A taste for fairness

Cristina Bicchieri

cb36@sas.upenn.edu
University of Pennsylvania

1. Introduction

One of the most important concepts in social exchanges and interactions is that of fairness. We can come to accept the most onerous tasks if we are convinced that the decision procedure was fair and, conversely, we may reject even a profitable exchange if we feel treated unfairly. Since the dawn of philosophy, a concern with fairness, what is it, how to define it, has been central to the philosopher's quest. Philosophers' concern, however, is more with finding reasons to justify and lend consistency to our intuitions about fairness than with the actual fairness judgments that people express. My interest here is rather to understand how people form fairness judgments, and what are the cognitive dynamics involved in the process. Within a given culture, there is usually a great deal of agreement as to how given goods, positions and opportunities ought to be allocated, and what properties of the claimants matter to the allocation. Every culture has developed a number of shared scripts about the fair allocation and distribution of various goods in different circumstances. Norms of fairness, in turn, are just an essential part of such shared scripts. Our fairness judgments are thus never completely subjective, independent of what our group or society consider fair given the circumstances. When we assess a situation, judge or decide about the fairness of an allocation, we apply scripts and obey norms that successfully coordinate our expectations and behaviors with the expectations and behaviors of other people. This does not mean that people always *agree* upon what a fair distribution is, given a set of circumstances. Disagreement, however, typically occurs when the situation is ambiguous and open to different interpretations. A typical example is the tension between equality and equity concerns in deciding how a given good or opportunity should be allocated. Interestingly enough, within a culture there usually exist agreed-upon justifications for deviating from equality. Reasons of merit or need are, depending on the nature of what has to be allocated, collectively judged to be acceptable or unacceptable, and typically disagreement occurs on the weight that different parties are prepared to give to some acceptable justifications.[1]

To say that people follow shared scripts and obey fairness norms differs from assuming that they have a 'preference' for fairness (Bicchieri 2000, 2006). To follow a fairness norm, one must have the right kind of expectations. One must expect others to follow the norm, too, and also believe that there is a generalized expectation that one will obey the norm in the present circumstances. The preference to obey a norm is *conditional* upon such expectations.[2] Take away some of the expectations, and behavior will significantly change. A conditional preference will thus be stable under certain conditions, but a change in the relevant conditions may induce a predictable preference shift. The predictions of a norm-based theory, as we shall see, are thus testable and quite different, at least in some critical instances, from the predictions of other theories that postulate a preference for fairness or a concern for reciprocity.

When economists postulate fairness preferences, they make two related, important assumptions. The first is that what matters to an agent is the final distribution, not the way the distribution came about (Fehr et al. 1999)). This is a *consequentialist* assumption. The second assumption is that preferences are *stable*. Both assumptions are easy to test. When falsified, however, it is less clear who the culprit is. For example, if a person has a stable preference for fair outcomes, we would expect her cross-situational behavior to be consistent and insensitive to the circumstances surrounding the specific distributive situation. Whether you are the Proposer in an Ultimatum or a Dictator game should not matter to your choice of how much money to give to a Responder. Similarly, information about *who* the Proposer is -- a real person or a random device -- should not have an effect on one's propensity to accept or reject its offer. What is observed instead is cross-situational inconsistency. The reason for this inconsistency is not obvious. It is possible that people do care about how a distribution came about, that the process itself matters. For example, one might accept an unequal share of the pie if it comes from a lottery, but reject it if it results from an auction. Preferences could still be assumed to be stable, but in this case what one would prefer is a combination of goods and processes to distribute/allocate those

[1] In our culture, for example, it is usually agreed that transplant organs should not be allocated according to the merit or income of the potential recipient, whereas we tend to believe that merit is the most important criterion in allocating scholarships.

[2]The conditions for following a norm are formally described in Chapter 1 of Bicchieri (2006) and in Appendix 1 here.

goods. On the other hand, we may take a different direction and go as far as saying that preferences are context-dependent. Change the context, or the context's description, and you have a noticeable preference shift. In the latter case, however, to be able to make any prediction we would need a mapping from contexts to preferences. No such mapping has ever been provided.

In what follows I will examine some of the most common games studied by experimental economists. Ultimatum and Dictator games come in many flavors and variants, but the simplest, bare versions of both games are in some sense ideal, since they offer a very simplified allocation problem. The good to be allocated (or divided) is money, and the situation is such that most familiar contextual clues are removed. It is thus possible to introduce in this rarefied environment simple contextual information and control for its effects on the perception of what constitutes a fair division. The results of such experiments consistently defy the predictions of traditional rational choice models. Agents are clearly not solely concerned with their monetary payoffs: they care about what other agents get and how they get it. The big challenge has been to enrich traditional rational choice models in such a way that they can explain (and predict) behavior that is not just motivated by material incentives in a variety of realistic contexts. I will compare some of the most interesting and influential new models with my norm-based approach and show that the hypothesis that people obey fairness norms offers a more complete explanation for the phenomena we observe. Where my predictions differ from those of alternative theories, the data seem to vindicate my model. However, we need many more experiments to test the effects that manipulating expectations (and thus norm-compliance) has on behavior.

2 The Ultimatum Game

In 1982, Guth, Schmittberger and Schwarze published a seminal study in which they asked subjects to play what is now known as an Ultimatum bargaining game. Their goal was to test the predictions of game theory about equilibrium behavior. Their results instead showed that subjects consistently deviate from what game theory predicts. To understand what game theory predicts, and why, let us look at a typical Ultimatum game (Figure 3.1)

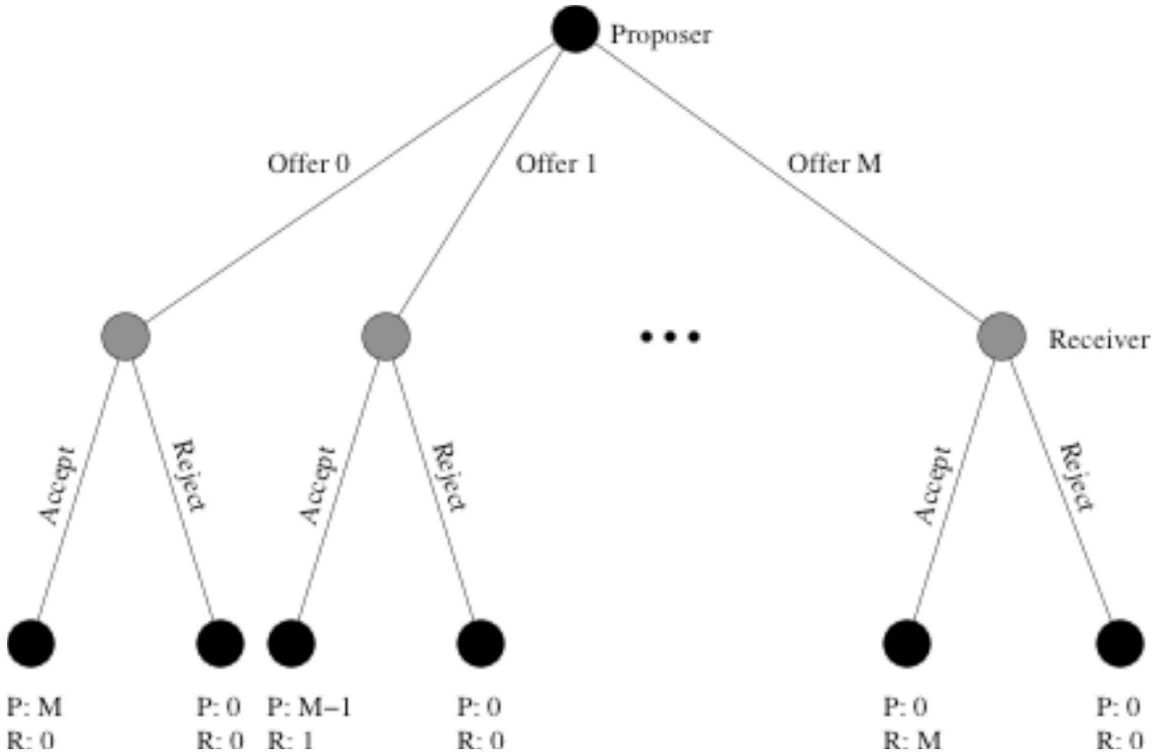


Figure 3.1

The structure of this game is fairly simple. Two people must split a fixed amount of money M according to the following rules: the Proposer (P) moves first and offers a division of M to the Responder (R), where the offer can range between M and zero. The Responder has a binary choice in each case: to accept the offer or to reject it. If the offer is accepted, the Proposer receives M-x and the Responder receives x, where x is the offer amount. If the offer is rejected, each player receives nothing. If rationality is common knowledge, the Proposer knows that the Responder will always accept any amount greater than zero, since Accept dominates Reject for *any* offer greater than zero. Hence P should offer the minimum amount guaranteed to be accepted, and R will accept it. For example, if M = $10 and the minimum available amount is 1 cent, the Proposer should offer it and the offer should be accepted, leaving the Proposer with $9.99.

When experiments are conducted, what one finds is that nobody offers 1 cent or even 1 dollar. Note that such experiments are always one-shot and anonymous. That is, subjects play the game only once with an anonymous partner and are guaranteed that their choice will not be disclosed. The absence of repetition is important to distinguish between generous behavior that is dictated by a rational, selfish calculation and genuine generosity. If an Ultimatum game is repeated with the same partner, or if one suspects that future partners will know of one's past behavior, it may be perfectly rational for a player who is only interested in his material payoff to give generously, if he expects to be on the receiving side at a future time. On the other hand, a Receiver who might accept the minimum in a one-shot game might want to reject a low offer at the beginning of a repeated game, in the hope of convincing future Proposers to offer more.

In the United States, as well as in a number of other countries, the modal and median offers in one-shot experimental games are 40 to 50 % of the total amount, and the mean offers are 30 to 40%. Offers below 20 % are rejected about half the time.[3] These results are robust with respect to variations in the amount of money that is being split, and cultural differences (Camerer 2003). For example, we know that raising the stake from $10 to $100 does not decrease the frequency of rejections of low offers (those between 10 and 20 dollars), and that in experiments run in Slovenia, Pittsburgh, Israel and Tokyo the modal offers were in the range of 40 to 50 percent (Hoffman et al. 1998; Roth et al. 1991).

