An Experimental Study of Asymmetric Reciprocity

Omar Al-Ubaydli and Min Sok Lee

July 2008

Discussion Paper

Interdisciplinary Center for Economic Science
4400 University Drive, MSN 1B2, Fairfax, VA 22030
Tel: +1-703-993-4850 Fax: +1-703-993-4851
ICES Website: www.ices-gmu.org
ICES RePEc Archive Online at: http://edirc.repec.org/data/icgmuus.html

An experimental study of asymmetric reciprocity[1]

Omar Al-Ubaydli and Min Sok Lee[2]

July 2008

Abstract

Do people have a stronger propensity to reward or punish? When reacting to intentions, Offerman (2002) concluded that people punish more. Using the Falk and Fischbacher (2006) model, we extend Offerman's design in two ways. First, we control for the strength of the positive/negative intentions to which an individual reacts when rewarding/punishing. Second, we can precisely compare the strength of intention- and distribution-based motives for reward/punishment. Doing so requires measuring second-order expectations of subjects' own behavior, i.e., what a subject predicts that other subjects predict that he will do. Second-order expectations can be elicited directly or they can be induced by telling a subject what others expect him to do. Under elicited second-order expectations, we find that negative reciprocity is stronger than positive reciprocity, though if we isolate the distributional motive for reciprocity, then we find that positive reciprocity is stronger than negative reciprocity. Under induced second-order expectations, positive distributional reciprocity is stronger than negative distributional reciprocity while other forms of reciprocity are equally strong.

JEL code: C9 Keywords: reciprocity, reward, punishment

[1] We wish to thank Dan Houser, John List and Nat Wilcox for advice on the experimental design, and Jason Aimone for help in running the experiment. We wish to thank Lint Barrage and Theo Offerman for helpful comments.
[2] Al-Ubaydli (corresponding author): Department of Economics and Mercatus Center, George Mason University, 4400 University Drive MSN 3G4 Fairfax, VA 22030 USA. Email: omar@omar.ec. Tel: +1-703-459-5675. Fax: +1-703-993-1133. Lee: University of Chicago.

"*Let the punishment match the offence.*" Cicero

I. Introduction

Do people have a stronger propensity to reward or punish? Consider the following example. Tom can reward Jane by giving her $10 or punish her with a $10 fine. Both cost Tom $2. Suppose that Jane recklessly destroys Tom's car (*unkind*) vs. Jane gives Tom a ride to work (*kind*). The graveness of the unkind offence seems to be larger than the generosity of the kind gesture. Consequently, we would expect a higher likelihood of punishment in response to *unkind* than we would reward in response to *kind*. For the same reason, we would not interpret such a bias towards punishment as being indicative of a stronger intrinsic propensity to punish.

This example illustrates two important points. First, there is a sense in which we can objectively gauge how kind or unkind an action is. Second, it is important to control for kindness when attempting to infer reward and punishment propensities.

Offerman (2002) found that, controlling for distributional concerns, negative intentions are more likely to induce punishment than positive intentions are to bring about reward ("hurting hurts more than helping helps"). However Offerman's design did not permit him to either control for kindness or ensure that it was exogenous. We seek to extend his design by controlling for kindness.

There are several models of social preferences that objectively define kindness.[3] The Falk and Fischbacher (2006) model is particularly attractive because it combines both intention- and distribution-based motives for reward and punishment and does so in a tractable way.

We collect data in the spirit of Offerman (2002) and use the Falk and Fischbacher (2006) model to control for kindness. We run an experiment where subjects play a variant of the trust game and the dictator game. This permits recovery of the parameters that govern both intention- and distribution-based motives for reciprocity. A key control is subjects' second-order expectations of their own actions, i.e., what they think that others think that they will do. These expectations can either be elicited directly or induced by telling subjects what others think that they will do. We collect data using both.

Our results differ to some extent by the method of controlling for second-order expectations. Under elicitation, like Offerman (2002), we find that negative intentional reciprocity is stronger than positive intentional reciprocity. In addition, we find that positive distributional reciprocity is stronger than negative distributional reciprocity, but that in net terms, negative reciprocity is still stronger than positive reciprocity. Under inducement, the only asymmetry we detect is that positive distributional reciprocity is stronger than negative distributional reciprocity.

This paper's contributions are three-fold. First, we highlight a potential problem in simply comparing the incidence of reward and punishment to infer people's relative propensities for each. Second, by specifying a structural model, we are able to parse different forms of reciprocity. For example, Tom's propensity to punish Jane for destroying his car is likely to be lower if he found out that she was forced to do so at gunpoint, i.e., when her intentions do not

[3] For example, Fehr and Schmidt (1999), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006)

seem as bad. We are able to estimate positive and negative distributional, intentional and net reciprocity. Offerman's (2002) design only permits him to infer the relative size of positive and negative intentional reciprocity and not the exact size of either. Finally, to our knowledge, this is the first attempt at estimating the parameters of the Falk and Fischbacher (2006) model using second-order expectations.

The remainder of this paper is organized as follows. Section II is the background and existing literature. Section III is the experimental design. Section IV is the results. Section V is the summary and discussion.

II. Background and existing literature

A. The judgment game

Figure 1 is the judgment game in Offerman (2002). The key features are that it is one-shot and that player 2 plays after seeing player 1's move. Like a trust game, the efficient outcome requires player 1 to trust player 2. Like the ultimatum game, player 2 can punish player 1. This game is a simple illustration of what are formally known as reward and punishment.

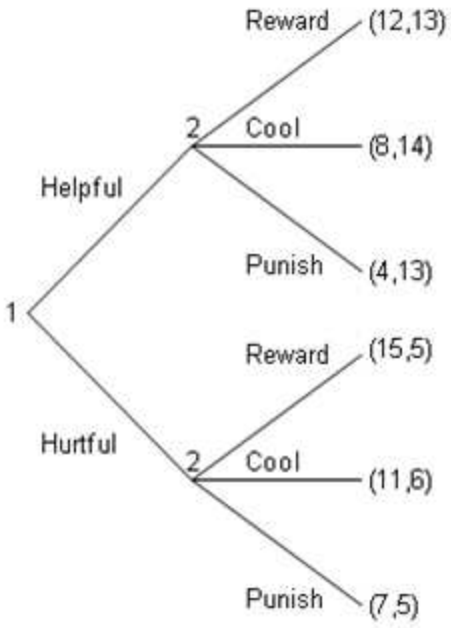

Figure 1: The judgment game in Offerman (2002)

Figures in parentheses are payoffs of player 1 and player 2, respectively. Figures without parentheses indicate which player makes a decision at that node.

Definition: *rewards* are deviations from 2's best response that increase 1's payoff. *Punishments* are deviations from 2's best response that decrease 1's payoff.[4]

Under neoclassical preferences, the game's equilibrium is for player 2 to always play cool and for player 1 to play hurtful. Offerman (2002) found that only 2 out of 28 pairs playing the game actually played the equilibrium. Moreover, 10 out of 12 pairs saw player 2 punishing an unkind move, and 12 out of 16 saw 2 rewarding a kind move. The neoclassical model is not a satisfactory description of the data.

[4] The layman's term 'punishment' potentially has two applications in this game. First, the profitable denial of a gain to 1; second, the costly imposition of a loss on 1. Under our definition, only the latter is referred to as a punishment. Positive transfers are rewards and negative transfers are punishments. In an employer-worker gift-exchange game (e.g., Fehr et al. (1993)), 2 can only reward 1 or play his best response (zero effort). 2 may describe supplying zero effort as a punishment in response to an unsatisfactory wage offer, but this is not a punishment in our terms. The same is true in the classic trust game (Berg et al. (1995)). Conversely, in the ultimatum game, 2 can only punish 1 or play his best-response. He may consider accepting a derisory offer as some kind of concession towards 1, but it does not constitute a reward under our definition.

