Reciprocating to Strategic Kindness[*]

A. Yeşim Orhun

Ross School of Business, University of Michigan

October 15, 2015

Abstract

This article examines how people reciprocate to a helpful action that is potentially motivated by strategic considerations. Both genuine kindness and self-interested material gain may drive decision-making in situations where individuals or firms expect a reciprocal reaction to their decision. Experimental results show that the degree of positive reciprocity second-movers display towards the same helpful action is lower when the likelihood that the first-mover was motivated to help in order to avoid potential punishment is higher. Moreover, the decline in the degree of positive reciprocity is associated with the deterioration of the degree of altruism inferred regarding the helpful first-movers. The results cannot be rationalized by reciprocity to outcomes or to perceived intentions, and highlight the distinct role of perceived motives in reciprocal decision-making.

Keywords: Reciprocity, Motives, Beliefs, Intentions, Social Preferences.

[*] I thank Gary Bolton, Jonathan Carmel, Yan Chen, James Cox, Emel Filiz-Ozbay, Aradhna Krishna, Erin Krupka, Steve Leider, Yusufcan Masatlioglu, Neslihan Uler, Erkut Ozbay, Tanya Rosenblat and seminar participants at New York University, University of Texas at Dallas and University of Michigan for their comments and suggestions. I thank Arun Varghese, Roshni Kalbavi, Valerie Laird and Catherine Dolan for their help in conducting experimental sessions. I thank University of Michigan's Ross Behavioral Lab and School of Information Behavioral Lab for providing excellent resources. I am grateful for the special research grant provided by the Ross School of Business, University of Michigan.

1 Introduction

Kindness perceptions are central to reciprocal decision-making. We treat others more generously than we otherwise would if we think they have been kind to us. Our assessment of how kind or unkind an action is may depend not only on the consequence the action had on our well-being, but also on our beliefs regarding what consequence the person intended his action to have, i.e, his intent, and the reason why he wanted to achieve this consequence, i.e., his motive. The importance of intentions and motives for determining the appropriate judgment of an action is immediately apparent in criminal law. Consider the trial of an alleged murderer. The court must investigate the person's intent behind the action that resulted in murder: Did the person shoot the victim by mistake while cleaning the gun? The court must also establish the person's motives: Was the action in self-defense or fueled by revenge? Similarly, people may rely on attributions of intentions and motives in assessing the kindness of a helpful action and determining the appropriate response to it. Previous research emphasized the impact of perceived intentions on reciprocal decision-making[1]. Notwithstanding its central role in understanding reciprocal decision-making across a wide array of economic interactions, the impact of perceived motives has been largely unexplored, possibly due to difficulties in isolating its role.

This article examines the impact perceived motives has on the degree of positive reciprocity a helpful action triggers, as distinct from the impact of it consequences and perceived intentions. People who are aware of the reciprocal nature of the interaction can be motivated to be helpful due to i) an intrinsic desire to benefit others, or ii) strategic motives, such as hopes of future material gains, or fears of future losses, because reciprocal interactions invariably feature rewards and/or punishment (Gneezy et al., 2000; Sobel, 2005; Segal and Sobel, 2007; 2008; Cabral et al. 2014). Therefore, the motive behind a helpful action in reciprocal interactions is often ambiguous. However, the incentives inherent in a particular interaction may shape perceptions of motives. Consider an employee asking her boss for a week off due to personal reasons right before the deadline of an important project. If the boss agrees, it would greatly help the employee, but would jeopardize the completion of the project. In one situation, if the boss does not agree, the employee may choose to badmouth him to his clients. In the other situation, the employee holds no such power over him. In both cases, the boss agrees. If perceived motives matter, the employee may judge the same decision to be kinder in the situation where it could not have been motivated by fear of retaliation.

We test for the role of perceived motives on positive reciprocity by varying the type of strategic incentives to help in order to manipulate the proportion of helpfulness motivated by genuine kindness versus strategic incentives. In particular, we study the degree of positive reciprocity the second-mover displays towards the first-mover who intentionally increased the second-mover's material well-being, either when the first-mover could have been motivated to help in order to avoid future

[1] Previous experimental literature shows that the same helpful action leads to more positive reciprocity if the decision maker intended the action to produce the actualized consequence (Blount, 1995; Offerman, 2002; Charness and Haruvy, 2002; Charness, 2004; Charness and Levine, 2007; Falk et al., 2008; Klempt, 2012), and if the decision maker chose the best he could among the alternatives offered to him (Brandts and Sola, 2001; Nelson, 2002; McCabe et al., 2003).

losses, or when such a motivation was not possible. Our experimental design features a two-stage reciprocal interaction where the first-mover decides between a selfish option that maximizes his payoffs and a helpful action that transfers some of his payoffs to the second-mover. We manipulate whether the second-mover could decrease the payoffs of the first-mover at a cost to herself in case the first-mover was not helpful. This approach allows us to vary the composition of motives for helping among the helpful first-movers, and consequently, second-movers' perceptions regarding the motive of a helpful first-mover.

The relevance of perceived motives for kindness inferences was pointed out in the earlier years of reciprocity research. For example Rabin (1998, p.22) remarks: "if you think somebody has been generous to you solely to get a bigger favor from you in the future, then you do not view his generosity to be as pure as if he had expected no reciprocity from you." However, the distinct role of motives has been elusive, primarily because in games that are commonly used to elicit positive reciprocity, such as the gift exchange game, the influence of perceived intentions and perceived motives may go hand in hand (Stanca et al., 2009; Netzer and Schmutzler, 2014). More generally, however, attributions of intentions and motives are driven by different considerations. In order to glean the first-mover's intentions, one needs to consider what outcome the first-mover expected his action to produce for the second-mover. Astutely, intention-based reciprocity theories define the relative kindness of perceived intentions by comparing the outcome the benefactor expected the beneficiary to obtain to a fair benchmark (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). On the other hand, in order to gauge the first-mover's motives, one also needs to consider what the thought he would gain or lose if he acted differently. Therefore, intention-based reciprocity models are not meant to capture a broad role of perceived motives in reciprocal decision making.

In order to help distinguish the role of perceived motives from the role of perceived intentions, we keep the agency, volition and the choice-set of the first-mover constant across the sequential reciprocity interactions we study. We further restrict the payoffs in the sub-game reached after the first-mover chooses to help to be identical across treatments. As we explain in the discussion section, the consideration of perceived motives in the reciprocal interactions we study leads to predictions that contradict both the predictions based on outcome-based reciprocity models (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), and the predictions of intentions-based reciprocity models (Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). As a result, our experimental design offer a strong test for the role of perceived motives as distinct from the role of perceived intentions and outcomes.

We find that when the benefactor could not have been motivated by fear of punishment, more beneficiaries choose to reward the same helpful action. In addition, when the reciprocal interaction presents weaker strategic incentives to be helpful, beneficiaries infer a larger difference between the altruism among the benefactors who were helpful in the reciprocal interaction and the altruism in the general population. Moreover, within-person changes in beneficiaries' perceptions regarding the degree to which a helpful first-mover is altruistic are associated with changes in their reciprocal behavior, providing additional causal evidence for kindness inferences driven by perceived motives influencing reciprocal decisions. These results support the conjecture that reciprocity towards a helpful action is not just a function of the beneficiary's perception as to whether the benefactor intended to make a sacrifice, but also hinges crucially on whether she believes that the benefactor made the sacrifice out of strategic motives or out of genuine care for others.

Rabin (1998) notes that "a crucial feature of the psychology of reciprocity is that people determine their dispositions toward others according to motives attributed to these others." Our results corroborate this early foresight. Moreover, they demonstrate that the consideration of motives may overturn what would have been expected based on intention-based reciprocity models alone. The results also lend support to the philosophical approach of formalizing reciprocity as a response to the altruism of the benefactor as revealed by his actions (Levine 1998, Cox et al., 2008a; Gl and Pesendorfer, forthcoming). This approach has been used to develop a tractable way to distinguish between genuine and strategic kindness in the context of a gift-exchange game (Arbak and Kranich, 2005; Dur, 2009; Non, 2012). We hope that the experimental design and results presented in this article are useful for future theoretical developments that define and generalize the role of perceived motives in reciprocal decision-making.

2 Experimental Investigation

In order to glean the first-mover's intentions, one needs to consider what outcome the first-mover expected his action to produce for the second-mover. Astutely, intention-based reciprocity theories construe perceived intentions in sequential interactions based on the outcome the first-mover expected the second-mover to obtain as a result of his action (Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). In contrast, in order to gauge the first-mover's motives, one needs to consider what he thought he would gain or lose if he acted differently. While in certain reciprocal interactions, such as the gift exchange game, the influence of perceived intentions and perceived motives may go hand in hand (Stanca et al., 2009; Netzer and Schmutzler, 2014), in general, consideration of intentions does not encapsulate the consideration of the role of motives.

In a two-stage reciprocal interaction where the first-mover decides between a selfish option that maximizes his payoffs and a helpful action that transfers some of his payoffs to the second-mover, we manipulate the perceived motive of the helpful first-mover by varying whether the second-mover could decrease the payoffs of the first-mover at a cost to herself in case the first-mover was not helpful. In addition, we keep the agency, volition and the choice-set of the first-mover constant. We further restrict the payoffs in the sub-game reached after the first-mover chooses to help to be identical across treatments

Identifying the role of motives is not without its challenges. The ideal experiment needs to jointly vary the motives of the first-mover and the second-mover's perceptions about these motives, without relying on any surprises about the true nature of the interaction. We present two experiments that manipulate perceived motives of the first-mover by varying the second-mover's access to different response options outside of the sub-game of interest. The first experiment features a between-subject design that tests the effect of perceived motives on reciprocal decision-making. In particular, it compares positive reciprocity of second-movers in response to a helpful action when the first-mover could have been motivated to help by fear of punishment, with their positive reciprocity in response to the same helpful action when the punishment-avoidance motive was absent.

The second experiment explores the relationship between perceived motives and reciprocity in greater depth. It elicits second-movers' demand for rewards and their inferences of altruism regarding a helpful first-mover across three within-subject treatments where first-movers i) have no strategic incentives to help, or ii) may be motivated to help by the hope of rewards, or iii) may be motivated to help by the fear of punishment. As a result, it is able to explore two additional paths of inquiry. First, it separately identifies the influence of perceptions regarding two different strategic motives, punishment-avoidance and reward-seeking, on reciprocal decision-making. Second, this experiment explores the mechanism by which the existence of strategic incentives may influence reciprocity. Specifically, it explores whether an increase in the strength of strategic incentives to help leads to a decline perceptions regarding the degree to which a helpful first-mover is altruistic, and whether these perceptions in turn shape reciprocal responses.

Previous exprimental literature shows that the same helpful action is seen as kinder if the first-mover had agency in his decision (Blount, 1995; Offerman, 2002; Charness and Haruvy, 2002; Charness, 2004; Charness and Levine, 2007; Falk et al., 2008; Klempt, 2012), and if the first-mover chose the best he could among his choice alternatives (Brandts and Sola, 2001; Nelson, 2002; McCabe et al., 2003). In our experiments, we keep agency and the choice set of the first-mover constant across treatments. In particular, the first-mover is given the same two choices, these choices are common information to all players, and his choice is directly implemented. Therefore, the design does not allow for any ambiguity regarding the choice the first-mover wanted to implement, and does not present any variation across treatments in what else he could have chosen.

