

# Fairness and Trust in Game Theory

Daniel M. Hausman

(This paper was written in 1998-99 and never published.)

## Abstract

In order to apply game theory to interactions among people, one needs to know what game people are playing, and to know what game people are playing, one needs to know their preferences. If the individuals care only about their own material payoff, it may be easy to know their preferences. But other things matter to people, and to apply or test game theory, some account of what those other things are and how they matter may be needed.

This essay examines the dependence of utility on material self-interest, altruism, concerns about fairness, and concerns about trust. It begins with Matthew Rabin's proposal for incorporating fairness into game theory (1993), points out inadequacies, and sketches remedies. The complexities that arise in even the simplest two-person interactions are daunting.

Keywords: Fairness, Game Theory, Reciprocity, Trust, Trustworthiness

### Fairness and Trust in Game Theory

When game theorists attempt to explain and predict how people behave or to advise them on what strategy to employ, they must first determine what game people are playing. Suppose, for example, that experimental subject 1 has to decide whether to play up or down, and subject 2 has to decide whether to play left or right. The subjects choose independently and are not known to one another. They will never interact again. The following payoff matrix is common knowledge (or so the experimenter hopes):

	left	right
up	\$2,\$2	\$0,\$3
down	\$3,\$0	\$1,\$1

If the players care only about their own monetary returns and believe what the experimenter hopes they do, then they are playing a prisoner's dilemma. If, for example, they are both altruists or if both seek to maximize the experimenter's financial costs, they are not playing a prisoner's dilemma. When subjects “cooperate” in such experiments, one possible explanation is that they are not playing a prisoner's dilemma. In some cases – though certainly not in all – it may be plausible to question the "obvious" materially self-interested construal of the game.

To know what game is being played, one needs to know the player's preferences. If there is no way to know the preferences apart from observing how people play, then game theory will have no role in predicting or advising, and the only words of explanation game theorists will be able to offer is that people chose as they preferred. Some framework is needed relating preferences to observable features of games. One such framework is established by the hypothesis that players care about nothing except their own payoffs. This hypothesis is convenient but not always true. Sometimes game theorists need a framework that relates preferences to additional properties of games, if only to know when they can

safely disregard all the determinants of preferences except a player's own material payoffs.

People obviously differ, but it may be possible to generate a short list of features of games upon which their preferences depend and to model the differences among people by the differences in the weights they place on difference factors. For example, in the special case of economic experimentation in which the subjects are anonymous, it seems a plausible first approximation to take people's ranking of outcomes to depend on material payoff to oneself (material self-interest), material payoff to others (altruism), reciprocation or sharing of benefits and costs ("fairness"<sup>1</sup>), and trustworthiness. If one boldly (and almost certainly wrongly) assumes that these factors are additive, one can represent a person's overall utility function as a weighted sum of utilities due to these separate factors.

The factors such a model incorporates are ill-defined, difficult to measure, and not exhaustive. Furthermore, one suspects that with so many weights to fiddle with, one could model an interaction between  $n$  persons as virtually any  $n$ -person game. On the other hand, the only alternatives are to cling to the simple hypothesis that only material self-interest motivates or to make ad hoc adjustments to incorporate other factors – a bit of altruism here, or some trustworthiness there. This essay explores the possibility of constructing a substantive general framework in which to conceptualize the dependence of preferences in simple games on matters of fairness and trust.

My treatment will build on Matthew Rabin's article, "Incorporating Fairness into Game Theory and Economics" (1993). In that essay, Rabin sketches a way in which the beliefs of individuals concerning how "kindly" their opponents are treating them and about how kindly they are treating their opponents influence preferences. Responding to kindness with kindness and to unkindness with unkindness captures shows one sort of fairness, and Rabin in this way incorporates fairness into game theory. Concerns about fairness cause net utility to diverge from "material utility," and can add or delete equilibria. This paper reformulates Rabin's path-breaking account in notation that I find more perspicuous, shows its limitations, and, building on his techniques, presents a schema for incorporating

---

<sup>1</sup>This is only one of many possible construals of fairness. Others involve equality and a consideration of factors that are usually missing from the experimental circumstances. Later I shall emphasize a different notion of fairness.

trust as well as fairness into game theory. This paper is less ambitious than Rabin's in one important regard. Whereas (following Geanakoplos, et al. 1989) Rabin suggests a way of reformulating game theory, this paper assumes a standard formulation of game theory and merely investigates how a player's preferences depend on features of game forms in addition to his or her own material payoff.

### 1. Rabin's theory reformulated

Consider a game between two people,  $A$  and  $B$ .  $A$  has a choice of strategy  $s$  from strategy set  $S$ , while  $B$  plays  $t$  from strategy set  $T$ . Let  $v_A: S \times T \rightarrow \Re$  be the "material payoffs" to  $A$  and similarly for  $v_B$ .<sup>2</sup>

Suppose that  $A$  chooses a particular strategy  $s$ . Let  $v^h(s)$  be the highest payoff  $A$  can receive when  $A$  plays  $s$ . Let  $v_A^{\min}(s)$  be the lowest payoff  $A$  can receive when  $A$  plays  $s$ . Let  $v_A^l$  be the lowest payoff among points that are Pareto-efficient in the set of payoffs for  $s$ . Rabin defines the "equitable" payoff,  $v_A^e(s) = \frac{1}{2}[v_A^h(s) + v_A^l(s)]/2$ , but many of his results hold for any arbitrary  $v^e$  provided that  $v^h \geq v^e \geq v^l$ . The equitable payoff to  $A$  depends solely on the set of optimal payoffs, given that  $A$  plays  $s$ . This departs from common-sense notions of fairness, in which  $A$ 's initial state and the extent of sacrifice shown by  $B$  matter. More on this point later.

Rabin holds that  $A$  is kind to  $B$ , when  $v_B(s, t^b) > v_B^e(t^b)$ , where  $t^b$  is the strategy that  $A$  believes that  $B$  is playing. When  $v_B(s, t^b) < v_B^e(t^b)$ ,  $A$  is unkind to  $B$ . One cannot, however, define  $A$ 's kindness to  $B$  as  $v_B(s, t^b) - v_B^e(t^b)$ , because this difference changes with positive affine transformations of  $v_B$ . Rabin thus normalizes by dividing this difference by  $[v_B^h(t^b) - v_B^{\min}(t^b)]$  and defines the kindness of  $A$  to  $B$ ,  $k_{AB}(s, t^b)$ , as  $[v_B(s, t^b) - v_B^e(t^b)] / [v_B^h(t^b) - v_B^{\min}(t^b)]$ , unless the denominator is zero, in which case  $k_{AB}(s, t^b) = 0$ .  $k_{BA}(s^b, t)$  is defined analogously.  $k_{AB}(s, t^b)$  assumes its maximum value of one-half if  $A$  gives  $B$   $v_A^h(t)$ , and  $v_A^l(t) = v_A^{\min}(t)$ . If  $A$ 's strategy provides  $B$ 's minimum payoff, then  $k_{AB} = [v_B^{\min} - \frac{1}{2}v_B^h - \frac{1}{2}v_B^l] / (v_B^h - v_B^{\min}) = -1 + \frac{1}{2}(v_B^h - v_B^l) / (v_B^h - v_B^{\min})$  which reaches its minimum value of  $-1$ , when there is only one Pareto optimal point and  $v_B^h = v_B^l$ . When  $v_B^l = v_B^{\min}$ , the minimum value of  $k_{AB}$  is  $-1/2$ .

