj (2002b) Treatment C data automate first movers, and the second movers' choices then are generally consistent with $\theta = 0$, suggesting a dominant interaction $r \times s$ with $s = 0$.[3] Zizzo and Oswald (2001) found that subjects with low status

[3] Alternatively, one could simply define m_0 as the automated choice of the first mover and obtain $r = 0$ directly.

are particularly eager to "burn" the payoffs of players with large unearned payoffs.

No doubt there are many other existing data sets to which the model can be fit. The model also suggests new experimental designs for sharper tests and further development. In particular, consider two player extensive form game experiments that elicit 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. The data would allow sharp estimates for the impact of each motive.

To illustrate, continue to hold s constant and take the linear Taylor series approximation of the systematic portion of the emotional state, $\theta = a + b\,r$, 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

$$\left(\frac{m}{y}\right)^{1-\alpha} = 1 + (1-\alpha)\,\frac{m-y}{y} + \frac{\alpha^2-\alpha}{2}\,\left(\frac{m-y}{y}\right)^2 + O\left(\left(\frac{m-y}{y}\right)^3\right).$$

Use the reciprocity variable $r(x) = m(x) - m_0$; 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 $WTP = \theta\,\left(\frac{m}{y}\right)^{1-\alpha}$ from Section 3 and use the Taylor series approximation of θ from Section 3 to obtain

$$\begin{aligned} WTP \;=\; & a \,+\, b\,(m(x) - m_0) \,+\, (a \,+\, b\,(m(x) - m_0))\,(1-\alpha)\,\tfrac{m-y}{y} \\ & +\,(a \,+\, b\,(m(x) - m_0))\,\tfrac{\alpha^2-\alpha}{2}\,\left(\tfrac{m-y}{y}\right)^2 \,+\, (a \,+\, b\,(m(x) - m_0))\,O\left(\left(\tfrac{m-y}{y}\right)^3\right). \end{aligned}$$

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

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 θ_i towards each player i, and my utility function is simply

$$u(x,\, y_1,\, \ldots,\, y_n) = \frac{1}{\alpha}\,(x^\alpha + \theta_1\, y_1^\alpha + \ldots + \theta_n\, y_n^\alpha).$$

Dependence of θ_i on the motives r and s is the same as in the two player case. Of course, in games where players can't separately identify the other players, there is only one θ. For games in which each player can observe the individual history of every other player, 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 Stackelberg duopoly. The MUG data consist of binary choices from the second mover following binary choices by a first mover. We derived and estimated a probit model that accounted for the data rather well and that produced parameter estimates consistent with a priori predictions (assumptions A.3 and A.4). The Stackelberg duopoly data consist of a range of choices by a second mover following a range of choices by a first mover. Again we derived and estimated a model (this time using non-linear least squares regression) that accounts for the data and produces parameter estimates that strongly support reciprocal behavior.

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. The data shown in figure 4 and other evidence suggests that people may differ in their responsiveness b to given reciprocity and status motives.

Therefore future work should consider estimation using random coefficient models.

The definitions presented here extend directly to extensive form games in which some players have several moves, to normal form games, and to some other games of incomplete information. Future empirical work will show how successful such extensions are relative to available alternatives. Our approach has several advantages that might survive beyond the current implementation. First of all, it uses a model of preferences and choice, not equilibrium, and so is 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.

References

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

[2] Ahlert, Marlies, Arwed Crüger, and Werner Güth (2001). "How Paulus Becomes Saulus: An Experimental Study of Equal Punishment Games," Homo Oeconomicus **18**, 303-318.

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

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

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

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

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

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

[9] Brandts, Jordi and Gary Charness (2000). "Retribution in a Cheap Talk Experiment," UPF Barcelona manuscript.

[10] Cason, Timothy and Daniel Friedman (2003). "Buyer Search and Price Dispersion: A Laboratory Study.," Journal of Economic Theory **112**, 232-60.