Among the previous experimental work of the role of intentions, only Stanca et al. (2009) provide a manipulation that varies the perceptions of motives and intentions at the same time. In particular, they compare the degree of positive reciprocity in the second-stage of a constituent game across two between-subject treatments. In treatment 1, the first-mover decides on a transfer that gets multiplied before being given to the second-mover, and the second-mover in return decides on a transfer that gets multiplies before being given to the first-mover. In treatment 2, the first-mover makes the same transfer decision as the first-stage of treatment 1 in the context of a modified dictator game, without knowing that there will be a second-stage. This decision is followed by a surprise, where the second-mover makes a transfer decision in a modified featuring the same decision as the second-stage of treatment 1. The results show that the slope of the second-movers' rewards with respect to transfers from the first-mover is steeper when first-stage choices are made in absence of knowledge of the second stage.

The results seem to support the conjecture that the second-movers are more likely to positively reciprocate to a helpful action if that action is motivated by genuine kindness rather than strategic motives. However, even though one may be tempted to ascribe the role of perceived motives to explain the results, the results can also be explained by intention-based reciprocity. In particular, according to the Falk and Fischbacher (2006) model, the first-mover is perceived to have kinder intentions in treatment 2 because he expected the second-mover to obtain a higher payoff as a result of not having the opportunity to exercise the costly reward option. Consequently the authors rely on the Falk and Fischbacher (2006) intention-based reciprocity model to explain their results.

The current paper builds on this attempt to study the role of motives in several important ways. The experiments in this paper i) provide evidence for differences in the overall *level* of reciprocal responses across treatments, ii) isolate the role of motives from the role of perceived intentions, and, iii) do so without misleading subjects about the nature of the interaction.

The experiments are composed of four parts. Part 1 elicits other-regarding preferences in the absence of strategic considerations, part 2 elicits expectations of other-regarding preferences in the given session, part 3 elicits first- and second-movers' choices in a sequential reciprocity game. Bolton et al. (1998) voice a fair warning regarding models of reciprocity where the perceived kindness of actions depend on beliefs: "Beliefs about intentions are not directly observable, and hence the evidence on the intentions hypothesis is particularly susceptible to confounding with other strategic issues." Therefore, part 4 elicits several belief measures regarding the sequential reciprocity game. The focal interest is to provide evidence for differences in the reciprocal behavior in the second-stage of the interaction presented in part 3 as we manipulate the strength of strategic incentives to help in the fisrt-stage. Therefore, for both experiments, we present the reciprocal interaction before we detail the entire protocol.

2.1 Experiment 1

The reciprocal interaction

Experiment 1 presents a two-stage reciprocity game (Game Γ_1) depicted in Figure 1. In the first-stage, player A makes a choice between (S) and (H). If he chooses (S), he receives \$4 and the second-mover receives \$1. If he chooses (H), he sacrifices \$1 in order to increase the payoff of player B by \$1.

In the second-stage, having observed the decision of player A, player B in turn makes a decision that affects player A's material payoffs. If player A chooses (H), player B chooses between (l), whereby she leaves the distribution that player A delivered unchanged, and a costly reward option (r), whereby she can increase player A's payoffs by \$2.50 by sacrificing \$0.50 from her own payoffs. The outcome (H, l) yields (\$3, \$2) and the outcome (H, r) yields (\$5.50, \$1.50). If player A chooses (S), player B chooses between (l), whereby she leaves the distribution that player A delivered unchanged, and a costly option whereby she sacrifices \$0.50 from her payoffs to change the payoff of player A to the amount m. Game Γ_1 takes on very different natures depending on the value of m.

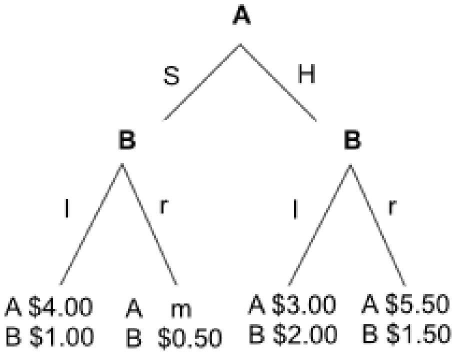

Figure 1: Game Γ_1

The experiment manipulates the first-mover's motives in a between-subject design by setting m=\$6.50 in Treatment RO, and m=\$1.50 in treatment RP. The treatments are symmetric: treatment RP (RO) gives player B a costly punishment (reward) option if player A chooses (S), whereby she can decide to sacrifice \$0.50 in order to decrease (increase) player A's payoff by \$2.50. In treatment RO, Game Γ_1 is a simple trust game, offering first-movers the potential of being rewarded for choosing (H). Therefore, among the first-movers who choose (H), there could be a mix of people motivated by other-regarding preferences (altruistic motive) and/or hope of rewards (reward-seeking motive). In treatment RP, Game Γ_1 is a judgment game that also offers the potential of being punished for not choosing (H)[2]. Clearly, treatment RP gives stronger strategic incentives for the first-mover to choose (H): the first-movers who choose (H) could be motivated by altruism, hope of rewards and/or fear of punishment (punishment-avoidance motive). Given the stronger incentives to choose (H), the proportion of people motivated by altruistic motives among the helpful first-movers should be lower.

Note that the material payoffs of player B are the same across the two treatments. The design also keeps the action space of the first-mover, and that of the second-mover conditional on the first-mover choosing (H) constant. The two treatments only vary in the action space of the second-mover conditional on the first-mover choosing (S). The payoffs of the second-mover are never higher than that of the first-mover, which guarantees that 1) players' relative standing is kept stable, and 2) punishing player A always decreases the magnitude of the inequality of material payoffs between player A and player B, and rewarding him always increases it.

[2] Co-existence of rewards and punishments in the second-stage of a reciprocal interaction are not very common in the literature. Abbink et al. (2000) presented a "moonlighting" game where kind and unkind actions were available to both the first- and second-movers. Offerman (2002) presented a "hot response" game where rewarding and punishing were available options for the second-mover, regardless of first-mover's choice. Experiment 1 design is closer to Al-Ubaydli and Lee (2012) in presenting a "judgment" game where a reward option is available if the first-mover has been helpful, and a punishment option is available if the first-mover has chosen selfishly.

The Protocol

A total of 258 participants (recruited through ORSEE) participated in eighteen 60-minute sessions at the [blinded for review] Lab during November 2014. In each session, an even number of participants (10 to 20 participants per session) interacted using the software Z-Tree (Fischbacher, 2007) in a double-blind payoff protocol. Only one treatment was implemented for all subjects in a session, and subjects could only participate in one session. The participants were told that the session would last 60 minutes and had 4 parts. Each part was introduced with its own set of instructions to all subjects at the same time. Subjects were informed that their payments from each part were independent of their choices in the future or previous parts of the experiment. All identities and choices were kept anonymous throughout the experiment.

Subjects earned a fixed participation fee of $5. They also earned additional payments from each of the four parts. If the parts included more than one task, one task was selected at random from each part to determine additional payments. Subjects learned the randomly selected tasks and their earnings at the end of the study. The average total earnings were $15.14. Experimental instructions, questions and detailed protocol are included in the Experimental Instructions Appendix.

In part 1, half the participants were randomly and anonymously assigned the role of player A and the rest were assigned the role of player B. The participants kept these roles throughout the experiment. Player As made decisions in six binary modified dictator games, while player Bs waited. Player As were asked to choose between ($4.50, $1.50) and ($4, $4); ($2.50, $0) and ($2, $1.50); ($4, $1) and ($3, $2); ($5, $2) and ($4, $4); ($1, $4) and ($0.50, $6.50); ($2, $3) and ($1.50, $5.50) where the first amount denotes the payoff of player A and the second denotes that of player B. Note that these choices included some of the same binary options player A and player B would choose between later in the context of Game Γ_1. All participants were told that one game from part 1 would be randomly chosen, and player A's choices in that game would determine payments for that player A and a randomly matched player B at the end of the experiment.

In part 2, four of the modified dictator games from part 1 were presented to all the participants. Participants were incentivized to predict the percentage of player As in that session who had chosen each option in the following decision tasks presented as modified dictator games: ($2.50, $0) and ($2, $1.50); ($4, $1) and ($3, $2); ($1, $4) and ($0.50, $6.50); ($2, $3) and ($1.50, $5.50). They earned $4 if they guessed the proportion of player As who picked each option correctly, and their earnings declined quadratically as a function of their inaccuracy. They were informed that one question would be chosen at random at the end of the experiment to determine their earnings from part 2.

In part 3, all participants in a given session made decisions in either treatment RO or treatment RP versions of Game Γ_1. All the details of the game were explained to all participants at the same time. The program matched player As and player Bs randomly and handled communication of choices anonymously.

In part 4, player A's first-order beliefs about player B's responses, player B's first-order beliefs about player A's choices and player B's second-order beliefs (expectations regarding player A's first-order beliefs) were elicited. The participants were again incentivized for accuracy in the same manner and were informed that one question would be chosen at random at the end of the experiment to determine their accuracy payments from part 4. At the end of the experiment, the program displayed the earnings to each participant, and explained how these earnings were achieved by going over their decisions in the tasks that were selected from each part. Each participant was paid privately.

We briefly explain our motivation for including all four parts in the experimental design. Asking Player As to make choices in modified dictator games in part 1 allows us to learn about their other-regarding preferences (Charness and Rabin, 2002)[3]. Note that part 1 included a game that presented the same choice options as (S) and (H) in Game Γ_1, as well as games that presented the same choice options that player Bs in the two sub-games of Game Γ_1 would face. Knowing their choices in a situation where player Bs cannot respond provides information regarding how much of the helpful behavior in Game Γ_1 results from strategic considerations. The predictions elicited in part 2 serve as baseline beliefs about the degree of altruism in the population of participants in a given session. This information is useful in determining whether the beliefs elicited in part 4 reflect an understanding of strategic considerations on the part of the first-movers, as well as an understanding of the mental model of the second-movers. Therefore, results from the first two parts of the experiment can provide baseline of behavior and expectations when reciprocal or strategic considerations are absent. Finally, the beliefs elicited in part 4 are useful in establishing internal validity of the experiment, exposing the mental models of players regarding the game and the other player, identifying beliefs regarding reciprocity separately from beliefs regarding distributional preferences, and for ruling out alternative explanations based on second-order beliefs.

Hypotheses

The central hypothesis of Experiment 1 is that player Bs are less likely to reward (H) if they perceive it to be more likely to be motivated by strategic considerations. Given that previous literature showed that sanctions in combination with rewards are more motivating than rewards alone (Andreoni et al. 2003), the manipulation in Experiment 1 should motivate more player As to choose (H) in treatment RP, and do so for fear of punishment. Therefore we propose the following hypothesis to establish the intended manipulation:

> *H0: More player As choose (H) in treatment RP than in treatment RO.*

If *H0* holds, the central hypothesis of this experiment can be stated as:

> *H1: In response to player A choosing (H), a higher proportion of player Bs will choose (r) in treatment RO.*

[3] One may be concerned that telling the subjects that there will be another decision task following the dictator games may make the dictators more generous. If such a bias existed, we would underestimate the degree of strategic considerations in the reciprocal interaction. Cox et al. (2008b) investigated this methodology question. They found that tests for trust, fear and reciprocity using data from a within-person experiment that involves the moonlighting game as well as dictator games imply the same conclusions as tests using data from across-subjects experiments where different groups of people play these games. Also note that even if such a bias exists, it would not confound our hypothesis tests, since both treatments share the same structure.