B. Existing literature

Numerous psychology and economics papers have discussed the factors that affect reward/punishment decisions in one-shot environments such as the judgment game above.[5] Figure 2 below summarizes the components of the decision:

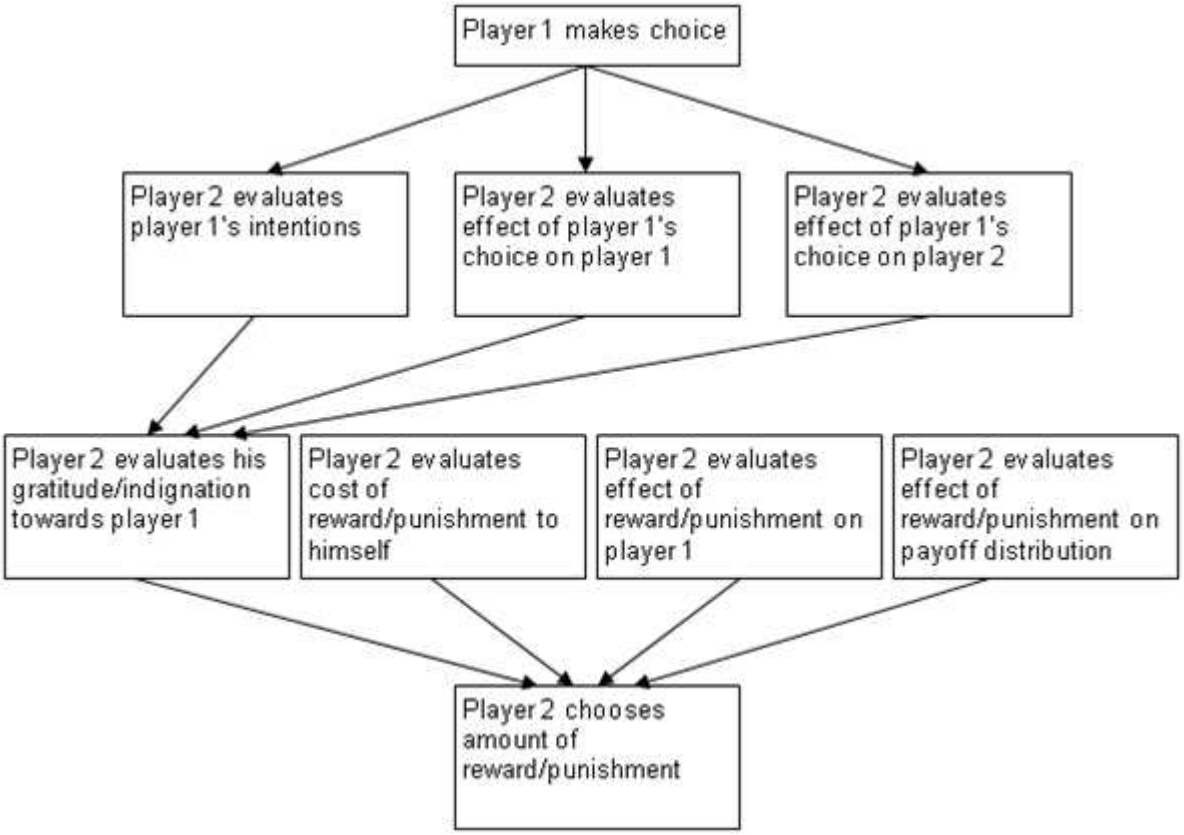

Figure 2: A model of reward/punishment decisions

After player 1 makes his choice, player 2 takes an emotional stance towards player 1. In the case of a kind gesture, player 2 would take into account the cost incurred by player 1 and the effect of the choice on player 2's payoff. Less obvious is the importance of intentions. Several studies

[5] From the psychology literature, see for example, Tesser et al. (1968), Greenberg and Frisch (1972), Blount (1995), McCullough et al. (2001), Ames et al. (2004). From the economics literature, see Geanakoplos et al. (1989), Rabin (1993), Levine (1998), Fehr and Schmidt (1999), Bolton and Ockenfels (2000), Charness and Rabin (2002), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006), Cox et al. (2007).

(Blount (1995), Offerman (2002)) have demonstrated that whether 1's action was accidental or fully intentional affects how grateful or indignant player 2 will be in response.

After establishing his emotional stance, player 2 then takes into account the cost of reward/punishment and its effect. Distributional considerations can enter the final decision too (Fehr and Schmidt (1999), Bolton and Ockenfels (2000)). Player 2 then makes his reward/punishment choice.

Offerman (2002) wants to answer the question: what has a larger effect on reward/punishment out of positive and negative intentions in the judgment game? There are two treatments: in the 'flesh and blood' treatment (henceforth the human treatment), both players make their own choices, while in the 'nature' treatment, it is common knowledge that a device makes player 1's choice on his behalf randomly.

Offerman then identifies the causal effect of intentions by comparing player 2's reward/punishment decisions across treatments. For example, he finds that 83% punish a play of hurtful in the human treatment vs. 17% in the nature treatment (p-value < 1%). In contrast, people reward a play of helpful with the same frequency across the two treatments (75% in human vs. 50% in nature; statistically insignificant).[6]

Offerman concludes that negative intentions have a stronger effect on the probability of reward/punishment than positive intentions. His explanation is based on the idea of self-serving expectations. People hold higher opinions of themselves than others do. They feel as though they are more deserving of kind gestures – and less deserving of unkind gestures – than others. This generates an asymmetric response to positive and negative intentions.

[6] He also compares emotional states (elicited verbally) across treatments.

C. Extending the existing literature

Offerman implicitly acknowledges that there are factors outside intentions that determine reward/punishment. These include how much player 1 has given up to be nice, or the benefit to player 2 of player 1's kind gesture. However he does not model them explicitly. He assumes that either they are constant or exogenous with respect to intentions when comparing choices across treatments.

In other words, if player 2 is angrier and more likely to punish player 1 when player 1 plays hurtful on purpose vs. accidentally, then is it only because the intent has changed? Is there a systematic change in one of the other determinants of player 2's anger?

There are several models of emotional state, and when they are applied to the current game they cast doubt on Offerman's exogeneity assumption.[7] We focus on the Falk and Fischbacher (2006) model (henceforth FF) for two reasons: first, it allows for both intention- and distribution-based motivation for reciprocity. Second, it has a specific enough functional form for estimation.

In FF, player 2's preferences in the judgment game take the following form:

$$u_2 = \pi_2 + G(\pi_2'' - \pi_1'')(\pi_1 - \pi_1'')\,, G \geq 0$$

π_i is the material payoff of player i. G is the reciprocity parameter – it drives reward and punishment: the larger G, the larger player 2's propensity to reward and punish. π_i'' is player 2's second-order expectation (henceforth SOE) of i's payoff, i.e., it is what he thinks that player 1 thinks that i will earn.

[7] Rabin (1993), Dufwenberg and Kirchsteiger (2004), Falk and Fischbacher (2006).

Definition: in the FF model, player 2's perception of player 1's kindness towards him at player 2's decision node is $k_{21} = \pi_2'' - \pi_1''$.

If player 2 thinks that player 1 expects both their payoffs to be equal, then player 2 will regard this as neutral behavior and will optimize neoclassically. If player 2's SOE implies a higher payoff for player 2 than player 1, player 2 regards player 1's choice as kind. The term $(\pi_1 - \pi_1'')$ implies that at the margin, player 2 wants to reward kind behavior $(\pi_2'' > \pi_1'')$. Conversely, player 2 wants to punish unkind behavior $(\pi_2'' < \pi_1'')$.

Note that the absolute size of positive vs. negative kindness is not arbitrarily defined in the FF model – it is based on principles external to the model. Reciprocity is meaningfully symmetric in the following sense: suppose that reward and punishment cost the same and their effect on player 1's payoff are equal in size and opposite in direction (which occurs in the judgment game). For example if it costs player 2 $1 to increase player 1's payoff by $2 or to decrease player 1's payoff by $2. Let the kindness in situation A be $k^A > 0$ and in situation in B be $k^B < 0$, and let $k^A = -k^B$. Then player 2 will reward in situation A with the same likelihood as he will punish in situation B.

Distributional concerns in the FF model are reflected in the definition of kindness, i.e., the relative payoff of the two players. Intentions affect utility via SOE. They are also reflected by the reciprocity parameter G which actually takes one of two values $\{\gamma, \tilde{\gamma}\}$, where $\gamma \geq \tilde{\gamma} \geq 0$, depending on the choices available to player 1. Loosely speaking, $\tilde{\gamma}$ reflects the propensity to reciprocate when player 1 does not even have a choice, like the nature treatment of the judgment game. On the other hand, γ reflects the propensity when player 1 has a choice, like the human treatment of the judgment game.