---

<sup>2</sup>The exposition follows Rabin 1993, except that I have changed the notation and have introduced a weight attached to fairness rather than a scale factor attaching to material payoffs.

In Rabin's view,  $A$ 's utility depends on how kind  $A$  is being to  $B$ , given how kind  $A$  believes  $B$  is being to  $A$ . How kind  $A$  believes that  $B$  is being to  $A$  is defined analogously to kindness itself as  $k_{BA}^b = [v_A(s^{bb}, t^b) - v_A^e(s^{bb})] / [v_A^h(s^{bb}) - v_A^{\min}(s^{bb})]$ , where the superscript  $b$  in the kindness function indicates that this is how kind  $A$  believes that  $B$  is being, and the superscript  $bb$  indicates that this is the strategy that  $A$  believes that  $B$  believes that  $A$  is playing.

Rabin then suggests that  $A$ 's utility depends on  $A$ 's material payoffs, and the *product* of  $A$ 's kindness to  $B$  and how kind  $A$  believes that  $B$  is being to  $A$ . Rabin adopts the following form:

$$u_A(s, t^b, s^{bb}) = v_A(s, t^b) + \alpha_A k_{BA}^b(s^{bb}, t^b) [1 + k_{AB}(s, t^b)].$$

He does not explicitly include  $\alpha$ , the parameter that weights fairness. Instead he considers games in which the material payoffs can be scaled up or down. One advantage of the formulation here is that it leaves open the possibility that the players place different weights on fairness. Rabin employs the form above rather than  $u_A(s, t^b, s^{bb}) = v_A(s, t_b) + \alpha_A k_{BA}^b(s^{bb}, t^b) \cdot k_{AB}(s, t^b)$  on the grounds that if  $A$  believes that  $B$  is being unkind, then  $u < v$ , regardless of whether  $k_{AB} > 0$ . As Rabin notes, this form implies that we prefer it if those we're unkind to are nicer to us. Given these utility functions, concerns about fairness will be swamped by sufficiently large material payoffs.

Rabin defines a fairness equilibrium as a pair of strategies  $(s, t)$  such that, given  $t$ ,  $s$  maximizes  $u_A(s, t)$ , and given  $s$ ,  $t$  maximizes  $u_B(s, t)$  and  $s = s^b = s^{bb}$ , and  $t = t^b = t^{bb}$ . Much of Rabin's essay is concerned with properties of fairness equilibria, which I shall not explore here.

## 2. A successful application

Consider an illustration Rabin provides of how his construction could be used to model labor relationships involving gift exchange. Workers can choose either a high or a low level of effort,  $H$  or  $L$ . High effort provides the firm with revenue  $R > 0$  and causes the worker disutility  $d$ . The firm provides a benefit level  $b$  to the workers ( $0 \leq b \leq R$ ). Rabin does not discriminate between what benefit level firms supply and what strategy firms employ, nor between the level of effort workers choose and what strategy

they employ.<sup>3</sup>

The worker's "material" utility function,  $v_w$  is  $b^{1/2} - d$  if the worker chooses a high effort level, and  $v_w = b^{1/2}$ , if the worker chooses a low effort level. The firm's material utility,  $v_F = (R - b)^{1/2}$ , if the worker exerts high effort (and  $b \leq R$ ) and zero otherwise. Ignoring fairness, the unique Nash equilibrium involves low effort and a zero benefit level.

One can then calculate the kindness of firms to workers as follows:

$$\begin{aligned} v_w^h(H) &= R^{1/2} - d & v_w^l(H) &= v_w^{\min}(H) = -d & v_w^e(H) &= (R^{1/2}/2) - d \\ v_w^h(L) &= R^{1/2} & v_w^l(L) &= v_w^{\min}(L) = 0 & v_w^e(L) &= 1/2 R^{1/2} \end{aligned}$$

So  $k_{Fw}(H, b) = (b/R)^{1/2} - 1/2$  and  $k_{Fw}(L, b) = 1/2$ . The kindness of workers to firms is then calculated from:

$$\begin{aligned} v_F^h(b) &= (R - b)^{1/2} & v_F^l(b) &= v_F^{\min}(b) = 0 & v_F^e(b) &= 1/2(R - b)^{1/2} \\ v_F^h(0) &= R^{1/2} & v_F^l(0) &= v_F^{\min}(0) = 0 & v_F^e(0) &= 1/2 R^{1/2}. \end{aligned}$$

Thus  $k_{wF}(H, b) = k_{wF}(H, 0) = 1/2$  and  $k_{wF}(L, b)$  and  $k_{wF}(L, 0) = -1/2$ .

In calculating the net utilities of firms and workers, Rabin takes the weight of fairness concerns to be unity. So  $u_F(H, b) = (R - b)^{1/2} + 1/2[(b/R)^{1/2} + 1/2]$ . The firm will maximize its utility when  $u'_F [= -1/2(R - b)^{-1/2} + .25(b/R)^{-1/2}] = 0$ . So  $b^*$ , the level of benefits that maximizes utility for the firm is  $R/(1 + 4R)$ . Workers will supply a high level rather than a low level of effort if and only if  $u_w(H, b) > u_w(L, b)$ .  $u_w(H, b) = b^{1/2} - d + 1.5[(b/R)^{1/2} - 1/2]$  and  $u_w(L, b) = b^{1/2} + 1/2[(b/R)^{1/2} - 1/2]$ .<sup>4</sup> So  $u_w(H) > u_w(L)$  if and only if  $-d + (b/R)^{1/2} - 1/2 > 0$ . So  $b/R > (d + 1/2)^2$ . If  $b = R/(1 + 4R)$ , workers prefer to provide higher effort if and only if  $[1/(1 + 4R)] > (d + 1/2)^2$ . If  $R$  is positive, the left-hand side must be less than 1 and so a high-effort equilibrium is possible only if  $d < 1/2$  and  $R < .25[(d + 1/2)^{-2} - 1]$ .

---

<sup>3</sup>If one elaborates Rabin's example so that firms pay wages in advance and workers then decide what level of effort to exert, workers can employ a strategy  $s$  of supplying high effort if  $b^{1/2} > d$  and low effort otherwise. The kindness of firms to workers who employ such a strategy is different than the kindness of firms to workers who always supply low or high effort, but when  $b^{1/2} > d$ ,  $k_{wF}(s, b) = k_{wF}(H, b)$ , and when  $b^{1/2} < d$ ,  $k_{wF}(s, b) = k_{wF}(L, b)$ . This conflicts with the intuition that the kindness of strategy  $s$  should be between the kindness of  $H$  and the kindness of  $L$ . In defense of Rabin's account, one could argue that our intuitions confuse kindness and fairness.

<sup>4</sup>Rabin apparently uses the formula  $u_A = v_A + k_{BA} \cdot k_{AB}$ , rather than  $u_A = v_A + k_{BA} \cdot (1 + k_{AB})$  and so the last term in  $u_w(H, b)$  in Rabin's text is  $1/2$  rather than  $1.5$ , and the last term in  $u_w(L, b)$  in his text is  $-1/2$  rather than  $+1/2$ . Since the difference is unaffected, this error has no consequences for Rabin's calculations. Rabin also inconsequentially miscalculates the utility difference as  $.25[(d + 1/2)^{-1/2} - 1]$ .

