

Incorporating Fairness Into Game Theory and Economics: Comment

by

Tanga McDaniel

E. Elisabet Rutström

Melonie Williams*

March 1994

* University of South Carolina. We are grateful to Glenn W. Harrison for useful comments, and to Clayton Lesley and Tara Nuss for assistance. Funding was provided by the Dewey H. Johnson Chair in Economics at the University of South Carolina.

In "Incorporating Fairness into Game Theory and Economics" in this *Review*, Rabin [1993] proposes a useful distinction between behavior based on altruism and behavior based on fairness. The distinction rests on whether deviations from individually rational behavior, defined within the narrow confines of material payoffs in the immediate game, are *conditional* on beliefs about other agents. "If somebody is being nice to you, fairness dictates that you be nice to him. If somebody is being mean to you, fairness allows - and vindictiveness dictates - that you be mean to him" (p 1281). Pure altruism, on the other hand, would imply unconditionally nice behavior, and reciprocal altruism would only imply conditionally *nice* behavior, not conditionally *mean* behavior.¹

In this comment we show how incomplete information weakens the case for fairness and strengthens the case for individual rationality in many games. We also show how a *general* model of fairness that incorporates low and zero subjective fairness values can be difficult to distinguish from random play. We do this based on predictions and experimental tests of the game of Chicken.

Rabin presents three stylized facts from the experimental economics literature in support of this type of *fair* behavior. These stylized facts are that people cooperate to a greater degree than would be implied by narrow self-interest (in public goods experiments), that people will sacrifice to hurt others who are being unfair (in ultimatum bargaining experiments), and that people are willing to sacrifice less in the name of fairness when the material stakes are high (see Leventhal and Anderson [1970]).² Although these stylized facts have *inspired* the theory of fairness, they cannot be viewed

¹ The idea that outcomes can be rationalized conditional on beliefs is not new. The novelty rather lies in the combination of conditionality and subjective altruistic valuations.

² Although the first stylized fact could be support for either pure or reciprocal altruism, as well as for fairness, the second fact could not. However, this second stylized fact is based on experiments where the respondent knows the proposer's offer with certainty, and where beliefs about other respondents actions are also potentially important due to the common knowledge aspect of these experiments. Indeed, Harrison and McCabe [1992] find that when manipulating beliefs about other respondents actions in the experimental lab (by employing computerized players), fewer respondents sacrifice to hurt proposers that are being unfair, which is not consistent with the fairness equilibrium model by Rabin. This type of behavior on behalf of respondents is instead consistent with group reputation effects, as discussed by Andreoni and Miller [1993].

as a *test* of the hypothesis that behavior is conditional of beliefs in the way proposed.

A natural way to directly test these hypotheses is by means of controlled laboratory experiments, where it is possible to collect information not only on actions but also on beliefs. We propose, and execute, such a design below.

Rabin suggests three games where Fairness Equilibria (FE) can be sharply distinguished from Nash Equilibria (NE): Prisoner's Dilemma (PD), Battle-of-the-Sexes (BOS), and Chicken. We focus on experiments using the game of Chicken.³

The game of Chicken is the most interesting game to study experimentally, as fairness play could *rule out* strict NE if everyone is fair in some strong sense⁴. However, as is the case in the BOS game, the existence of multiple equilibria for this game will imply that observations of actions alone can neither distinguish between NE and fair actions when beliefs do not match actions, nor between FE and individually rational (IR) actions when beliefs do not match actions. To properly test the theory of fairness requires a design that elicits observations of actions *and* beliefs.

Rabin's approach is based on conditions of complete information, leading him to compare FE to NE, rather than to the less restrictive Rationalizable Equilibrium (RE) based on conditions of incomplete information. We show that support for fairness and altruism is considerably weaker under the latter conditions. This result is general in the sense of applying to *all* games that have no dominant strategies, and also to many games that do.

Rabin's discussion of fairness treats the hypotheses of IR and fairness as mutually exclusive.

³ In the PD game it is only conditionally kind behavior that can be distinguished from NE. However, such behavior cannot readily be distinguished from reciprocal altruism. Andreoni and Miller [1993] test the hypothesis of reciprocal altruism and finds support for it both in repeated and one-shot PD games. In the BOS game Rabin identifies potential FE in addition to the ones that are also NE. The FE that are distinguishable from NE are all based on behavior that punishes when the opponent is believed to be unkind. There are no FE based on conditionally nice behavior in this game. More importantly, for our purposes, when allowing for incomplete information conditions there are no strategies that distinguishes fair play from individually rational play.