$\tilde{\gamma}$ is a measure of the strength of distributional reciprocity. To see why suppose Fred and Wilma are playing a dictator game where Wilma is the dictator and she has $\tilde{\gamma} > 0$. Fred starts with \$0 and Wilma starts with $\$Z$. She may well choose to give more to Fred when $Z = 100$ than when $Z = 10$. This differential reciprocity cannot be the result of differential intentions since Fred's intentions are constant across the two scenarios. It is Wilma's concern over the distribution of material outcomes that drives the difference in behavior.

As such, γ is a measure of the total effect of distributional and intentional reciprocity, i.e., net reciprocity. We can therefore think of $(\gamma - \tilde{\gamma})$ as a measure of intentional reciprocity.[8]

To generate asymmetric reciprocity, we allow the values of $\{\gamma, \tilde{\gamma}\}$ to depend on the sign of k_{21}:

$$u_2 = \pi_2 + \begin{cases} A(\pi_2'' - \pi_1'')(\pi_1 - \pi_1'') \; if \; \pi_2'' < \pi_1'' \text{ (negative reciprocity)} \\ B(\pi_2'' - \pi_1'')(\pi_1 - \pi_1'') \; if \; \pi_2'' > \pi_1'' \text{ (positive reciprocity)} \end{cases}$$

Where $A \in \{\alpha, \tilde{\alpha}\}, B \in \{\beta, \tilde{\beta}\}$.

In light of FF's model and Offerman's explanation of asymmetric reciprocity, figure 3 is a refinement of the psychological model presented in figure 2. Like the model from figure 2, reward/punishment decisions are driven by player 2's subjective emotional state. This is now mediated by an objective evaluation of player 1's kindness combined with potentially self-serving expectations.

[8] The choice of distance metric is arbitrary; we use the absolute difference for simplicity. We experiment with other metrics in the empirical section below.

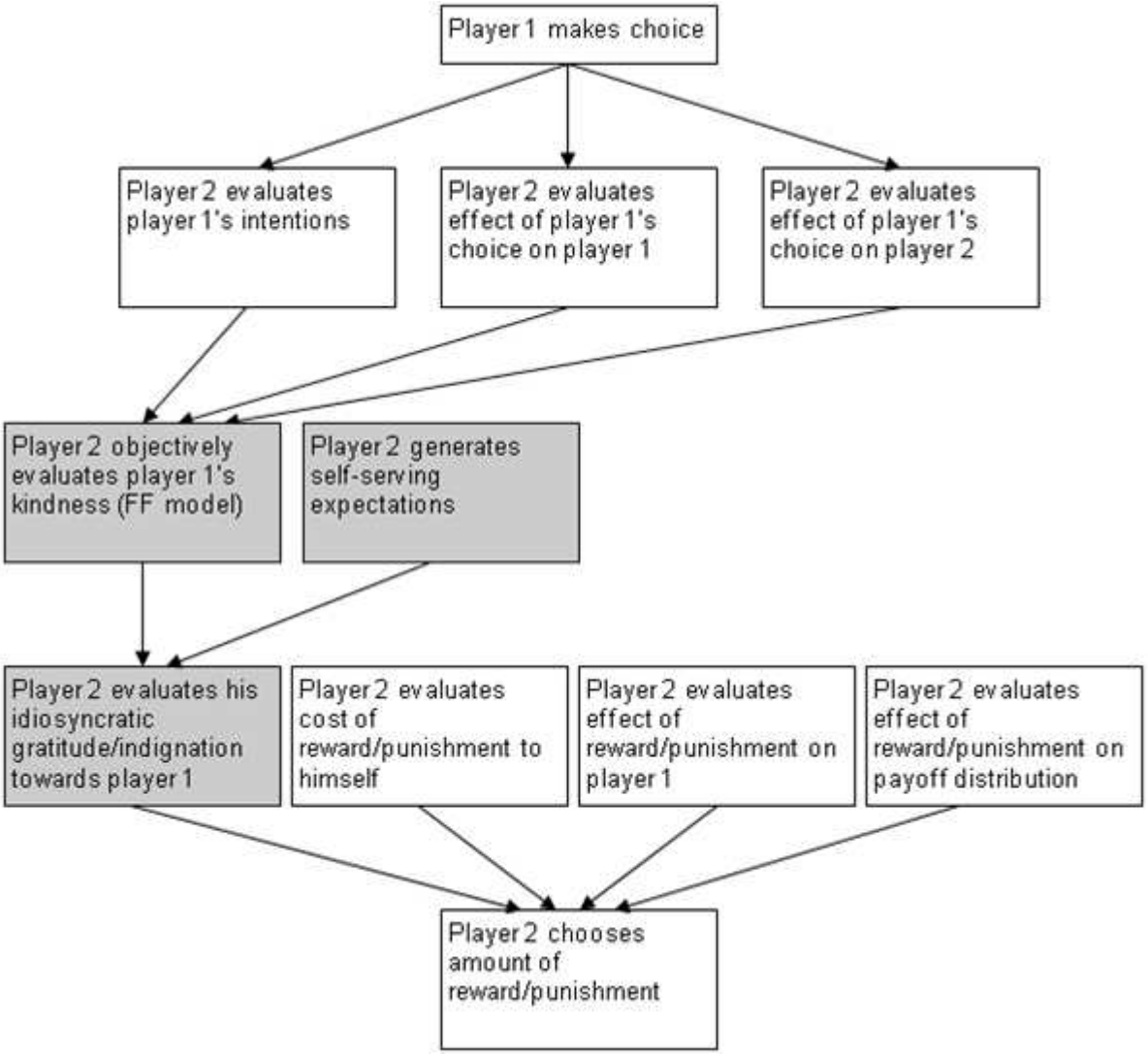

Figure 3: Refined model of reward/punishment decisions

Gray boxes are differences to figure 2

Using the FF model, one can examine three issues of asymmetric reciprocity.

Definition: asymmetric reciprocity

- *Asymmetric net reciprocity* occurs when $\alpha \neq \beta$.
- *Asymmetric distributional reciprocity* occurs when $\tilde{\alpha} \neq \tilde{\beta}$.
- *Asymmetric intentional reciprocity* occurs when $\alpha - \tilde{\alpha} \neq \beta - \tilde{\beta}$.

The decision to reward/punish is driven by several factors, many of which do not represent what we would regard as an intrinsic bias towards reward vs. punishment. The goal is to design an experiment that allows us to attribute asymmetric reward/punishment behavior to asymmetries in the way that an individual interprets objective kindness as opposed to asymmetries in the objective kindness itself. The Tom and Jane example presented in the introduction suggests that there is such a thing as objective kindness, and, among others, FF give us a formal way of modeling it. The distinction between objective and subjective kindness is not a vacuous one.

We can apply the FF model to Offerman's judgment game (shown in figure 1). Player 1 can either play helpful (e) or hurtful (u); denote this choice by $x \in \{e, u\}$. This choice can be done either by a human player 1 in the human (H) treatment or by a computer in the nature (N) treatment; denote the treatment by $y \in \{H, N\}$. Let $p_{x,y}$ denote the probability that player 2 plays cool given node x in treatment y, and let $p''_{x,y}$ denote player 2's SOE of this probability.[9] Accordingly, player 2 will reward helpful if and only if:

$$B > \left[4\left(1 + 5p''_{e,y}\right)\right]^{-1}$$

And player 2 will punish hurtful if and only if (see the appendix for details):

$$A > \left[4\left(2 + 3p''_{u,y}\right)\right]^{-1}$$

One interpretation of Offerman (2002) is that $\alpha - \tilde{\alpha} > \beta - \tilde{\beta}$ in the FF model. In the appendix, we show that Offerman's exogeneity assumption is arbitrary. While this does not imply that his findings are incorrect, it is interesting to control for the potential endogeneity.

[9] For expositional simplicity, we abstract from the possibility that player 2 rewards a play of hurtful or punishes a play of helpful. The analysis does not depend upon these assumptions.

It should be noted that Offerman designed and presented the judgment game in a way that attempted to generate symmetric objective kindness. Table 1 below shows his presentation of the game:

Choices	Player 1 payoff	Player 2 payoff
Choice of player 1		
Helpful	8	4
Hurtful	11	-4
Choice of player 2		
Reward	4	9
Cool	0	10
Punish	-4	9

Table 1: Presentation of judgment game in Offerman (2002)

Total payoffs of player i are calculated by summing player i's payoff across the two choice cells

In this presentation, choosing helpful gives player 2 $+4$ while choosing hurtful results in -4. Presumably, Offerman is trying to make the objective kindness of helpful equal to the objective unkindness of hurtful. Moreover, the cost/effect of reward/punishment is clearly symmetric. However as demonstrated in the appendix, under the FF model this does not suffice due to the role of SOE.