The larger the disutility of working,  $d$ , the *smaller* must be  $R$  (the greater output owing to high effort). Rabin justifies this implication by arguing that if  $d$  is large, “the worker is very tempted to ‘cheat’ the firm by not working hard. The only way he will not cheat is if the firm is being very kind. But the firm’s material costs of yielding a given percentage of profits to the worker increases as  $R$  increases...” (1993, p. 1294).

Increasing the weight given to fairness loosens the constraint that  $d < \frac{1}{2}$ . If both workers and firms weight fairness with some positive number  $\alpha$ , then the firm maximizes utility when  $b^* = \alpha^2 R / (\alpha^2 + 4R)$ , and workers will provide high effort if and only if  $(b/R)^{1/2} > (d + \frac{1}{2}\alpha)/\alpha$ . Since  $b \leq R$ ,  $[(d + \frac{1}{2}\alpha)/\alpha] \leq 1$ , and  $d \leq \frac{1}{2}\alpha$ . As  $\alpha$  increases, the limits on  $R$  are also relaxed.

### 3. An Unsuccessful Application<sup>5</sup>

In a recent experiment, Berg, Dickhaut, and McCabe (1995; BDM henceforth) give each subject in one room (the investors) an endowment  $e$  and the opportunity to send some investment  $i$  ( $0 \leq i \leq e$ ) to an anonymous subject in a second room (the responders). The subjects in the second room begin with the same endowment,  $e$ . The amount invested is tripled by the experimenters, and each responder can then, if he or she chooses, send some of the windfall ( $3ri$ ) back to the investor ( $0 \leq r \leq 1$ ). All of this is common knowledge, but the players are anonymous, even to the experimenters. This experiment apparently has much in common with the gift exchange example just considered.

BDM found that nearly 90% invested something, and in roughly half the cases there were positive returns to investment ( $r > 1/3$ ). A substantial minority invested their entire endowment, and the average investment was somewhat more than half of the endowment. Larger investments generally earned larger paybacks. These results have been replicated in an unpublished study carried out in France (Meidinger, Robin, and Ruffieux; MRR henceforth). The unique subgame perfect equilibrium is of course no investment and no return. Let us consider whether Rabin’s framework can be applied to

---

<sup>5</sup>The discussion that follows is influenced by Meidinger, Robin, and Ruffieux (forthcoming), who also show that Rabin’s model fails to apply to a replication of Berg, Dickhaut, and McCabe.

explain the experimental findings.

Let  $A$  be an investor and  $B$  a responder and suppose that the utility function  $v$  stating the "material payoffs" as a function of money is linear, so that the money amounts will serve as utility amounts. I will individuate the pure strategies of the investors by the *amount* invested,  $i$ , and the pure strategies of the responders by the *proportion*  $r$  of the windfall they return. Since investors do not know  $r$ , conditional strategies are not feasible for them. Responders can adopt strategies whereby  $r$  depends on  $i$ , but I shall for the moment ignore that possibility as well as mixed strategies. Since  $v_A^h(i) = e + 2i$  and  $v_A^l(i) = e - i$ ,  $v_A^e(i) = e + i/2$ .  $v_B^l(r)$  is either  $4e - 3re$  if  $r > 1/3$ , or  $e$  if  $r < 1/3$ , while  $v_B^h(r) = 4e - 3re$ . So  $v_B^e(r)$  is  $4e - 3re$  for  $r > 1/3$  or  $(5e - 3re)/2$  for  $r < 1/3$ .  $v_A^{\min}(i) = e - i$ , and  $v_B^{\min}(r) = e$ .

Since the most responders can do is to return the whole windfall and the least is to return nothing, the kindness of responders to investors,  $k_{BA}(i, r) = [e - i + 3ri - e - i/2]/3i = r - 1/2$ . For  $r > 1/3$ ,  $k_{AB}(i, r) = [e + 3i - 3ri - 4e + 3re]/[4e - 3re - e] = (i/e) - 1$  and for  $r \leq 1/3$   $k_{AB} = 2(i/e) - 1$ . For  $r > 1/3$ ,  $k_{AB}$  is never positive and is only zero when  $i = e$ . Not to be unkind, investors must invest their entire endowment. Since greater investment brings a larger material payoff if  $r > 1/3$ , this implication seems not reasonable, but the implication that investors are unkind if they do not invest at least half their endowment, even if responders return nothing is unacceptable. As mentioned before, kindness should not be defined without some reference to the costs to the benefactor as well as to the benefits to the recipient.<sup>6</sup> One could, if one wanted, simply stipulate that if  $r < 1/3$ ,  $v_B^e = 0$ , in which case  $k_{AB}(i, r)$  for  $r < 1/3$  would be  $i/e$ , but my goal, like Rabin's, is to sketch a general framework within which to describe the role of fairness and trust rather than to admit such ad hoc accommodations.

For  $r > 1/3$ ,  $u_B(i, r) = e + 3i - 3ri + \alpha[(i/e) - 1][r + 1/2]$ , where  $\alpha$  is the weight that responders place on fairness. Since both the coefficients on  $r$  are negative, responders maximize utility by setting  $r$

---

<sup>6</sup>Suppose a beggar stops at the home of a poor family and asks for food. That evening it happens that the family has enough food to provide a feast for the beggar, though it will mean that family members will starve afterwards. If the amount they give provides the beggar with a material utility less than the midpoint between where he would be with nothing or with their whole stock, they are, in terms of Rabin's framework, unkind. If the beggar stops at the home of a rich family, whose larder happens to be empty after a large dinner, and the rich family gives the few crumbs it has left, then (assuming that the beggar can only make use of food) it is very kind. Both the actual sacrifice and the beggar's benefit in the first case may be much larger, but Rabin's account of kindness forces us to judge the rich family to be kinder.

as low as possible. When  $r < 1/3$ ,  $u_B(i, r) = e + 3i - 3ri + \alpha[2(i/e) - 1](r + 1/2)$ , and so  $u_B$  is increasing in  $r$  when  $r < 1/3$  and  $\alpha[2(i/e) - 1] > 3i$ . For  $r < 1/3$ ,  $u_A(i, r) = e - i + 3ri + 2\alpha(r + 1/2)(i/e)$ , which for  $\alpha > 0$  is a decreasing function of  $i$ . Thus, when  $r > 1/3$ , *regardless of the weight they place on fairness*, responders will prefer to return less, and when  $r \leq 1/3$ , investors will, *regardless of the weight they place on fairness* prefer to invest nothing. So only the subgame perfect no-investment equilibrium remains. Rabin's framework cannot account for BDM's data.

Suppose that the parties both placed a *negative* weight on fairness. Then for  $r > 1/3$ ,  $u_B(i, r) = e + 3i - 3ri - \alpha[(i/e) - 1](r + 1/2)$ , where  $\alpha$  is the magnitude of the negative weight. This is increasing in  $r$  if and only if  $\alpha[1 - (i/e)] > 3i$ . For a sufficiently large negative weight on fairness it will maximize utility for  $B$  to set  $r = 1$ . For  $r > 1/3$ ,  $u_A(i, r) = e - i + 3ri - \alpha[r - 1/2][i/e]$ . Unless the negative weight place on fairness is too large (for  $r = 1$ ,  $\alpha < 4e$ ), this will reach a maximum at  $i = e$ . So there will be an (un)fairness equilibrium at  $i = e$  and  $r = 1$  if the parties place a small negative weight on fairness!

#### 4. More problems

What exactly has gone wrong? Consider the following monetary payoff matrix. Row goes first and has three choices, top, middle or bottom. It is common knowledge that Column sees how Row plays before responding with left or right.

