Incorporating Fairness into Game Theory and Economics: Comment

By William Robert Nelson, Jr.*

This note proposes one adjustment to Matthew Rabin's (1993) insightful model, which defines a "fairness equilibrium" and demonstrates some general implications of fairness on game theory and economics. One of the general implications of his model is "Proposition 5 states essentially that as material payoffs become large, the players' behavior is dominated by material self-interest. In particular, players will play only Nash equilibria if the scale of payoffs is large enough." As the utility function is written, he is correct. However, in the present article it is shown that if the relationship between material- and fairness-based utility is generalized, fairness becomes a normal good as stakes move from *very high* to *very very high.* When stakes are high enough and players' wealth is high regardless of the outcome of the game, agents demand more fairness and are willing to sacrifice material utility in its pursuit.

I. Rabin's Model

Throughout Rabin (1993), utility inputs are provided in two forms: material-based and fairness-dependent. These sum into total utility such that

$$V_i(a_i, b_i, c_i) \equiv v_i[\pi_i(a_i, b_j) + f_j^*(b_j, c) \cdot f_i(a_i, b_j)],$$

where V_i represents total utility and π is the utility from material benefits. The utility input derived from the fairness of the outcome is represented by the product of f_j^* and f_j, where the two represent how kind i thinks j is being to him and how kind i is actually being to j, respectively. Both f_j^* and f_i can be of either sign. If i expects j to act kindly, then f_j^* is positive; if i is kind to j, then f_i is positive. Therefore, the impact of fairness on i's total utility can also either be positive or negative. Fairness considerations increase total utility if the signs match. But if both f_j^* and f_i are negative, material losses usually outweigh psychic gains.

Intuitively, if i thinks j is going to act fairly toward him, he is more likely to act fairly in return. When both act fairly, both derive positive psychological utility from the exchange in addition to any material utility. But if i thinks j is going to act unfairly toward him, this will increase his incentive to act unfairly in return. Acting fairly toward a partner who is unfair reduces the actor's materially based and psychologically based utility. A player also receives psychological utility by acting selfishly toward one that acts selfishly toward him. The equilibrium where both agents act fairly is generally preferred.[1]

II. Augmentation of the Utility Function

According to Rabin's (1993) assumptions, the utility derived from material payoffs in a game monotonically increases along with the stakes, but the utility derived from the fairness of the outcome does not increase with the stakes of the game being played. Thus, as stakes increase, players' strategies are dominated by material outcome and the impact of fairness on the players' strategies diminishes. Accordingly, as the stakes increase, fairness and Nash equilibria become indistinguishable. But if the utility

* University at Buffalo School of Management, Jacobs Management Center, Buffalo, NY 14260 (e-mail: econrob@hotmail.com). I appreciate financial support from the Mercatus Center. I thank Bryan Caplan, Tyler Cowen, Rois Beal, and Erik Tallroth for their useful comments. I am grateful for helpful comments from the editor and the anonymous referees. Any errors are my own.

[1] Graciela Chichilnisky (1994) modeled bilateral international trade and found that if otherwise identical countries hold natural resources under different property-rights regimes (private versus common pool), the free-trade equilibrium is not always preferred to the no-trade equilibrium.

function is kept more general by not specifying how the material and fairness inputs interact while providing total utility, it is possible that, as payoffs become *very very large*, players' behavior will be dominated by fairness. Such a utility function is

$$V_i(a_i, b_i, c_i) \equiv v_i[\pi_i(a_i, b_j), f_j^*(b_j, c) \cdot f_i(a_i, b_j)].$$

The only difference between the two utility functions is the replacement of the "+" with a ",". This form allows for diminishing marginal returns of material-based utility on total utility. If the return to additional quantities of material-based utility becomes low as the stakes, and thus the wealth of players, rise, then fairness considerations may, at *very very high* stakes, dominate players material considerations.