⁴ If all players are strongly fair then material payoffs do not dominate psychological payoffs.

We believe that it is more natural to think of the fairness model as more *general* than the rationality model or the altruistic model, since the latter two emerge as special cases of the former. The test is then to see if fairness *adds* explanatory power to the other models. However, such a general model will often be consistent with play of a large proportion of strategies. If, in addition, the strategies that are consistent *only* with strong fairness are played with low frequency, fairness might be indistinguishable from random play. We encounter this problem in the Chicken game.

In the following we introduce an experimental design for testing behavior *conditional* on beliefs. We will then analyse our experimental results from the Chicken game, first based on IR and fairness as mutually exclusive hypotheses and with complete information beliefs. We then compare this analysis to one based on incomplete information beliefs, and finally discuss our results based on fairness as a general model, allowing both fair, altruistic, and rational players in the population.

Experimental Design

Controlled laboratory experiments are uniquely suited to test for the subtle differences in behavior between hypotheses suggesting rational, altruistic, and fair behavior. The important distinction between these hypotheses *behaviorally* does not lie with the equilibrium concepts but rather with the consistency between actions and beliefs. The fact that the Chicken game has multiple NE implies a coordination problem that *ex post* will lead to outcomes that might *appear* to be fair or altruistic in action space even though players are individually rational, but that are inconsistent with the belief structure of fair or altruistic behavior.

We therefore propose a design that collects data not only on actions, but also on beliefs. Our proposed design is salient in that subjects are rewarded when the reported beliefs are confirmed by

the other player's actions, as explained below.

Row Player's Payoffs			Column Player's Payoffs		
	UP (chicken)	DOWN (dare)		UP (chicken)	DOWN (dare)
UP (chicken)	0.75	0.50	UP (chicken)	0.75	1.00
DOWN (dare)	1.00	0.00	DOWN (dare)	0.50	0.00

Figure 1: Payoff matrix for the experimental game

Figure 1 presents the payoff matrix for the experimental game. We labelled the strategies up and down, rather than chicken and dare, so as not to bias subject behavior. Up therefore corresponds to chicken and down to dare. Following Figure 4 in Rabin we assume a value of X equal to 25 cents, but we also transform our payoff matrix to avoid negative entries.⁵ We believe that this value of X is small enough not to prohibit potentially fair behavior. Indeed, this value for X makes the foregone profits for players acting in a fair manner no larger than ones reported in the experimental outcomes cited by Rabin.

Subjects were also paid according to how well they could predict an opponent's choices. They were asked to report both their beliefs about the opponent's choice of actions, and their beliefs about the opponent's predictions of their own actions. Subjects were paid an additional 25 cents for each belief that was confirmed. As this payment does not involve any strategic interaction between subjects it does not change the predicted equilibria or pattern of play based on either rationality, fairness, or altruism.

⁵ This transformation was employed in order to maintain the credibility in the experiment, as (dare, dare) could not be sustained in the experiment without subjects having to pay the experimenter at the end. Although there is nothing wrong with allowing this possibility in an experimental design, it invites credibility problems. Some subjects will simply not believe that they really have to pay up at the end, unless the design in some way manage to enforce this payment. There is no generally accepted way of dealing with this problem, and to keep our design as simple as possible we decided to avoid the issue by simply transforming the payoffs. Alternatively, we could have opted for giving all subjects a sufficiently deep pocket to sustain (dare, dare) outcomes. This, however, would have increased the cost of the experiment dramatically.

Twenty subjects for the experiment were recruited in a variety of undergraduate classes in the College of Business Administration at the University of South Carolina. All interactions between subjects were handled by a computer so that anonymity of opponents could be maintained. No verbal interaction was allowed. Subjects were instructed that they would play the game ten times, each time with a new, randomly selected opponent. To facilitate their understanding of the game a set of written instructions was handed out and read to them.⁶ The instructions included some examples on how to read a payoff matrix, and information on the simultaneous move nature of the game. Before starting the ten periods for profit they played two periods without monetary payoffs to familiarize themselves with the computer program and the payoff matrix.