<i>game 1</i>	left	right
top	\$3,0	\$3,0
middle	0,\$12	0,0
bottom	\$2,\$6	\$5,\$5

Row has only three pure strategies, which I shall call T, M, and B. Column has eight. I shall focus on only two of these: (L) "always play left" and (R\*) "play right if row plays bottom, otherwise left." If utility is an increasing function of money and of nothing else, (T, L) is a Nash equilibrium, and (B, R\*) is not. The kindness of Row to Column,  $k_{RC}(B, L)$ , is zero, while  $k_{CR}(B, L)$  is  $-1/2$ .  $k_{RC}(B, R^*)$  is  $-1/12$  and

$k_{CR}(B, R^*)$  is  $\frac{1}{2}$ . Since  $u_c(\cdot) = v_c + k_{RC} \cdot (1 + k_{CR})$ ,  $k_{RC}(B, L) = 0$ , and  $k_{RC}(B, R^*) < 0$ ,  $(B, R^*)$  is never a kindness equilibrium. ( $u_c(B, L) = 6$ , while  $u_c(B, R^*) = 5 - 1/8$ .)

The kindness of row to column of  $(B, R^*)$  should not be negative, though if Row believes that Column's strategy is  $R^*$ , then B maximizes Row's return and perhaps should not be regarded as kind, either.<sup>7</sup> Row's playing B is better for Column than Row's (materially self-interested) Nash equilibrium move, and places Row at risk of doing worse. Moreover, compare game 1 with game 2:

<i>game 2</i>	left	right
top	\$3,\$12	\$0,0
middle	\$0,0	0,0
bottom	\$2,\$6	\$5,\$5

Because  $(M, L)$ ,  $(B, L)$ ,  $(T, R)$  and  $(M, R)$  are all Pareto inefficient,  $k_{RC}(B, L)$  is now  $-\frac{1}{4}$  rather than zero and  $k_{RC}(B, R^*)$  is  $-\frac{7}{24}$ , while as before  $k_{CR}(B, L) = -\frac{1}{2}$  and  $k_{CR}(B, R^*) = \frac{1}{2}$ . In this case, intuitively, there is nothing kind (or sensible) about Row's playing B, but the small difference between the two games in Rabin's framework fails to capture the sharp divide between games 1 and 2.

## 5. Rabin's theory revised

Suppose for the moment that we hold to Rabin's view of fairness as reciprocation of benefits or harms and see whether we can avoid the unacceptable implications of his account by measuring kindness differently. Suppose that the players cared only about their own payoffs and suppose that in the resulting game there were a unique Nash or subgame perfect equilibrium.<sup>8</sup> Let the strategies in that equilibrium be  $s^*$  and  $t^*$ . I propose using the payoffs to these strategies as the benchmark for the measurement of

---

<sup>7</sup> The fact that Row is trusting Column should, however, matter, too. See §8 below.

<sup>8</sup>When there are multiple equilibria, substitute the average payoff in all the equilibria for the payoff of the unique equilibrium. It simplifies the exposition in the text to address only the special case involving a unique equilibrium.

kindness. If  $v_B(s, t) > v_B(s^*, t^*)$ , then  $A$  is being kind to  $B$ . If  $v_B(s, t) < v_B(s^*, t^*)$ , then  $A$  is unkind to  $B$ . Why? If any rationale can be given for playing Nash equilibrium strategies (which is not my subject here), then there is something intuitively natural in regarding the outcome of the game that would be played by materially self-interested players as a zero point against which kindness and malevolence can be measured. Later, in section 7, I shall suggest a different rationale.<sup>9</sup>

Once again we need to normalize so that kindness will not vary with a positive affine transformation of material utility, and for the moment, let us normalize in Rabin's way. Thus we define a revised kindness measure,  $k'_{RC}(s, t)$ , as  $[v_C(s, t) - v_C(s^*, t^*)]/[v_C^h(t) - v_C^{\min}(t)]$ , where  $v_C(s^*, t^*)$  is the material equilibrium payoff to Column. If the denominator is zero, let  $k'_{RC}(s, t)$  be zero.  $k'$  varies between  $-1$  and  $1$ . If  $v_C(s, t) - v_C(s^*, t^*) > 0$ , one can think of it as a measure of  $C$ 's share of the surplus that arises from the fact that  $R$  and  $C$  do not play their materially self-interested strategies.

On this alternative account of kindness, Row and Column are neither kind nor unkind, if they play  $s^*$  and  $t^*$ . But a player can be kind and materially self-interested if the other players are not materially self-interested. For example, if Row believes that Column will play  $R^*$ , then the payoff from  $B$  is larger than the payoff from  $L$ , and Row has a materially self-interested reason for playing  $B$ . Yet  $k'_{RC}(B, R) = 5/12$ .<sup>10</sup> The kindness of a strategy depends on the strategies chosen by the other players. Instead of comparing the payoff a player gets to an "equitable payoff," as Rabin proposes, I am suggesting that one compare it to the payoff achieved by a materially self-interested player when playing materially self-interested opponents. If you provide a benefit to me in playing your materially self-interested equilibrium strategy, then you are not being kind, and there is nothing unfair if I pursue my own material self-interest.

In game 1,  $(s^*, t^*)$  is  $(T, L)$ , and the payoff to Column of  $(T, L)$  and  $(T, R^*)$  is  $0$ . So  $k'_{RC}(B, L) =$

---

<sup>9</sup>Taking  $v^e$  to be the Nash or subgame perfect equilibrium implies in game 2 that if Row believed that Column plays strategy  $R$  (always right), then no matter what Row does, Row is unkind to Column. Perhaps one should take  $v_B^e(t)$  to be  $v_B^l(t)$  if the Nash equilibrium payoff to  $B$  is less than  $v_B^l(t)$ .

<sup>10</sup>One may question whether this accords with Rabin's basic intuition that fairness is reciprocation of benevolence, and below in section 7, I will suggest an alternative construal of fairness.

$\frac{1}{2}$  and  $k'_{RC}(B, R^*) = 5/12$ .  $k'_{RC}(B, L) = -1$  and  $k'_{RC}(B, R^*) = 2/3$ . Assuming that material utility is proportional to money payoff,  $u_C(B, L) = 6$  and  $u_C(B, R^*) = 5 + 25\alpha/36$ . So if  $\alpha > 36/25$ , a fairness equilibrium in Rabin's sense is possible. In game 2,  $k'_{RC}(B, L) = -\frac{1}{2}$  and  $k'_{RC}(B, R^*) = -7/12$ , while  $k'_{RC}(B, L) = -1$  and  $k'_{RC}(B, R^*) = 2/3$ . So  $u_C(B, L) = 6$  and  $u_C(B, R^*) = 5 - 35\alpha/36$ . Regardless of the (positive) value of  $\alpha$ ,  $(B, R^*)$  is not a kindness equilibrium in game 2.

This revision not only helps explain the difference in kindness in playing bottom in games 1 and 2. It also helps explain the difference in kindness between playing bottom in game 2 and game 3:

<i>game 3</i>	left	right
top	\$0,0	\$0,0
middle	\$-1,\$12	\$-1,0
bottom	\$2,\$6	\$5,\$5

In game three, because  $(T, L)$ ,  $(T, R)$  and  $(M, R)$  are Pareto inefficient,  $k_{RC}(B, L) = -\frac{1}{4}$  and  $k_{RC}(B, R^*) = -7/24$ , just as in game 2. Yet intuitively there is a big difference between games 2 and 3. In game 3, unlike game 2, there is a clear self-interested reason for Row to play B, regardless of whether Column is moved by considerations of fairness. If unfair at all, it is much less unfair for a Column player to play left in response to bottom in game 3 than in game 2. On the measure of kindness defined above,  $k'_{RC}(B, L) = 0$  (since  $(B, L)$  is the Nash equilibrium) and  $k'_{RC}(B, R^*) = -1/6$ .