In addition to the main hypothesis, elicited beliefs allow us to investigate ancillary hypotheses regarding the sophistication of player Bs regarding player As motives in the reciprocal interaction, player As understanding of player Bs reciprocal feelings, and player Bs expectation of this understanding.

> *H2a: Player Bs expect more player As to choose (H) in treatment RP than in treatment RO.*
>
> *H2b: Player As expect more positive reciprocity from player Bs in treatment RO than in treatment RP.*
>
> *H2c: Player Bs think that the expectations of player As regarding the likelihood of player Bs choosing (r) in response to (H) are higher in treatment RO than in treatment RP.*

Results

The first column of Table 1 displays the number and percentage of player As' choosing the option that gives them the higher payoff (option 1) in the modified dictator games presented in part 1. The second and third columns respectively report player As' and player Bs' average beliefs regarding the proportion of player As choosing option 1 in the four modified dictator games presented in part 2. In line with early findings of Charness and Rabin (2002), dictators are more likely to sacrifice their own payoffs to help another person when their payoffs are higher than the other person, and when the sacrifice produces a larger gain on the part of the other person. Beliefs reflect an understanding of these preferences, as they follow the ordering of choice proportions. However, subjects seem to be averse to reporting beliefs close to the extremes (0% or 100%), thus displaying some conservatism bias.

Table 1: Behavior and Beliefs regarding Behavior in Modified Dictator Games

Choice Question	N		Player A's	Player B's
(Option 1) vs. (Option 2)		Option 1 Choice	Option 1 Beliefs	Option 1 Beliefs
		(1)	(2)	(3)
($4.50, $1.50) vs. ($4.00, $4.00)	129	37 (29%)		
($2.50, $0) vs. ($2.00, $1.50)	129	37 (29%)	43%	37%
($4.00, $1.00) vs. ($3.00, $2.00)	129	92 (71%)	60%	54%
($5.00, $2.00) vs. ($4.00, $4.00)	129	71 (55%)		
($1.00, $4.00) vs. ($0.50, $6.50)	129	89 (69%)	71%	77%
($2.00, $3.00) vs. ($1.50, $5.50)	129	95 (74%)	70%	77%

Remember that the first-stage of Game Γ_1 presents a choice between {$4 for player A, $1 for player B} and {$3 for player A, $2 for player B} to player As. When player As were asked to choose between the same options in part 1 where player Bs could not respond in any way (Table 1, row 3), 92 of the player As (71% of them) chose to keep $4 to themselves. On average, player As thought that 60% of other player As would choose this option. Similarly, on average, player Bs believed that 54% of player As would choose this option.

We expect a larger fraction of player As to sacrifice \$1 from their payoffs in the first-stage of Game Γ_1 than they did in part 1, as Game Γ_1 offers strategic incentives for doing so. Table 2 summarizes the choices observed in Game Γ_1 across the two conditions. Indeed, we see that player As are more willing to sacrifice \$1 to help player B in the first-stage of Game Γ_1 than in part 1, both in treatment RO (29% vs. 66%, McNemar test, $\chi^2(1) = 23.15$, $p = 0.000$) and in treatment RP (29% vs. 93%, McNemar test, $\chi^2(1) = 39$, $p = 0.000$). Comparing the proportion of player As who chose (H) across the two treatments, we find that more player As choose (H) in treatment RP than in treatment RO, in support of hypothesis *H0* (93% vs. 66%, Chi-square test, $\chi^2(1) = 14.25$, $p = 0.000$). Presumably, fear of punishment motivated player As to be more helpful in treatment RP, as Player As reported a relatively high potential punishment expectation in this treatment RP (41% on average, one sample t-test, $t = 36.52$, $p = 0.000$).

Table 2: Behavior in Game Γ_1

	# sessions	N	A Choice		B Response			
			S	H	l|S	r|S	l|H	r|H
Treatment RO	10	140 (70 pairs)	24	46	23	1	20	26
Treatment RP	18	118 (59 pairs)	4	55	3	1	36	19

Experiment 1 is designed to test the hypothesis that the same helpful action (H) will trigger a higher degree of positive reciprocity in treatment RO than in treatment RP (hypothesis $H1$). In support of this hypothesis, only 20 out of 56 (36%) of player Bs rewarded (H) in treatment RP, whereas 26 out of 45 (58%) of player Bs rewarded (H) in treatment RO (Chi-square test, $\chi^2(1) = 4.90$, $p = 0.027$). This result suggests that second-movers are less likely to positively reciprocate to the same helpful action when the reciprocal interaction provides stronger strategic incentives for the first-movers to be helpful.

Comparing second-movers' first-order expectations across the two treatments can give us a sense of whether they are cognizant the incentives each treatment presents to the first-movers. Table 3 summarizes the beliefs elicited in part 4 of the experiment. Player Bs expected meaningful differences in the extent to which player As were willing to choose (H) in treatment RP (41%) versus in treatment RO (62%) (Two-sample Wilcoxon rank-sum (Mann-Whitney) test, $z = 5.78$, $p = 0.000$), providing support for hypothesis *H2a.* This result gives further support to the conjecture that second-movers contemplate first-mover's motives.

Table 3: Beliefs about Game Γ_1 play across two conditions

	N[4]	B FOE		A FOE				B SOE			
		S	H	l|S	r|S	l|H	r|H	l|S	r|S	l|H	r|H
Treatment RO	70	59%	41%	81%	19%	60%	40%	84%	16%	57%	43%
Treatment RP	59	28%	62%	59%	41%	70%	30%	46%	54%	73%	27%

If player As have a good mental model of player Bs, we would expect to see lower expectations of positive reciprocation for choosing (H) in treatment RP, as stated by hypothesis *H2b.* On average, player As expected 40% of player Bs in treatment RO, and 30% of player Bs in treatment RP to reward (H) (Two-sample Wilcoxon rank-sum (Mann-Whitney) test, $z = -1.74$, $p = 0.082$). Note that in part 2, player As reported an average expectation of only 24% of people sacrificing 50 cents in order to help another participant by \$2.50 in the modified dictator game that corresponded to the sub-game player B faces. Therefore, expectations of positive reciprocity can be identified by comparing how likely a player A thought the general population of participants were to choose (R) in a modified dictator game to how likely they thought this choice was when the person was responding to (H) in Game Γ_1. In support of hypothesis *H2b,* we find that their expectation of costly rewards from player Bs reflect an expectation of positive reciprocity over and beyond an expectation of altruism in treatment RO, (Wilcoxon sign-ranked test, $z = -4.25$, $p = 0.000$), but not in treatment RP (Wilcoxon sign-ranked test, $z = -1.01$, $p = 0.311$).

Finally, player B's second-order beliefs allow us to investigate whether they expected player As to have an understanding of their reciprocal feelings (hypothesis *H2c*). On average, player Bs thought player As expected an average of 43% of player Bs to reward (H) in treatment RO, and an average of 27% of player Bs to reward (H) in treatment RP (Two-sample Wilcoxon rank-sum (Mann-Whitney) test, $p = 0.006$), providing support for this hypothesis.

A striking feature of the results presented in Table 3 is how aligned Player As' FOEs and Player Bs' SOEs are. For example, player As expect 40% of player Bs on average to reward (H) in treatment RO, and player Bs think player As expect 43% of player Bs to do so. Similarly, player As expect 30% of player Bs on average to reward (H) in treatment RP, and player Bs think player As expect 27% of player Bs to do so. Moreover, these expectations are also reflective of actual behavior. These data suggest that the participants were sophisticated about the second-stage behavior, and each others' expectations.

Summary Experiment 1 compares the level of positive reciprocity second-movers display towards a helpful first-mover in a situation where the first-mover could have been motivated to help because he feared punishment, with the level of positive reciprocity to the same helpful action in a situation where the punishment option in the second-stage was absent. The main finding of Experiment 1 is that second-movers are less likely to positively reciprocate to the same helpful action when the game form provides stronger strategic incentives for the first-movers to be helpful. This finding presents novel evidence that the second-mover's reciprocity is influenced by her perceptions of the first-mover's motives, above and beyond any outcome of his actions.

2.2 Experiment 2

Experiment 2 extends Experiment 1 in several aspects. First, it explores the mechanism by which the existence of strategic incentives may influence reciprocity. We expect there to be a close relationship between perceptions of motives and perceptions of altruism regarding a helpful person. Therefore, Experiment 2 elicits beliefs about the altruism of the person taking the helpful action in a within-subjects design. Second, Experiment 2 allows for comparing the role of perceived reward-seeking motives and the role of perceived punishment-avoidance motives separately to the no-incentive benchmark, while Experiment 1 features a potential for the first-mover to be motivated by reward-seeking across all treatments.

The reciprocal interaction

Experiment 2 investigates positive reciprocity in the context of a probabilistic sequential game where the same helpful action could be motivated by punishment-avoidance, reward-seeking, and/or altruism. Consider Game Γ_2 depicted in Figure 2. First, Player A chooses between (S), which pays him \$4.50 and player B \$2.50, and (H), which pays both players \$4. Therefore, in the first-stage player A decides whether to take the option that pays him more, or the option where he sacrifices 50 cents to increase player B's earnings by \$1.50. Then, nature chooses either 0, 1 or 2. If Nature chooses 0, the game ends, and the option player A chose determines both players' final payments. If nature chooses 1, the game ends if player A chose (S). But if player A chose (H), then player B gets to choose between (N) and (R). The choice of (N) leaves the allocation player A chose unaltered. The choice of (R) costs player B 50 cents and *increases* player A's earnings by \$1.50. The outcome of (S,1) yields (\$4.50, \$2.50), the outcome of (H,1,N) yields (\$4, \$4) and the the outcome of (H,1,R) yields (\$5.50, \$3.50). If nature chooses 2, the game ends if player A chose (H). But if Player A chose (S), then player B chooses between not altering the allocation player A choice (N) or paying 50 cents to *decrease* player A's earnings by \$1.50 (P). The outcome of (H,2) yields (\$4, \$4), outcome of (H,2,N) yields (\$4.50, \$2.50) and the the outcome of (S,2,P) yields (\$3, \$2).

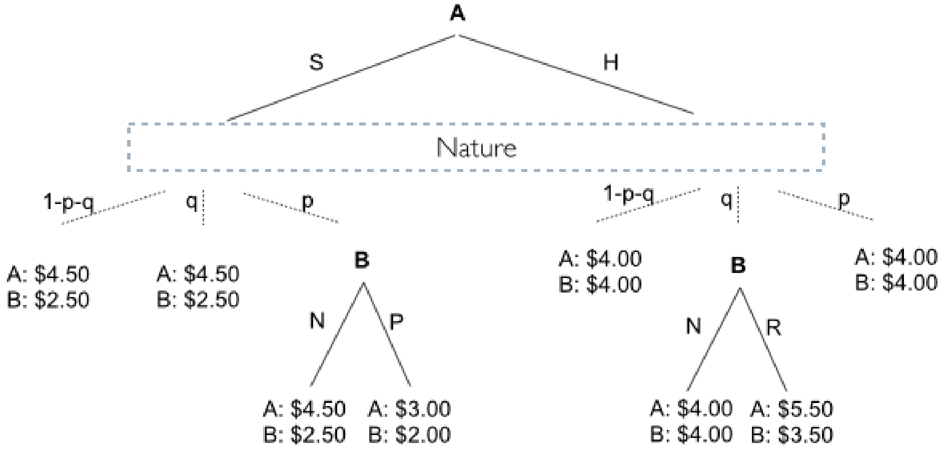

Figure 2: Game Γ_2

Let q be the probability that nature chooses 1, and p be the probability that nature chooses 2. Consider how changing p and q may affect player A and player B behavior. First, consider q approaching 1. Then if player A chooses (S), player A will get \$4.50. But if player A chooses (H) and nature chooses 1, Person B may choose (R), giving player \$5.50. Depending on his belief about how likely player B is to choose (R), player A may be inclined to choose (H) even if he puts no weight on player B's well-being. Thus, as q gets larger, there will be more player As who are primarily motivated by reward-seeking motives among those who choose (H). Similarly, consider p approaching 1. Then if player A chooses (H) and nature chooses 2, player A will earn \$4. But if player A chooses (S) and nature chooses 2, player B may choose (P), giving player A only \$3. Therefore, player A may be inclined to choose (H) in order to avoid potential punishment. Thus, as p gets larger, there will be more player As who are primarily motivated by punishment-avoidance among player As who choose (H). In contrast, when p+q is close to zero, player A would only choose (H) if he genuinely prefers the more equitable allocation (\$4, \$4) to the more profitable allocation (\$4.50, \$2.50).