If by rationality we mean that subjects maximize expected utility *and* that they only value their monetary outcomes, then we must conclude that a subject who rejects a nonzero offer is acting irrationally. However, individuals' behavior across games suggest that money is not the sole consideration, and instead there is a concern for fairness, so much so that subjects are prepared to punish those that behave in inequitable ways at a cost to themselves.[4]

A concern for fairness is just one example of a more general fact about human behavior: we are often motivated by a host of factors of which monetary incentives are one, and often not the most important. We act out of love, envy, spite, generosity, desire to imitate, sympathy or hatred, to name just few of the 'passions' and desires that move us to act. When faced with different possible distributions, we usually care about how we fare with respect to others, how the distribution came about, and who and why implemented it. Experiment after experiment have demonstrated that individuals care about others' payoffs, that they may want to spend resources to increase or decrease such payoffs, and that what they perceive to be the (good or bad) intentions of those they interact with weigh in their decisions. Unfortunately, the default utility function in game theory is a narrowly selfish one: it is selfish because it depicts people who care only about their own outcomes, and it is narrow because motivations like altruism, benevolence, guilt, envy or hatred are kept out of the picture. Such motives, however, can and should be incorporated into a utility

[3] Guth et al. (1982) were the first to observe that the most common offer by Proposers was to give half of the sum to the Responder. The mean offer was 37% of the original allocation. In a replication of their experiments, they allowed subjects to think about their decision for one week. The mean offer was 32% of the sum, which is still very high.

[4] We know that Responders reject low offers even when the stakes are as high as three months' earnings (Cameron 1995). Furthermore, experiments in which third parties have a chance to punish an 'unfair' Proposer at a monetary cost to themselves show that (moderately) costly punishment is frequent (Fehr and Fishbacher 2000).

function, and economists have recently started to develop richer, more complex models of human behavior that try to explain what we have always known: people care about other people's outcomes. Thus a better way to explain what is observed in experiments (and real life) is to provide a richer definition of rationality: people still maximize their utilities, but the arguments of their utility functions include other people's utilities.

The obvious risk of such models is their *ad hocness*: one may easily explain any data by adjusting the utility function to reflect what looks like envy, or altruism, or a preference for equal shares. What we need are utility functions that are general enough to subsume many different experimental phenomena, and specific enough to make falsifiable predictions. In what follows I will look at some proposed explanations for the generous distributions we observe in Ultimatum games, and test them against some interesting variations of the game. Such testing is not always easy to conduct. The problem is that we still have quite rudimentary theories of how motives affect behavior. And to test a hypothesis about what sort of motives induce us to act one way or another, we have to be very specific in defining such motives, and the ways in which they influence our choices. Let me clarify this statement with an example.

Observing the results of Ultimatum games, someone might argue that subjects in the Proposer's role are behaving altruistically. Others would deny that, saying that people like to give because of the "warm glow" their actions induce in them (Andreoni 1990, 1995), and yet others would say that what we observe is just benevolence, nothing else. Now, to make sense at all, such concepts need to be made as specific as possible, and operational. Take for example a distribution (x_1, x_2) of, say, money between two people. Being an altruist would mean that 1's utility is an increasing function of 2's utility, i.e. $U_1 = f(x_2)$ and $\delta U_1/\delta x_2 > 0$.
That is, a true altruist would not care about his own share but he would only care about how much the other gets (and the more, the better). A Proposer who is a pure altruist would 'donate' all the money to the Responder, provided he believes the Responder only cares about money. [5] Being benevolent instead means that one cares about one's own payoff *and* the other's, that is, $U_1 = f(x_1, x_2)$. In this case, the first partial derivatives of $U_1 = f(x_1, x_2)$ with respect to x_1, x_2 are strictly positive, meaning that the utility of a benevolent player 1 increases as the utility of player 2 increases. Depending on one's degree of benevolence, one will turn out to be more or less generous, but a benevolent attitude on the part of the Proposers might explain, *prima facie*, the results of experimental Ultimatum games.

The results of typical Ultimatum games eliminate the 'pure altruist' hypothesis, since people almost never give more than 50%, but do not eliminate the benevolence hypothesis. If benevolence is a stable character disposition, however, we would

[5] I am not sure such characters exist, and if they do how much liked they would be. In the Cloven Viscount (1951) Calvino depicts the whereabouts of a totally virtuous half nobleman who, because he took virtue to the extreme, was feared and disliked as much as his totally evil half counterpart.

expect a certain behavioral stability or consistency in any situation in which a benevolent Proposer has to offer a division of money to an anonymous Responder. A variant of the Ultimatum game is the Dictator game, in which the Proposer receives a sum of money from the experimenter and decides to split the money anyway in which she chooses; her decision is final in that the Responder cannot reject whatever is offered. If we hypothesize that the Ultimatum game results reveal that a certain percentage of the population has a benevolent disposition, we should expect to observe roughly the same percentage of generous offers in all those circumstances in which one of the parties, the Proposer, is all-powerful. In most of the experiments, however, the modal offer is one in which the Proposer keeps all the money to himself, and in double-blind experiments 64% of the participants give nothing. Still, it must be mentioned that although the most frequent offer is zero, the mean allocation is 20% (Forsythe et al., 1994). These results suggest that people are not totally selfish, but it would be hard to argue they are benevolent, unless we are prepared to presume that benevolence is a changeable disposition, as mutable as the circumstances which we encounter.

3 Social Preferences

Altruism or benevolence are just two examples of *social preferences*, whereas by social preference I refer to how people rank different allocations of material payoffs to self and others. If we stay with the Ultimatum game as an example, we can think of other, slightly more complex ways to explain the results we discussed before. The uniformity of Responders' behavior suggests that people do not like being treated unfairly. That is, if subjects perceive an offer of 20 or 30 percent of the money as unfair, they may reject it to "punish" the greedy Proposer, even at a cost to themselves. It is important to repeat that the experiments I am referring to were all one-shot, which means that the participants were fairly sure of not meeting again; therefore, punishing behavior cannot be motivated as an attempt to convince the other party to be more generous the next time around. Similarly, Proposers could not be generous because they were expecting reciprocating behavior in future interactions. One possibility is to assume that both Proposers and Responders are showing a preference for fair outcomes, or an aversion to inequality. We can thus try to explain the experimental results with a traditional rational choice model, where the agents' preferences take into account the payoffs of others.

In models of inequality aversion, players prefer both more money and that allocations be more equal. Though there are several models of inequality aversion, perhaps the best known and most extensively tested is the model of Fehr and Schmidt (1999). This model intends to capture the idea that people may be uneasy, to a certain extent, about the presence of inequality, even if they benefit from the unequal distribution. Given a group of L persons, the Fehr-Schmidt utility function of person i is

$$U_i(x_1,...,x_L) = x_i - \frac{\alpha_i}{L-1}\sum_j \max(x_j - x_i, 0) - \frac{\beta_i}{L-1}\sum_j \max(x_i - x_j, 0)$$

where x_j denotes the material payoff person j gets. α_i is a parameter that measures how much player i dislikes disadvantageous inequality (an 'envy' weight), and β_i measures how much i dislikes advantageous inequality (a 'guilt' weight).[6] One constraint on the parameters is that $0 < \beta_i < \alpha_i$, which indicates that people dislike advantageous inequality less than disadvantageous inequality. The other constraint is $\beta_i < 1$, so that an agent does not suffer terrible guilt when she is in a relatively good position. For example, a player would prefer getting more without affecting other people's payoff even though that results in an increase of the inequality.

Applying the model to the game in Figure 1, the utility function is simplified to

$$U_i(x_1, x_2) = x_i - \begin{cases} \alpha_i(x_{3-i} - x_i) & if\ x_{3-i} \geq x_i \\ \beta_i(x_i - x_{3-i}) & if\ x_{3-i} < x_i \end{cases} \qquad i = 1,2$$

Obviously if the Responder rejects the offer, both utility functions are equal to zero, that is, $U_{1reject} = U_{2reject} = 0$. If the Responder accepts an offer of x, the utility functions are as follows:

$$U_{1accept}(x) = \begin{cases} (1+\alpha_1)M - (1+2\alpha_1)x & if\ x \geq M/2 \\ (1-\beta_1)M - (1-2\beta_1)x & if\ x < M/2 \end{cases}$$

$$U_{2accept}(x) = \begin{cases} (1+2\alpha_2)x - \alpha_2 M & if\ x < M/2 \\ (1-2\beta_1)x + \beta_2 M & if\ x \geq M/2 \end{cases}$$

The Responder should accept the offer if and only if $U_{2accept}(x) > U_{2reject} = 0$. Solving for x we get the *threshold for acceptance*: $\boldsymbol{x > \alpha_2 M/(1+2\alpha_2)}$. Evidently if α_2 is close to zero, which indicates that player 2 (R) does not care much about being treated unfairly, the Responder will accept very mean offers. On the other hand, if α_2 is sufficiently big, the offer has to be close to half to be accepted. In any event, the threshold is not higher than $M/2$, which means that hyper-fair offers (more than half) are not necessary for the sake of acceptance.

Note that for the Proposer, the utility function is monotonically decreasing in x when $x \geq M/2$. Hence a rational Proposer will not offer more than half of the money. Suppose $x \leq M/2$; two cases are possible depending on the value of β_1. If $\beta_1 > 1/2$, that is, if the Proposer feels sufficiently guilty about treating others unfairly, the utility is monotonically increasing in x, and his best choice is to offer $M/2$. On the other hand, if $\beta_1 < 1/2$, the utility is monotonically decreasing in x, and hence the best offer for the Proposer is the minimum one that would be accepted, i.e. (a little bit more than) $\alpha_2 M/(1+2\alpha_2)$. Lastly, if $\beta_1 = 1/2$, it does not matter how much the Proposer offers, as long
as it is between $\alpha_2 M/(1+2\alpha_2)$ and $M/2$. Note that the other two parameters, α_1 and β_2, are not identifiable in Ultimatum games.

---

6 The term max ($x_j - x_i$, 0) denotes the maximum of $x_j - x_i$ and 0; it measures the extent to which there is disadvantageous inequality between i and j.