Predictions

Figure 2 illustrates the eight possible strategy choices and how they correspond to six

	{c,c,c}	{c,c,d}	{c,d,c}	{c,d,d}	{d,c,c}	{d,c,d}	{d,d,c}	{d,d,d}
NE			X			X		
RE	X	X	X		X	X		
AL $\hat{a} > \hat{a}^*$	X							
AL	X	X	X		X	X		
FE	X							X
FAIR	X	X	X	X	X	X	X	X

c = chicken, d = dare, NE - nash equilibrium, RE = rationalizable equilibrium, AL = altruism, FE = fairness equilibrium, FAIR = general fairness model, \hat{a} = altruism parameter, \hat{a}^* = critical altruism value

Figure 2: Correspondence between strategies and behavioral type

⁶ The instructions, as well as the data, are available on request.

behavioral hypotheses. The strategies are shown in columns as $\{x, y, z\}$ ⁷, where x denotes the action choice, y denotes the belief about the opponent's action, and z denotes the belief about the opponent's expectation of your own play. Each of these three can take on the value c for "chicken" or d for "dare".

First assume that all players are IR. Such players expect their opponents to be playing their constituent element of a NE. The problem for them in Chicken is that there are two NE in pure strategies (we will also focus solely on pure strategy predictions). So there is some strategic uncertainty as to what the opponent will do in terms of picking one or other NE.

Model this uncertainty in the form of a parameter p_c , the probability that the other player will play the chicken strategy. The payoff matrix for our basic game is shown in Figure 1. The expected value to an IR players from playing chicken is therefore a function of the expected value of p_c . Our experimental design gives us information on the latent variable p_c in the form of beliefs that the opponent will play chicken or dare. So we have to be careful in deriving our hypotheses to keep these two separate: the underlying latent probability which is presumed to motivate actions, and the observed binary counterpart.

It is easy to show that if p_c exceeds \mathbf{b} then the expected value to an IR player of dare will exceed the expected value of chicken. Hence we should expect to see the experimental outcome $\{d, c\}$ ⁸. In other words, this subject will state a belief that the other player will play chicken (the second entry in this predicted outcome) and will choose the action chicken for himself since this is the IR response (the first entry). Thus we would have the expected NE outcome, at least from the subjective

⁷ This notation differs from the notation in Rabin. x refers to the action choice a , y to belief b , and z to belief c .

⁸ For convenience we refer to the outcome $\{x, y\}$ when we do not restrict the third element of the strategy $\{x, y, z\}$. Thus, $\{d, c\}$ denotes either $\{d, c, c\}$ or $\{d, c, d\}$ in this case.

perspective of the player we are considering. What his opponent *actually* chooses as his action is another matter.

Similarly, for values of p_c less than \mathbf{b} and greater than $\frac{1}{2}$ the expected value to an IR player of chicken will exceed the expected value of dare. Hence we should expect to see the experimental outcome $\{c, c\}$. Note that this is not an *expected* NE from the perspective of the player being studied, although it is IR given these beliefs and it is a RE.

Finally, for values of p_c less than $\frac{1}{2}$ the expected value to an IR player of chicken will continue to exceed the expected value of dare. Hence we should expect to see the experimental outcome $\{c, d\}$.

Our RE hypothesis therefore predicts five choices: $\{c, c, c\}$, $\{c, c, d\}$, $\{c, d, c\}$, $\{d, c, c\}$, and $\{d, c, d\}$. The strategy $\{c, d, d\}$, cannot be consistent with RE as $x = d$ and $y = d$ is not predicted and it is not rational then to hold beliefs $y = d$ and $z = d$.

The complete information approach would eliminate the second possibility here, since it restricts attention to certain beliefs about opponent behavior. In other words, he assumes that the player either holds a belief that $p_c = 1$ or that $p_c = 0$. These are special cases of those considered above, and imply that we would only observe the outcomes $\{d, c\}$ or $\{c, d\}$. This can be seen as a restriction that is consistent with the NE way of modelling IR, but that violates the weaker RE way of modelling IR.

Now assume that players are *reciprocally* altruistic, in the specific sense that they enjoy extra non-material payoffs from playing the chicken strategy when the opponent is expected to do likewise. Use the parameter α to denote these extra payoffs. The relevant payoff matrix is shown in Table 1.