Experiment 2 features three treatments: 1) Treatment N, where 1-p-q=0.98, and the first-stage player therefore expects the second-stage player to be a passive recipient most the time, 2) Treatment RN, where where q=0.98, and the first-stage player therefore expects the second-stage player (almost always) to have the option to reward a helpful action, and 3) Treatment PN, where p=0.98, and the first-stage player therefore expects the second-stage player (almost always) to have the option to punish a selfish action. In treatment N, player A is mainly motivated by altruism, however in treatments PN and RN, he can also be motivated by punishment-avoidance and reward-seeking respectively.

We want to compare how likely player Bs are to reward player A for choosing (H) across the three treatments. The probabilistic design allows us to elicit reward demand from player Bs even in cases where player Bs were expected to be able to reward player As, without misleading participants about the nature of the interaction. In each of the three treatments, player B is asked to designate her response for each contingency using the strategy method.[5] In particular, player Bs are asked to indicate, in each treatment, whether they would choose (R) or (N) if player A chose (H) and Nature chose 1. They are also asked to indicate whether they would choose (P) or (N) if player A chose (S) and Nature chose 2 in each treatment. For example, in treatment N, player Bs will not get the chance to act 98% of the time, and therefore player As are mostly motivated by altruism, and player Bs know that. We ask player Bs whether they would choose (N) or (R) if they faced with these options. In this manner, we learn about the reciprocal preferences across all treatments.

Is player B more likely to want to reward (H) when player A expected her not to be able to respond or when he expected her to be able to punish? Note that unlike Experiment 1, Experiment 2

[5] Investigating differences between the direct-response and the strategy methods, Brandts and Charness (2011) showed that any treatment effect demonstrated using the direct-response method could also be demonstrated using the strategy method. Also, Charness and Levine (2007) noted that the strategy method should be innocuous if it does not interact with the treatment status and tests changes in the rate of positive responses, rather than the level of the rate.

focuses on one strategic motivation in each treatment, minimizing any expectations of the alternative strategic motivation. As a result, it permits comparing reciprocity to a helpful action in a case where a punishment-avoidance motive is likely to drive helpfulness to reciprocity to the same action in a case where strategic motivations are not likely to drive the first-mover's choice.

Comparing reciprocity differentials between treatments N and NR is interesting, however, as we dicussed before, such differences can also be driven by intention-based reciprocity[6]. Therefore, while we also analyze and present the response differences across treatments N and NR, we rely on the comparison of reciprocal differences across treatments N and NP to provide evidence for the distinct role of motives. While the comparison of N and NR treatments are not our main focus, let us briefly comment on some of the relations this comparison has to the previous experimental literature. The comparison of positive reciprocity across the two treatments tested by Stanca et al. (2009) and treatments N and NR in Experiment 2 both address the following question: "How much positive reciprocity do we observe for the same helpful action when the helpful action was motivated by altruism versus when it could also be motivated by reward-seeking?" While Stanca et al. (2009) use a between-subject design, Experiment 2 presents within-person variation and does not rely on an element of surprise. Strassmair (2009) also compares two treatments similar to the N and NR treatments in Experiment 2 by varying the probability with which the second mover in a trust game could reciprocate to the first-mover's choice among five possible transfers. She does not find any differences in the second-mover's reciprocal choices to a given level of transfer. Experiment 2's design differs from Strassmair (2009) in multiple ways, making direct comparisons of results difficult.

The Protocol

A total of 176 participants participated in eleven 45-minute sessions conducted at the [blinded for review] Lab. Each session had 12-20 subjects. Subjects earned a \$5 participation fee and up to \$5.50 in additional earnings.

In part 1, all participants made decisions in eight modified dictator games, each of which presented two options where the payoffs were denoted in tokens. The conversion rate was 200 tokens = \$1. In particular, all participants were asked to choose between (800, 800) and (700, 1100); (800, 200) and (600, 400); (900, 500) and (800, 800); (500, 900) and (400, 1200); (500, 0) and (400, 300); (900, 0) and (800, 200); (400, 600) and (300, 1100); (500, 900) and (400, 600).

The choices in part 1 elicited altruistic preferences of all the participants. Knowing each person's other-regarding preferences allows us to separately identify transfers resulting from reciprocity and the role of strategic incentives in Game Γ_2. Among these dictator games, player As were presented

[6] Moreoever, note that Nature is equally unlikely to choose 1 in treatments N and NP, but is very likely to choose 1 in treatment NR. We may worry about comparing responses elicited across sub-games that have drastically different probabilities of being carried out. For example, when the probability of implementation is low, the participants may have other objectives, such as seeming nice to the experimenter. We tried to eliminate such concerns by making all actions and all pairings anonymous. More importantly, any bias generated by the low probability of the event should be common to both treatments N and NP. As a result, we do not expect such potential biases to impact the difference in reciprocal responses between treatments N and NP - the main comparison of interest in Experiment 2.

with games that presented the same choice options as (S) and (H) in Game Γ_2 as well as games that presented the same choice options that player Bs faced in some the sub-games of Game Γ_2. Elicitation of preferences in these games allows us to incentivize reporting of truthful expectations regarding the genuine kindness of helpful first-movers in Game Γ_2, as we explain below.

In part 2, participants were asked to predict the percentage of participants in that session who chose each option across four of the modified dictator games from part 1. The participants were incentivized for accuracy. These predictions served as baseline beliefs about the degree of genuine kindness in the population of participants in a given session.

Part 3 presented the subjects three within-person treatments of Game Γ_2.[7] The payoffs were denoted in tokens, where 200 tokens=\$1. Participants were randomly assigned to the role of player A and player B. Each player A was randomly and anonymously matched with one player B for each treatment. As player As made a choice between (S) and (H) in each treatment, player Bs were asked to indicate their preferred choices for each contingency using the strategy method in that treatment.

In part 4, player A's first order beliefs about player B's responses[8] and player B's first order beliefs about player A's choices for each treatment were elicited using accuracy incentives. Part 4 also elicited player Bs' altruism inferences regarding player As who were helpful in each treatment. We wanted to know what proportion of the helpful player As player Bs thought would have behaved similarly if it weren't for the reciprocal nature of each treatment of Game Γ_2. In particular, player Bs were asked "Only consider the group of player As who chose H in (a given treatment). Among these player As, what percentage chose each of the following options presented to them in Part 1 of the study? Option 1. 500 tokens for him/herself, 0 for the other participant ____% , Option 2. 400 tokens for him/herself, 300 for the other participant _____%." Note that both in the first-stage of Game Γ_2 and in this modified dictator game, player As decide whether they want to sacrifice 100 tokens (50 cents) in order to increase the payoff of player B by 300 tokens (\$1.50). Therefore, player Bs' beliefs regarding helpful player As' choices in this modified dictator games gives us an idea of their beliefs regarding how they would choose in the first-stage of Game Γ_2 if it were not for strategic considerations. We employ a within-subjects design in Experiment 2 in order to test whether the heterogeneity in these altruism inferences across subjects explain the heterogeneity they display in their reciprocal responses.[9] At the end of the experiment, one question was chosen at random to determine payments of all participants. Further details of the instructions, questions and protocol of Experiment 2 are included in the Experimental Instructions Appendix.

[7] The order of the three treatments were counterbalanced among the following three sequences (PN-N-RN, N-RN-PN, RN-N-PN) to randomize the first treatment that participants in each session see. The ordering of the treatments did not affect any of the results.

[8] Note that because these beliefs are incentivized with respect to the actual proportions in the session, we could only ask for the FOE's of player As concerning the reaction of player Bs in the sub-game that was implemented with 98% chance.

[9] Charness et al. (2012) discusses the advantages and disadvantages of within and between subject designs. This article establishes the influence of perceived motives on reciprocity with both designs. In both designs, we minimize experimenter demand effects by keeping choices anonymous. In Experiment 2, we vary the order of treatments to control for anchoring, but do not find any order effects.

Hypotheses

If player As are motivated by strategic considerations above and beyond altruism, and if punishment is a stronger motivator than rewards, we expect

H0: The percentage of player As choosing (H) is the greatest in treatment NP, followed by in treatment NR and the least in treatment N.

In line with player A behavior, we expect player B's to intuit the differences in willingness to choose the helpful action across treatments when strategic incentives exist:

> *H1: Player Bs believe that the percentage of player As choosing (H) is the greatest in treatment NP, followed by in treatment NR and the least in treatment N.*

If player B's are sophisticated about the selection of player As induced by strategic motivations, we would expect them to report higher expectations of altruistic player As among H-choosers in treatments where strategic incentives are weaker:

> *H2: The altruism inferences regarding player As who chose (H) are the highest in treatment N, followed by in treatment NR and the lowest in treatment NP.*

If the player B cares about why player A was helpful, we expect the following to hold true:

> *H3: The intended positive reciprocity in response to (H) decreases with the strength of the strategic motivation. (N>NR>NP if H1 holds).*

H3 is the central hypotheses Experiment 2 is designed to test. Finally, if kindness inferences moderate the degree of reciprocity towards a helpful action within-person, we would expect that

> *H4: A within-person increase (deterioration) of kindness inference about helpful player As from one treatment to another is associated with an increase (decrease) in player B's propensity to reward player A for being helpful.*

Results

Table 4 summarizes the choices of player As and player Bs in the dictator games presented in part 1 and the beliefs regarding behavior in these games as elicited in part 2. Again, we see that dictators are more likely to sacrifice their own payoff to help the other participant when their payoffs are larger than that of the other person's and when the sacrifice leads to a larger gain. Expectations regarding choices are mostly in line with observed behavior, albeit slightly conservative.

Table 4: Behavior and Beliefs regarding Behavior in Modified Dictator Games

Choice Question	N	Option 1	Option 1	Option 1 Beliefs	Option 1 Beliefs
(Option 1) vs. (Option 2)		Player A	Player B	Player A	Player B
(800, 800) vs. (700, 1100)	88	70%	80%	75%	76%
(800, 200) vs. (600, 400)	88	60%	49%		65%
(900, 500) vs. (800, 800)	88	44%	41%		46%
(500, 900) vs. (400, 1200)	88	69%	73%	78%	
(500, 0) vs. (400, 300)	88	29%	20%	45%	44%
(900, 0) vs. (800, 200)	88	31%	27%		
(400, 600) vs. (300, 1100)	88	69%	72%	66%	
(500, 900) vs. (400, 600)	88	81%	79%		

Some of these dictator games represent the same choices presented in Game Γ_2 and therefore can provide a baseline of behavior and expectations when reciprocal or strategic considerations are absent. For example, when faced with the same two options in the first-stage of Game Γ_2, 44% of player As chose the option {900 tokens for me, 500 tokens for another participant} over the option {800 tokens for me, 800 tokens for another participant} in part 1 (Table4 , row 3). On average player Bs expected 46% of player As to do so. Also, when faced with the same options the sub-game of interest in game Γ_2 presented player Bs, 80% of player Bs choose {800 tokens for me, 800 tokens for another participant} over {700 tokens for me, 1100 tokens for another participant} in part 1 (Table 4 , row 1) and on average As expected 75% of Bs to do so. Clearly, we would expect both the behavior in and expectations regarding these choices to be different in Game Γ_2.