Several other studies generate reward and punishment data that can in principle be compared to evaluate differences in reward/punishment propensities.[10] However were one to use the data for such a purpose (which none of them do, except for Pereira et al. (2006) in a passing comment (p149)), one would encounter the endogeneity problems discussed above.

[10] For example: Sefton et al. (2001), Andreoni et al. (2003), Walker and Halloran (2004), Pereira et al. (2006) and Sutter et al (2006).

III. Experimental design

A. Null hypotheses

Consider a 2-player game of perfect information where player 1 plays first, making at most one move, and player 2 plays second making exactly one move. If player 1 does not make a move, then it is a dictator game; otherwise it is a non-dictator game. Note that in our dictator games, we assign player 2 the role of dictator.

In a non-dictator game, player 2's preferences are:

$$u_2 = \pi_2 + \begin{cases} \alpha k_{21}\pi_1 \ if \ k_{21} < 0 \\ \beta k_{21}\pi_1 \ if \ k_{21} > 0 \end{cases}$$

In a dictator game player 2's preferences are:

$$u_2 = \pi_2 + \begin{cases} \tilde{\alpha} k_{21}\pi_1 \ if \ k_{21} < 0 \\ \tilde{\beta} k_{21}\pi_1 \ if \ k_{21} > 0 \end{cases}$$

Recall that $\alpha \geq \tilde{\alpha} \geq 0, \beta \geq \tilde{\beta} \geq 0$.

Hypothesis 1 (symmetric net reciprocity): for any player 2 i, α_i and β_i are identically distributed.

Hypothesis 2 (symmetric distributional reciprocity): for any player 2 i, $\tilde{\alpha}_i$ and $\tilde{\beta}_i$ are identically distributed.

Hypothesis 3 (symmetric intentional reciprocity, Offerman (2002)): for any player 2 i, $(\alpha_i - \tilde{\alpha}_i)$ and $\left(\beta_i - \tilde{\beta}_i\right)$ are identically distributed.

To test these hypotheses, we need an estimate for $(\alpha_i, \tilde{\alpha}_i, \beta_i, \tilde{\beta}_i)$ for each person i. When i makes a decision based on the parameter $G_i \in \{\alpha_i, \tilde{\alpha}_i, \beta_i, \tilde{\beta}_i\}$, denote this decision by x_i^G. Let k_i^G denote i's perception of his partner's kindness towards him when i is choosing x_i^G.

B. Procedure

Subjects were recruited by email using a campus database at George Mason University. Each session had 18 or 20 subjects. Upon arrival, subjects drew a colored chip from a container that determined their role (blue or red; equal numbers). Blues and reds were then led to separate rooms where they were informed that they would be playing three different games. In each game, each blue was anonymously and randomly matched with a unique red. No two players were matched together for more than one game (perfect stranger). Blues always took the role of player 1 and reds always took the role of player 2. There was no communication between subjects. We ran eight such sessions in total.[11]

The first game (a non-dictator game) is a continuous, non-linear version of Offerman's (2002) judgment game. It is designed to elicit $\left(x_i^\alpha, x_i^\beta\right)$. The strategy sets are W for player 1 and X for player 2, and elements of X are (x^α, x^β). The payoffs and strategy sets are:

$$W = \{a, b\}, \qquad X = [0,12] \times [0,12]$$

$$\pi(t; x^\alpha, x^\beta) = \left(10 + x^\beta, 24 - \frac{7}{100}\left(x^\beta\right)^2\right), \qquad \pi(b; x^\alpha, x^\beta) = \left(16 - x^\alpha, 12 - \frac{7}{100}(x^\alpha)^2\right)$$

[11] The recorded behavior of players 1s was not used for any of the identification in this paper. Thus even if 20 subjects participated in a session, only the 10 player 2s' decisions were used.

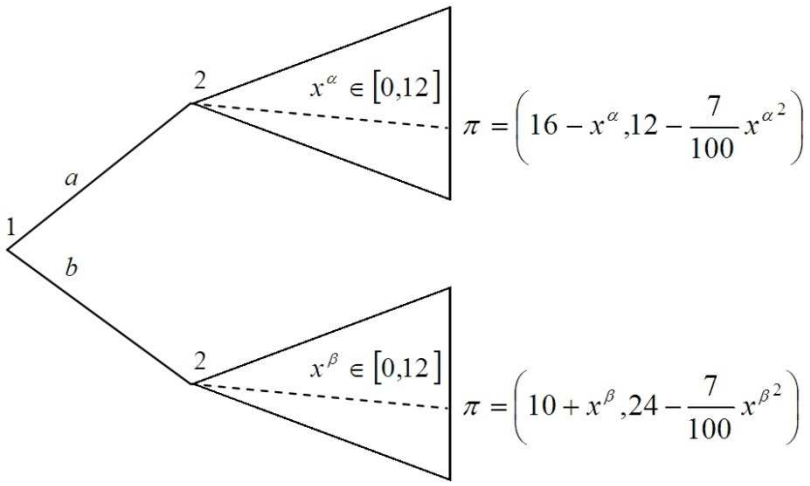

Figure 4: The non-linear judgment game in extensive form

Figure 4 shows the game in extensive form. The quadratic cost of reward/punishment implies that the marginal cost of reward/punishment goes from 0 to 1.7.[12] Given its comparative complexity, subjects were given a diagram (see the instructions) of the game's payoffs and were required to complete a short quiz to confirm their ability to locate payoffs on the diagram.[13]

Player 2s selected their move using the strategy method. This was necessary for obtaining an estimate of (α, β) for each subject, which in turn was necessary for a high power test of our hypotheses. The interchangeability of hot and cold decisions remains an open empirical question.[14] There are no cases of a treatment effect being found using a strategy method that does not arise when the game is played in dynamic form (Charness and Dufwenberg (2006)), though admittedly our requirements are more stringent since we interested in precise absolute estimates rather than the sign of treatment effects. Player 1s' actual decisions were not revealed to their partners until the end of the session.

[12] We chose a quadratic payoff structure to ensure interior solutions under the FF model.
[13] In the instructions, (t, b) are referred to as $(dash, solid)$. This allowed a visual distinction in the payoff figures.
[14] Schotter et al. (1994), Blount and Bazerman (1996), Rapoport (1997) and Brosig et al. (2003) find a difference while Cason and Hui (1998), Brandts and Charness (2000) and Solnick (2007) do not.

The second game is a non-linear dictator game where the dictator must end up with more than his partner. It is designed to elicit $x^{\tilde{\beta}}$. Player 2 is the dictator. Elements of X are $x^{\tilde{\beta}}$. The strategy sets and payoffs are:

$$W = \emptyset, \qquad X = [0,12]$$

$$\pi\left(x^{\tilde{\beta}}\right) = \left(x^{\tilde{\beta}}, 24 - \frac{7}{100}\left(x^{\tilde{\beta}}\right)^2\right)$$

The third game is another non-linear dictator game but this time the dictator must end up earning less than his partner. Also after a point, closing the gap between the players' payoffs requires the dictator to decrease both their payoffs. It is designed to elicit $x^{\tilde{\alpha}}$. Player 2 is the dictator. Elements of X are $x^{\tilde{\alpha}}$. The strategies and payoffs are:

$$W = \emptyset, \qquad X = [0,20]$$

$$\pi\left(x^{\tilde{\alpha}}\right) = \left(24 - x^{\tilde{\alpha}}, 9 + x^{\tilde{\alpha}} - \frac{7}{100}\left(x^{\tilde{\alpha}}\right)^2\right)$$

To obtain the perceived kindness vector $\left(k_i^{\alpha}, k_i^{\beta}, k_i^{\tilde{\alpha}}, k_i^{\tilde{\beta}}\right)$, we measured SOE. Recall that $k_{21}^{\gamma} = \pi_2'' - \pi_1''$, i.e., it is the difference between player 2's SOE of his own-payoff at the decision node and player 2's SOE of his partner's payoff.

We measured SOE in two ways. Our first method was to directly elicit them from each player 2, incentivizing them using a quadratic scoring rule. To reward the player 2s' accuracy in their SOE, we elicited the player 1s' first-order predictions. This procedure was similar to that employed by Dufwenberg and Gneezy (2000) and Charness and Dufwenberg (2006).