As noted by Fehr and Schmidt, the model allows for the fact that individuals are heterogeneous. Different α's and β's correspond to different types of people. Although the utility functions are common knowledge, the exact values of the parameters are not. The Proposer, in most cases, is not sure what type of Responder she is facing. Along the Bayesian line, her belief about the type of the Responder can be formally represented by a probability distribution P on α_2 and β_2. When $\beta_1 > 1/2$, the Proposer's rational choice does not depend on what P is. When $\beta_1 < 1/2$, however, the Proposer will seek to maximize the expected utility:

$$EU(x) = P(\alpha_2 M / (1 + 2\alpha_2) < x) \times ((1 - \beta_1)M - (1 - 2\beta_1)x)$$

Therefore, the behavior of a rational Proposer in the Ultimatum game is determined by her own type (β_1) and her belief about the type of the Responder. The experimental data suggest that for many Proposers, either β is big ($\beta > 1/2$), or they estimate the Responder's α to be large. The choice of the Responder is only determined by his type (α_2) and the offer. Small offers are rejected by Responders with a positive α..

The positive features of the above-described utility function are that it can rationalize both positive and negative outcomes, and that it can explain the observed variability in outcomes with heterogeneous types. One of the major weaknesses of this model, however, is that it has a consequentialist bias: players only care about final distributions of outcomes, not about how such distributions come about.[7] As we shall see, more recent experiments have established that how a situation is framed matters to an evaluation of outcomes, and that the same distribution can be accepted or rejected depending on 'irrelevant' information about the players or the circumstances of play. Another difficulty with this approach is that, if we assume the distribution of types to be constant in a given population, then we should observe, overall, the same proportion of 'fair' outcomes in Ultimatum games. Not only this does not happen, but we also observe individual inconsistencies in behavior across different situations in which the monetary outcomes are the same. If we assume, as is usually done in economics, that individual preferences are stable, then we would expect similar behaviors across Ultimatum games. If instead we conclude that preferences are context-dependent, then we should provide a mapping from contexts to preferences that indicates in a fairly predictable way how and why a given context or situation changes one's preferences. Of course, different situations may change a player's expectation about another player's envy or guilt parameters, and we could thus explain why a player may change her behavior depending upon how the situation is framed. In the case of Fehr and Schmidt's utility function, however, experimental evidence that I shall discuss later implies that a player's *own* β (or α) changes value in

[7] This is a *separability* of utility assumption: what matters to a player in a game is her payoff at a terminal node. The way in which that node was reached, and the possible alternative paths that were not taken are irrelevant to an assessment of her utility at that node. Utilities of terminal node payoffs are thus separable from the path through the tree, and from payoffs on unchosen branches.

different situations. Yet nothing in their theory explains why one would feel consistently more or less guilty (or envious) depending on the decision context.

4 Reciprocity

Theories of inequality aversion only include other players' material payoffs in the calculation of utility. What other players did, and why they did it, does not play any role in a player's utility. Yet we tend to take into account what we believe are the intentions of those we interact with, and respond accordingly. Reciprocity is a common phenomenon in human interaction: we tend to be kind to kind persons and punish the mean. By leaving reciprocity out, the previous model gains simplicity and tractability. But, as Matthew Rabin (1993) forcefully argued, in order to model reciprocity we have to include beliefs and intentions in our models. In Rabin's model, utilities do not just depend on terminal-node payoffs but also on players' beliefs. As a result, his model builds upon the framework of what is called psychological game theory (Geanakoplos et al. 1989).

Consider a two-person game of complete information. According to Rabin's model, a player's utility is not only determined by the actions taken, but it also depends on the player's beliefs (including second-order beliefs, viz. beliefs about beliefs). Specifically, player *i* will evaluate her "kindness" to the other player, f_i, by the following scheme:

$$f_i(a_i, b_j) = \begin{cases} \dfrac{\pi_j(b_j, a_i) - \pi_j^c(b_j)}{\pi_j^h(b_j) - \pi_j^{\min}(b_j)} & \text{if } \pi_j^h(b_j) - \pi_j^{\min}(b_j) \neq 0 \\ 0 & \text{otherwise} \end{cases} \qquad i = 1,2; \;\; j = 3 - i$$

where a_i is the strategy taken by player *i*, b_j is the strategy that player *i believes* is chosen by player *j*. π_j is *j*'s material payoff that depends on both players' strategies. $\pi_j^h(b_j)$ is the highest material payoff and $\pi_j^{\min}(b_j)$ the lowest payoff that player *j* can potentially get by playing b_j. In other words, they denote respectively the highest and lowest payoffs player *i* can grant player *j* given the latter is playing b_j. A key term here is $\pi_j^c(b_j)$, which represents a "fair" material payoff player *j* "should" get by playing b_j, and is defined by Rabin as:

$$\pi_j^c(b_j) = \frac{\pi_j^h(b_j) + \pi_j^l(b_j)}{2}$$

$\pi_j^l(b_j)$ is the worst payoff player *j* may incur given that players do not play Pareto dominated strategies. Obviously we have $\pi_j^l(b_j) \geq \pi_j^{\min}(b_j)$. Thus a positive $f_i\,(a_i, b_j)$ means player i has been kind to j, because j got a payoff higher than the fair one, and a negative value signifies i was mean to j, who got a lower than fair payoff.

Similarly, player *i* can estimate player *j*'s kindness towards her, denoted by

$$\tilde{f}_j(b_j, c_i) = \begin{cases} \dfrac{\pi_i(c_i, b_j) - \pi_i^c(c_i)}{\pi_i^h(c_i) - \pi_i^{\min}(c_i)} & \text{if } \pi_i^h(c_i) - \pi_i^{\min}(c_i) \neq 0 \\ 0 & \text{otherwise} \end{cases}$$

where c_i is *i*'s belief about *j*'s belief about the strategy taken by *i*, a second-order belief. The meanings of other terms are obvious given the previous explanations. Clearly this estimated kindness is just a conjecture about player *j*'s intentions, and to form this conjecture player *i* must make a guess about what player *j* believes that *i* will do.
Finally, the utility function of player *i* depends on her strategy, (first-order) belief, and second-order belief:

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

The first term tells us that player *i* cares about her material payoff, the second term tells us that it matters to *i* whether she is treated nicely or not, and reciprocity lies in the last interaction term (the product of the kindness *i* expects and of her own kindness). Intuitively, it satisfies a player to be kind to kind players and tough to tough ones. An equilibrium of the game, called a *fairness equilibrium*, occurs when every belief turns out to be correct and each player's utility is maximized.

A problem with this approach is that there can be many fairness equilibria, depending upon the beliefs the players happen to have. Since this model applies no constraint on possible beliefs, it becomes impossible to predict which equilibrium will be played. Furthermore, though it is certainly more realistic to assume that players care about other players' intentions, we do not attribute good or bad intentions in a vacuum. An intention is only good or bad against a background of expectations. Such expectations are often dictated by the situation one is in, and are thus quite homogeneous among players. As we shall see, a theory of social norms can predict the beliefs and expectations the players will have in a particular setting, and thus predict that a specific equilibrium will obtain.

5. Norms matter

Rule-based approaches are not completely new. Guth (1995) for example interpreted the results of the Ultimatum game as showing that people have rules of behavior such as sharing money equally, and they apply them when necessary. The problem with such solutions is that we need a plausible story about how people change their behavior in response to changes in payoffs and framing. If rules are inflexible, but we observe flexible compliance, there must be something wrong with a rule-based approach. Indeed, a common understanding of norms, one that I have tried to dispel in my definition (see Appendix 1), is that they are inflexible behavioral rules that one would apply in any circumstance that calls for them. Nothing could be farther from the truth. To be effective, norms have to be *activated* by salient cues.[8] As I explain in (Bicchieri 2000, 2006), a norm may exist but it may not be followed simply because the relevant expectations are not there, or because one might be unaware of being in a situation to which the norm applies. I have argued that people have *conditional preferences* for conformity to a norm, in that they would prefer to follow it on condition that (a) they expect others to follow it and (b) they believe that, in turn, they are expected by others to abide by the norm (see Appendix 1 and Bicchieri 2006). Both conditions have to be present in order to generate conformity. Indeed, there is plenty of evidence that manipulating people's expectations has an effect on norm compliance (Cialdini 1990). Thus I would argue that belief-elicitation in experiments is crucial to determine whether a norm will be perceived as relevant and then followed. We already know, for example, that telling subjects how others have behaved in a similar game has a profound effect on their choices, and that allowing people to communicate before playing the game often results in a cooperative outcome.[9]

Furthermore, some norms are more *local* than others, in the sense that their interpretation is highly context-dependent. Fairness is a case in point. To be fair means different things in different contexts. In some situation being fair means sharing equally, in others it may mean giving more to the needy or to the deserving. Ultimatum games are in some sense ideal, since they offer a very simplified allocation problem. The good to be allocated (or divided) is money, and the situation is such that most familiar contextual clues are removed. It is thus possible to introduce in this rarefied environment simple contextual information and control for its effects on the perception of what constitutes a fair division. In the Ultimatum game, the salience of the equal split solution is lost if subjects are told that offers are generated by a random device (Blount 1995), or if it is believed that the Proposer was otherwise constrained in her decision. In both cases, Responders are willing to accept lower offers. This phenomenon is well known to consummate bargainers: if an unequal outcome can be credibly justified as a case of *force majeure*, people can be convinced to accept much less than an equal share. Also, variations in the strength of 'property rights' alter the shared expectations of the two players regarding the norm that determines the appropriate division. In the original Ultimatum game, the Proposer receives what amounts to a monetary *gift* from the experimenter. As a consequence, he is perceived as having no special right to the money, and is expected (at least in our culture) to share it equally with the Responder. Since the fairness norm that is activated in this context dictates an equal split, the Proposer who is offering little is perceived as mean, and consequently he gets punished. Note that the Proposer who was constrained in his decision is not seen as being intentionally mean, since intentions do matter only when the choice is perceived as being freely made. To infer another person's intention or motive, we consider not only the action chosen, but also the actions that were not chosen but, as far as we know, *could* have been chosen.

---

[8] Cues that activate or 'bring to mind' a norm may involve a direct statement or reminder of the norm, observing others' behavior, similarity of the present situation to others in which the norm was used, as well as how often or how recently one has used the norm.