The sign of the revised  $k'$  measure depends exclusively on whether the material outcome is better or worse than the materially self-interested equilibrium, while the magnitude of  $k'$  depends not only on how much better or worse it is, but on how this difference compares to the difference between the best and worst payoffs to the particular strategy. If  $A$  plays some strategy  $s$  other than  $s^*$  and achieves an outcome that  $A$  prefers to  $(s^*, t^*)$ , then (by the definition of  $(s^*, t^*)$ ),  $B$  is not making the best materially self-interested response to  $s$ . So if  $k'$  is positive, the giver must be sacrificing something. The  $k'$  measure does not, however, take into account how *much* the giver is sacrificing. If in game one, row could have

had \$6 rather than \$3 for sure by playing top, playing bottom would intuitively be a kinder thing to do, but on the measure of kindness proposed here, it would be no kinder or less kind than is playing bottom in game 1. More on this issue later.

This revision retains the successes Rabin's model registered. In the firm-worker game, for example,  $v_F(s^*, t^*)$  and  $v_W(s^*, t^*) = 0$ , and one can calculate that  $k'_{WF}(L, 0)$ ,  $k'_{FW}(L, 0)$ ,  $k'_{WF}(L, b)$ , and  $k'_{FW}(L, b)$  are all zero, and  $k'_{WF}(H, b)$  and  $k'_{WF}(H, 0)$  are 1.  $k'_{FW}(H, b) = (b^{1/2} - d)/R^{1/2}$  and  $k'_{FW}(H, 0) = -d/R^{1/2}$ . Assuming that firms and workers place an equal weight,  $\alpha$ , on fairness,  $u_F(H, b) = (R - b)^{1/2} + \alpha[1 + (b^{1/2} - d)/R^{1/2}]$ . So  $dU/dr = 0$  iff  $b = \alpha^2 R / (\alpha^2 + R)$ . For that level of benefits,  $u_w(H, b) > u_w(L, b)$  if and only if  $d < 2\alpha^2 R^{1/2} / [(R^{1/2} + 2\alpha)(\alpha^2 + R)^{1/2}]$ . So there is a fairness equilibrium, and, as in Rabin's own calculation, the disutility of work must be smaller as  $R$  increases.

## 6. Explaining Berg, Dickhaut, and McCabe's Data

In BDM's experiment,  $v_A(s^*, t^*)$  and  $v_B(s^*, t^*)$  are both  $e$ , the initial endowment. So  $k_{AB}(i, r) = [e + 3i - 3ri - e] / [4e - 3re - e] = i/e$ , and  $k_{BA}(i, r) = r - 1/3$ .  $u_B(i, r) = e + 3i - 3ri + \alpha(i/e)(r + 2/3)$ .  $u_B$  is thus an increasing function of  $r$  if  $\alpha > 3e$ .  $u_A(i, r) = e - i + 3ri + \alpha(r - 1/3)(1 + i/e)$ , which is increasing for any positive value of  $\alpha$  provided that  $r > 1/3$ . So one reaches the plausible conclusion that a fairness equilibrium is possible, but only if the responders place a sufficient weight on fairness.

This model does not, however, explain BDM's data, because it predicts that investors will either invest nothing if  $r < 1/3$  or their whole endowment if  $r > 1/3$ , and that responders will either return nothing or the entire amount they are given. Regardless of their risk aversion or the weight they place on fairness, investors who assess gains and losses symmetrically should always invest either nothing or their whole endowment.<sup>11</sup> So it seems that the failure of the prediction that investors will invest either nothing or their whole endowment does not reflect any deficiency within this account of fairness itself. I conjecture that one should explain why subjects invest part of their endowment by supposing that they

---

<sup>11</sup> $u(i) = u(e - i + 3ri)$ , so  $u' > 0$  for all values of  $i$  if  $r > 1/3$ , and  $u' < 0$  for all values of  $i$  if  $r < 1/3$ .

“anchor” on their endowment of  $e$  and evaluate losses and gains from this base line asymmetrically.<sup>12</sup>

Behavior of this kind is well established in the experimental literature. So let us suppose that the material utility of monetary gains is a linear function of the monetary gains, while the material utility of losses is more than a linear function of the monetary losses. Investors will maximize their net utility when  $v'_A(i) = 3r_e + (\alpha/e)(r_e - 1/3)$ , where  $v_A(i)$  is the appropriate material utility function when an agent is considering losses. Provided that  $v'_A(i) > 1$ , the utility-maximizing investment will be positive only for an expected value of  $r$  greater than  $1/3$ . The maximizing amount to invest will increase when the weight placed on fairness,  $\alpha$ , and the expected value of  $r$  increase. For the purposes of a simple illustration, suppose that  $v_A(i) = 3i^2/e$ . Then an investor maximizes utility by investing  $i = re/2 + \alpha r/6 - \alpha/18$ . If we suppose that investors weight fairness as heavily as responders so that  $\alpha = 3e$ , then the utility-maximizing amount of investment is  $e(r - 1/6)$ . Even for low weights on fairness and low expectations of return, people should still invest something. For example, if  $r = 4/9$ , and  $\alpha = e/9$ , the utility-maximizing investment is between a fourth and a fifth of the endowment.

In this way one might explain why most investors invest something, but typically only a portion of their endowment. To explain why responders send back only a portion of their windfall and why they send back a larger percentage when more is invested requires a modification of this account of fairness. Most people would judge that responders are behaving unfairly if they return less than a third of the windfall, but that they have no duty to return more than  $2/3$  of it.  $r = 2/3$  is a salient point, because it splits the gain evenly between  $A$  and  $B$ .<sup>13</sup> In BDM's experiment,  $r > 2/3$  is very rare, and it never occurs in the experiments of MRR. For all values of  $r$ ,  $B$ 's *benevolence* increases with  $r$ , but is this a case where reciprocation of kindness by kindness constitutes fairness? If  $A$  and  $B$  are both maximally kind (as measured by  $k'$ ), all the gain goes to  $A$ . Is that fair, as a proportional-reciprocation-of-benefits theory of

---

<sup>12</sup>MRR in contrast explain the data by setting the equitable investment and rate of return as some fraction respectively of total endowment and one. This enables one to derive a fairness equilibrium for some investment between 0 and  $e$ , but it seems to me ad hoc. MRR give no argument why fairness should be calculated their way.

<sup>13</sup>One sees this salience in the data, which is striking given that it takes some sophistication to see that  $r = 2/3$  leads to an even division. In MRR's replication,  $2/3$  is the modal value of  $r$ , and in the experiment that followed involving “cheap talk”  $2/3$  was by far the most common return offered.

fairness implies? Moreover, responders might be motivated by concerns about *trust*, rather than concerns about fairness, and most people would say that a fully *trustworthy* person would return only 2/3 of the gain to *A*.