[11] Cason, Timothy N., Tatsuyoshi Saijo, and Takehiko Yamato (2002). "Voluntary Participation and Spite in Public Good Provision Experiments: An International Comparison," Experimental Economics **5**, 133-53.

[12] Charness, Gary (2002). "Attribution and Reciprocity in an Experimental Labor Market," Working paper, University of California, Santa Barbara,

[13] Charness, Gary and Matthew Rabin (2002). "Understanding Social Preferences with Simple Tests," Quarterly Journal of Economics **117**, 817-69.

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

[15] Cox, James C. (2004). "How to Identify Trust and Reciprocity," Games and Economic Behavior **46**, 260-81.

[16] Cox, James C. and Cary A. Deck (2002). "On the Nature of Reciprocal Motives," University of Arizona, Department of Economics Working Paper 02-01.

[17] Cox, James C. and Daniel Friedman (2002). "A Tractable Model of Reciprocity and Fairness," University of Arizona, Department of Economics Working Paper 02-04.

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

[19] 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.

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

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

[22] Cox, James C., Klarita Sadiraj, and Vjollca Sadiraj (2002a). "A Theory of Competition and Fairness for Egocentric Altruists," University of Arizona discussion paper.

[23] Cox, James C., Klarita Sadiraj, and Vjollca Sadiraj (2002b). "Trust, Fear, Reciprocity, and Altruism," University of Arizona Working Paper 01-06.

[24] Croson, Rachel T. (1999). "Theories of Altruism and Reciprocity: Evidence from Linear Public Goods Games," Wharton manuscript.

[25] Davidson, Russell, and James G. MacKinnon (1993). Estimation and Inference in Econometrics. Oxford University Press, Oxford U.K.

[26] Dufwenberg, Martin and Georg Kirchsteiger (2004). "A Theory of Sequential Reciprocity." Games and Economic Behavior **47**, 268-98.

[27] Falk, Armin, Ernst Fehr, and Urs Fischbacher, (2003). "On the Nature of Fair Behavior," Economic Inquiry **41**, 20-26.

[28] Falk, Armin and Urs Fischbacher (2001). "A Theory of Reciprocity," CEPR Discussion Paper no. 3014, University of Zurich.

[29] Fehr, Ernst and Simon Gächter (2000). "Fairness and Retaliation: The Economics of Reciprocity," Journal of Economic Perspectives **14**, 159-81.

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

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

[32] Gächter, Simon and Arno Riedl (2003). "Moral Property Rights in Bargaining with Infeasible Claims," Tinbergen Institute Discussion Paper 03-055/1.

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

[34] Geanakopolos, John, David Pearce, and Ennio Stacchetti (1989). "Psychological Games and Sequential Rationality," Games and Economic Behavior **1**, 60-79.

[35] Gibbons, Robert and Leaf Van Boven (2001). "Contingent Social Utility in the Prisoners' Dilemma," Journal of Economic Behavior and Organization **45**, 1-17.

[36] Güth, Werner, and Judit Kovács (2001). "Why do people veto? An experimental ananlysis of the valuation and the consequences of varying degrees of veto power." Homo Oeconomicus **18**, 277-302.

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

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

[39] Harrison, Glenn W. and Peter Morgan (1990). "Search Intensity in Experiments," Economic Journal **100**, 478-86.

[40] Huck, Steffan, Wieland Müller, and Hans-Theo Normann (2001). "Stackelberg beats Cournot: On collusion and efficiency in experimental markets," Economic Journal **111**, 749-766.

[41] Kagel, John H., Chung Kim and Donald Moser (1996). "Fairness in Ultimatum Games with Asymmetric Information and Asymmetric Payoffs," Games and Economic Behavior **13**, 100-110.

[42] Kagel, John H. and Katherine Wolfe (2001). "Tests of Fairness Models Based on Equity Considerations in a Three Person Ultimatum Game," Experimental Economics **4**, 203-220.