We elicited the expectations in each game immediately after the players made their choices and before the start of the next game. Eliciting expectations after players make their choices carries the risk of consistency bias: subjects may attempt to report expectations that ex post rationalize their choices in an attempt to look consistent to the experimenter. As a simple check on the presence of such a bias, we conducted an additional session where we had observers. The observers did everything identically to player 2s except that they only reported SOE and did not make any choices.

As an alternative to direct elicitation, our second method of measuring SOE was to directly induce them. We were interested in seeing if this made an impact on the results. Moreover an additional problem with expectations elicited after decisions have been made is that they may not reflect the expectations held at the time of the decision.

In these inducement treatments, after eliciting the player 1s' expectations, we showed each player 2 their partner's expectations. We also told player 2s that their partners were being rewarded for the accuracy of their expectations and that their partners were not told that the player 2s would see their expectations (which was true). This prevented the player 1s from being strategic in their reported expectations.

Subjects were given no feedback between tasks. Once the subjects had finished making their choices and predictions, they were paid in private.

IV. Empirical results

A. Overview of data

Table 2 shows the sessions we ran. We obtained 77 observations from subjects actually playing the three games, in addition to 28 observations from observers. The data was collected over nine sessions during March and April 2008 in George Mason University. Average earnings were $14 for a session that lasted about an hour.

Session	Treatment	Date	# game-playing subjects	# observers
1	Elicitation	3/26/2008	10	5
2	Elicitation	3/27/2008	10	0
3	Elicitation	3/28/2008	10	0
4	Elicitation	3/28/2008	9	0
5	Inducement	3/28/2008	9	0
6	Inducement	3/31/2008	10	0
7	Inducement	3/31/2008	10	0
8	Inducement	3/31/2008	9	0
9	Elicitation	4/17/2008	0	23

Table 2: Sessions

Table 3 is the sample statistics for the identifying variables.

Variable (symbol)	Variable (description)	Mean	SD
x_i^{α}	Player 2's choice of *x* in response to player 1 choosing unkind (b).	Elicitation: 3.28 Inducement: 2.61	Elicitation: 3.15 Inducement: 2.98
k_i^{α}	Kindness of player 1 choosing unkind (b) based on player 2's elicited second-order expectations.	Elicitation: -2.11 Inducement: -0.97	Elicitation: 2.54 Inducement: 4.16
x_i^{β}	Player 2's choice of *x* in response to player 1 choosing kind (t)	Elicitation: 3.33 Inducement: 3.76	Elicitation: 2.72 Inducement: 2.98
k_i^{β}	Kindness of player 1 choosing kind (t) based on player 2's elicited second-order expectations.	Elicitation: 4.92 Inducement: 4.24	Elicitation: 6.31 Inducement: 6.68
$x_i^{\widetilde{\alpha}}$	Player 2's choice of *x* in reverse dictator. Note that 7 is the neoclassical optimum.	Elicitation: 7.22 Inducement: 6.82	Elicitation: 2.87 Inducement: 3.59
$k_i^{\widetilde{\alpha}}$	Kindness of player 1 based on player 2's elicited second-order expectations in reverse dictator.	Elicitation: -3.15 Inducement: -3.30	Elicitation: 6.32 Inducement: 5.38
$x_i^{\widetilde{\beta}}$	Player 2's choice of *x* in dictator.	Elicitation: 1.95 Inducement: 3.47	Elicitation: 2.02 Inducement: 3.34
$k_i^{\widetilde{\beta}}$	Kindness of player 1 based on player 2's elicited second-order expectations in dictator.	Elicitation: 16.57 Inducement: 14.64	Elicitation: 10.68 Inducement: 10.15

Table 3: Sample statistics

Unless otherwise stated, the neoclassical optimum for all choice variables is zero.

All the means have the expected sign. If we compare the distributions of the variables across measures of SOE (elicited vs. induced) using Kolmogorov-Smirnov tests, we fail to reject equality in every case except $x_i^{\widetilde{\beta}}$ (p-value = 9%).

B. Hypothesis tests

To investigate hypothesis 1, we conduct a Wilcoxon signed-rank test (paired-values) on the estimates of (α_i, β_i) obtained from the judgement game. Using elicited expectations, we reject the null hypothesis (Z = 3.26; p-value < 1%). Using induced expectations, we fail to reject the null hypothesis (Z = 0.414; p-value = 68%). Figure 5 shows histograms of our recovered values of $(\alpha_i - \beta_i)$ across the two treatments.

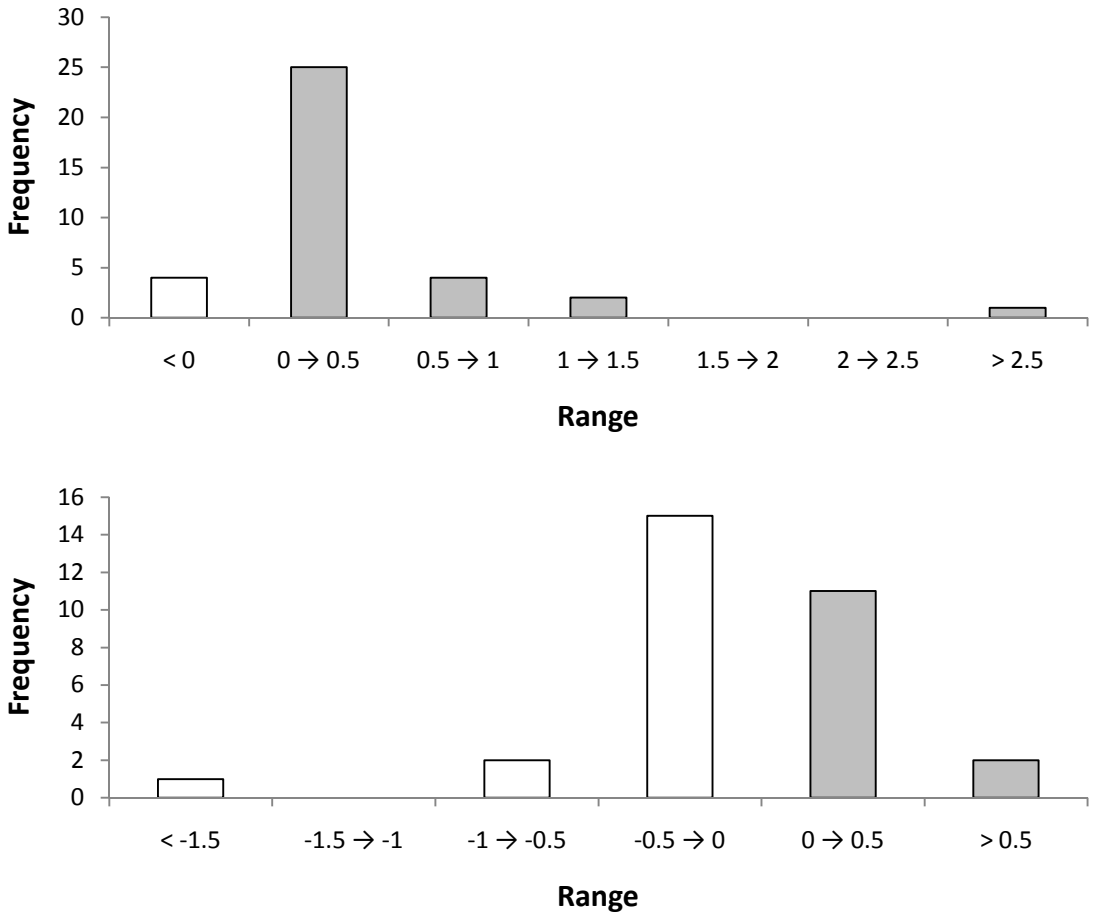

Figure 5: Difference between negative and positive net reciprocity $(\alpha_i - \beta_i)$ by individual

Top histogram is elicitation, bottom histogram is inducement. White bars are negative, gray are positive.

Result 1: using elicited expectations, negative net reciprocity is stronger than positive net reciprocity. Using induced expectations, positive and negative net reciprocity are equally strong.