To accommodate these facts, one might incorporate conventions into the analysis. For example, let  $v_A^c(i)$  be the material utility of the conventionally accepted kindest reply ( $r = 2/3$  in this case). Then let  $k'_{BA}(i, r) = (v_A(i, r) - v_A(s^*, t^*)) / [v_A^c(i) - v^{\min}(t)]$ , if  $v_A(i, r) < v_A^c(i)$  or 1 otherwise.  $u_B$  will then diminish for  $r > 2/3$ . This seems reasonable, but less informative than one might wish, since it relies on the conventional (and unexplained) kindest reply. Alternatively, one might attempt to explain *B*'s behavior not merely in terms of *B*'s concern to be fair, but also in terms of *B*'s concern to be *trustworthy*. I will explore this alternative in section 8. But perhaps instead it is time to jettison the notion of fairness as proportional reciprocation of benefits.

## 7. Fairness as a just division of the surplus

Even though an *A* player who invests his or her entire endowment is being as kind to a *B* player as it is possible to be, there is nothing at all unfair about the *B* player returning only 2/3 of the windfall and thus not being maximally kind in return. Furthermore, one might question whether “investing” one’s whole endowment is, after all, (intuitively) “kind.” Regardless of whether *A* players place any weight on fairness, they should invest only if they believe that  $r > 1/3$ . But then they might as well be self-interested. On the other hand, the revised apparatus does seem to fit a good deal of behavior.

I suggest that players take fairness to be a matter of justly sharing the benefits or costs (compared to the materially self-interested equilibrium) rather than a matter of proportional reciprocation of benefits. One can take the players to be playing an implicit bargaining game and to be adopting an analogue to the Nash bargaining solution (with the materially self-interested equilibrium as an analogue to the non-agreement point). Insofar as they are concerned with fairness, the agents seek to maximize the product of their shares of the surplus -- that is, of their gains as compared to their payoffs in the material

self-interested equilibrium. Some sort of normalizing is necessary, however, or else fairness will play a disproportionately large role in determining  $A$ 's preferences when  $A$  has much less to gain from cooperation than  $B$ . These considerations apparently suggest that the  $k'$  measure ought to work. Apart from the fact that the revision of Rabin's account takes  $u_A$  to depend on  $k'_{AB} \cdot (1 + k'_{BA})$ , rather than  $k'_{AB} \cdot k'_{BA}$ , which is not what gives rise to the problems, the revised version of Rabin's theory seems equivalent to such an analogue to a Nash bargaining solution.

The solution to the puzzle lies in recognizing that Rabin's normalization is inappropriate in this context. The numerators of  $k'_{AB}$  and  $k'_{BA}$  are  $A$ 's and  $B$ 's shares of the surplus, but the denominators introduce factors that are irrelevant to whether these shares are fairly distributed. One needs some other normalization. One possibility is to compare the shares of the surplus (the difference between the actual payoffs and the payoffs of the materially self-interested equilibrium) to the greatest surplus a player could achieve. (Section 9 below suggests another possibility.) Instead of normalizing by taking  $[v_A^h(s) - v_A^{\min}(s)]$  as the denominator, which depends on the particular strategy,  $s$ ,  $A$  chooses, one should normalize by dividing each player's surplus by the player's maximum surplus -- that is, by the difference between the largest amount the player  $A$  could possibly get from *any* set of strategies,  $v_A^{\max}$ , and  $v_A(s^*, t^*)$ . The "fairness factor" in each individual's utility will be the product of (the expectation of)  $k''_{BA} = [v_A(s, t) - v_A(s^*, t^*)]/[v_A^{\max} - v_A(s^*, t^*)]$  (or 0, if the denominator is zero) and  $k''_{AB} = [v_B(s, t) - v_B(s^*, t^*)]/[v_B^{\max} - v_B(s^*, t^*)]$ , weighted by a factor reflecting how much each individual values fairness.  $k''$  differs from  $k'$  only with respect to the normalization. In this example,  $v_A^{\max} = 3e$ ,  $v_A(s^*, t^*) = e$ ,  $v_B^{\max} = 4e$ , and  $v_B(s^*, t^*) = e$ . So the fairness factor is  $[\frac{1}{2}(i/e)^2(1 - r) \cdot (r - 1/3)]$ , which is positive for  $r > 1/3$  and which reaches a maximum when  $i = 1$  and  $r = 2/3$ . There can be no fairness equilibrium with  $r > 2/3$ , because  $B$ 's utility diminishes as  $r$  increases past  $2/3$ .  $u_B$  reaches a maximum when  $r = 2/3 - e^2/\alpha i$ . This construal of fairness explains why  $r$  is an increasing function of  $i$ , and why  $r > 2/3$  is not found.  $r$  should be less than  $2/3$  if  $B$ 's choice represents a tradeoff between fairness (which is maximized at  $r = 2/3$ ) and self-interest. This model does not explain why a considerable minority of responders return a full  $2/3$  of their windfall. I

shall have more to say about this phenomenon in the next section.

This revision still captures Rabin's successes, although at some cost in algebraic complications.

In the case of the workers and firms, for example,  $u_F(H, b) = (R - b)^{1/2} + \alpha(R - b)^{1/2} \cdot (b^{1/2} - d) / [R^{1/2} \cdot (R^{1/2} - d)]$ .

This reaches a maximum when  $2\alpha b + (R - dR^{1/2} - \alpha d)b^{1/2} - \alpha R = 0$  and has no simple analytic solution. Let

$\alpha = (R - dR^{1/2})/d$ , which assumes that the weight firms and workers place on fairness is reasonably high

and that the disutility of work is reasonably low. Then  $u_F$  reaches a maximum when  $b = 1/2R$ . For this

value of  $b$ ,  $u_w(H, b) > u_w(L, b)$  iff  $-d + R/d - (1/2R)^{1/2} > 0$ , which will hold when  $d < b^{1/2}$ . For a constant

value of  $\alpha$  (rather than a value proportional to  $R$  and inversely proportional to  $d$ ),  $d$  must diminish as  $R$

increases.

Look back now to games 1-3. In game one,  $v_R(s^*, t^*) = 3$ ;  $v_R^{\max} = 5$ ;  $v_C(s^*, t^*) = 0$ ; and  $v_C^{\max} = 12$ . So  $u_C(B, R^*) = 5 + \alpha(5/12)(1)$ , which is greater than  $u_C(B, L)$  whenever  $\alpha > 12/5$ . As before, there is a fairness equilibrium. In game two,  $v_R(s^*, t^*) = 3$ ;  $v_R^{\max} = 5$ ;  $v_C(s^*, t^*) = 12$ ; and  $v_C^{\max} = 12$ . Since  $u_C(B, R^*) = 5 - \alpha(7/12)(1)$ ,  $u_C(B, R^*) < u_C(B, L)$ , and there is no fairness equilibrium. In game 3,  $v_R(s^*, t^*) = 2$ ;  $v_R^{\max} = 5$ ;  $v_C(s^*, t^*) = 6$ ; and  $v_C^{\max} = 12$ .  $u_C(B, R^*) = 5 - \alpha/6$ , and once again there is no fairness equilibrium.