[43] Konow, James (2000). "Fair Shares: Accountability and Cognitive Dissonance in Allocation Decisions," American Economic Review **90**, 1072-1091.

[44] Konow, James (2001). "Fair and Square: The Four Sides of Distributive Justice," Journal of Economic Behavior and Organization **46**, 137-164.

[45] Konow, James (2003). "Which Is the Fairest One of All? A Positive Analysis of Justice Theories," Journal of Economic Literature **61**, 1188-1239.

[46] Levine, David K. (1998). "Modeling Altruism and Spitefulness in Experiments," Review of Economic Dynamics **1**, 593-622.

[47] Offerman, Theo (2002). "Hurting Hurts More than Helping Helps," European Economic Review **46**, 1423-37.

[48] Rabin, Matthew (1993). "Incorporating Fairness into Game Theory and Economics," American Economic Review **83**, 1281-1302.

[49] Slonim, Robert and Alvin E. Roth (1998). "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic," Econometrica **66**, 569-96.

[50] Smith, Adam (1759). *The Theory of Moral Sentiments*. Reprinted by Indianapolis: Liberty Classics, 1976.

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

[52] Sobel, Joel (2001). "Interdependent Preferences and Reciprocity," University of California, San Diego manuscript.

[53] Zizzo, Daniel J. and Andrew J. Oswald (2001). "Are People Willing to Pay to Reduce Others' Income?" Annales d'Economie et de Statistique **63-64**, 39-65.

Appendix A.1: Utility function for $\alpha = 0$

Let

$$u(m,\, y\,;\, \alpha) = \begin{cases} \frac{1}{\alpha}\,(m^{\,\alpha} + \theta\, y^{\,\alpha}), & \alpha \in (-\infty,\, 0) \cup (0,\, 1], \\ m\, y^{\,\theta}, & \alpha = 0. \end{cases}$$

We want to show that for $\alpha \neq 0$, the indifference curves of $u(m,\, y\,;\, \alpha)$ converge to indifference curves of $u(m,\, y\,;\, 0) = m\, y^{\,\theta}$. Fix a point $(m_{_0},\, y_{_0})$ with $m_{_0} > 0$ and $y_{_0} > 0$. For every $\alpha \in (-\infty,\, 0) \cup (0,\, 1]$, the set $\{(m, y) : u(m,\, y\,;\, \alpha) = u(m_{_0},\, y_{_0}\,;\, \alpha)\}$ is the indifference curve for the given α that passes through the point $(m_{_0},\, y_{_0})$. On this indifference curve, y can be written as a function of m:

$$y(m\,;\, \alpha) = \left(\frac{m_{_0}^{\,\alpha} + \theta\, y_{_0}^{\,\alpha} - m^{\,\alpha}}{\theta}\right)^{1/\alpha}.$$

It suffices to show for each fixed $\bar{m} > 0$ that $y(\bar{m}\,;\, \alpha)$ converges pointwise to $y(\bar{m}\,;\, 0) = m_{_0}^{1/\theta}\, y_{_0}\, \bar{m}^{-1/\theta}$ as $\alpha \to 0$.

The limit of $y(\bar{m}\,;\, \alpha)$ as $\alpha \to 0$ can be determined by applying L'Hospital's rule to $\ln y(\bar{m}\,;\, \alpha)$:

$$\begin{aligned} \lim_{\alpha \to 0} \ln y(\bar{m}\,;\, \alpha) &= \lim_{\alpha \to 0} \frac{1}{\alpha} \ln\left(\frac{m_{_0}^{\,\alpha} + \theta\, y_{_0}^{\,\alpha} - \bar{m}^{\,\alpha}}{\theta}\right) \\ &= \lim_{\alpha \to 0} \frac{m_{_0}^{\,\alpha} \ln m_{_0} + \theta\, y_{_0}^{\,\alpha} \ln y_{_0} - \bar{m}^{\,\alpha} \ln \bar{m}}{\theta}. \end{aligned}$$