Table 1: Payoffs to reciprocally altruistic players

		Column	
		Chicken	Dare
Row	Chicken	$\frac{3}{4} + \alpha, \frac{3}{4} + \alpha$	$\frac{1}{2}, 1$
	Dare	$1, \frac{1}{2}$	$0, 0$

The only change from the IR case is the addition of the altruistic payoffs parameter to the top left cell. The issue here is whether the incorporation of positive values of this parameter adds any explanatory power in terms of our experiments.

If, however, players are *purely* altruistic, in the specific sense that they enjoy extra non-material payoffs from playing the chicken strategy irrespective of whether the other player does likewise. Again using the parameter α to denote these extra payoffs, the relevant payoff matrix is

Table 2: Payoffs to purely altruistic players

		Column	
		Chicken	Dare
Row	Chicken	$\frac{3}{4} + \alpha, \frac{3}{4} + \alpha$	$\frac{1}{2} + \alpha, 1$
	Dare	$1, \frac{1}{2} + \alpha$	$0, 0$

shown as Table 2. The only difference from the reciprocal altruism case is that the psychological payoff appears in all of the chicken cells for the player choosing chicken.

It can be shown that predictions about strategy choices and beliefs are the same for both pure and reciprocal altruism, and that they are also independent of information conditions. However, the critical value of \hat{a} that determines whether a player chooses chicken independent of beliefs, will not be the same across these cases. The critical \hat{a} level is not an operationally meaningful distinction, however, as we do not observe \hat{a} in our experimental design.

If all players in the population have an \hat{a} above the critical value we get the prediction of Figure 2 that everyone will play chicken. However, if we allow a more general distribution of \hat{a} values across the population, such that the IR model is a special case of altruism with $\hat{a} = 0$, we cannot distinguish the altruistic model from the individually rational model with incomplete information.

Finally let us examine predictions for fair players. If all players have kindness functions that dominate the influence from material payoffs on strategy choices we get the predictions of Figure 2 that everyone will play either $a = c$ based on beliefs ($b = c, c = c$), or $a = d$ based on beliefs ($b = d, c = d$). These correspond to FE strategies.

It can be shown that the critical level of kindness necessary to exhibit conditionally unkind behavior is higher than the critical level that is needed to exhibit conditionally kind behavior. If kindness can vary across players, such that some play altruistically and others individually rational (for low enough kindness values), then any strategy choice is consistent with the fairness model. First, it is possible for a strongly fair player to hold any beliefs about opponent's play, as they could be of any of the three behavioral types depending on the strength of their kindness function. Second, given any set of beliefs ($b = [c, d], c = [c, d]$) it is possible to see a player pick any action ($a = [c, d]$)

depending on the strength of their kindness function. Take the case of $\{c, d, d\}$ play, for instance. Beliefs $(b = d, c = d)$ are consistent with the player believing that at least 50% of players are strongly fair and exhibit conditionally unkind behavior. Play of $a = c$ conditional on belief $b = d$ is consistent with the player himself not valuing fairness (even though he recognizes that others do) and therefore playing IR.

The problem with testing this general fairness model for the Chicken game is that it exhausts the strategy space and therefore *any* pattern of play is consistent with fairness. A test of whether play of strategies other than those consistent with IR and altruism is based on fairness or random errors is not operational.

Results

Table 3: Raw data from experiment

period	{c,c,c}	{c,c,d}	{c,d,c}	{c,d,d}	{d,c,c}	{d,c,d}	{d,d,c}	{d,d,d}
3	8	4	2	0	2	4	0	0
4	6	0	0	6	4	4	0	0
5	12	0	0	2	4	0	0	2
6	4	0	0	4	8	4	0	0
7	12	0	2	2	4	0	0	0
8	10	0	2	0	2	2	2	2
9	8	0	2	0	4	2	2	2
10	8	0	2	0	4	0	6	0
11	12	0	2	0	0	0	4	2
12	12	4	0	0	0	0	4	0
subject								
1	2	1	1	0	6	0	0	0
2	4	0	1	1	0	3	0	1
3	7	0	0	1	1	1	0	0
4	1	2	0	4	0	1	1	1
5	5	0	0	1	0	2	2	0
6	7	1	0	0	0	2	0	0
7	2	1	1	1	1	2	2	0
8	10	0	0	0	0	0	0	0
9	6	0	0	0	2	1	1	0
10	3	1	1	0	4	1	0	0
11	6	1	1	0	0	0	1	1
12	2	0	2	0	5	0	1	0
13	5	0	0	1	2	0	2	0
14	8	0	1	0	1	0	0	0
15	2	0	2	0	2	2	1	1
16	1	0	0	1	3	1	3	1
17	5	1	1	1	1	0	0	1
18	4	0	0	0	2	0	4	0
19	6	0	0	2	1	0	0	1
20	6	0	1	1	1	0	0	1
total	92	8	12	14	32	16	18	8