Table 5 presents the behavior and expectations in Game Γ_2 across the three treatments. A total of 81% of player As chose (H) in treatment NP, followed by 72% in treatment NR and 48% in treatment N (matched-pairs sign test, N<NR: $p = 0.000$; N<NP: $p = 0.000$; and NR<NP: $p = 0.048$). This data gives support to *H0*, suggesting both rewards and punishment are successful motivators, with punishment being the stronger of the two. Clearly, player As would not have been motivated by the mere existence of reward and punishment options in player B's disposal, if they did not think that player Bs were likely to use them when they had access to these options. Indeed, player As on average expected 40% of player Bs to choose R in treatment NR if they chose (H), and they on average expected 43% of player Bs to choose P in treatment NP if they chose (S).

Table 5: Observed Behavior and First-Order Beliefs in Game Γ_2

Treatment	N	% As	B's FOE	% Bs	A's FOE	% Bs	A's FOE
		choosing H	of H	choosing R | H	of R | H	choosing P | S	of P | S
Treatment N	88	48%	51%	63%		43%	
Treatment NR	88	72%	67%	52%	40%	42%	
Treatment NP	88	81%	75%	42%		50%	43%

In accordance with player As' choices, player Bs expected the highest proportion of player As (75%) to choose (H) in treatment NP, followed by player As in treatment NR (67%), and the lowest proportion of player As in treatment N (51%) (matched-pairs sign test, N<NR: $p = 0.000$; N<NP:

$p = 0.000$; and NR<NP: $p = 0.008$). This result provides evidence for hypothesis *H1*. It suggests that player Bs understood the differences in motivations across treatments and expected meaningful differences in the extent to which player As were willing to choose (H).

How did player Bs respond to helpful player As across the three treatments? Remember that in treatment N player As are not likely to be motivated by strategic motives. The percentage player B's choosing R in response to (H) in treatment N is 63%. Given that player Bs have to move away from an equal distribution of 800 tokens for each player to 700 tokens for themselves and 1100 tokens for player A in order to reward the choice of (H), and only that 20% would do so in part 1, 63% is a substantially positive reciprocal response. As hypothesized, player Bs were less reciprocal when (H) is chosen in the other two treatments. A total of 52% of player B's indicated that they would choose R if player A chose (H) in treatment NR and 42% of player B's indicated that they would choose R if player A chose (H) in treatment NP. The differences in reciprocal response rates are consistent with the ranking proposed by the main hypothesis *H3* (matched-pairs sign test. N>NR: $p = 0.047$; N>NP: $p = 0.000$; and NR>NP: $p = 0.047$). This result indicates that player Bs reciprocate more to the same helpful action the weaker the strategic incentives the situation presents for taking that helpful action.

We propose that the perceived motive behind a helpful action is closely tied to the perceived altruism of the person taking the action. Would player A have chosen the same action if he did not have any strategic incentives to do so? Experiment 2 directly elicits these inferences. Player Bs predicted on average 73% of player As who choose (H) in treatment N to choose {400 tokens for player A, 300 tokens for another participant} over {500 tokens for player A, 0 tokens for another participant}. However, they predicted an average of 65% of player A's who chose (H) in treatment NR, and 54% of player As who chose (H) in treatment NP to make the same choice (matched-pairs sign test, N>NR: $p = 0.061$; NR>NP: $p = 0.001$ and N>NP: $p = 0.000$). These results suggest that player Bs believed that strategic incentives to avoid punishment lead to a lower proportion of truly generous people among those who choose the helpful action than reward incentives do, in line with *H2* presented above[10]. In addition, player Bs also inferred that the H-choosers in the NP condition are not kinder than the population of player As in general, since they had reported an expectation (elicited in part 2) of 56% of player As choosing {400 tokens for player A, 300 tokens for another participant} over {500 tokens for player A, 0 tokens for another participant} in part 1 (Table 4, row 5).

Although, overall, player Bs think that a lower proportion of the H-choosers in treatments NP and NR are altruistic than in treatment N, they display considerable heterogeneity in the degree to which they the existence of each strategic motives to be implicating. For example, some players think that the H-choosers in treatment NR are much less likely to be motivated by altruism than

[10] Even though player B's are correct about the nature of type selection each treatment induces, they are pessimistic and conservative in their beliefs about the generosity of player As in this question. Looking at player A's actual choices in part 1 on this question, we see that 71% of all player As, 72% of those who chose (H) in treatment NP, 86% of those who chose (H) in treatment NR and 93% of those who chose (H) in treatment N chose {400 tokens for player A, 300 tokens for another participant} over {500 tokens for player A, 0 tokens for another participant} in part 1.

the H-choosers in treatment N, whereas others do not infer such a big difference. The within-person design of Experiment 2 allows us to ask whether differences in individual player B's altruism inferences across treatments are associated with changes in their responses.

<table>
<tr><th colspan="5">Among the 55 player B's with response (R | H) in treatment N</th></tr>
<tr><th>player B choice</th><th>N</th><th colspan="2">Altruism belief</th><th>Difference</th></tr>
<tr><td></td><td></td><td>(treatment N)</td><td>(treatment NR)</td><td></td></tr>
<tr><td>(R | H) in NR</td><td>39</td><td>79.8%</td><td>70.2%</td><td>-9.6%</td></tr>
<tr><td>(N | H) in NR</td><td>16</td><td>79.6%</td><td>57.7%</td><td>-21.9%</td></tr>
<tr><td></td><td></td><td>(treatment N)</td><td>(treatment NP)</td><td></td></tr>
<tr><td>(R | H) in NP</td><td>34</td><td>76.7%</td><td>60.1%</td><td>-16.6%</td></tr>
<tr><td>(N | H) in NP</td><td>21</td><td>84.7%</td><td>52.8%</td><td>-31.9%</td></tr>
</table>

Table 6: Within-person changes in kindness inferences and reciprocity towards H

Table 6 presents the altruism inferences of the fifty-five player Bs who reward H-choosers in treatment N. The first two rows in Table 6 split these player Bs based on whether they also reward H-choosers in treatment NR. We want to see whether those who withdrew rewards vary in the change in their altruism inferences from those who continue to reward H-choosers in treatment NR.

We find that using within-person difference approach eliminates the potential confounds arising from possible correlation of preferences and beliefs. In line with previous literature on projection bias, we see that player Bs who are more altruistic have higher expectations of altruism given helpful behavior[11]. If we simply test whether individuals with high kindness inferences are more likely to reciprocate to helpful behavior, we would be confounding the causal impact of kindness inferences with their baseline willingness to help in the given sub-game. Instead, we take an approach that controls for differences across individuals in order to infer a meaningful relationship between kindness inferences and concern withdrawal. In particular, we relate changes in kindness inference to changes in recirpcal behavior. The identifying assumption that player Bs who are more altruistic do not have lower degrees of deterioration in kindness inference across treatments. Our data supports this assumption[12].

The first column reports the percentage of H-choosers each group believes is altruistic in treatment N. The second column reports their average beliefs concerning the percentage of altruistic H-choosers in treatment NR, and the last column reports the difference. We see that player Bs who rewarded action (H) in treatment N but stopped rewarding it in treatment NR perceive a larger

[11] Looking at player Bs' choices in Part 1 and their kindness inferences in Part 4, we find a significant correlation between choosing ($3.50, $5.50) over ($4, $4) in Part 1, and beliefs concerning the percentage of altruistic H-choosers in treatment N ($p = 0.087$) and in treatment NP ($p = 0.033$). Similarly, we find a significant correlation between choosing ($4, $4) over ($4.50, $2.50) in Part 1, and beliefs concerning the percentage of altruistic H-choosers in treatment N ($p = 0.000$), in treatment NR ($p = 0.000$) and in treatment NP ($p = 0.037$).

[12] We do not find any siginificant correlation between choosing ($3.50, $5.50) over ($4, $4) in Part 1, and the degree to which beliefs concerning the percentage of altruistic H-choosers decline between treatment N and treatment NP ($p = 0.539$), or between treatment N and NR ($p = 0.426$). Similarly, we do not find any siginificant correlation between choosing ($4, $4) over ($4.50, $2.50) in Part 1, and the degree to which beliefs concerning the percentage of altruistic H-choosers decline between treatment N and treatment NP ($p = 0.138$), or between treatment N and NR ($p = 0.942$).

difference in the altruism of helpful player As , compared to those who continue to reward action (H) (Wilcoxon rank-sum (Mann-Whitney) test: $z = -2.12$, $p = 0.034$)[13].

The last two rows of Table 6 split the player Bs who rewarded H-choosers in treatment N based on whether they also reward H-choosers in treatment NP. Again, the player Bs who withdraw rewards for helpful behavior show a larger decrease in their altruism inferences regarding the H-choosers in treatment NP (Wilcoxon rank-sum (Mann-Whitney) test: z=2.31, p=0.017)[14]. In sum, the results show that a within-person increase (deterioration) of kindness inference about helpful player As from one treatment to another is associated with an increase (decrease) in player B's propensity to reward player A for being helpful. These results provide evidence for hypothesis *H4*.

Summary Experiment 2 compares the degree of reciprocity towards a helpful action when the second-mover knows that the first-mover mostly expected the second-mover not to be able to respond in the second-stage, with the degree of reciprocity towards the same action when the second-mover knows that the first-mover either mostly expected the second-mover to be able to reward a helpful action, or mostly expected her to be able to punish an unhelpful action. The results show that second-movers positively reciprocate to a helpful action more when the strategic incentives to be helpful are weaker. Results also reveal a direct association between inferences of altruism and reciprocal choices. We see that altruism inferences regarding player As who chose to be helpful decreases with the strength of the strategic motivation the game form presents to be helpful. Importantly, individual heterogeneity in kindness inferences explains individual differences in how much withdrawal of concern player Bs display when strategic motives to be helpful are present in the game form.

[13] Using a complementary data analysis, we can also look at the subset of player Bs who did not reward H-choosers in treatment NR and test whether within-person differences in inferences can predict which ones are likely to reward H-choosers in treatment N. Among the forty-two player Bs who did not reward H-choosers in treatment NR, twenty-six of them also did not reward H-choosers in treatment N. These player Bs did not see any difference in the composition of genuinely kind player As among H-choosers (reported average beliefs of 56.1% of genuinely kind player As among H-choosers in treatment N and average beliefs of 56.3% of genuinely kind player As among H-choosers in treatment NR). Compared to the sixteen player Bs (in second row of Table 6) who chose to reward H-choosers in treatment N even though they did not reward them in treatment NR, the average inference deterioration of these twenty-six player Bs is significantly lower (Wilcoxon rank-sum (Mann-Whitney) test: z=-3.43, p=0.001).