From this it follows that

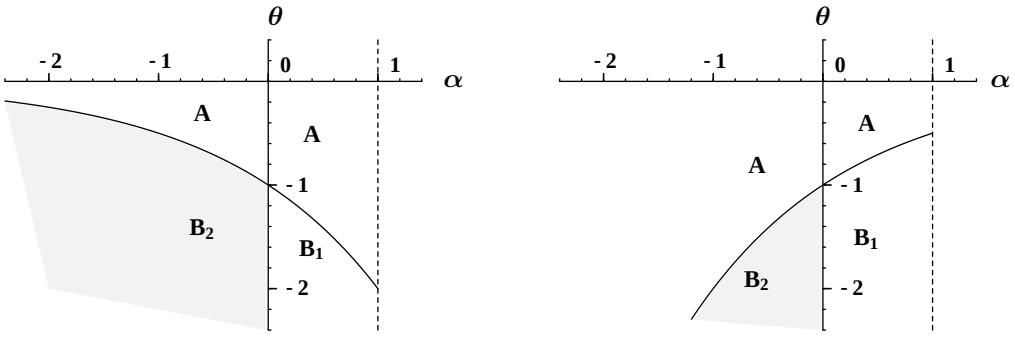
$$\ln y(\bar{m}\,;\, 0) = \frac{\ln m_{_0} + \theta\, \ln y_{_0} - \ln \bar{m}}{\theta}$$

so along the indifference curves of $u(m,\, y\,;\, 0)$, $m\, y^{\,\theta} = m_{_0}\, y_{_0}^{\,\theta}$, which is the required result.

Appendix A.2: Definition of the best-response function $q^*(x\,;\, \theta,\, \alpha)$

Theorem 1: For each $x \in (0,\, 24)$ and each $(\theta,\, \alpha) \in (-\infty,\, \infty) \times (-\infty,\, 1]$ there is a unique $q^* \in (0,\, 24 - x]$ that maximizes the utility function $U(x,\, q) = \frac{1}{\alpha}\,(24 - x - q)^{\alpha}\,(q^{\alpha} + \theta\, x^{\alpha})$.

Proof: We partition the space of values for θ, α, and x into a connected (relatively) open set A with a unique interior solution to the utility maximization problem and into a connected closed set B with a boundary solution to the utility maximization problem. The boundary between sets A and B, which depends on θ, α, and x, is characterized by the function $\theta(x,\, \alpha) = -\left(\frac{24-x}{x}\right)^{\alpha}$. Region B is subdivided into a region $B_{_1}$ where the utility function is bounded, and a region $B_{_2}$ where the utility function has an asymptote as $q \to 24 - x$. Figure A.2.1 (a) shows these three regions in a cross section for $x = 8$; figure A.2.1 (b) shows a cross section for $x = 16$.



(a) $\theta(x)$ for $x = 8$ (b) $\theta(x)$ for $x = 16$

Figure A.2.1: Regions evaluated to characterize the best response function

In order to evaluate properties of $U(x, q)$, it is useful to represent the output q of S as $q = c\,(24 - x)$ with $c \in [0, 1]$. Define $\tilde{U}_{x}(c) \equiv U(x, c\,(24 - x))$. Then

$$\tilde{U}_{x}(c) = \frac{1}{\alpha}\,(1 - c)^{\alpha}\,(24 - x)^{\alpha}\,(c^{\alpha}\,(24 - x)^{\alpha} + \theta\,x^{\alpha}). \tag{A.2.1}$$

The derivative of this utility function is

$$\tilde{U}_{x}'(c) = (1 - c)^{\,\alpha - 1}\,(24 - x)^{\,\alpha}\,((1 - 2\,c)\,c^{\,\alpha - 1}\,(24 - x)^{\,\alpha} - \theta\,x^{\,\alpha}). \tag{A.2.2}$$