Table 3 presents the raw data on behavior by subject and period in the experiment. The strategy choice $\{c, c, c\}$, corresponding to a player choosing the chicken action conditional on the

belief that the opponent does the same *and* that the opponent believes the player will also choose chicken, is by far the most popular. It is chosen 46% of the time. 11 out of 20 subjects played this strategy at least 50% of the time. There does not appear to be any strong trend in play over time.

Table 4: Behavior by type

Subject	NE	RE	AL $\hat{a} > \hat{a}^*$	FE
1	1	10	2	2
2	4	8	4	5
3	1	9	7	7
4	1	4	1	2
5	2	7	5	5
6	2	10	7	7
7	3	7	2	2
8	0	10	10	10
9	1	9	6	6
10	2	10	3	3
11	1	8	6	7
12	2	9	2	2
13	0	7	5	5
14	1	10	8	8
15	4	8	2	3
16	1	5	1	2
17	1	8	5	6
18	0	6	4	4
19	0	7	6	7
20	1	8	6	7
total	28	160	9	100

Table 4 presents the individual subjects' play classified by behavioral hypothesis: NE, RE, strongly altruistic, and FE. We know that the general fairness model is consistent with 100% of play in a trivial sense.

Comparing the frequency of play of NE and FE we find overwhelming support for fairness. 50% of play is consistent with FE and only 14% with NE. However, when we look at the less restrictive case of RE the story is quite different. 80% of play is consistent with RE. We can therefore conclude that results are very sensitive to assumptions about beliefs. Incomplete information strengthens the case for rationality.

Another important qualifier regarding the strength of fairness play is the fact that $\{c, c, c\}$ - the most popular strategy choice - is also consistent with an altruistic model. The only strategy choice that distinguishes fair players $\{d, d, d\}$ is played only 4% of the time, which is considerably less than the 12.5% expected if *all* play were completely random.

Classifying the subjects according to frequency of play, we find that 19 out of 20 subjects play RE at least 50% of the time, while only 12 play FE with that frequency. We observe 8 players play RE at least 90% of the time, a frequency that for FE only applies to one subject.

Conclusions

We find the distinction between fairness and altruism conceptually useful. We raise, however, some serious concerns about the ability of testing fairness theory. We point to two important assumptions that are pivotal when testing fairness vs. rationality. These are the assumptions of complete information beliefs and mutually exclusive behaviors. When we relax these assumptions we find that it is very difficult to confirm any systematic impact on behavior from subjective values based

on fairness. Our discussion has focussed on the game of Chicken, but many other games suffer from the same problems. Although this conclusion does not invalidate the fairness hypothesis, it does suggest that it has a limited operational significance.

REFERENCES

- Harrison, Glenn W., and McCabe, Kevin,** "Expectations and Fairness in Simple Bargaining Experiments", *Economics Working Paper B-92-10*, Division of Research, College of Business Administration, University of South Carolina, September 1992.
- Andreoni, James, and Miller, John H.,** "Rational Cooperation in the Finitely Repeated Prisoner's Dilemma: Experimental Evidence," *The Economic Journal*, May 1993, *103(418)*, 570-585.
- Leventhal, Gerard, and Anderson, David,** "Self-Interest and the Maintenance of Equity," *Journal of Personality and Social Psychology*, May 1970, *15*, 57-62.
- Rabin, Matthew,** "Incorporating Fairness into Game Theory and Economics", *American Economic Review*, December 1993, *83(5)*, 1281-1302.

APPENDIX (NOT FOR PUBLICATION)

TANGE WE NEED TO INSERT OUR INSTRUCTIONS HERE.