To investigate hypothesis 2, we conduct a Wilcoxon signed-rank test (paired-values) on the estimates of $(\tilde{\alpha}_i, \tilde{\beta}_i)$ obtained from the dictator games. Using elicited expectations, we reject the null hypothesis (Z = 3.39; p-value < 1%). Using induced expectations, we also reject the null hypothesis (Z = 3.65; p-value < 1%). Figure 6 shows histograms of our recovered values of $(\tilde{\alpha}_i - \tilde{\beta}_i)$ across the two treatments.

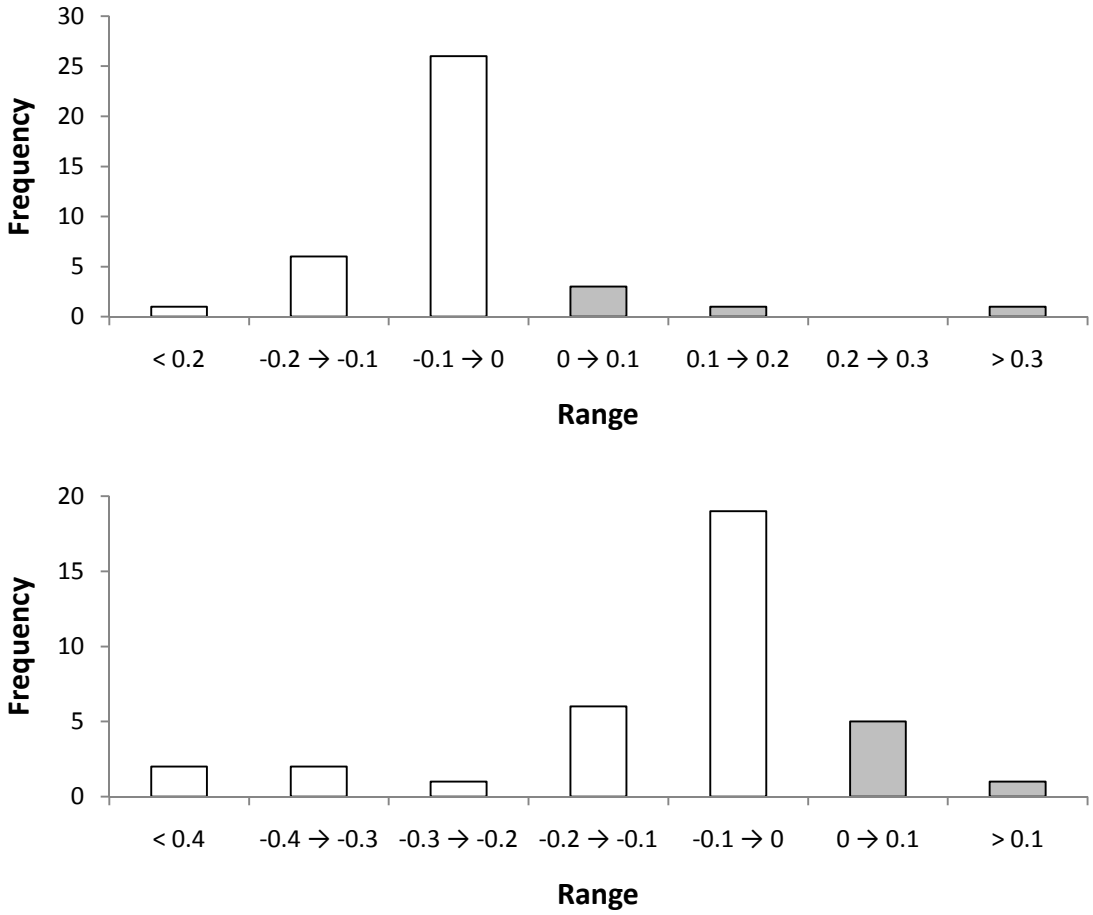

Figure 6: Difference between negative and positive distributional reciprocity $(\tilde{\alpha}_i - \tilde{\beta}_i)$ by individual

Top histogram is elicitation, bottom histogram is inducement. White bars are negative, gray are positive.

Result 2: positive distributional reciprocity is stronger than negative distributional reciprocity according to both measures of expectations.

To investigate hypothesis 3, we conduct a Wilcoxon signed-rank test (paired-values) on the estimates of $(\alpha_i - \tilde{\alpha}_i, \beta_i - \tilde{\beta}_i)$ obtained from the three games. Using elicited expectations, we reject the null hypothesis (Z = 3.44; p-value < 1%). Using induced expectations, we fail to reject the null hypothesis (Z = 1.48; p-value = 14%). Figure 7 shows histograms of our recovered values of $\left[(\alpha_i - \tilde{\alpha}_i) - \left(\beta_i - \tilde{\beta}_i\right)\right]$ across the treatments.

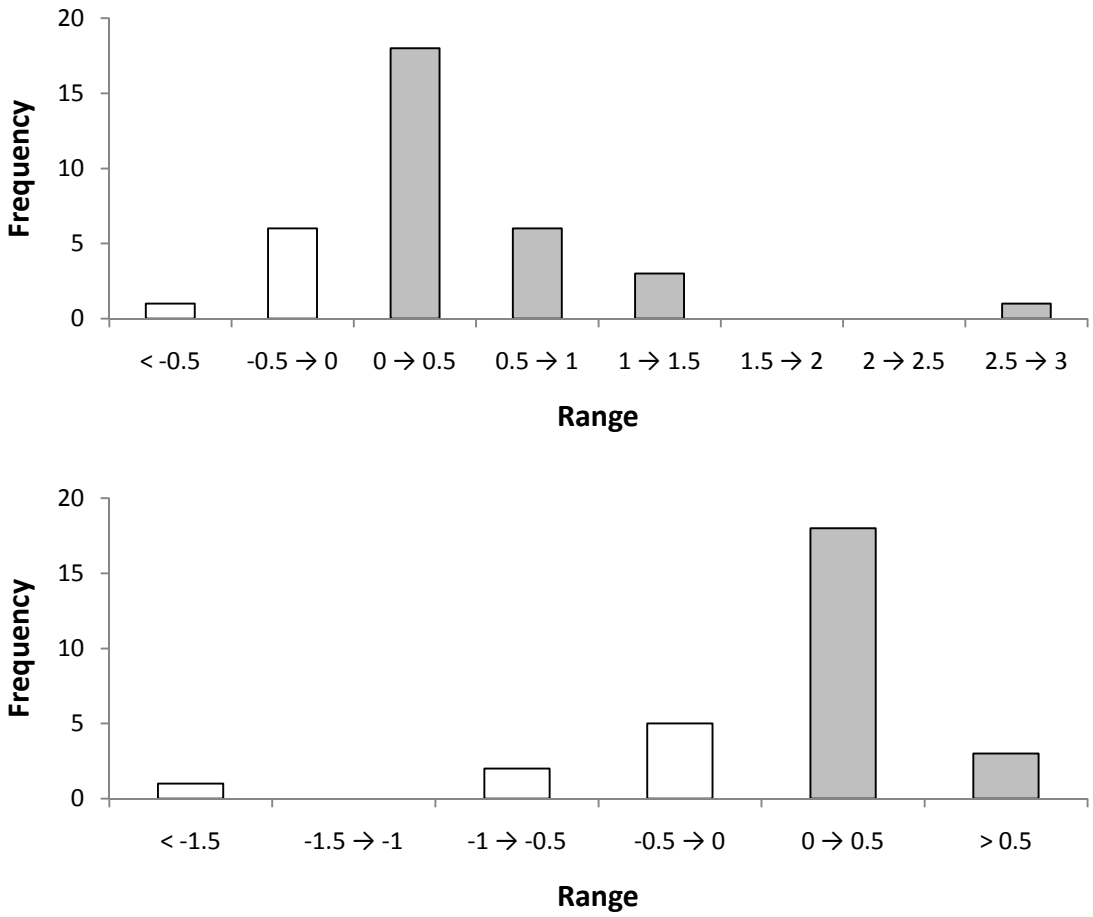

Figure 7: Difference between negative and positive intentional reciprocity $\left[(\boldsymbol{\alpha_i} - \tilde{\boldsymbol{\alpha}}_{\boldsymbol{i}}) - \left(\boldsymbol{\beta_i} - \tilde{\boldsymbol{\beta}}_{\boldsymbol{i}}\right)\right]$ by individual

Top histogram is elicitation, bottom histogram is inducement. White bars are negative, gray are positive.