For $c \in (0, 1)$, $\tilde{U}_{x}'(c) = 0$ if and only if $f_{x}(c) \equiv (1 - 2\,c)\,c^{\,\alpha - 1}\,(24 - x)^{\,\alpha} - \theta\,x^{\,\alpha}$ is zero. Lemma 1 shows that for $\alpha \in [-2, 0) \cup (0, 1)$, $f_{x}(c)$ is a convex function, so that $f_{x}(c)$ has at most two roots. Lemma 2 shows that for $\alpha \in (-\infty, -2)$, $f_{x}(c)$ is convex on an interval $(0, c')$ and monotonically increasing on $(c', 1)$, so that it again has at most two roots. The two lemmas are used subsequently to prove Claims 1 through 3, which show that there is a unique maximizer of $\tilde{U}_{x}(c)$ for regions A, B_{1}, and B_{2}. Claims 4 and 5 treat the cases $\alpha = 0$ and $\alpha = 1$ separately, but shows that they are consistent with the other cases. Theorem 1 follows from Claim 1 through 5. ■

Lemma 1: For $\alpha \in [-2, 0) \cup (0, 1)$, $f_{x}(c)$ is a convex function of c, for all $c \in (0, 1)$. Therefore the first-order condition for a local maximum, $\tilde{U}_{x}'(c) = 0$, has at most two roots in $(0, 1)$ for these values of α.

Proof: For $\alpha \in (-\infty, 0) \cup (0, 1]$, $f_{x}''(c)(\alpha - 1)\,c^{\alpha - 3}\,(24 - x)^{\alpha}\,(\alpha - 2\,\alpha\,c - 2)$. For $\alpha < 1$, this has the opposite sign from the last term, so $f_{x}(c)$ is convex when $\alpha - 2\,\alpha\,c - 2 < 0$. For $\alpha \in (0, 1)$ and for $\alpha \in [-2, 0)$ this inequality holds for all $c \in (0, 1)$. Hence for $\alpha \in [-2, 0) \cup (0, 1)$, $f_{x}(c)$ is convex. ■

Lemma 2: For $\alpha \in (-\infty, -2)$, $f_x(c)$ is a convex function of c for $c \in \left(0, \frac{\alpha-2}{2\alpha}\right)$, and it is a strictly increasing function for $c \in \left(\frac{\alpha-2}{2\alpha}, 1\right)$. Therefore, $f_x(c)$ has at most two roots on $(0, 1)$.

Proof: As noted in the proof of Lemma 1, $f_x(c)$ is convex only if $\alpha - 2\alpha c - 2 < 0$. For $\alpha < 0$ this is equivalent to the inequality $c < \frac{\alpha-2}{2\alpha}$, and if $\alpha < -2$, then $\frac{\alpha-2}{2\alpha} < 1$ so that convexity only holds for $c \in \left(0, \frac{\alpha-2}{2\alpha}\right)$. Since $f_x'(c) = c^{\alpha-2}(24-x)^{\alpha}(\alpha - 2\alpha c - 1)$ is positive for $c > \frac{\alpha-1}{2\alpha}$, and $\frac{\alpha-1}{2\alpha} < \frac{\alpha-2}{2\alpha}$ when $\alpha < -2$. The conclusion of the lemma follows from this observation. ■

Claim 1: In region A, with $\theta > \theta(x, \alpha)$, there is a unique value $c^* \in (0, 1)$ where $\tilde{U}_x(c)$ takes on its maximum value.

Proof: As $c \to 0$, $\tilde{U}_x'(c) \to \infty$, so the value of c that maximizes $\tilde{U}_x(c)$ is in the interval $(0, 1]$. As $c \to 1$, the first term in equation (A.2.2) approaches ∞, the second term is finite, and the last term has the finite limit $g(x) \equiv -(24-x)^{\alpha} - \theta x^{\alpha}$. Whether $\tilde{U}_x'(c)$ approaches $+\infty$, 0, or $-\infty$ as $c \to 1$ therefore depends on the sign of the last term, which is $f_x(c)$.