The implication of this minor change can be demonstrated by taking the derivative of the new value function with respect to a_i.

$$\frac{\partial Vi}{\partial a_i} = \frac{\partial V_i}{\partial \pi_i}\frac{\partial \pi_i}{\partial a_i} + \frac{\partial V_i}{\partial f_i}\frac{\partial f_i}{\partial a_i} f_j(b_j, c_i) = 0$$

at the maximum. As π_i approaches infinity, V_i/π_i will approach zero, given diminishing marginal returns to material gains. This will in turn increase the relative importance of fairness to player one. Rather than predicting the importance of fairness monotonically decreasing in wealth, the adjusted model predicts a U shape in impact of fairness on players' strategies. If payoffs are *very very large*, players' strategies will be dominated by fairness.

It is possible that material utility is not a substitute for fairness utility, and thus their interaction is not characterized by the standard assumption of diminishing marginal rates of substitution. But the applicability is broadened and the predictions are more accurate if the model is generalized by including the possibility of diminishing returns to material wealth. Intuitively, if all material wants are fulfilled, players will begin to indulge their preference for fairness to eliminate some of the guilt they might suffer because of their good fortune.

III. Empirical Evidence

A combination of experimental ultimatum and dictator-game results and charitable-giving data is used to show how the augmented form of Rabin's (1993) value function is consistent with empirical evidence. The ultimatum game is played by two people (player one and player two) who are placed in separate rooms so that they remain anonymous throughout the experiment. Player one is told to propose the division of a sum of money (M), often ten dollars, between himself (O) and player two (T) so that $T = M - O$. Player two has the choice either of accepting or rejecting the offer. If player two accepts the offer, both players receive payment of the sum of money allocated to them by player one. If two rejects the offer, both receive zero. The subgame-perfect Nash equilibrium is for player one to offer player two the smallest available positive allocation possible and for two to accept it. This equilibrium is rarely observed. In the many ultimatum games that have been run with stakes in the $10 range, the modal offer is one-half. Offers of less than $5 are rejected frequently enough so that an even split is the wealth-maximizing offer in the United States (see Alvin E. Roth et al., 1991, for an international comparison). Apparently, with small stakes, fairness or some other-regarding behavior plays a significant role in determining subjects' strategies.

Several experiments have been run with high stakes. This has been tried in the United States by Elizabeth Hoffman et al. (1996), in which M equals $100. It has also been tried with subjects in countries where income is low and the dollar is strong so that M will equal as much as several months' wages. For the purpose of the present paper, these stakes are considered *high.* Convergence toward subgame-perfect equilibrium is observed, but offers remain significantly above zero and positive offers are regularly rejected. In Robert Slonim and Roth (1998), subjects in the Slovak Republic participated in ultimatum games with different partners, in which the stakes varied by a multiple of 25. Initially, the offers were similar at all stakes but rejections were rarer at high stakes. After playing the game several times, the offerers learned that lower offers in percentage terms were accepted at higher stakes. Then subjects' offers

converged toward this lower level—toward the subgame-perfect equilibrium. Lisa A. Cameron (1999) used Indonesian subjects in an ultimatum game in which M equaled three times the average participant's monthly expenditures. Her results are consistent with Slonim and Roth's (1998). Offers remained similar to those in small-stake experiments but rejections were less frequent for similar percentage splits. Both experimental results are consistent both with Rabin's (1993) original utility function and the one proposed.

Dictator-game results also deviate from subgame-perfect equilibrium. The dictator game is similar to the ultimatum game except player two does not have the option of rejecting the offer. Player one's division stands, regardless. Subgame-perfect equilibrium is thus for player one to keep all of the money and give none to player two. The typical distribution of the divisions is bimodal, with about 21 percent offering either zero or half (Robert Forsythe et al., 1988). The rest of the offers are scattered between zero and half.

The dictator game has only been played with small stakes in experimental settings, but it is constantly played with *very high* and *very very high* stakes outside of the laboratory. Essentially, anyone with any wealth is playing the dictatorship game at all times. There is always a person or cause willing to accept gifts. The income range across the United States provides a range of natural dictator games varying from *very high* to *very very high* stakes. The percentage of income given to charity is interpreted similarly to the percentage offered in laboratory dictator games. High percentages given to charity are interpreted as indicating that fairness has a large impact on behavior.