[9] I discuss these results and the relevant literature in (Bicchieri 2006).

Since what counts as fair is highly context-dependent, a specific context simultaneously gives reasons to expect behavior appropriate to the situation, and a clue as to the Proposer's intention, especially when the offer is different from what is reasonably expected in that context. Subjects approach resource-sharing or, for that matter, any other situation with implicit knowledge structures (scripts) that detail conditions that are prototypically associated with sharing tasks. Once we have categorized the particular decision task we face, we enact scripts that tell us how people typically behave and what they expect others to do. However, it must be emphasized that people will display expected, appropriate behavior to the extent that crucial environmental cues match those of well-known prototypical scripts. An interesting question to ask is thus under which conditions an equal sharing norm will be violated. I shall discuss more extensively this point later on, but for now let me say that my hypothesis is that a deviation from equal sharing will be mainly due to (a) the presence of appropriate and acceptable justifications for taking more than an equal share or (b) the shift to a very different script that involves different roles and expectations. An example of the second reason is when the Proposer is labeled "seller", and the Responder "buyer"; in this case the Proposer offers a lower amount that in the control, and Responders readily accept (and expect) less than an equal share (Hoffman et al. 1994). In this case, the interaction is perceived as being market-like, and in a market script it is deemed equitable that a seller earns a higher return than a buyer. An example of the first reason is when the Proposer has "earned" the right to the money by, for example, getting a higher score on a general knowledge quiz (Frey and Bohnet 1995; Hoffman and Spitzer 1985). In this case the Proposer has an available, acceptable justification for getting more than the equal share. Doing better than someone else in a test is a common and reasonable mechanism, at least in our society, for determining differential access to a shared resource. It thus seems appropriate to many Proposers to choose equity versus equality in such conditions even if, as we shall see, this self-serving rule is not shared by the Responder.[10]

There is continuity between real life and experiments with respect to how 'rights' and 'entitlements', considerations of merit, need, desert or sheer luck shape our perception of what is fair and what kind of reasons count as acceptable justifications for violating a fairness norm. Cultures differ in their reliance on different allocative and distributive rules, since such rules depend on different forms of social organization. Within a given culture, however, there usually is a consensus about how different goods and opportunities should be allocated or distributed. Cross-cultural studies of Ultimatum and Dictator games in fifteen small-scale societies show quite convincingly that the behavior displayed in such games was highly correlated with the economic organization and social structure of each society (Henrich et al. 2004). Furthermore, since experimental play is presumably categorized according to the specific socio-cultural patterns of each society, the experimental results showed much greater variability than the results of typical Ultimatum and Dictator games played in

[10] Kahneman et al. (1986) describe different norms of fairness, including situations in which unfair behavior is commonly accepted and "excused".

modern Western (or westernized) societies.[11] These results lend even more support to the hypothesis that social norms, and the accompanying shared expectations, play a crucial role in shaping behavioral responses to experimental games.

A norm-based explanation of the results of experiments with Ultimatum and Dictator games predicts that -- whenever Proposers are focused upon the relevant expectations -- they will behave in a norm-consistent way. In the traditional Ultimatum game, the expected opportunity cost of not following an equal division rule may be enough to elicit fair behavior. In asking herself what the Responder would accept, the Proposer is forced to look at the situation and categorize it as a case in which an equality rule applies. This does not mean that the person that follows the norms is in fact fair, or casts a high value on equitable behavior. As I made plain in my definition of what it takes to follow an existing norm, if a player assesses a sufficiently high probability to her opponent's following the norm, and expects to be punished for non-compliance, she will prefer to conform to a norm even if she has no interest in the norm itself.

The general utility function I introduced in (Bicchieri 2006) can now be applied to the Ultimatum game. Let π_i be the payoff function for player *i*. Recall that the norm-based utility function of player *i* depends on the strategy profile *s*, and is given by

$$U_i(s) = \pi_i(s) - k_i \max_{s_{-j} \in L_{-j}} \max_{m \neq j} \{\pi_m(s_{-j}, N_j(s_{-j})) - \pi_m(s), 0\}$$

Where $k_i \geq 0$ is a constant representing i's sensitivity to the relevant norm. Such sensitivity may vary with different norms; for example, a person may be very sensitive to equality and much less so to equity considerations. The first maximum operator takes care of the possibility that the norm instantiation (and violation) might be ambiguous in the sense that a strategy profile instantiates a norm for several players simultaneously (as would be the case, for example, in a Social Dilemma with three players). The second maximum operator ranges over all the players other than the norm violator. In plain words, the discounting term (multiplied by k_i) is the maximum payoff deduction resulting from all norm violations.

The model is motivated by people's apparent respect (or disregard) for social norms regarding fairness. In the traditional Ultimatum game, the norm usually prescribes a 'fair' amount the Proposer ought to offer. The norm functions that represent this norm are the following: N_1 is a constant N function, and N_2 is nowhere defined.[12] If the Responder (player 2) rejects, the utilities of both players are zero.

$$U_{1reject}(x) = U_{2reject}(x) = 0$$

[11] In some groups, rejections were extremely rare, even when offers were very low, whereas in other groups 'hyper-fair' offers were frequently rejected, pointing to very different (but inter-culturally shared) interpretations of the experimental situation.

[12] Intuitively, N_2 should proscribe rejection of fair (or hyper-fair) offers. The incorporation of this consideration, however, will not make a difference in the formal analysis.

Given that the Proposer (player 1) offers x and the Responder accepts, the utilities are the following:

$$U_{1accept}(x) = M - x - k_1 \max(N_1 - x, 0)$$
$$U_{2accept}(x) = x - k_2 \max(N_2 - x, 0)$$

where N_i denotes the amount player i thinks he should get/offer according to some social norm applicable to the situation, and k_i is non-negative. Note that k_1 measures how much player 1 dislikes to deviate from what he takes to be the norm. To obey a norm, 'sensitivity' to the norm need not be great nor due to an appreciation for what the norm stands for. Fear of retaliation may make a Proposer with a 'low' k behave according to what fairness dictates but, absent such risk, his attitude to deviations may lead him to be unfair. For the moment, I assume it is common knowledge that $N_1 = N_2 =$ N, which is not too unreasonable in the traditional Ultimatum game. Again, the Responder should accept the offer if and only if $U_{2accept}(x) > U_{2reject} = 0$, which implies the following *threshold for acceptance*: $\boldsymbol{x > k_2N/(1+k_2)}$. Notice that an offer larger than the norm dictates is not necessary for the sake of acceptance.

For the Proposer, the utility function is decreasing in x when $x \geq N$, hence a rational Proposer will not offer more than N. Suppose $x \leq N$. If $k_1 > 1$, the utility function is increasing in x, which means that the best choice for the Proposer is to offer N. If $k_1 < 1$, the utility function is decreasing in x, which implies that the best strategy for the Proposer is to offer the least amount that would result in acceptance, i.e. (a little bit more than) the threshold $k_2N/(1+k_2)$. If $k_1 = 1$, it does not matter how much the Proposer offers provided the offer is between $k_2N/(1+k_2)$ and N.

It should be noted that k_1 plays a very similar role as that of β_1 in the Fehr-Schmidt model. In fact, if we take N to be $M/2$ and k_1 to be $2\beta_1$, the two models agree on what the Proposer's utility is. It is equally apparent that k_2 in this model is analogous to α_2 in the Fehr-Schmidt model. There is, however, an important difference between these parameters. The α's and β's in the Fehr-Schmidt model measure people's degree of averseness towards inequality, which is a very different disposition than the one measured by the k's, i.e. people's sensitivity to different norms. The latter may be a stable disposition, and behavioral changes may be due to changes in focus or in expectations. A theory of norms can explain such changes, whereas a theory of inequity aversion does not. I will come back to this point later.

It is also the case that the Proposer's belief about the Responder's type figures in her decision when $k_1 < 1$. The belief can be represented by a joint probability over k_2 and N_2, if the value of N_2 is not common knowledge. The Proposer should choose an offer that maximizes the expected utility

$$EU(x) = P(k_2N_2/(1+k_2) < x) \times (M - x - k_1(N_1 - x))$$

As will become clear, an advantage this model has over the Fehr-Schmidt model is that it can explain some variants of the traditional Ultimatum game more naturally. However, it shares a problem with the Fehr-Schmidt model: they both entail that fear of rejection is the only reason why people offer almost-fair amounts rather than lower sums. This prediction, however, could be easily refuted by a parallel Dictator game where rejection is not an option.

6 Variation on the Ultimatum Game

So far I only considered the basic Ultimatum game, which is not the whole story. There have been a number of interesting variants of the game in the literature, to some of which I now apply the models to see if they can tell reasonable stories about what happens in those experiments.

Ultimatum Game with Asymmetric Information and Payoffs

Kagel, et al. [1996] designed an ultimatum game in which the Proposer is given a certain amount of chips. The chips are worth either more or less to the Proposer than they are to the Responder. Each player knows how much a chip is worth to her, but may or may not know that the chip is worth differently to the other. Participants play an Ultimatum game over 10 rounds with changing opponents, and this is public knowledge. The particularly interesting setting is one in which the chips have higher (three times more) values for the Proposer, and only the Proposer knows it. It turns out that in this case the offer is (very close to) half of the chips and the rejection rate is low. A popular reading of this result is that people merely prefer to *appear* fair, as a really fair person is supposed to offer about 75% of the chips. As Figure 2 shows, Proposers offered close to 50% of the chips, and very few such offers were rejected.[13]

[13] In the condition in which only Proposers know the chips value, when Proposers' chips were worth less, offers declined to mean 31.4 chips and rejections increased to 21%. Note that in the condition in which only Responders know the chips value, Proposers who had a higher chip conversion offered a mean of 45.7 chips over 10 bargaining periods. However, the chips are worth less to the Responder, who knows both values, hence the rejection rate was 34%. When the Responder had a high conversion rate, Proposers offered less (mean = 29.7%). The authors of the study did not report if the offers lower than 50% were those rejected, but they concluded that rejections were due to inequality aversion. Rejections were 21%. Note that the mean offer is close to the money equalizing split. When both players knew the chips value, Proposers with a high conversion rate offered a mean of 54.4 chips over the first three rounds. Rejections in these rounds were high at 52%. This brought the offer up to 63.7 chips by round 10, and the overall rejection rate lowered to 39%. Unfortunately, the overall mean offer was not reported in the paper. When the chips were worth more to the Responder, mean offers stayed close to 255, the money-equalizing split, throughout the game. Rejection rates were low at 14%.