Since $\lim_{c\to 0} f_x(c) = \infty$ and $\lim_{c\to 1} f_x(c) = g(x)$ is negative for $\theta > \theta(x, \alpha)$, $f_x(c)$ changes sign on $(0, 1)$ at least once. By Lemma 1, $f_x(c)$ changes sign at most twice in $(0, 1)$ for $\alpha \in [-2, 0) \cup (0, 1)$ (and hence $\tilde{U}_x'(c)$ changes sign at most twice). By Lemma 2, $f_x(c)$ changes sign at most twice in $(0, 1)$ for $\alpha \in (-\infty, -2)$. Therefore there are at most two roots of $f_x(c) = 0$ in $(0, 1)$ (and equivalently, there are at most two roots of the first order condition $\tilde{U}_x'(c) = 0$). As $c \to 1$, $f_x(c) \to g(x)$ and $g(x) < 0$ in region A. Since $f_x(c)$ approaches a negative limit as $c \to 1$, it has a unique root in $(0, 1)$, which demonstrates that $\tilde{U}_x'(c) = 0$ has a unique root in $(0, 1)$.

Claim 2: In region B_1, with $\theta \leq \theta(x, \alpha)$ and $\alpha \in (0, 1)$, we show that $\tilde{U}_x'(c) > 0$ for all $ac \in (0, 1)$, so that there is a boundary maximum of $\tilde{U}_x(c)$ at $c = 1$, i.e., $q^* = 24 - x$.

Proof: The sign of $\tilde{U}_x'(c)$ is the same as the sign of $f_x(c)$, so it is sufficient to show that $f_x(c) > 0$ at its minimum on $(0, 1)$. The argument below demonstrates first that $f_x(c)$ is decreasing on $(0, 1)$ so that it takes on its minimum at $c = 1$ and then shows that $f_x(1) > 0$ so that $\tilde{U}_x'(c) > 0$ for all $c \in (0, 1)$.

Since $f_x'(c) = c^{\alpha-2}(24-x)^{\alpha}(c - 2c\alpha - 1)$, $f_x'(c) < 0$ if and only if $c - 2c\alpha - 1 < 0$. The last inequality is equivalent to the inequality $c > \frac{\alpha-1}{2\alpha}$ for $\alpha < 0$. Since this inequality holds

for all $\alpha \in (0, 1)$, $f_x(c)$ is decreasing on $(0, 1)$. Since $f_x(1) = g(x)$, and $g(x)$ is positive in region B_1, it follows that $\tilde{U}_x'(c) > 0$ for all $c \in (0, 1)$.

Claim 3: In region B_2, with $\theta \leq \theta(x, \alpha)$ and $\alpha < 0$, we show that for any $c' < 1$, $\tilde{U}_x(c)$ is bounded for $c \in [0, c']$ and $\tilde{U}_x(c) \to \infty$ as $c \to 1$ so that there is an asymptote of the utility function at $c = 1$. Consequently, there is a boundary maximum of $\tilde{U}_x(c)$ at $c = 1$, i.e., $q^* = 24 - x$.

Proof: It is clear from equation (A.2.1) that $\tilde{U}_x(c)$ is bounded for $c \in [0, c']$. As $c \to 1$, the term $(1 - c)^{\alpha} \to \infty$ for $\alpha < 0$, and the first and third terms are both finite, so $\tilde{U}_x(c) \to \infty$ if the last term, $c^{\alpha}(24 - x)^{\alpha} + \theta x^{\alpha}$, tends to a negative limit as $c \to 1$. Since $\theta < \theta(x, \alpha)$ in region B_2 and this expression is equivalent to $(24 - x)^{\alpha} + \theta x^{\alpha}$, the claim follows.

Claim 4: For $\alpha = 0$, there is a unique maximum of $U(x, q)$ at $q^* = \frac{24-x}{2+\theta}$ when $\theta > -1$ and there is a unique maximum of $U(x, q)$ at $q^* = 24 - x$ when $\theta \leq -1$.