Intentional reciprocity is constructed from the difference between net and distributional reciprocity. Given results 1 and 2, with sufficient data, one ought to reject equality of positive

and negative intentional reciprocity in both treatments. This is especially true given the high significance level of the test statistics for distributional reciprocity. However this only ends up being true for the elicited expectations. We explain this apparent incongruence by the relatively large standard deviation of the differences in net reciprocity (S.D. = 0.48) compared to distributional reciprocity (S.D. = 0.14).

Result 3: using elicited expectations, negative intentional reciprocity is stronger than positive intentional reciprocity. Using induced expectations, positive and negative intentional reciprocity are equally strong.

The result under elicited expectations is consistent with Offerman (2002).

C. Robustness

As described in the design section, in the elicitation treatment, we elicit data from subjects who take the role of player 2 observers. This is to see if the act of eliciting expectations after making a decision biases the reported expectations. We compare the data from the 28 observers to the 39 decision-making player 2s from the elicitation treatment using Kolmogorov-Smirnov tests. For each of the eight variables which we elicit expectations on, we fail to reject the null hypothesis of the two distributions being the same (p-value = 46%, 63%, 61%, 45%, 21%, 27%, 16%, 33%).

A potential problem with the data on SOE is that some of the observations contain errors. A subject may predict a SOE of an impossible payoff because it is outside the feasible range of payoffs or because it is not congruent with the another SOE. For example, a player 2 may report SOE of both players earning 15 points in the judgment game conditional on the player 1

selecting b. If we exclude such observations (be they elicited or induced), none of our results are affected.

An additional issue is the arbitrary functional form assumptions associated with structurally estimating the FF model. For example, kindness is specified as the absolute difference between the SOE of the players' payoffs, $(\pi_2'' - \pi_1'')$, but one could plausibly use $(\ln \pi_2'' - \ln \pi_1'')$ or a host of alternatives. We use two alternatives: logarithmic, $k_{21} = (\ln(1 + \pi_2'') - \ln(1 + \pi_1''))$, and exponential, $k_{21} = (\exp(\pi_2'') - \exp(\pi_1''))$. Under elicited SOE, all results are unaffected. Under induced SOE, results 1 and 2 are also unaffected, but result 3 exhibits some sensitivity. When kindness is specified exponentially, negative intentional reciprocity becomes stronger than positive intentional reciprocity.

Along similar lines, as mentioned in footnote 6, our choice of the absolute distance metric for inferring intentional reciprocity $d = \gamma - \tilde{\gamma}$ is also arbitrary. Using logarithmic and exponential distance metrics instead does not affect any of the results.

A final robustness check concerns the presence of an order-effect in the games played. Subjects always played the judgment game first, but we randomized what came next out of the standard non-linear dictator and the reverse non-linear dictator. Results 2 and 3 are unaffected by the order of the two games.

V. Summary and discussion

The decision to reward/punish is driven by several factors, many of which do not represent what we would regard as an intrinsic bias towards reward vs. punishment. The example of Jane and

Tom in the introduction is a simple demonstration: if Tom punishes Jane with higher likelihood when she destroys his car than he rewards her when she gives him a ride to work, this does not represent a bias towards punishment.

The FF model of reciprocity formalizes how indignant or grateful Tom should feel about Jane's actions, i.e., how kind Tom thinks Jane was. Tom's SOE of his own actions are a subtle but important determinant of his evaluation. His final reward/punishment decision is based on an emotional composite of Jane's objective kindness and his residual subjective emotions. It is at this point that a meaningful asymmetry can arise.

If the goal is to use reward/punishment data to infer a bias in residual subjective emotions, then one needs to control for objective kindness. The FF model tractably permits this, and decomposes net reciprocity into two parts: distributional and intentional. Offerman (2002) tried to isolate the intentional component and concluded that negative intentional reciprocity is stronger than positive intentional reciprocity, but his design did not allow him to control for objective kindness. Consequently, Offerman could not be sure that he was identifying an intrinsic asymmetry.

We conduct an experiment where we control for objective kindness by measuring SOE. We do this in two ways: by directly eliciting each player 2's SOE, and by inducing them. This allows us to recover the parameters of the FF model and to compare positive vs. negative (1) net, (2) distributional, and (3) intentional reciprocity. Interestingly, our conclusions differ by treatment: under elicitation, net and intentional reciprocity are stronger in the negative domain than the positive domain, respectively, while positive distributional reciprocity is stronger than negative distributional reciprocity. Under inducement, positive distributional reciprocity is stronger than

negative distributional reciprocity while the other forms of reciprocity are equally strong. However we believe that the even under inducement, with sufficient data, we would find that negative intentional reciprocity is stronger than positive intentional reciprocity.

The elicitation results are the same as Offerman (2002), suggesting that his results were not driven by endogeneity bias. As such, we subscribe to the self-serving expectations explanation that he offered. People like to maintain a positive self-image, attributing good events to themselves and bad events to others. They will feel they are more deserving of objectively kind behavior than others, and less deserving of objectively unkind behavior. At this point, subjective emotions enter the decision calculus and an asymmetric intrinsic bias towards punishment is generated.

This explains why negative net and intentional reciprocity may be stronger, but not why positive distributional reciprocity is stronger.[15] Perhaps the most convincing explanation for this is the fundamental inefficiency of negative reciprocity vis-à-vis positive reciprocity: punishment decreases the payoff of both players and is therefore Pareto-dominated by no punishment. When residual inequity is not the fault of the initial mover (e.g., when the initial mover didn't even have a choice), it seems particularly wasteful to punish. Reward, on the other hand, is always efficient because one of the parties gains.

Why do our elicitation and inducement results differ? Ideally we would be able to elicit and induce expectations at the decision-making stage without any cognitive disruption. This is not possible practically. The act of either eliciting or inducing expectations may well alter the way in which a subject makes his decision, and they need not alter the decision-making process in the

[15] Danneberg et al (2007) find a result that is similar in spirit, though they are estimating the Fehr and Schmidt (1999) model.

same way. Social psychologists have established how forcing individuals to process information that is cognitively-relevant but not payoff-relevant can generate anchoring affects: the mind takes advantage of the cognitively accessible information (Mussweiler and Strack (1999)). Moreover if, as we did, we elicit expectations after the decision has been made, there is no way to be certain that the expectations elicited are the same as the expectations held at the point that the decision was made. The FF model is not designed for modeling these differences beyond making $(\alpha, \tilde{\alpha}, \beta, \tilde{\beta})$ contingent on the environment.

If we believe that symmetric net reciprocity is in some sense a rational benchmark, one might speculate that the extra reflection time generated by inducing expectations makes the rational outcome more likely. The self-serving bias may be more likely under instinctive or less measured decision-making. Explaining these differences remains an open empirical question and a promising avenue for future research.

This paper's contributions are three-fold. First, we highlight a potential problem in comparing the incidence of reward and punishment to infer individuals' propensities to reward and punish. Depending on the way in which we measure SOE, addressing the potential problem may alter Offerman's (2002) conclusion. Second, by specifying FF's structural model, we are able to estimate positive and negative distributional, intentional and net reciprocity. Offerman's (2002) design only permitted him to infer the relative size of positive and negative intentional reciprocity and not the exact size of either. Finally, to our knowledge, this is the first attempt at estimating the parameters of the FF model.

References

Ames, D., F. Flynn and E. Weber (2004). "It's the thought that counts: on perceiving how helpers decide to lend a hand," Personality and Social Psychology Bulletin. 30, p461-74.

Andreoni, J., W. Harbaugh and L. Vesterlund (2003). "The carrot or the stick: rewards, punishments and cooperation," American Economic Review. 93(3), p893-902.

Berg, J., J. Dickhaut and K. McCabe (1995). "Trust, reciprocity and social norms," Games and Economic Behavior. 10(1), p122-42.

Blount, S. (1995). "When social outcomes aren't fair: the effect of causal attributions on preferences," Organizational Behavior and Human Decision Processes. 63(2), p131-44.

Blount, S. and M. Bazerman (1996). "The inconsistent evaluation of absolute vs. comparative payoffs in labor supply and bargaining," Journal of Economic Behavior and Organization. 30, p227-40.

Bolton, G. and A. Ockenfels (2000). "ERC: a theory of equity, reciprocity and competition," American Economic Review. 90(1), p166-93.

Brandts, J. and G. Charness (2000). "Hot vs. cold: sequential responses and preference stability in experimental games," Experimental Economics. 2, p227-38.

Brosig, J., J. Weimann and C-L Yang (2003). "The hot versus cold effect in a simple bargaining experiment," Experimental Economics. 6, p75-90.