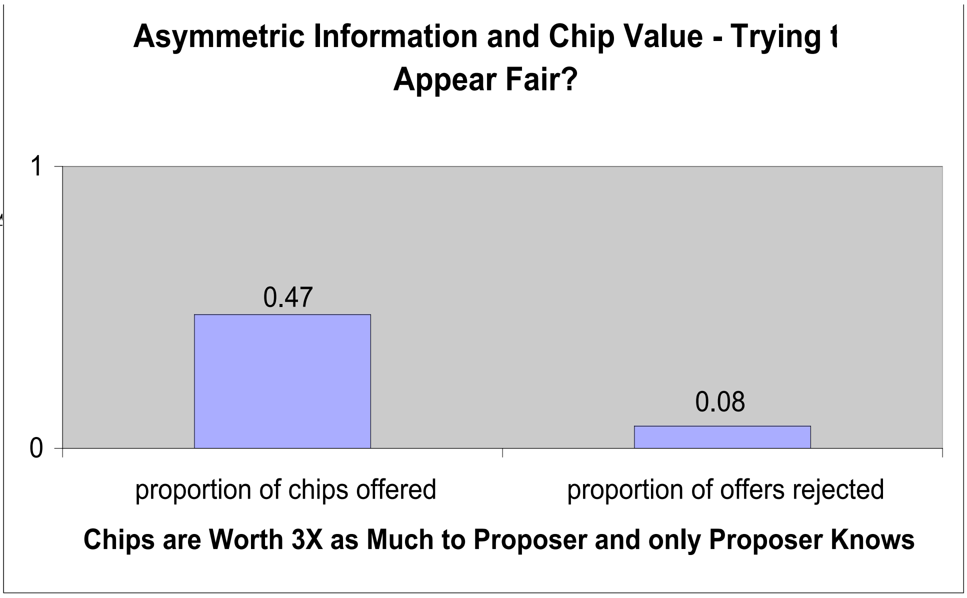


Figure 3.2

To analyze this variant formally, we only need a small modification to our original setting. That is, if the Responder accepts an offer of x, the Proposer actually gets $3(M-x)$ though, to the Responder's knowledge, she only gets $M-x$. In the Fehr-Schmidt model, the utility function of player 1 (the Proposer), given the offer gets accepted, is now the following:

$$U_{1accept}(x) = \begin{cases} (3+3\alpha_1)M - (3+4\alpha_1)x & \text{if } x \geq 3M/4 \\ (3-3\beta_1)M - (3-4\beta_1)x & \text{if } x < 3M/4 \end{cases}$$

The utility function of the Responder upon acceptance does not change, as to the best of his knowledge, the situation is the same as in the simple Ultimatum game. Also, if the Responder rejects the offer, both utilities are again zero. It follows that the Responder's threshold for acceptance remains the same, he accepts the offer if $x > \alpha_2 M/(1+2\alpha_2)$. For the Proposer, if $\beta_1 > 3/4$, her best offer is $3M/4$, otherwise her best offer is the minimum amount above the threshold. An interesting point is that even if someone offers $M/2$ in the simple Ultimatum game, which indicates that $\beta_1 > 1/2$, she may not offer $3M/4$ in this new condition. This prediction is consistent with the observation that almost no one offers 75% of the chips in the real game.

At this point, it seems the Fehr-Schmidt model does not entail a difference in behavior in this new game. But Proposers in general do offer more in this new setting than they do in the usual Utimatum game, which naturally leads to the lower rejection rate. Can the Fehr-Schmidt model explain this? One obvious way is to adjust α_2 so that the predicted threshold increases. But there is no reason in this case for the Responder to change his attitude towards inequality. Another explanation might be that under this new setting, the Proposer believes that the Responder's distaste for inequality increases, for after all it is the Proposer's belief about α_2 that affects the offer. This move sounds as questionable as the last one, but it does point to a reasonable explanation. Since the Proposer is uncertain about what kind of Responder she is facing, her belief about α_2 should be represented by a non-degenerate probability distribution. She should choose an offer that maximizes her expected utility, which in this case is given by the following:

$$\mathrm{EU}(x) = \mathrm{P}\,(\alpha_2 < x / (M - 2x)) \times ((3 - 3\beta_1)\, M - (3 - 4\beta_1)\, x)$$

The main difference between this expected utility and the one in the simple Ultimatum game is that it involves a bigger stake. Hence it is likely to be maximized at a bigger x unless the distribution (her belief) over α_2 is sufficiently odd. Thus the Fehr-Schmidt model can explain the phenomenon in a reasonable way.

If we apply my model to this new setting, again the utility function of player 2 does not change. The utility function of player 1 (the Proposer) given acceptance is changed to

$$U_{1accept}(x) = 3(M - x) - k_1 \max(N_1' - x, 0)$$

I use N_1' here to indicate that the Proposer's perception of the fair amount, or her interpretation of the norm, may have changed due to her awareness of the informational asymmetry.[14] My model behaves quite similarly to the previous one. Specifically, the Responder's threshold for acceptance is still $k_2N_2/(1+k_2)$. The Proposer will/should offer N_1' only if $k_1 > 3$, so people who offer the "fair" amount in the simple Ultimatum game ($k_1 > 1$) may not offer the "fair" amount under the new setting. That means that even if $N_1' = 3M/4$, the observation that few people offer that amount does not go against my model. The best offer for most people ($k_1<3$) is the least amount that would be accepted. However, since the Proposer is not sure about the Responder's type, she will choose an offer to maximize her expected utility, and this in general leads to an increase of the offer given an increase of the stake. Although it is not particularly relevant to the analysis in this case, it is worth noting that N_1' is probably less than $3M/4$ in the situation as thus framed. This point will become crucial in games with obvious framing effects.

The Rabin model, as it stands, has several difficulties. The primary trouble still centers on the kindness function. It is not hard to see that according to Rabin's definition of kindness, the function that measures the Proposer's kindness to the Responder does not change at all, while the function that measures the other way around does change[15]. This does not sound plausible. Intuitively, other things being equal, the only thing that may change is the Proposer's measure of her kindness to the other. There is no reason to think that the Responder's estimation of the other player's kindness to him will change, as the Responder does not have the relevant information. Strictly speaking, Rabin's original model cannot be applied to the situation where asymmetric information is present, because his framework assumes the payoffs being common knowledge.

[14]It is important to note that since norms are very dependent on expectations, informational asymmetries will almost certainly affect norm-following behaviors.

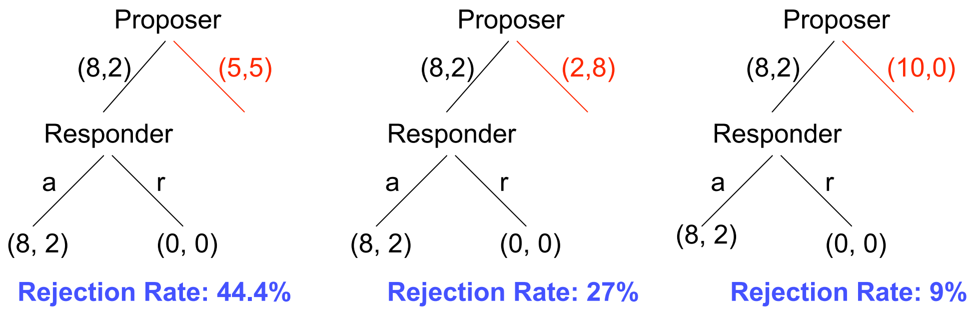
[15] By Bicchieri and Zhang's definition of kindness, both functions remain the same as in the simple setting.

It is, however, worth noting that if the kindness functions all remain the same (as it is the case under the definition of kindness in Bicchieri and Zhang, 2004), the arguments available to Rabin to address the new situation are very similar to the ones available to the previous models. One move is to manipulate α's, which is unreasonable as already pointed out. Another move is to represent beliefs with more general probability distributions (than a point mass distribution), and to look for Bayesian equilibria. The latter move will inevitably further complicate the already complicated model, but it does seem to match the reality better.

Ultimatum Game with Different Alternatives

There is also a very simple twist to the Ultimatum game, which turns out to be quite interesting. Falk et al. (2000) introduced a simple Ultimatum game where the Proposer has only two choices: either offer 2 (and keep 8) or make an alternative offer that varies across treatments in a way that allows the experimenter to test the effect of reciprocity and inequity aversion on rejection rates. The alternative offers in four treatments are (5/5), (8/2), (2/8) and (10/0). As Figure 3 shows, when the (8/2) offer is compared to the (5/5) alternative, the rejection rate is 44.4%, and it is much higher than the rejection rates in each of the alternative three treatments. In fact, it turns out that the rejection rate depends a lot on what the alternative is. The rejection rate decreases to 27% if the alternative is (2/8), and further decreases to 9% if the alternative is (10/0).[16]

Figure 3.3



[16] Note that 30% of the subjects proposed (8,2) when the alternative was (5,5), 70% proposed (8,2) when the alternative was (2,8) and ()% proposed (8,2) when the alternative was (10,0). Each player played 4 games, presented in random order, in the same role.

It is hard for the Fehr-Schmidt model to explain these results? In this consequentialist model there does not seem to be any role for the available alternatives to play. As the foregoing analysis shows, the best reply for the Responder is acceptance if $x > \alpha_2 M/(1+2\alpha_2)$. That is, different alternatives can affect the rejection rate only through their effects on α_2. It is not entirely implausible to say that "what could have been otherwise" affects one's attitude towards inequality. After all, one's dispositions are shaped by all kinds of environmental or situational factors, to which the 'path not taken' seem to belong. Still it sounds quite odd that one's sensitivity to fairness changes as alternatives vary, and in particular, it is not compatible with the assumption of independence of irrelevant alternatives, a common assumption in decision theory.
The norm-based model, by contrast, seems to have an easier time. For one thing, my model can explain the data by telling a story about how the norm's perception might change, and the story, unlike the previous case, can be quite plausible. Recall that my definition of what it takes to follow a norm relies heavily on expectations, both empirical and normative. As I discussed in (Bicchieri 2000, 2006), how we decide and act in a situation depends upon how we interpret, understand and encode it. Once a situation is categorized as a member of a particular class, a schema (or script) is invoked. Such script allows us to make inferences about unobservable variables, predict other people's behavior, make causal attributions and modulate emotional reactions. The script we invoke is the source of both projectible regularities and the legitimacy of our expectations. If, as I argued, social norms are embedded into scripts, then the particular way a situation is framed will have a large effect on our expectations about others' behavior and what they expect from us. Thus a change in the way a situation is framed will induce a change in expectations and have an immediate effect on our focusing (or not focusing) on the norm that has (or has not) been elicited.