Proof: This follows immediately from the utility maximization problem for $\alpha = 0$.

Claim 5: For $\alpha = 1$, there is a unique maximum of $U(x, q)$ at $q^* = 12 - \frac{1+\theta}{2} x$ when $\theta > \theta(x, 1)$ and there is a unique maximum of $U(x, q)$ at $q^* = 24 - x$ when $\theta \leq \theta(x, 1)$.

Proof: This follows immediately from the utility maximization problem for $\alpha = 1$.

Appendix A.3: Calculation of $q^*(x\,;\,\theta,\,\alpha)$

Claim 1 in Appendix A.2 demonstrates that for all $(x, \theta, \alpha) \in B_1 \cup B_2$ (where $\theta < \theta(x, \alpha)$), $U(x, q)$ takes on its maximum at $q = 24 - x$. Claim 1 also demonstrates that (1) if $(x, \theta, \alpha) \in A$, then $U'(x, 0) = \infty$ and $U'(x, 24-x) < 0$ and (2) $U'(x, q)$ has a single root in $(0, 24-x)$. We use (1) and (2) to calculate $q^*(x\,;\,a,\,b,\,\alpha)$. Since the derivative is infinite at $q = 0$, we start by evaluating $U'(x, 1)$. If $U'(x, 1) > 0$ we use the secant method with $U'(x, 1)$ and $U'(x, 24 - x)$ to find q^* such that $U'(x, q^*) = 0$. If $U'(x, 1) < 0$, we bisect the interval until we find 2^{-k} such that $U'(x, 2^{-k}) > 0$, and then apply the secant method to identify q^* such that $U'(x, q^*) = 0$.

Appendix A.4: Gauss-Newton non-linear regression

The general form of the Gauss-Newton non-linear regression is

$$\beta^{(j+1)} = \beta^{(j)} - c^{(j)} \left(D^{(j)}\right)^{-1} g^{(j)}, \tag{A.4.1}$$

where $\beta^{(j)}$ is the parameter estimate after j iterations of the Gauss-Newton algorithm, $D^{(j)}$ is an approximation to the Hessian matrix of the regressors, $g^{(j)}$ is the gradient of $SSR(\beta^{(j)})$, and $c^{(j)}$

is a constant that is chosen to assure convergence. (In this application, the regressor functions are $q^*\left(x\,;\,\beta^{(j)}\right)$. See Davidson and MacKinnon [1993, pp. 201-5] for a general formulation of the Gauss-Newton method in non-linear least squares.) This appendix describes the choices of $D^{(j)}$ and $c^{(j)}$ used to find parameter estimates for the Stackelberg game. For notational convenience, the parameter triple $(a,\, b,\, \alpha)$ and β are used interchangeably throughout this appendix.

The matrix $D^{(j)}$ is constructed from the Jacobian matrix $J(\beta)$ of the regressors, which in this case is the best-response function $q^*(x\,;\,a,\,b,\,\alpha)$. The j^{th} row of $J(\beta)$ is the derivative of the regressor $q^*(x\,;\,a,\,b,\,\alpha)$ evaluated at the j^{th} observation x_j, i.e.,

$$(J_{n,\,1}(\beta),\, J_{n,\,2}(\beta),\, J_{n,\,3}(\beta)) \left(\frac{\partial q^*(x_n\,;\,a,\,b,\,\alpha)}{\partial a},\, \frac{\partial q^*(x_n\,;\,a,\,b,\,\alpha)}{\partial b},\, \frac{\partial q^*(x_n\,;\,a,\,b,\,\alpha)}{\partial \alpha} \right).$$

We take $D^{(j)} = 2\,J\left(\beta^{(j)}\right)^{\top}\,J\left(\beta^{(j)}\right) + I$. With $\beta^{(j+1)}$ defined as in equation (A.4.1), $SSR\left(\beta^{(j+1)}\right)$ may be greater than $SSR\left(\beta^{(j)}\right)$ for many values of $c^{(j)}$. The value of $c^{(j)}$ is selected so that $SSR\left(\beta^{(j+1)}\right) < SSR\left(\beta^{(j)}\right)$. In the algorithm we employ, $c^{(j)} = 2^{-k}$, where k is the first value from the set $k \in \{0,\, 1,\, 2,\, \ldots,\, 20\}$ such that $SSR\left(\beta^{(j+1)}\right) < SSR\left(\beta^{(j)}\right)$.