[14] We can also compare the changes in the kindness inferences of player Bs who did not reward H-choosers in treatment NP based on whether they rewarded them in treatment N. Among the fifty-one player Bs who did not reward H-choosers in treatment NP, thirty of them also did not reward H-choosers in treatment N. These player Bs on average reported a 12.1% decline in the composition of genuinely kind player As among H-choosers (reported average beliefs of 60.5% of genuinely kind player As among H-choosers in treatment N and average beliefs of 48.3% of genuinely kind player As among H-choosers in treatment NP). Compared to the twenty-one player Bs (in fourth row of Table 6) who chose to reward H-choosers in treatment N even though they did not reward them in treatment NR, the average inference deterioration of these twenty-six player Bs is significantly lower (Wilcoxon rank-sum (Mann-Whitney) test: z=3.39, p=0.0001).

3 Discussion and Conclusion

3.1 Relation of results to reciprocity models

The reciprocity literature has proposed three main determinants of reciprocal decision-making. Outcome-based models of altruism and reciprocity (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) model the positive relationship between a helpful action and a reaction based only on preferences over payoff allocations. Because our experiments compare differences in reciprocal reactions to the same helpful action, and keep the payoff structure in the sub-game of interest constant, the results cannot be rationalized by outcome-based models.

A second class of models emphasize people's desire to punish hostile intentions and reward kind intentions (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). In these models, beliefs regarding what the first-mover intended for the second-mover are central in evaluating the kindness of an action: the second-mover considers an action to be relatively kind if she believes that the first-mover intended the second-mover's material payoff to be larger than a fair benchmark as a result of his action[15]. While the Falk and Fischbacher (2006) (FF) model defined kindness of an action based on comparisons of outcomes across players, the Dufwenberg and Kirchsteiger (2004) (DK) model defines it based on comparisons between potential outcomes for the second-mover. Here we provide an intuitive discussion of how our results relate to these models, and refer the technicalities to the Appendix.

In the context of our experiments, the DK model defines the relative kindness of (H) from the perspective of player B based on the difference between player B's expected payoffs in the sub-game reached after (H) to the average of her expected payoffs across both sub-games. Since Experiment 1 keeps all of player Bs payoffs constant across the two treatments, the difference in the perceived kindness of (H) can only arise from the differences in player B's second-order beliefs across the two treatments. Contrary to our findings, the DK model predicts player Bs to perceive the choice of (H) as kinder in treatment RP than in treatment RO. This prediction is driven by two factors. First, the payoffs that player Bs are expected to obtain if player A chooses (S) are lower in treatment RP, because both second-order expectations of player B and her behavior indicate that more player Bs sacrifice 50 cents to punish (S) in treatment RP than they reward (S) in treatment RO. As a result, the benchmark to which the expected outcome of (H) is compared to is lower in treatment RP. Second, the expected payoff of (H) is lower in treatment RO than in treatment RP, because both second-order expectations of player B and her behavior indicate more player Bs would sacrifice 50 cents to reward (H) in treatment RO than in treatment RP. Similarly, in Experiment 2, the DK model predicts a higher perceived kindness of (H) in treatment NP than in treatment N, because

[15] The second-mover's beliefs regarding what the first-mover intended for the second-mover are central to this definition of kindness. Belief-driven intention-based models (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006; Sebald, 2010) therefore rely on psychological game theory (see Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009 for an overview). Cox et al. (2007) capture similar drivers of intention-based reciprocity without relying on beliefs. They model reciprocal preferences as a function of gratitude and resentment emotions that are driven by the alternative action space of the first-mover and the maximum payoff that the second-mover can guarantee for herself.

the reference point of payoffs associated with choosing (S) is lower in treatment NP, since player As believes that a positive proportion of player Bs would pay to punish (S). Therefore, the DK model predicts the opposite of our findings: a higher degree of positive reciprocity to (H) when the punishment option is expected to be available to player B.

In the context of our experiments, the FF model defines the relative kindness of (H) from the perspective of player B based on the difference between player B's beliefs about player A's intended outcome for player B versus player A's intended outcome for himself as a result of choosing (H). A strict interpretation of the FF model would not predict any positive reciprocity to (H) in either of the experiments because player B's payoffs are always lower than those of player A. Since Falk and Fischbacher (2006) note that the central element of the FF model is the kindness term, we focus on discussing under what conditions the FF model would predict (H) to look less unkind in treatment RO than in treatment RP. Experiment 1 restricts payoffs in the final nodes of Game Γ_1 such that rewarding A increases the inequality of material payoffs between player A and player B. Therefore, the only way (H) can be perceived as less unkind in treatment RO than in treatment RP is if player As expected fewer player Bs to positively reciprocate in treatment RO, decreasing the expected level of inequality between player A and player B. Clearly, this condition produces a misalignment between equilibrium beliefs and behavior. Therefore, the kindness specification in the FF model is at odds with the results of Experiment 1. In Experiment 2, The FF model predicts equal perceived kindness of (H) in treatments N and NP, because player A expects player B not to have a choice if player A chooses H, and the resulting payoffs are the same in both treatments. In sum, the FF model is also not able to rationalize the data from either experiment.

A third class of models promote the idea that people's reciprocal behavior is driven by their inferences of kindness regarding the person taking a helpful or hurtful action (Cox et al., 2008a; Gül and Pesendorfer, forthcoming). Levine (1998) introduced the idea that a person's concern for another person's well-being increases in relation to how altruistic the other person is. Building on this idea, the Gül and Pesendorfer (forthcoming) model predicts higher rewards for the same helpful action when the person taking the action is perceived to have a higher degree of altruism. In a similar spirit, but without having to rely on beliefs, Cox et al. (2008a) model predicts higher rewards for the same helpful action if it helps the benefactor more than it helps the person who took the action.

Given the close relationship between perceptions of motives and perceptions of altruism of the person taking the helpful action, the results presented by Experiment 1 can potentially be explained by the GP model, presuming that the second-movers' altruism inferences regarding helpful first-movers are more positive in treatment RO then in treatment RP. This seems plausible, since second-movers expect more strategically motivated first-movers in treatment RP. Similarly, a random player A who chooses (H) in treatment N of Experiment 2 is on average a more altruistic person than a random player A who chooses (H) in treatment NP is. In fact, the average type choosing (H) across treatments in Experiment 2 can be ordered as being the kindest in treatment N, followed by in treatment NR and the least kind in treatment NP. This ordering corresponds to the ordering of the degree of positive reciprocity for the same helpful action, as the GP model predicts. In addition, the relationship between kindness inferences and reciprocal behavior across treatments provide further support for the mechanism proposed in the GP model.

The Cox et al. (2008) (CFS) model proposes that (H) would be perceived as a more generous action than (S) if (i) the maximum payoff player B can get if player A chooses (H) is greater than or equal to the maximum payoff player B can get if player A chooses (S), and (ii) the maximum payoff player A can get by choosing (H) minus what he can get by choosing (S) is (at least weakly) less than the maximum payoff player B can get if player A chooses (H) minus what she can get if player A chooses (S). In other words, (H) is more generous than (S) if it can help player B, and if it can help player B more than it can help player A[16]. This model is not immediately applicable to comparing the perceived generosity of (H) across the two treatments, because in each treatment, the payoff (at least weakly) satisfy both (i) and (ii) and thus (H) is considered to be more generous than (S). Therefore, a strict interpretation of this model would not produce any differences in the degree of positive reciprocity in either experiment. However, because this model considers what player A can obtain, it can be particularly suitable for thinking about the role of motives. Let us consider a simple extension that defines the generosity differential between (H) and (S) as the difference between how much choosing (H) over (S) helps player B minus how much it helps player A. Across the two treatments in Experiment 1, the payoffs of player B are fixed and the payoffs of player A are the same if player A chooses (H). In treatment RO, the maximum payoff of player A can get if he chooses (S) is $6.50 and in treatment RP, it is only $4. Therefore, the extended Cox et al. (2008) model would predict that choosing (H) rather than (S) in treatment RO looks more generous than choosing (H) rather than (S) in treatment RP, thus capturing the differences in positive reciprocity we document in Experiment 1. However, the extended CFS model cannot rationalize the results of Experiment 2. The maximum payoff player A can get if he chooses (H) in treatments N and NP are both equal to $4. Since choosing (H) over (S) also helps player B by the same amount in treatments N and NP, choosing (H) leads to the same generosity differential and thus reveal the same degree of generosity, and should be rewarded equally in treatments N and NP. This is not in line with the the prediction and findings of this paper.

3.2 Summary, Implications and Future Directions

Making a relational investment often brings benefits in the future, since additional incentives (rewards or punishment) are implicitly or explicitly inherent to many professional and personal reciprocal relationships. These incentives are shown to motivate socially desirable, helpful actions (Andreoni et al., 2003) and can result in large efficiency gains by enforcing these actions (Fehr et al. 1997). By virtue of being successful, however, the existence of such incentives obscures the motives of people who act generously in these interactions.

[16] We hold the default option across all treatments the same for player A, therefore the treatments do not present any differences in whether choosing (H) over (S) can be considered an omission or a commission. Thus, any reciprocity differences across treatments in these experiments can only be driven by the perceived generosity of choosing (H), as presented in Axiom R of Cox, Friedman, Sadiraj (2008).

This paper presents data from two experiments designed to isolate the role of perceived motives on reciprocal behavior. Evidence from both a between-subjects and a within-subject designs show that positive reciprocity declines in the perceptions regarding the degree to which a helpful action is strategically motivated. These results suggest that people are quite sophisticated about others' mental models and contemplate their motives when deciding on the appropriate reciprocal response.

The finding that perceived motives play an important role in shaping reciprocal decisions paves the way for several future research directions. Our findings shed some light on the type of reciprocity models that can incorporate the role of perceived motives. The results suggest at least two directions. One possibility is to allow perceptions of what the first-mover expected to gain or lose as a result of his action to influence the perceived kindness of an action. Another possibility is to model reciprocity as a response to the revelead altruism of the first-mover, taking care to specify the equilibrium properties under which the first-mover's actions are informative regarding his altruism (Gül and Pesendorfer, forthcoming). Recent work has proposed models exploring reciprocal behavior in gift-exchange games where i) individuals are care about others to the extent that others are altruistic, and ii) altruism is private information (Arbak and Kranich, 2005; Dur, 2009; Non, 2012). We hope that the experimental design and data presented in this paper are useful for spurring an interest in future work in this area that considers sensitivity to outcomes, intentions and motives in explaining reciprocal decision-making.

Our findings also suggest several avenues for future empirical investigation. For example, the current research does not compare the importance of perceived motives in relation to distributional preferences or perceived intentions. Assessing the relative importance of each of these factors is important, especially because such research may also help clarify seemingly contradictory results in the literature. For example, Bolton and Ockenfels (1998) showed that in the context of the Güth van Damme's three person bargaining game, people positively reciprocate to the slice of the pie given to themselves, but don't care about the slice of the pie given to a 3rd party who cannot respond. This evidence may seem contradictory to the idea that people are kinder to those who are genuinely kind, since the slice of the pie given to the third person can be a signal of kindness. However, given that the Güth van Damme game is a zero-sum game, signals of kindness come at the cost of how big a slice can be offered to the responding party. Therefore, in order to test whether people are kinder in response to non-strategic kindness, future research needs to compare the relative importance of outcomes and perceived motives in a design that varies these factors independently.

The central hypothesis tested in this paper is related to a broader question that has been pivotal in the recent research on reciprocity: How to evaluate kindness. It is our hope that the experiments and results presented in this article add to this discourse. This question is important to answer across many domains that involve reciprocal considerations. In recent work, Celen et al. (2014) offer a definition of kindness based on a notion of blame, similar to the notion of relative kindness of players in the GP model. Future research can further this inquiry by testing different notions of kindness in the laboratory.