As the possible alternatives vary, the player may no longer believe that the same norm applies and it is quite reasonable to conjecture that different alternatives point the Responder to different norms (or lack thereof). In the (8,2), (5,5) situation, players are naturally focused on the equal split. The Proposer who could have chosen it but did not is sending a clear message about his disregard for fairness. If the expectation of a fair share is violated, the Responder will probably feel outraged, attribute a mean intention to the Proposer, and punish him accordingly. If the alternatives are (8,2) or (2,8), few people would expect a Proposer to 'sacrifice' for the Responder. In real life, situations like this are decided with a coin toss. In the game context, it is difficult to see that any norm would apply to the situation. This is why 70% of the subjects choose the (8,2) split, and only 27% reject it. Finally, the choice of (8,2) when the alternative is (10,0) appears quite nice, and indeed the rejection rate is only 9%. When the alternative for the Proposer is to offer the whole stake, there is little reason for the Responder to think that the norm is still (50%, 50%) or something close to this. Thus a natural explanation given by my model is that N_2 changes (or may be empty) as the alternative varies.

The results of this experiment tell us that most people do not have selfish material preferences, in which case they would always accept the (8,2) division. But they also tell us that people are not simply motivated by a dislike for inequality, otherwise we would have observed the same rejection rate in all contexts.

Ultimatum Game with Framing

Framing effects, a topic of continuing interest to psychologists and social scientists, have also been investigated in the context of Ultimatum games. Hoffman et al.(1992), for example, designed an Ultimatum game in which groups of twelve participants were ranked on a scale 1-12 either randomly or by superior performance in answering questions about current events. The top six were assigned to the role of "seller" and the rest to the role of "buyer". They also ran studies with the standard Ultimatum game instructions, both with random assignments and assignment to the role of Proposer by contest. The exchange and contest manipulations elicited significantly lowered offers, but rejection rates were unchanged as compared to the standard Ultimatum game.[17]

Figure 3.4

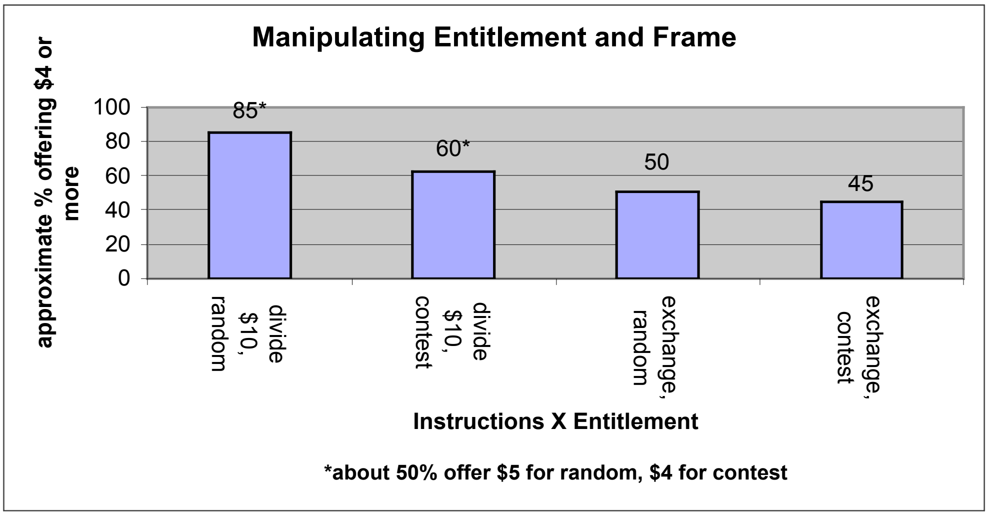


Figure 3.4 shows that the ‘exchange’ framing significantly lowered offers, but also the fact of being the winner of a contest in the traditional Ultimatum game had an effect on the Proposers’ offers. Several other experiments have consistently shown that when the Proposer is a ‘contest winner’ (Frey and Bohnet 1995), or has ‘earned the right’ to that role (Hoffman and Spitzer 1985) offers are lower that in the traditional Ultimatum game. As I suggested before, in the presence of prototypical,

[17] Rejections remained low throughout, about 10%. All rejections were on offers of $2 or $3 in the exchange instructions, there was no rejection in the contest entitlement/divide $10, and 5% rejection of the $3 and $4 offers in the random assignment/divide $10.

acceptable justifications for deviating from equality, subjects will be induced to follow an equity principle. Framing in this case provides salient cues suggesting that an equity rule is appropriate to the situation.

Since, from a formal point of view, these situations are not different from that of a traditional Ultimatum game, the previous analysis remains the same. Hence, within the Fehr-Schmidt model, one has to argue that the framing of the game decreases α_2. In other words, the role of a "buyer" or the knowledge that the Proposer was a superior performer or had simply earned the right to his role lowers the Responder's concern for fairness. This does not sound intuitive, and demands some explanation. In addition, the Proposer has to actually *expect* this change in order to lower his offer. It is equally, if not more difficult, to see why the framing can lead to different beliefs the Proposer has about the Responder.

In my model, the parameter N plays a vital role again. Although we need more studies about how and to what extent framing affects people's expectations and perception of what norm is being followed, it is intuitively clear that framing, like the examples mentioned above, will change the players' conception of what is fair. The 'exchange' framework is likely to elicit a market script where the seller is expected to try to get as much money as possible, whereas the entitlement context has the effect of focusing subjects away from equality in favor of an equity rule. In both cases, what has been manipulated is the perception of the situation, and thus the expectations of the players. An individual's sensitivity and concern for norms may be unchanged, but the relevant norm is clearly different from the usual 'fairness as equality' rule.

Games with Computers

To better understand the impact of norms on behavior in Ultimatum games, it is useful to look at experiments in which expectations are irrelevant. Such games are typically played against a computerized opponent. Blount (1995) performed a one-shot Ultimatum game experiment in which Responders played against a computer making random offers as well as against human Proposers. In these games, subjects knew when they were paired with computers or humans. Subjects rejected, as usual, low offers from humans, but rarely rejected low offers coming from the computer. The computer has no expectation that its human opponent will follow a norm, and the player has no reason to expect that the computer will follow a norm, be fair, or have any intention whatsoever. As a result, human players quickly begin to play as predicted by the standard theory.

Dictators with Uncertainty

In a theory of norms, the role of expectations is crucial. Norms and expectations are part of the same package. Focusing people on a norm usually means eliciting certain expectations and, in turn, when people have the right empirical and normative expectations they will tend to follow the relevant norm. In the traditional Ultimatum game, at least in Western societies, the possibility of rejection forces the Proposer to focus upon what is expected of her.[18] In the absence of information about the Responder, and without a history of previous games and results as a guide, equal (or almost equal) shares become a focal point. Eliminate the possibility of rejection, and equality becomes much less compelling: for example, we know that when the Dictator game is double blind, 64% of the Proposers keep all the money. The Dictator game is particularly interesting as a testing ground for the study of how norms influence behavior, since it illustrates in a clear manner how sensitive we are to the presence, reminder, or absence of others' expectations.

Because I always thought that it was not at all obvious what one should choose in a Dictator game, and I did not find an 'equal share' compelling, I was curious to know what people perceive as the 'normal' thing to do in such games, and whether it is different from what they think the 'right' thing to do is. I thus ran a questionnaire on 126 undergraduate students at Carnegie Mellon University (see Appendix 2). The students were all enrolled in Philosophy 80-100, a course that almost every student takes, irrespective of his or her major. What the answers would do, I thought, was to define a baseline perception of the Dictator game. Interestingly, there was no overall consensus about what to do and what 'most people' were expected to do. Almost 56% of the students thought that 10-0 would be the most common allocation, and only 13% thought 5-5 to be the norm. Furthermore, 46 of the 70 (65.7%) who thought that 10-0 would be the most common allocation also felt that such allocation was not unfair. When explicitly asked about the 'fair' allocation, 68% felt that 5-5 was fair, but a sizable 21.4% thought 10-0 to be fair. As for the unfair question, almost 56% felt that nothing was unfair or greedy. The Dictator game seems thus to be a situation in which there is no obvious norm to follow and, because of that, it is an excellent testing ground for the role expectations (and their manipulation) can play in the emergence of a consensual script and, consequently, a social norm.

A recent experiment done by Dana, Weber and Kuang (2003) enlightens this point. The basic setting is a Dictator game where the allocator has only two options. The game is played in two very different situations. Under the "Known Condition" (KC), the payoffs are unambiguous, and the allocator has to choose between option A: (6, 1) and option B: (5, 5), where the first number in the pair is the allocator's payoff, and the second number is the receiver's payoff. Under the "Unrevealed Condition" (UC), the allocator is to choose between option A: (6, ?) and option B: (5, ?), where the receivers's payoff is 1 with probability 0.5 and 5 with probability 0.5 (Figure 5). Before the allocator makes a choice, however, she is given the option to privately find out at no cost which game is being played and thus know what the receiver's payoff is.

[18] I do not want to imply that sanctions are crucial to norm-following. They may just reinforce a tendency to obey the norm and serve the function -- together with several other indicators -- of focusing individuals' attention on the particular norm that applies to the situation.

It turns out that 74% of the subjects choose B: (5, 5) in KC, and 56% choose A:(6, ?) without revealing the actual payoff matrix in UC.