## 8. Kindness and trust

Consider game 4:

<i>game 4</i>	left	right
top	\$3,0	\$3,0
middle	\$0,\$12	\$0,\$12
bottom	\$-5,\$6	\$5,\$5

Row is, I would suggest, no *kinder* in playing bottom in game 4 than in game 1, but Row is taking a much greater risk. As in game 1, if Row believes that Column will play R or R\*, then Row is acting in his or

her self-interest. But in choosing bottom in this game, Row is taking a bigger risk and is placing more trust in Column. One would judge a Column player who played left in response to bottom more harshly in game 4 than in game 1, because Column would be betraying this greater *trust*. If these normative factors matter to people and one models how they matter by their influence on preferences, then trustworthy column players should more strongly prefer to play  $R^*$  in game 4 than in game 1.<sup>14</sup>

Considerations concerning trust should be kept separate from considerations concerning reciprocation of kindness, but they should be included, too. There are at least two ways of thinking about the trust shown by a first-mover, such as Row. The simplest involves looking at the lowest payoff from playing strategy  $s$  compared to the lowest payoff of  $s^*$ , the materially self-interested equilibrium strategy --  $v_A^{\min}(s^*) - v_A^{\min}(s)$ . This is a paranoid way of assessing how great a risk one is taking, because the responder(s) would have to make a stupid or malevolent sacrifice in order to secure an outcome for  $A$  with a material utility below  $v_A^l$ . So one might instead compare  $v_A^l(s^*)$  and  $v_A^l(s)$ . In either case, one can normalize by dividing by  $v_A^h(s) - v_A^{\min}(s)$ .

In the laboratory, one should probably discount the possibility of suboptimal responses and employ

the second measure, but whether this is correct depends on how subjects think about their interactions. One could also take the measure of trust to be an average of these two measures. I will define  $T_{AB}$ , the extent to which  $A$  trusts  $B$ , as  $[v_A^l(s^*) - v_A^l(s)]/[v_A^h(s) - v_A^{\min}(s)]$  or zero if the denominator is zero or the numerator is negative. I am not sure what one should say about  $T_{AB}$  if  $v_A(s^*, t^*) > v_A^h(s)$ . Probably it is best to regard it as zero in that case, too.  $T_{AB}$  will range between 0 and 1.

Although trusting may have its own rewards, in this construction, it is (other things being equal) undesirable to trust others. Trusting necessarily involves taking greater risks. These may nevertheless be worth taking if others are *trustworthy* – that is, if they will fulfill the trust placed in them. Indeed it may be advantageous to trust even if the best result from trusting is no better than the best result from

---

<sup>14</sup>I do not know of any experiments that investigate whether players in fact play right more often in response to bottom in games like 4 than in games like 1.

choosing a safer alternative, if the trust that is placed in others motivates them to fulfill that trust.

Obviously, something needs to be said about trustworthiness. One could measure how trustworthy  $B$ 's strategy  $t$  is as a response to  $A$ 's strategy  $s$  by comparing  $v_A(s, t)$  to  $v_A^{\min}(s)$ , or one could compare  $v_A(s, t)$  to  $v_A^l(s)$ . Since (pending further research), it seems reasonable in an experimental context to discount the risk that the other player will be malevolent, one might tentatively measure the extent of trustworthiness by comparing  $v_A(s, t)$  to  $v_A^l(s)$ . But, as we have seen, complete trustworthiness need not mandate that  $B$  provide  $A$  with the largest possible payoff. When I trust you with my wallet, I don't expect you to return it with the addition of all the money you have in your pockets, and I don't count your action as less than maximally trustworthy if you don't. There seems to be no reasonable way to avoid invoking norms and context-dependent expectations. Let  $v_A^w(s)$  be the utility payoff to  $A$  from the "trustworthy" response to  $s$  ( $r = 2/3$  in the case of BDM's experiment). One might then propose the following measure of trustworthiness.  $w_{BA}(s, t) = [v_A(s, t) - v_A^l(s)]/[v_A^w(s) - v_A^{\min}(s)]$  for  $v_A(s, t) < v_A^w(s)$  and 1 otherwise. This measure of trustworthiness ranges between  $-1$  (when  $v_A(s, t) = v_A^{\min}(s)$  and  $v_A^l(s) = v_A^w(s)$ ) and 1.

One cannot, however, take the contribution of trust and trustworthiness to preference to be a weighted linear function of the product of  $T_{AB}$ , the extent to which  $A$  trusts  $B$ , and  $w_{BA}$ , the extent of  $B$ 's trustworthiness with respect to  $A$ , because there is a very great difference between someone who returns my wallet with everything still in it and someone who returns it with only one dollar missing, while there is comparatively little difference between the trustworthiness of someone who returns the wallet with one dollar missing and someone who returns it with two dollars missing. Furthermore, someone can fail to be trustworthy, even if  $v_A(s, t) \geq v_A^w(s)$ . A friend who carelessly fails to water my plants while I am on vacation is untrustworthy, even if he or she replaces them with new plants that I prefer to the old ones. Although this second factor may not arise in many economic experiments, it can be accommodated in the same way as the first factor.

Let  $\Psi(s)$  be the set of all "fully trustworthy responses to  $s$ " (so that  $v_A(s, t) \geq v_A^w(s)$  for all  $t$  in

$\Psi(s)$ ). To model the effect of trust and trustworthiness on preference requires that one add *two* factors to the utility of  $B$  (a player who has been trusted). The effect of trustworthiness on  $B$ 's utility will depend on two terms,  $\theta X T_{AB}(s)$  and  $\gamma T_{AB}(s) \cdot w_{BA}(s, t)$  where  $X$  is a dummy variable equal to 1 if  $B$ 's strategy is fully trustworthy and -1 otherwise, and  $\theta$  and  $\gamma$  are weights placed on these two components of trustworthiness. Notice that in this formulation  $B$  has a stronger motive to choose a trustworthy strategy when  $A$  is more trusting. If one pretends that the different factors influencing preferences are additive, then a player's utility will be a sum of material utility, utility due to fairness as discussed above, and utility due to trustworthiness. Utility due to trustworthiness has two components, one which depends on the extent of the trust placed in a player and on whether or not the player's strategy completely fulfills the trust and one which depends on both the extent of the trust and the extent to which the player fulfills the trust. The first component in trustworthiness may explain why an appreciable number of people in BDM's and MRR's experiments return 2/3 of the windfall they receive to the investor.<sup>15</sup>

These measures capture the difference between games 1 and 4. In game 4  $T_{RC}(T) = 0$ , and  $T_{RC}(B) = 8/10$ , while  $w_{BA}(B, L) = 0$  and  $w_{BA}(B, R^*) = 1$ . In game 1, in contrast,  $T_{RC}(B) = 1/3$ . The influence of trustworthiness on preference is much stronger in game 4. I have defined trust and trustworthiness only for games in which the trustor plays before the trustee, but there is no reason why the definitions cannot be extended to simultaneous play games in which each of the players is simultaneously trustor and trustee.

## 9. Still More Difficulties

---

<sup>15</sup>MRR have provided some interesting evidence against this conjecture and in support of Rabin's view of fairness as reciprocation of kindness. MRR added an initial step to BDM's experiment, permitting each  $B$  player (each responder) to send a non-binding message to the investor ( $A$  player) stating what percentage of the windfall the  $B$  player will return. The  $A$  players then make their investment and the experiment continues exactly as BDM's experiment. Although the effect was not large, MRR found that the cheap talk led to increased investment and *diminished* returns. If trustworthiness were what motivated  $B$  players to share the windfall they receive, then the "cheap talk," the non-binding initial promise, should make  $B$  players *more* likely to make the return they promised to make (which was, incidentally, most often 2/3 of the windfall). If, on the other hand, some notion of "repaying kindness" is what motivates  $B$  players, then one can explain these results by pointing out that, given the initial promise of a return, the investment may appear more self-interested and less "kind." Since the effect MRR found was small and can be explained in other ways, too, I would not place too much weight on this argument.