The results presented in this paper may also have implications for the so-called positive reciprocity puzzle. Previous research noted an emerging consensus that the propensity to punish harmful behavior is stronger than the propensity to reward friendly behavior (for example, Fehr and Gächter, 2000; Cox and Deck, 2005; Charness and Rabin, 2002, 2005; Offerman, 2002). Offerman (2002) showed that negative intentionality is more likely to induce payoff decreases than positive intentionality induces payoff increases. They found that subjects are 67% more likely to reciprocate to an intentional hurtful choice over an unintentional hurtful choice. However, they are only 25% more likely to reciprocate to an intentional helpful choice over an unintentional helpful choice. Al-Ubaydli and Lee (2009) elicited second-order expectations and incorporated them in the Falk and Fischbacher (2006) model in order to tease out whether this asymmetry is a result of asymmetric intrinsic tendencies to reward or punish, or asymmetries in the extent to which rewards and punishments are objectively merited due to the differences in the perceived kindness of the first-mover given the game form that Offerman (2002) used. In light of the evidence presented in this paper, the reader may wonder whether the role of motive-attribution can also contribute to this asymmetry. In the case of intentional hurtful actions in a reciprocal context, the motives of the first-mover are unambiguously unkind and therefore deserve retribution. However, intentional helpful actions in a reciprocal context can be driven by kindness as well as self-interest. Since there is room for ambiguity in the casual attribution for these actions, the reciprocal response may not be as strong as it would have been if the helpful action were unambiguously driven by kindness. A closer look at the asymmetry between positive and negative reciprocity that disentangles these possible explanations would be worthwhile pursuing.

Finally, the results urge us to deliberate on seemingly contradictory predictions stemming from the literature on guilt aversion (see Dufwenberg and Gneezy, 2000; Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007 for an overview of guilt aversion theory, and Al-Ubaydli and Lee, 2009 for a more specific discussion regarding this contradiction.). Consider the investment-game where the first-mover makes a risky investment by trusting the second-mover to reciprocate. The guilt aversion literature would predict that the higher the second-order expectations are of the second-mover regarding what the first-mover expected of her, the more likely she is to reciprocate. If we think that the likelihood of the first-mover's being motivated by altruism is lower if his expectations of the second-mover are higher, we may conclude that the guilt aversion literature predicts the second-mover to reciprocate more positively towards the first-movers who are more strategically motivated. However, altruistic first-movers do not necessarily have lower expectations of the second-movers. For example, in Experiment 1, expectations of reward are 40% on average among the group of first-movers who would have been helpful even in the absence of strategic incentives, and 39% on average among those who are strategically motivated. Therefore, future research needs to isolate the second-mover's perceptions about the motives of the first-mover from her perceptions regarding his expectations from her.

REFERENCES

Abbink, Klaus, Bernd Irlenbusch, and Elke Renner. 2000. “The moonlighting game. An experimental study on reciprocity and retribution.” Journal of Economic Behavior & Organization, 42, 265-277.

Al-Ubaydli, Omar, and Min Sok Lee. 2012. "Do you reward and punish in the way you think others expect you to?" The Journal of Socio-Economics 41.3: 336-343.

Al-Ubaydli, Omar and Min Sok Lee. 2009. “An experimental study of asymmetric reciprocity.” Journal of Economic Behavior & Organization, 72, 738-749.

Andreoni, James and John Miller. 2002. "Giving according to GARP: An experimental test of the consistency of preferences for altruism." Econometrica, 70, 737-753.

Andreoni, James, William Harbaugh, and Lise Vesterlund. 2003. "The carrot or the stick: Rewards, punishments, and cooperation." American Economic Review, 893-902.

Arbak, Emrah and Laurance Kranich. “Can Wages Signal Kindness?” Working paper, Groupe d’Analyse et de Theorie Economique, University of Lyon.

Battigalli, Pierpaolo, and Martin Dufwenberg. 2007. "Guilt in games." The American Economic Review, 170-176.

Battigalli, Pierpaolo, and Martin Dufwenberg. 2009. "Dynamic psychological games." Journal of Economic Theory, 144.1, 1-35.

Berg, Joyce, John Dickhaut, and Kevin McCabe. 1995. "Trust, reciprocity, and social history." Games and Economic Behavior, 10, 122-142.

Blount, Sally. 1995. "When social outcomes aren´t fair: The effect of causal attributions on preferences." Organizational Behavior and Human Decision Processes, 63, 131-144.

Bolton, Gary E. and Axel Ockenfels. 1998. “Strategy and equity: An ERC-Analysis of the Güth-van Damme Game.” Journal of Mathematical Psychology, 42, 215-226.

Bolton, Gary E. and Axel Ockenfels. 2000. "ERC: A theory of equity, reciprocity, and competition." American Economic Review, 166-193.

Bolton, Gary E., Jordi Brandts, and Axel Ockenfels. 1998. “Measuring motivations for the reciprocal responses observed in a simple dilemma game.” Experimental Economics, 1, 207-219.

Brandts, Jordi and Carles Solà. 2001. “Reference points and negative reciprocity in simple sequential games.” Games and Economic Behavior, 36, 138-157.

Brandts, Jordi, and Gary Charness. 2011. "The strategy versus the direct-response method: a first survey of experimental comparisons." Experimental Economics 14.3: 375-398.

Cabral, L., Ozbay, E. Y., & Schotter, A. 2014. “Intrinsic and instrumental reciprocity: An experimental study.” Games and Economic Behavior, 87, 100-121.

Celen, Bogachan, Mariana Blanco and Andrew Schotter. 2014. “On blame and reciprocity: An experimental study.” Working paper.

Charness, Gary. 2004. "Attribution and reciprocity in an experimental labor market." Journal of Labor Economics, 22, 665-688.

Charness and Dufwenberg 2006. "Promises and partnership." Econometrica 74.6: 1579-1601.

Charness, Gary and David I. Levine. 2007. “Intention and stochastic outcomes: An experimental study.” The Economic Journal, 117, 1051-1072.

Charness, Gary and Ernan Haruvy. 2002. “Altruism, equity, and reciprocity in a gift-exchange experiment: an encompassing approach.” Games and Economic Behavior, 40, 203–231.

Charness, G., Gneezy, U., & Kuhn, M. A. 2012. “Experimental methods: Between-subject and within-subject design.” Journal of Economic Behavior & Organization, 81(1), 1-8.

Charness, Gary and Matthew Rabin. 2002. "Understanding social preferences with simple tests." Quarterly Journal of Economics, 817-869.

Charness, G., & Rabin, M. 2005. Expressed preferences and behavior in experimental games. Games and Economic Behavior, 53(2), 151-169.

Cox, James C. and Cary Deck. 2005. “On the Nature of Reciprocal Motives.” Economic Inquiry, Volume 43, Issue 3, pages 623–635, July 2005

Cox, James C., Daniel Friedman, and Steven Gjerstad. 2007 "A tractable model of reciprocity and fairness." Games and Economic Behavior 59.1: 17-45.

Cox, James C., Daniel Friedman, and Vjollca Sadiraj. 2008a. "Revealed Altruism." Econometrica, 76, 31-69.

Cox, James C., Klarita Sadiraj, and Vjollca Sadiraj. 2008b. "Implications of trust, fear, and reciprocity for modeling economic behavior." Experimental Economics, 11, 1-24.

Dufwenberg, Martin, and Uri Gneezy. 2000. "Measuring beliefs in an experimental lost wallet game." Games and Economic Behavior 30.2: 163-182.

Dufwenberg, Martin and Georg Kirchsteiger. 2004. "A theory of sequential reciprocity." Games and Economic Behavior, 47, 268-298.

Dur, Robert. 2009. “Gift Exchange in the Workplace: Money or Attention?” Journal of the European Economic Association. 7 (2-3): 550-560.

Dohmen, Thomas, Armin Falk, David Huffman, and Uwe Sunde. 2009. “Homo Reciprocans: Survey Evidence on Behavioral Outcomes.” Economic Journal, 119, 592–612.

Falk, Armin and Urs Fischbacher. 2006. "A theory of reciprocity." Games and Economic Behavior, 54, 293-315.

Falk, Armin, Ernst Fehr, and Urs Fischbacher. 2008. “Testing theories of fairness—Intentions matter.” Games and Economic Behavior, 62, 287-303.

Fehr, Ernst, Simon Gächter, and Georg Kirchsteiger. 1997. "Reciprocity as a contract enforcement device: Experimental evidence." Econometrica, Vol. 65, No. 4. p. 833-860.

Fehr, Ernst and Klaus M. Schmidt. 1998. "A theory of fairness, competition, and cooperation." Quarterly Journal of Economics, 817-868.

Fehr, Ernst and Simon Gächter. 2000. “Fairness and retaliation: The economics of reciprocity.” The Journal of Economic Perspectives, 159-181.

Fischbacher, Urs. 2007. “z-Tree: Zurich toolbox for ready-made economic experiments.” Experimental Economics, 10, 171-178.

Geanakoplos, John, David Pearce, and Ennio Stacchetti. 1989. "Psychological games and sequential rationality." Games and Economic Behavior 1.1: 60-79.

Gneezy, U., Güth, W., & Verboven, F. 2000. "Presents or investments? An experimental analysis." Journal of Economic Psychology, 21(5), 481-493.

Gül, Faruk, and Wolfgang Pesendorfer. "Interdependent preference models as a theory of intentions." Conditionally accepted by: Journal of Economic Theory (2010).

Güth, Werner and Eric Van Damme. 1998. "Information, strategic behavior, and fairness in ultimatum bargaining: An experimental study." Journal of Mathematical Psychology, 42, 227-247.

Heider, F. 1958. "The psychology of interpersonal relations." Wiley, New York. Kelley, Harold H. 1967. "Attribution theory in social psychology." Nebraska symposium on motivation. University of Nebraska Press.

Kelley, Harold H. 1973. "The processes of causal attribution." American psychologist 28.2.

Klempt, Charlotte. 2012 "Fairness, spite, and intentions: Testing different motives behind punishment in a prisoners' dilemma game." Economics Letters, 116/3: 429-431.

Levine, David K. 1998. "Modeling altruism and spitefulness in experiments." Review of Economic Dynamics, 1, 593-622.

McCabe, Kevin. A., Mary L. Rigdon, and Vernon L. Smith. 2003. "Positive reciprocity and intentions in trust games." Journal of Economic Behavior & Organization, 52, 267-275.

Nelson Jr, William Robert. 2002. "Equity or intention: it is the thought that counts." Journal of Economic Behavior & Organization, 48, 423-430.

Netzer, Nick, and Armin Schmutzler. 2014. "Explaining gift-exchange–the limits of good intentions." Journal of the European Economic Association, 12(6), 1586-1616.

Non, Arjan. 2012. "Gift-exchange, incentives, and heterogeneous workers." Games and Economic Behavior, 75, 319-336.

Offerman, Theo. 2002. "Hurting hurts more than helping helps." European Economic Review, 46, 1423-1437.

Rabin, Matthew. 1993. "Incorporating fairness into game theory and economics." The American Economic Review, 1281-1302.

Rabin, Matthew. 1998. "Psychology and economics." Journal of economic literature, Vol XXXVI, March, 11-46.

Rand, David G., Drew Fudenberg, and Anna Dreber. 2013. "It's the thought that counts: The role of intentions in reciprocal altruism." Working paper.