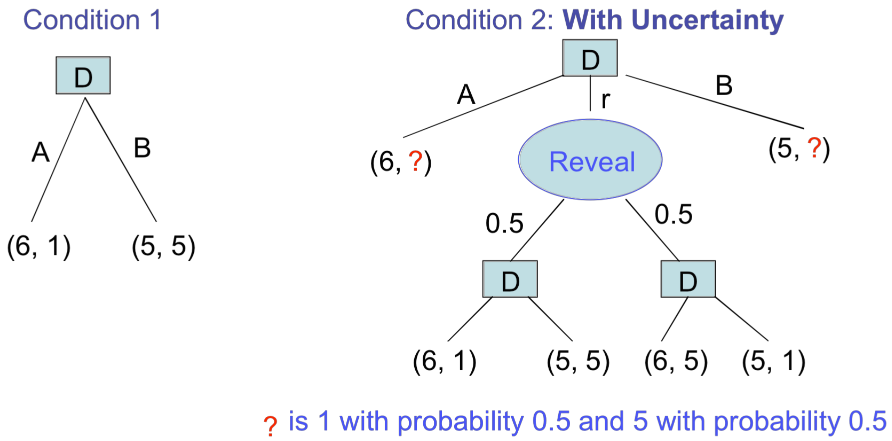


Figure 3.5

This result, as Dana et al. point out, stands strongly against the Fehr-Schmidt model. If we take the revealed preference as the actual preference, choosing (5, 5) in KC implies that $\beta_1 > 0.2$, while choosing (6, ?) without revealing in UC implies that $\beta_1 < 0.2$.[19] Hence, unless a reasonable story could be told about β_1, the model does not fit the data. If a stable preference for fair outcomes is inconsistent with the above results, can a conditional preference for following a norm show greater consistency? Note that, if we were to *assume that* N_i *is fixed* in both experiments, a similar change of k would occur in my model, too.[20] However, the norm-based model can offer a natural explanation of the data through an interpretation of N_i. In KC, subjects have only two, very clear choices. There is a 'fair' outcome (5,5) and there is an inequitable one (6,1). Choosing (6,1) entails a net loss for the receiver, and only a marginal gain for the allocator. A similar situation, and one that we frequently

---

[19] In KC, choosing option B implies that U_1 (5,5) >U_1 (6,1), or 5-α_1 (0) > 6-β_1 (5). Hence, 5 > 6-5-β_1 and therefore $\beta_1 > 0.2$. In UC, not revealing and choosing option A implies that U_1 (6, (.5(5), .5(1))) >U_1 (.5(5,5), .5(6,5)), since revealing will lead to one of the two 'nice' choices with equal probability. We thus get 6 -.3(β_1) > 2.5 + .5(6-β_1), which implies that $\beta_1 < 0.2$.

[20] According to my model, if we keep N_i constant, choosing option B in KC means that U_1 (5,5) >U_1 (6,1), hence 5 > 6-k_1 (4). It follows that $k_1 > 0.25$. In UC, not revealing and choosing option A implies that U_1 (6, (.5(5), .5(1))) >U_1 (.5(5,5), .5(6,5)), hence 6-k_1 (2) > 5.5, which implies that $k_1 < 0.25$.

encounter, is giving to the poor or otherwise disadvantaged. In Dana's example, what is $1 more to the allocator is $4 more to the receiver, mimicking the multiplier effect that money has for a poor person. In this experiment, what has probably been activated is a norm of beneficence, and subjects uniformly respond by choosing (5,5). Indeed, when receivers in Dana's experiment were asked what they would choose in the allocator's role, they unanimously chose the (5,5) split as the most appropriate. Interestingly, in a related experiment (Dana et al. 2003), in the presence of uncertainty all of the receivers believed that the most frequently chosen option would be the most unfavorable to them, indicating that there is a consensus about when equal shares are to be expected and when they are not.

A natural question to ask is whether we should hold *N* fixed, thus assuming a variation in people's sensitivity to the norm (*k*), of if instead what is changing here is the perception of the norm itself. I want to argue that what changes from the first to the second experiment is the perception that a norm exists and applies to the present situation, as well as expectations about other people's behavior and what their expectations about one's own behavior might be. Recall that in my definition of what it takes for a norm to be followed, a necessary condition is that a sufficient number of people expect others to follow it in the appropriate situations *and* believe they are expected to follow it by a sufficient number of other individuals. People will *prefer* to follow an existing norm *conditionally* upon entertaining such expectations. In KC, the situation is transparent, and so are the subjects' expectations. If a subject expects others to choose (5,5) and believes she is expected so to choose, she might prefer to follow the norm (provided her *k*, which measures her sensitivity to *N,* is large enough).[21] In UC, on the contrary, there is uncertainty as to what the receiver might be getting. To pursue the analogy with charitable giving further, in UC there is uncertainty about the multiplier ("am I giving to a needy person or not?") and thus there is the opportunity for *norm evasion*: the player can avoid activating the norm by not discovering the actual payoff matrix. Though there is no cost to see the payoff matrix, people will opt to not see it in order to avoid having to adhere to a norm that could potentially be disadvantageous. So a person who chooses (5, 5) under KC may

[21] A similar example of focusing on the 'fair' outcome is provided by a two-part experiment conducted by Kahneman et al. (1986). In the first task, subjects in a Dictator game had to choose between two possible allocations of $20: either one could keep $18 and give $2 to an anonymous Responder, or one could split $20 evenly. A lottery selected eight pairs (out of eighty) to actually be paid. Subjects chose to divide the money evenly 76% of the time. In the second part of the experiment, the same subjects were presented with another choice. This time subjects had to decide between splitting different amounts of money with one of two subjects who had previously played the game with somebody else. Either subjects could split $12 evenly with another subject, who had chosen to keep $18 in the first part of the experiment, or they could halve $10 with a subject who had divided equally the sum of money in the first part of the experiment. Most (74%) of the subjects preferred to split the money with the person who had previously acted fairly. They were clearly condemning unfair behavior by preferring to lose $1 rather than to split a greater sum with someone who had acted unfairly.

choose (6,?) under UC with the same degree of concern for norms. Choosing to reveal looks like what moral theorists call a *supererogatory* action. We are not morally obliged to perform such actions, but it is awfully nice if we do. Indeed, I believe few people would expect an allocator to choose to reveal, and similarly I would expect few people would be willing to punish an allocator who chooses to remain in a state of uncertainty.[22]

A very different situation would be one in which the allocator has a clear choice between (6,1) and (5,5), but she is told that the prospective receiver *does not even know* he is playing the game. In other words, the binary choice would focus the allocator, as in the KC condition, on a norm of beneficence, but she would also be cued about the absence of a crucial expectation. If the recipient does not expect her to give anything, is there any reason to follow the norm? This is a good example of what I have extensively discussed in (Bicchieri 2000, 2006). A norm exists, the subject knows it and knows she is in a situation in which the norm applies, but her preference for following the norm is conditional on having certain empirical and normative expectations (see Appendix 1). In our example, the normative expectations are missing, since the recipient does not know that a Dictator game is being played, and his part in it. In this case, I predict that a large majority of allocators will choose (6,1) with a clear conscience. This prediction is different from what a 'fairness preference' model would predict, but it is also at odds with theories of social norms as 'constraints' on action. One such theory is Rabin's (1995) model of moral constraints. Very briefly, Rabin assumes that agents maximize expected utility subject to constraints: Thus our allocator will seek to maximize her payoffs but experience disutility if her action is in violation of a social norm. However, if the probability of harming another is sufficiently low, a player may 'circumvent' the norm and act more selfishly. Since in Rabin's model the norm functions simply as a constraint, beliefs about others' expectations play no role in a player's decision to act. Because the (6,1) choice does in fact 'harm' the recipient, Rabin's model should

[22] It should be stressed here that there are several ways in which a person might be focused on expectations that induce more generous behavior. For example, we know that being able to look at one's partner, or to communicate with him, has an effect on how much is allocated. In experimental variations in which the allocator can look the prospective receiver in the face or is allowed to talk to him, an offer of half the money is the norm. For example, Frey and Bohnet (1993) describe three Dictator game experiments in which allocators were given CHF 13 to keep or share with a receiver. In each experiment, a different level of interpersonal identity was made salient. In the first experiment, allocator and receiver were unknown to one another and the mean amount of money given by the allocator to the receiver was CHF 3.38. In the second experiment, the partners faced one another, but were not allowed to communicate. Allocators gave an average of CHF 6.25 to recipients. In the third experiment, subjects were given the choice of whether or not to communicate with each other. The majority (75%) chose communication; the mean allocation from allocators in this majority was CHF 5.70. Interestingly, when allocators in the third experiment answered a questionnaire concerning the reasons for their decisions, most cited binding agreements or commitments with their partner as the reason for their choice of allocation. A face-to-face encounter, or the possibility of communication, evidently generates a cognitive and emotional shift of attention. As social distance between the parties dwindles, one is forced to focus upon the reasonable expectations of the other party.

predict that the number of subject who choose (6,1) is the same as in the KC of Dana's experiment. In my model, however, the choices in the second experiment will be significantly different from the choices we have observed in Dana's KC condition.

To summarize, the norm-based model explains the behavioral changes observed in the above experiments as due to a (potentially measurable) change in expectations. An individual's propensity to follow a given norm would remain fixed, as would her preferences. However, since preferences in my model are conditional upon expectations, a change in expectations will have a major, predictable effect on behavior.

• Appendix 1

Conditions for a social norm to exist

Let R be a *behavioral rule* for situations of type S, where S can be represented as a mixed-motive game. We say that R is a social norm in a population P if there exists a sufficiently large subset $P_{cf} \subsetneq P$ such that, for each individual $i \in P_{cf}$:

Contingency: i knows that a rule R exists and applies to situations of type S;
Conditional preference: i prefers to conform to R in situations of type S on the condition that:
(a) *Empirical expectations*: i believes that a sufficiently large subset of P conforms to R in situations of type S;
and either
(b) *Normative expectations*: i believes that a sufficiently large subset of P expects i to conform to R in situations of type S;
or
(b') *Normative expectations with sanctions*: i believes that a sufficiently large subset of P expects i to conform to R in situations of type S, prefers i to conform and may sanction behavior.

A social norm R is *followed* by population P if there exists a sufficiently large subset $P_f \subsetneq P_{cf}$ such that, for each individual $i \in P_f$, conditions 2(a) and either 2(b) or 2(b') are met for i and, as a result, i prefers to conform to R in situations of type S.