The alternative measures of kindness and the measures of trust do not avoid a further difficulty to which Rabin's construction is subject. Consider the following four game forms, in which, once again, Row plays first and Column knows how Row has played before moving.

<i>game a</i>	left	right	<i>game b</i>	left	right
up	\$3,\$2	\$1,\$3		\$3,\$2	\$-3,\$3
down	\$2,\$0	\$2,\$1		\$2,\$0	\$2,\$1
<i>game c</i>	left	right	<i>game d</i>	left	right
up	\$3,\$2	\$1,\$3		\$3,\$2	\$-3,\$3
down	\$2,\$-3	\$2,\$-2		\$2,\$-3	\$2,\$-2

In each game, Row has only two pure strategies, Up and Down (U and D), while Column has four. I shall once again consider only two of these, R (right) and L\* (left if Row plays up and right if Row plays down). In all four games  $k_{RC}(U, L^*) = k_{CR}(U, L^*) = 1/2$ . Yet intuitively, playing up is kinder in games *c* and *d* than in games *a* and *b*, and row is more trusting in games *b* and *d* than in games *a* and *c*. The reason why Rabin's account fails to capture these differences lies in the fact that the largest and smallest payoffs influence both  $v^*$  and the denominator. Since the measures of kindness are normalized separately in each game, the fact that choosing up benefits Column so much more in games *c* and *d* than in games *a* and *b* is divided out. If one wants to use Rabin's account to measure the extent of kindness of moves in different games or the extent of trustworthiness, one could employ as a common denominator the largest denominator in the games being compared. If one plays this trick, the kindness (calculated in Rabin's way) of playing down in all the games remains zero, and  $k_{RC}(U, L^*)$  in games *c* and *d* remains  $1/2$ .  $k_{RC}(U, L^*)$  in games *a* and *b*, in contrast, is now only  $1/5$ . Scaling across games is, however, likely to be extremely arbitrary.

Alternatively, and more plausibly, one might normalize by taking as the denominator (the

expected value of) the material utility of the wealth of the relevant player. How great a risk someone is taking, how generous someone is being, and how much sacrifice someone is making all depend on the difference their move makes compared to their own wealth or the wealth of the recipient. Normalizing with respect to upper and lower bounds within the game does not correctly capture kindness, sacrifice, or trust. In economic experiments, subjects are usually unknown to one another, and consequently they do not know one another's wealth. Nevertheless, they can make estimates. In many experiments, for example, subjects know that the other subjects are students, who are presumably not rich. The fact that they are willing to participate in the experiment in exchange for a moderate payment reinforces this presumption. The numerator of Rabin's kindness measure for playing  $U$  against  $L^*$  will be four times larger in games  $c$  and  $d$  than in games  $a$  and  $b$ .

The alternative measures of kindness proposed here show the same features as Rabin's measure.  $k'_{RC}(U, L^*) = 1$  in all four games, while  $k'_{CR}(U, L^*) = 1/2$  in games  $a$  and  $c$  and  $1/6$  in games  $b$  and  $d$ . The fairness factors,  $k''$  (which are 1 and  $1/2$ ) are also the same in all four games. So  $k'$  and  $k''$  are subject to the same difficulties, and the cures are the ones I just suggested for Rabin's measure. The measures of trust already discriminate games  $b$  and  $d$ , in which Row is intuitively more trusting, from games  $a$  and  $c$ .  $T_{RC}(U) = 1/2$  in games  $a$  and  $c$ , and  $5/6$  in games  $b$  and  $d$ . The trustworthiness of  $L^*$  is 1 in all four games.

## 10. Conclusions

Where do all these complexities lead? The most ambitious (but not necessarily the most reasonable) conclusion would be to suppose that theorists now have a general schema to be used to infer preferences from features of game forms. If one takes altruists to prefer that the material utility of others be larger, one can take the preferences of the subjects who play two-person games in experimental economics laboratories to have the following form:  $u_A(s, t) = v_A(s, t) + \beta_A v_B(s, t) + \alpha_A k''_{BA} \cdot k''_{AB} + \theta_A X T_{BA} + \gamma_A w_{AB} \cdot T_{BA}$  (where  $X$  is 1 or -1 depending on whether  $A$ 's strategy is completely trustworthy). If one pursues the proposal to normalize with respect to the utility of wealth instead, then the formula is even

nastier to write out. These forms generalize in an obvious way to  $n$ -person games.

This construal of the argument of this essay faces an obvious difficulty. Since the equations depend on four different parameters, which will not be the same for different people and different contexts, one wonders whether the algebra in this paper serves any purpose. Will a theory that asserts that people's preferences satisfy this relation have any content? Are not game theorists better off with the false but contentful view that people's preferences depend only on their own material payoffs?

If one could make no generalizations about the typical relative size of  $\alpha$ ,  $\beta$ ,  $\gamma$ , and  $\theta$ , then this modeling would serve only as a (hopefully useful) reminder of problems that cannot be solved. It is an empirical question whether these parameters have reasonably stable distributions and whether what economists can find out about their distributions will be of predictive value. Notice in particular that the distribution of these parameters can be investigated separately. For example, one can study the size of  $\beta$  by examining how experimental subjects would choose between pairs of payoffs for themselves and others, because such choices should not be influenced by trust or fairness. Given knowledge of  $\beta$ , one can study the effect of fairness by comparing how people play games such as game 1 with how they play games such as game 2, where there is no difference with respect to trust. Given knowledge of  $\beta$ , one can study the effect of trust separately by comparing games  $a$  and  $b$  or  $c$  and  $d$ , which are equivalent with respect to fairness though they differ with respect to trust.

The complexities involved in incorporating fairness and trust into even such extremely simple interactions are daunting, and one cannot be optimistic about the success of a framework like the one sketched in this paper. However complicated they may be, the phenomena treated here are real and important, but their importance is, of course, no guarantee that satisfactory theories of them can be created.

This thought leads to a more pessimistic conclusion. Rather than regarding the inquiry in this paper as providing the foundation for a general theory of preference, one can take it as highlighting features that invalidate a simple inference from payoffs to preferences and thereby as highlighting

difficulties in applying game theory to human interactions, even within the controlled circumstances of experimental economics. Understanding the factors besides dollar payoffs that influence preferences may help experimenters improve their controls, and it may help defuse jejune “refutations” of game theory. But it is often very hard to know people’s preferences, and consequently it is very hard to know what games they are playing.

#### Acknowledgments

I am indebted to Matthew Rabin, Richard Bradley, and audiences at the London School of Economics, the University of East Anglia, and the University of Paris (VII) for helpful criticisms.

## References

Berg, Joyce, John Dickhaut, and Kevin McCabe: (1995), "Trust, Reciprocity, and Social History."

*Games and Economic Behavior* 10: 122-42.

Geanakoplos, John, Pearce, David and Stacchetti, Ennio: (1989), "Psychological Games and Sequential

Rationality." *Games and Economic Behavior* 1: 60-79.

Meidinger, Claude, Stéphane Robin, and Bernard Ruffieux (unpublished) "Confiance, Réciprocité et

"cheap talk."

Rabin, Matthew: (1993), "Incorporating Fairness into Game Theory and Economics." *American*

*Economic Review* 83: 1281-1302.

University of Wisconsin-Madison

Madison, WI 53706

U.S.A.

tel. (608) 263-3700 fax (608) 265-3701

email: dhausman@facstaff.wisc.edu