In addition to the iterative Gauss-Newton parameter estimation procedure above, there are two other aspects of the algorithm that we should note. First, iterations continue so long as the maximum of the differences $\left|a^{(j+1)} - a^{(j)}\right|$, $\left|b^{(j+1)} - b^{(j)}\right|$, $\left|\alpha^{(j+1)} - \alpha^{(j)}\right|$, and $\left|SSR\left(\beta^{(j+1)}\right) - SSR\left(\beta^{(j)}\right)\right|$ is greater than 10^{-8}. Finally, once the adjustment of both the parameter estimates and the SSR is below the threshold 10^{-8} and a parameter estimate $\beta^{(j^*)}$ is obtained, we conduct a grid search over $SSR(\beta)$ in a region around $\beta^{(j^*)}$ to insure that $\beta^{(j^*)}$ minimizes $SSR(\beta)$.

Appendix A.5: Tests of hypotheses

H.1 *The parameter pair* $(a,\, b) = (0,\, 0)$.

When a and b are both restricted to be zero (so that the model is restricted to the standard model of individualistic preferences), $SSR = 1476.5$. The F statistic is

$$\begin{aligned} \frac{SSR(\tilde{\beta}) - SSR(\hat{\beta})/2}{SSR(\hat{\beta})/(n-k)} &= \frac{(1476.5 - 892.8)/2}{892.8/217} \\ &= 70.9, \end{aligned}$$

where $\tilde{\beta}$ is the parameter estimate pair for the restricted model and $\hat{\beta}$ is the parameter estimate pair for the unrestricted model. The cumulative distribution of $F(2, 217)$ at 70.9 is greater than $1 - 10^{-16}$, so we are able to reject the hypothesis that preferences are individualistic with a p-value of less than 10^{-16}.

H.2 *For $m_0 = 64$ (which is the Cournot-Nash profit level for S), the parameter a is positive.*

When a is restricted to be greater than or equal to zero, SSR is minimized at $a = 0$, $b = 1.14$, and $c = 0.78$. For these parameters, $SSR = 936.4$. The F statistic for the test of the hypothesis that $a > 0$ is

$$\begin{aligned} \frac{SSR(\tilde{\beta}) - SSR(\hat{\beta})/1}{SSR(\hat{\beta})/(n-k)} &= \frac{936.4 - 892.8}{892.8/217} \\ &= 10.6, \end{aligned}$$

where $\tilde{\beta}$ is the parameter estimate pair for the restricted model and $\hat{\beta}$ is the parameter estimate pair for the unrestricted model. The cumulative distribution of $F(1, 217)$ at 10.6 is 0.999, so the hypothesis that $a > 0$ can be rejected with a p-value of 0.001, i.e., we can reject the hypothesis that the second mover has a reference profit level that is at or below the Cournot-Nash profit level.

H.3 *The parameter b is negative.*

When b is restricted to be less than or equal to 0, SSR is minimized when $a = -0.30$, $b = 0$, and $c = 1.00$. For these parameters, SSR = 941.0. The F statistic for the test of the hypothesis that $b < 0$ is

$$\begin{aligned} \frac{SSR(\tilde{\beta}) - SSR(\hat{\beta})/1}{SSR(\hat{\beta})/(n-k)} &= \frac{941.0 - 892.8}{892.8/217} \\ &= 11.7. \end{aligned}$$

The cumulative distribution of $F(1, 217)$ at 11.7 is greater than 0.999, so the hypothesis that $b < 0$ can be rejected with a p-value of less than 0.001.