Cason, T. and V-L. Mui (1998). "Social influence in the sequential dictator game," Journal of Mathematical Psychology. 42, 248-65.

Charness, G. and M. Dufwenberg (2006). "Promises and partnership," Econometrica. 74(6), p1579-1601.

Charness, G. and M. Rabin (2002). "Understanding social preferences with simple tests," Quarterly Journal of Economics. 117, p817-69.

Cox, J., D. Friedman and S. Gjerstad (2007). "A tractable model of reciprocity and fairness," Games and Economic Behavior. 59, p17-45.

Dufwenberg, M. and U. Gneezy (2000). "Measuring beliefs in an experimental lost wallet game," Games and Economic Behavior. 30, 163-82.

Dufwenberg, M. and G. Kirchsteiger (2004). "A theory of sequential reciprocity," Games and Economic Behavior. 47, p268-98.

Falk, A., and U. Fischbacher (2006). "A theory of reciprocity," Games and Economic Behavior. 54, p293-315.

Fehr, E., G. Kirchsteiger and A.Riedl (1993) "Does fairness prevent market clearing? An experimental investigation," Quarterly Journal of Economics. 108(2)m p437-60.

Fehr, E. and K. Schmidt (1999). "A theory of fairness, competition and cooperation," Quarterly Journal of Economics. 114(3), p817-68.

Geanakoplos, J., D. Pearce and E. Stacchetti (1989). "Psychological games," Games and Economic Behavior. 1, p60-79.

Greenberg, M. and D. Frisch (1972). “Effect of intentionality on willingness to reciprocate a favor,” Journal of Experimental Social Psychology. 8, p99-111.

Levine, D. (1998). “Modeling altruism and spitefulness in experiments,” Review of Economic Dynamics. 1, p593-622.

McCullough, M., R. Emmons, S. Kilpatrick, and D. Larson (2001). “Is gratitude a moral effect?,” Psychological Bulletin. 127(2), p249-66.

Mussweiller, T., and F. Strack (1999). “Comparing is believing: a selective accessibility model of judgmental anchoring,” European Review of Social Psychology. 10(1), p135-67.

Offerman, T. (2002). “Hurting hurts more than helping helps,” European Economic Review. 46, p1423-37.

Pereira, P., N. Silva and J Silva (2006). “Positive and negative reciprocity in the labor market,” Journal of Economic Behavior and Organization. 59, p406-22.

Rabin, M. (1993). “Incorporating fairness into game theory and economics,” American Economic Review. 83(5), p1281-1302.

Rapoport, A. (1997). “Order of play in strategically equivalent games in extensive form,” International Journal of Game Theory. 26, 113-36.

Schotter, A., K. Weigelt and C. Wilson (1994). “A laboratory investigation of multiperson rationality and presentation effects,” Games and Economic Behavior. 6, 445-68.

Sefton, M., R. Shupp and J. Walker (2001). “The effect or rewards and sanctions in the provision of public goods,” Mimeo, University of Indiana.

Solnick, S. (2007). “Cash and alternate methods of accounting in an experimental game,” Journal of Economic Behavior and Organization. 62, 316-21.

Sutter, M., S. Haigner and M. Kocher (2006). “Choosing the carrot of the stick?,” Mimeo, University of Cologne.

Tesser, A. R. Gatewood and M. Driver (1968). “Some determinants of gratitude,” Journal of Personality and Social Psychology. 9(3), p233-6.

Walker, J. and M. Halloran (2004). “Rewards and sanctions and the provision of public goods in one-shot settings,” Experimental Economics. 7, p235-47.

Appendix: Kindness in Offerman (2002)

In Offerman’s (2002) version of the judgment game, we have the following payoffs:

$$\pi(help, reward) = \pi(e, r) = (12,13)$$

$$\pi(help, cool) = \pi(e, c) = (8,14)$$

$$\pi(hurt, cool) = \pi(u, c) = (11,6)$$

$$\pi(hurt, punish) = \pi(u, p) = (7,5)$$

Offerman realizes that there are both redistribution- and intentions-based motives for reward and punishment. Therefore the appropriate model is FF. Using the FF model, 2's utility is:

$$u_2 = \pi_2 + G(\pi_2'' - \pi_1'')\pi_1$$

$$\pi_2'' - \pi_1'' < 0 \Rightarrow G \in \{\alpha, \tilde{\alpha}\}, \pi_2'' - \pi_1'' > 0 \Rightarrow G \in \{\beta, \tilde{\beta}\}$$

Where $G = \gamma$ if a 1's action is chosen by a human, and $G = \tilde{\gamma}$ if 1's action is chosen by nature. Applying this utility to the above game, we have:

$$\pi''(e) = \left(12 - 4p_{e,Y}'', 13 + p_{e,y}''\right)$$

$$\pi''(u) = \left(7 + 4p_{u,Y}'', 5 + p_{u,y}''\right)$$

Where $p_{x,y}''$ is 2's SOE of his probability of playing c given that 1 has played x and that 1's choice was made by $y \in \{Human, Nature\}$. 2 will reward if and only if $B > \left[4\left(1 + 5p_{e,y}''\right)\right]^{-1}$. 2 will punish if and only if $A > \left[4\left(2 + 3p_{u,y}''\right)\right]^{-1}$.

Let $d(\gamma, \tilde{\gamma})$ be a distance metric. Offerman is hypothesising:

$$H_0: E[d(\alpha, \tilde{\alpha})] = E\left[d\left(\beta, \tilde{\beta}\right)\right]$$

$$H_1: E[d(\alpha, \tilde{\alpha})] > E\left[d\left(\beta, \tilde{\beta}\right)\right]$$

For each player 2 i, let $D_{x,y,i} = 1$ if i rewards or punishes and 0 otherwise. Then:

$$plim\left(\frac{1}{n_e}\sum_{i=1}^{n_e} D_{e,y,i}\right) = \Pr\left(B_i > \left[4\left(1+5p''_{e,y,i}\right)\right]^{-1}\right)$$

$$plim\left(\frac{1}{n_u}\sum_{i=1}^{n_u} D_{u,y,i}\right) = \Pr\left(A_i > \left[4\left(2+3p''_{u,y,i}\right)\right]^{-1}\right)$$

Offerman is effectively using a test statistic with the following probability limit:

$$T = \left[\Pr\left(\beta > \left[4\left(1+5p''_{e,H}\right)\right]^{-1}\right) - \Pr\left(\tilde{\beta} > \left[4\left(1+5p''_{e,N}\right)\right]^{-1}\right)\right] - \left[\Pr\left(\alpha > \left[4\left(2+3p''_{u,H}\right)\right]^{-1}\right) - \Pr\left(\tilde{\alpha} > \left[4\left(2+3p''_{u,N}\right)\right]^{-1}\right)\right]$$

In general, whatever distance metric we use, unless we make specific, arbitrary assumptions about the distributions of $\left(\beta, \tilde{\beta}, \alpha, \tilde{\alpha}, p''_{e,H}, p''_{e,N}, p''_{u,H}, p''_{u,N}\right)$, T will not be zero under the null hypothesis.

An example of sufficient conditions is as follows. The values of $\left\{\left[4\left(1+5p''_{e,H}\right)\right]^{-1}, \left[4\left(1+5p''_{e,N}\right)\right]^{-1}, \left[4\left(2+3p''_{u,H}\right)\right]^{-1}, \left[4\left(2+3p''_{u,N}\right)\right]^{-1}\right\}$ are all induced, *equal* constants, (α, β) have the same distribution, $\left(\tilde{\alpha}, \tilde{\beta}\right) = \left(\delta_\alpha \alpha, \delta_\beta \beta\right)$ where $\left(\delta_\alpha, \delta_\beta\right)$ are constants that are potentially unequal and $d(X,Y) = |X-Y|$. The null hypothesis then reduces to $\delta_\alpha = \delta_\beta$, and under the null, $T = 0$. Virtually any perturbation of these conditions will result in $T \neq 0$ under the null.

The reason is that Offerman is not controlling for important covariates – specifically kindness $(\pi''_2 - \pi''_1)$. One could assume that kindness is exogenous, but existing studies of kindness (e.g., FF) reject the exogeneity of kindness in this game because of the payoff structure. It is not even enough to assume that $p''_{x,y}$ is exogenous due to the non-linearity of the thresholds in $p''_{x,y}$.