Appendix 2

With the help of Jason Dana, I conducted the following survey in 5 sections of 80-100. The sections ranged in size from 20-30 respondents. We had 126 responses (one person did not answer what was the most common allocation, two did not answer what was the fairest allocation).

Imagine that the conductors of this survey give a survey respondent (call them **person A**) 10 dollar bills and the following instructions: Another survey respondent (**person B**) has been paired with you randomly. This pairing is anonymous, meaning that we will not inform you who you are paired with, nor will we inform the person you are paired with who you are. Your task is to distribute the 10 dollars between yourself and the person you are paired with in any way that you want. That means that you may keep or give away all of the bills, or take any action in between. Your choice would be final; you keep as many of the bills as you wanted and the rest are given to the other person.

What is the thing that person A ought to do in this situation? That is, what action would you consider fair and reasonable? Please indicate below:
Keep ___ bills and give away ___ bills
(these numbers should add up to 10)

Are there any actions that person A could take that you would consider excessively greedy or unfair? If so, how many bills would he/she have to keep to be greedy? Please indicate below:
Keep ___ bills and give away ___ bills or circle: any action is fair

What do you expect that most people in the position of person A would do? That is, which division would be most common? Please indicate below:
Keep ___ bills and give away ___ bills

Now consider two people involved in such a situation. Imagine that you, as a third party, are allowed to look at a proposed division of $10 like the one above, between two people anonymous to you. In this situation, you have the right to inspect the offer and "**accept**" or "**reject**" it. If you **accept** the offer, the two people get the amounts of money proposed by person A. If you **reject** the offer, *both* people will receive zero dollars. You neither gain nor lose money by accepting or rejecting. Thus, you have no personal monetary stake in the outcome. For the divisions listed below, indicate which, if any, you would reject.

(please circle accept or reject – accept means divide as proposed, reject means

both get zero)

Person A keeps 5 dollars, and gives 5 to B.	Accept	Reject
Person A keeps 2 dollars, and gives 8 to B.	Accept	Reject
Person A keeps 1 dollar, and gives 9 to B.	Accept	Reject
Person A keeps 10 dollars, and gives 0 to B.	Accept	Reject
Person A keeps 8 dollars, and gives 2 to B.	Accept	Reject
Person A keeps 6 dollars, and gives 4 to B.	Accept	Reject
Person A keeps 3 dollars, and gives 7 to B.	Accept	Reject
Person A keeps 9 dollars, and gives 1 to B.	Accept	Reject
Person A keeps 7 dollars, and gives 3 to B.	Accept	Reject
Person A keeps 4 dollars, and gives 6 to B.	Accept	Reject
Person A keeps 0 dollars, and gives 10 to B.	Accept	Reject

Here is a summary of the results of the survey.

For the "common" response, a strong mode of almost 56% felt that 10-0 would be the most common allocation. The rest of the responses were spread fairly evenly between 5-9, with the next biggest bump being almost 13% choosing 5-5 as the most common. 46 of the 70 (65.7%) who thought that 10-0 would be most common also felt that nothing was unfair.

For the "ought to" or fair question, a strong mode of 68.3% felt that 5-5 was fair. A smaller bump of 21.4% felt that 10-0 was what ought to be done; very few people indicated anything but these two responses.

As for the unfair question, almost 56% felt that nothing was unfair. Of these 70 who felt that nothing was unfair, 46 (65.7%) thought that 10 would be the most common answer. Almost 20% felt that 10-0 was unfair, and about 15% felt that keeping 6 or more was unfair. Of those 56 who did think that something was greedy, 43 (76.8%) went on to punish at least some allocation choice. Of the 70 who thought nothing was unfair, 24 (34%) still punished at least some allocation.

For the punishment option, 54% chose to punish at least one allocation. More than 48% would punish a 10-0 offer, 31.7% would punish a 9-1 offer, and 26.9% would punish an 8-2 offer. 19% displayed some pure inequity aversion, preferring to punish some offers keeping more than 5 as well as some keeping less than 5. 4 of the 126 respondents would punish only for giving money away, but not for keeping 10.

Whether or not some offers were judged greedy was most strongly related to punishment. Of interest is that a logistic regression of a dummy punish/no punish variable on both common and fairest variables shows that common is a significant predictor while fairest is not. Thus, people are more likely to punish an offer that they think violates a descriptive norm than one that violates their sense of what is fair.

References

Andreoni, J. (1990). "Impure Altruism and Donations to Public Goods: A Theory of Warm-Glow Giving." *The Economic Journal* **100**: 464-477.

Andreoni, J. (1995). "Cooperation in public-goods experiments: Kindness or confusion?" *American Economic Review* **90**: 166-193.

Bicchieri, C. (1999). "Local Fairness." *Philosophy and Phenomenological Research* **LIX**(1): 229-236.

Bicchieri, C. (2000). Words and Deeds: a Focus Theory of Norms. *Rationality, Rules and Structure*. J. Nida-Rumelin and W. Spohn. Dordecht, Kluwer Academic Publishers.

Bicchieri, C. and J. Zhang (2004). "Ultimatum Behavior: Fairness Preferences or Fairness Norms?" *working paper, Carnegie Mellon Department of Philosophy*.

Bicchieri, C. (2006). *The Grammar of Society: The nature and Dynamics of Social Norms*. Cambridge: Cambridge University Press.

Blount, S. (1995). "When Social Outcomes Aren't Fair: the effect of causal attributions on preferences." *Organizational Behavior and Human Decision Processes* **63**: 131-144.

Calvino, I. (1951). *The nonexistent knight and the cloven viscount*, Giulio Einaudi Editore.

Camerer, C. and R. H. Thaler (1995). "Anomalies: Ultimatums, dictators, and manners." Journal of Economic Perspectives **9**(2): 209-219.

Camerer, C. (2003). *Behavioral Game Theory: Experiments on Strategic Interaction*. Princeton, NJ, Princeton University Press.

Camerer, C., G. Loewenstein, et al. (2004). *Advances in Behavioral Economics*. Princeton, Princeton University Press.

Cameron, L. (1995). "Raising the Stakes in the Ultimatum Game: Experimental Evidence from Indonesia." *Working paper, Princeton Department of Economics - Industrial Relations Sections* **345**.

Cialdini, R., C. Kallgren, et al. (1990). "A focus theory of normative conduct: a theoretical refinement and reevaluation of the role of norms in human behavior." *Advances in Experimental Social Psychology* **24**: 201-234.

Dana, J., R. Weber, et al. (2003). "Exploiting Moral Wriggle Room: Behavior Inconsistent with a Preference for Fair Outcomes." *Carnegie Mellon Behavioral Decision Research Working Paper* **349**.

Falk, A., E. Fehr, et al. (2000). "Informal Sanctions." *IEER Working Paper No. 59*.

Fehr, E. and K. Schmidt (1999). "A Theory of Fairness, Competition, and Cooperation." *The Quarterly Journal of Economics* **114**(3): 817-868.

Fehr, E. and S. Gachter (2000). "Fairness and Retaliation: The Economics of Reciprocity." *Journal of Economic Perspectives* **14**(3): 159-181.

Fehr, E. and S. Gachter (2000). "Cooperation and Punishment in Public Goods Experiments." *American Economic Review* **90**(4): 980-994.

Fehr, E. et al. (2003). "Testing Theories of Fairness - Intentions Matter". Institute for Empirical Research in Economics, University of Zürich, Working Paper No. 63
Forsythe, R., J. L. Horowitz, et al. (1994). "Fairness in Simple Bargaining Experiments." *Games and Economic Behavior* **6**(3): 347-369.

Frey, B. and I. Bohnet (1995). "Institutions affect fairness: Experimental investigations." *Journal of Institutional and Theoretical Economics* **151**(2): 286-303.

Frey, B. and I. Bohnet (1997). "Identification in Democratic Society." *Journal of Socio-Economics* **XXVI**: 25-38.

Geanakoplos, J., D. Pearce, et al. (1989). "Psychological games and sequential rationality." *Games and Economic Behavior* **1**: 60-79.

Guth, W., R. Schmittberger, et al. (1982). "An experimental analysis of ultimatum games." *Journal of Economic Behavior and Organization* **3**: 367-388.

Guth, W. (1995). "On ultimatum bargaining experiments: a personal review." *Journal of Economic Behavior and Organization* **27**: 329-344.

Heinrich, J. (2000). "Does Culture matter in economic behavior? Ultimatum game bargaining among the Machiguenga." *American Economic Review* **90**(4): 973-979.

Henrich, J., R. Boyd, et al., Eds. (2004). *Foundations of Human Sociality: Ethnography and Experiments in 15 Small-scale Societies*. Oxford, Oxford University Press.

Hoffman, E. and M. Spitzer (1985). "Entitlements, rights, and fairness: An experimental examination of subjects' concept of distributive justice." *Journal of Legal Studies* **2**: 259-297.

Hoffman, E., K. A. McCabe, et al. (1994). "Preferences, Property Rights, and Anonymity in Bargaining Games." *Games and Economic Behavior* **7**: 346-380.

Hoffman, E., K. A. McCabe, et al. (1998). "Behavioral foundations of reciprocity: Experimental economics and evolutionary psychology." *Economic Inquiry* **36**: 335-352.

Kagel, J. H., C. Kim, et al. (1996). "Fairness in ultimatum games with asymmetric information and asymmetric payoffs." *Games and Economic Behavior* **13**: 100-110.

Kahneman, D., J. Knetsch, et al. (1986). "Fairness as constraint on profit seeking: Entitlements in the market." *American Economic Review* **76**(4): 728-741.

Rabin, M. (1993). "Incorporating Fairness into Game Theory and Economics." *American Economic Review* **83**: 1281-1302.

Rabin, M. (1995). “"Moral Preferences, Moral Constraints, and Self-Serving Biases," Berkeley Department of Economics Working Paper No. 95-241, August 1995

Roth, A. E., V. Prasnikar, et al. (1991). "Bargaining and Market Behavior in Jerusalem, Ljubljana, Pittsburgh, and Tokyo: An Experimental Study." *American Economic Review* **81**(5): 1068-1095.