The empirical evidence provided by charitable-giving research is flawed for two reasons. First, charitable gifts are tax deductible in the United States. Accordingly, as the marginal income tax rate of a donor increases, the marginal cost of charitable giving decreases. A combined state and federal marginal income tax rate of 50 percent is not uncommon in the United States. Such rates reduce the after-tax cost of charitable giving by half for upper-income individuals, relative to participants in standard dictator games in which game participants' gifts are not tax deductible.

Charitable donations might purchase status, creating a second flaw in the use of charitable-giving data as a natural dictator game. Philanthropists buy permanent status and a reputation for being "fair." Note the praise Ted Turner received upon promising $1 billion to the United Nations. The impact of reputation purchases on philanthropy is especially relevant for *very very large* gifts, i.e., in games with *very very high* stakes. Both of these flaws in using charitable-giving data as a natural dictator game increase with the income of the philanthropist, causing the lower-income data to be more compelling, but should not cause the evidence provided by charitable-giving studies to be discredited.

Paul G. Schervish and John F. Havens (1995, 1998) used data from the 1989 Federal Reserve Survey of Consumer Finances and the 1990, 1992, and 1994 Giving and Volunteering Surveys and concluded the following about charitable-giving patterns in games with *very high* and *very very high* stakes.

1. The percentage of family income donated to charities changes little among those with incomes below $100,000.
2. The percentage of households contributing rises steadily.
3. There is no trend in the percentage of income given to charity by households with incomes below $100,000.
4. The percentage of income given to charity increases from 1.75 for all categories from $0–124,999 to a mean of 3.3 for those with incomes of $125,000 and above.
5. The highest percentage of income given to charity is 4.9 percent, which is donated by households with incomes above $1,000,000.

Schervish and Havens' (1998) results contradict Rabin's (1993) original value function at *very very high* stakes but are consistent with the augmented form of the value function proposed in the present paper. From low to *very high* stakes, the impact of fairness on actual equilibrium diminishes; if stakes are *very very high*, the importance of fairness increases. Rather than fairness having a monotonically decreasing impact on behavior as stakes increase, a U curve describes the data more accurately. Bill Gates

has endowed his charitable foundation with $21 billion, approximately 25 percent of his wealth, providing an extreme observation that is consistent only with the augmented utility function proposed here. The original utility function is consistent with Bill Gates giving nothing to charity. It is worth generalizing the fairness utility function to account for the charitable behavior of high-wealth individuals.

In sum, there is considerable evidence that fairness is a normal good. The present paper examined fairness as a standard economic good. Like the backward-bending labor supply, fairness may have a U-shaped impact as a function of people's wealth.

REFERENCES

Cameron, Lisa A. "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia." *Economic Inquiry*, January 1999, *37*(1), pp. 47–59.

Chichilnisky, Graciela. "North-South Trade and the Global Environment." *American Economic Review*, September 1994, *84*(4), pp. 851–74.

Forsythe, Robert; Horowitz, Joel L.; Savin, N. E. and Sefton, Martin. "Replicability, Fairness and Pay in Experiments with Simple Bargaining." *Games and Economic Behavior*, May 1994, *6*(3), pp. 347–69.

Hoffman, Elizabeth; McCabe, Kevin A. and Smith, Vernon L. "On Expectations and the Monetary Stakes in Ultimatum Games." *International Journal of Game Theory*, Summer 1996, *25*(3), pp. 289–301.

Rabin, Matthew. "Incorporating Fairness into Game Theory and Economics." *American Economic Review*, December 1993, *83*(5), pp. 1281–302.

Schervish, Paul G. and Havens, John J. "Do the Poor Pay More: Is the U-Shaped Curve Correct?" *Nonprofit and Voluntary Secter Quarterly*, Spring 1995, *24*(1), pp. 79–90.

_____. Money and Magnanimity: New Findings on the Distribution of Income, Wealth and Philanthropy." *Nonprofit Management and Leadership*, Summer 1998, *8*(4), pp. 265–82.

Roth, Alvin E.; Prasnikar, Vesna; Okuno-Fujiwara, Masahiro and Zamir, Shmuel. "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *American Economic Review*, December 1991, *81*(5), pp. 1068–95.

Slonim, Robert and Roth, Alvin E. "Learning in High Stakes Ultimatum Games: An Experiment in the Slovak Republic." *Econometrica*, May 1998, *66*(3), pp. 569–96.