Ross, Michael, and Garth JO Fletcher. 1985. "Attribution and social perception." The handbook of social psychology 2: 73-114.

Sebald, Alexander. 2010. "Attribution and reciprocity." Games and Economic Behavior 68.1: 339-352.

Segal, U., & Sobel, J. 2007. "Tit for tat: Foundations of preferences for reciprocity in strategic settings." Journal of Economic Theory, 136(1), 197-216.

Segal, U., & Sobel, J. 2008. "A characterization of intrinsic reciprocity." International Journal of Game Theory, 36(3-4), 571-585.

Sobel, J. 2005. “Interdependent preferences and reciprocity.” Journal of Economic Literature, 392-43

Stanca, Luca. 2010. “How to be kind? Outcomes versus intentions as determinants of fairness.” Economic Letters, 106, 19-21.

Strassmair, Christina. 2009. “Can intentions spoil the kindness of a gift? An experimental sudy.” Working paper, University of Munich, Munich Discussion Paper No. 2009-4.

Toussaert, Séverine. 2014. “Intention-Based Reciprocity and Signalling of Intentions” Working paper, NYU.

Appendix

Predictions of Different Reciprocity Theories

In this Appendix, we discuss the predictions of the intention-based reciprocity models proposed by Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006) regarding Experiment 1 and Experiment 2. We simplify the discussion by considering slightly modified versions of the treatments in Experiment 2 by assuming that $p = q = 0$ in treatment N, $p = 1$ in treatment NP and $q = 1$ in treatment NR . This simplification greatly aids discussion without impacting the differences in the predictions of different theories.

For generality, we parameterize the payoffs in Game Γ_1 and Game Γ_2 to highlight the general features that allow us isolate the role of motives. Figure 3 below refers to the generalized version of Game Γ_1. The variable m is varied across treatments. In treatment RP, $m = x - 5k$, and in treatment RO $m = x + 5k$. Note that all of the payoffs of player A are (at least weakly) larger than the payoffs of player B ($x > y + 2t$, $x > y + 4k$), and we further restrict t and k such that all payoffs are positive ($y > k$, $x > 5k$, $5k > t$).

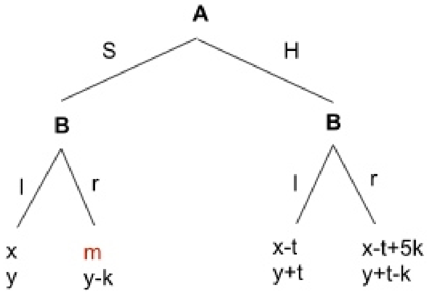

Figure 3: Generalized Game Γ_1

Figure 4 below refers to the generalized version of Game Γ_2. Note that all of the payoffs of player A are (at least weakly) larger than the payoffs of player B ($x > y$ and $x - y \geq 2t$), and we further restrict t and k such that all payoffs are positive. For simplicity of discussion, we also set $t = k$. Remember that Experiment 2 featured three within-person treatments: N ($q = p = 0.01$), NP ($q = 0.01$ and $p = 0.98$), and NR ($q = 0.98$ and $p = 0.01$). For the discussion of the predictions of different theories, we simplify the discussion by considering slightly modified versions of treatment N ($p = q = 0$), NP ($p = 1$) and NR ($q = 1$). This simplification greatly aids discussion without impacting the predictions of different theories.

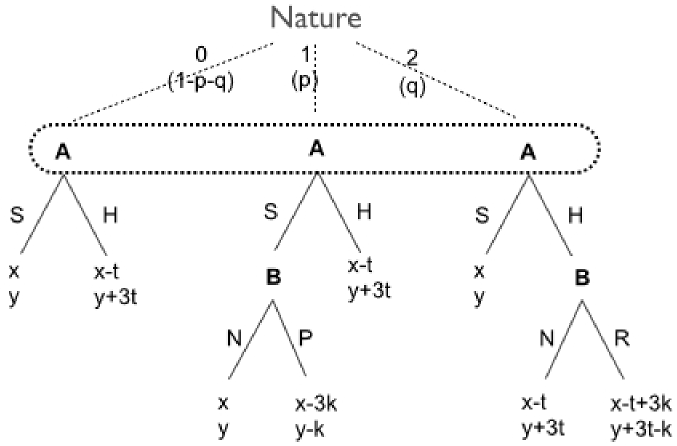

Figure 4: Generalized Game Γ_2

Dufwenberg and Kirchsteiger (2004)

The model respectively defines the perceived kindness of (H) and (S) from the perspective of B as $\kappa_B(H) = E_{BA}[\pi_B|H] - \frac{1}{2}\{E_{BA}[\pi_B|H] + E_{BA}[\pi_B|S]\}$ and $\kappa_B(S) = E_{BA}[\pi_B|S] - \frac{1}{2}\{E_{BA}[\pi_B|H] + E_{BA}[\pi_B|S]\}$, where $E_{BA}[\pi_B|S]$ denotes player B's beliefs regarding player A's expectations of player B's payoffs (π_B) if player A chooses (S).[17] The model posits that the degree of positive reciprocation to H increases in $\kappa_B(H)$ in the region where $\kappa_B(H) \geq 0$ and the degree of negative reciprocation to S increases in $|\kappa_B(S)|$ in the region where $\kappa_B(S) < 0$. The hypotheses in this paper are centered around the perceived kindness of (H), which depends on player B's second-order beliefs: what player B believes about what player A thinks player B will choose if player A chooses (H).

Experiment 1. Denote player B's second-order beliefs regarding the prevalence of (r) given action (H) as $b''_{RO}(r|H)$ and $b''_{RP}(r|H)$ in the two conditions. Similarly, denote player B's second-order beliefs regarding the prevalence of (r) given action (S) as $b''_{RO}(r|S)$ and $b''_{RP}(r|S)$ in the two conditions.

Then, in treatment RO, $E^{RP}_{BA}[\pi_B|H] = b''_{RO}(r|H)(y + t - k) + (1 - b''_{RO}(r|H))(y + t)$ and in treatment RP, $E^{RP}_{BA}[\pi_B|H] = b''_{RP}(r|H)(y + t - k) + (1 - b''_{RP}(r|H))(y + t)$. And the perceived kindness of (H) across two treatments are $\kappa^{RO}_B(H) = \frac{1}{2}\{b''_{RO}(r|H)(y + t - k) + (1 - b''_{RO}(r|H))(y + t)\} - \{b''_{RO}(r|S)(y - k) + (1 - b''_{RO}(r|S))(y)\}$ and $\kappa^{RP}_B(H) = \frac{1}{2}\{b''_{RP}(r|H)(y + t - k) + (1 - b''_{RP}(r|H))(y + t)\} - \{b''_{RP}(r|S)(y - k) + (1 - b''_{RP}(r|S))(y)\}$. Note that since the payoffs of player B are exactly the same across the two treatments, any differences in perceived kindness of (H) will stem from differences in second-order expectations. In order for the Dufwenberg and Kirchsteiger (2004) model to predict a higher degree of positive reciprocity to (H) in condition RO, we either need to maintain $b''_{RO}(r|H) < b''_{RP}(r|H)$, which contradicts the exact prediction we are trying to capture, or assume

[17] Both (H) and (S) are in the efficient set of actions for player A and are the only actions player A can take.

$b''_{RO}(r|S) > b''_{RP}(r|S)$ which contradicts the behavior and expectations in the data. Therefore, the Dufwenberg and Kirchsteiger (2004) model cannot explain the data from Experiment 1.

Experiment 2. In treatment N, player B gets y if player A chooses (S) and she gets $y + 3t$ if he chooses (H). The perceived kindness of choosing (H) in treatment N can be calculated by comparing $E_{BA}[\pi_B|S] = y + 3t$ to the midpoint of possible outcomes, which is $\frac{1}{2}\{E_{BA}[\pi_B|H] + E_{BA}[\pi_B|S]\} = y + 1.5t$. Therefore, the preceived kindness of (H) in treatment N are $\kappa_B^N(H) = 1.5t$.

In treatment NR, player B's beliefs about player A's expectations regarding player B's payoffs if player A chooses H are given by $E_{BA}[\pi_B|H] = b''(R|H)(y + 3t - k) + (1 - b''(R|H))(y + 3t)$ and player B's beliefs about player A's expectations regarding player B's payoffs if player A chooses S ($E_{BA}[\pi_B|S]$) are simply y. If $b''(R|H) = 0$, then the preceived kindness of (H) is the same in treatment N and NR. If, $b''(R|H) > 0$, then the preceived kindness of (H) is strictly lower in treatment NR than in treatment N, since $\kappa_B^{NR}(H) = 1.5t - 0.5b''(R|H)k$. Therefore the Dufwenberg and Kirchsteiger (2004) model would predict (at least weakly) higher level of positive reciprocity in treatment N than in treatment NR. This prediction is in line with the prediction in this paper and the results.

However, the model would produces a contradictory prediction of this paper in comparing the degree of positive reciprocity in treatment N compared to treatment NP. In treatment NP, player B's beliefs about player A's expectations regarding player B's payoffs if player A chooses S are given by $E_{BA}[\pi_B|S] = b''(P|S)(y - k) + (1 - b''(P|S))(y)$ and player B's beliefs about player A's expectations regarding player B's payoffs if player A chooses H are simply $y + 3t$. Therefore, perceived kindness of (H) in this treatment is $\kappa_B^{NP}(H) = 1.5t + 0.5b''(P|S)k$. If $b''(P|S) = 0$, then the perceived kindness of (H) is the same in treatment N and NP. If, $b''(P|S) > 0$, then perceived kindenss of (H) is higher in treatment NP than in treatment P, since choosing (S) may lead to player B sacrificing an amount k to punish player A in treatment NP. Therefore the modified Dufwenberg and Kirchsteiger (2004) model would predict (at least weakly) lower level of positive reciprocity in treatment N than in treatment NP. We do not focus on negative reciprocity in this paper. For completeness, the Dufwenberg and Kirchsteiger (2004) model produce $\kappa_B^N(S) = -1.5t$, $\kappa_B^{NR}(S) = -1.5t + 0.5b''(R|H)k$ and $\kappa_B^{NP}(S) = -1.5t - 0.5b''(P|S)k$, predicting that negative reciprocity in response to (S) should be highest in treatment NP, followed by in treatment N and the lowest in treatment NR.

Falk and Fishbacher (2006)

In Game Γ_1 and Game Γ_2, we keep most of the features that would impact the degree of intentionality in the Falk and Fishbacher (2006) model constant: Player A has the same choice set (S, H) and full control over his actions across all treatments. Having fixed these dimensions, we can investigate how perceived kindness of player A's actions differ across treatments. The model respectively defines the perceived kindness of (H) and (S) from the perspective of B as $\kappa_B(H) = E_{BA}[\pi_B|H] - E_{BA}[\pi_A|H]$ and $\kappa_B(S) = E_{BA}[\pi_B|S] - E_{BA}[\pi_A|S]$, where $E_{BA}[\pi_B|S]$ denotes player B's beliefs regarding player A's expectations of player B's payoffs (π_B) if player A chooses (S) and $E_{BA}[\pi_B|S]$ denotes player B's beliefs regarding player A's expectations of player A's payoffs (π_A) if player A chooses (S). Therefore the Falk and Fischbacher (2006) model determines the perceived kindness of an action based the difference between player B's beliefs about player A's intended outcome for player B versus player A's intended outcome for himself.

Experiment 1. In treatment RO, where $m = x + 5k$, the relative outcome kindness of (S) is $\kappa_B^{RO}(S) = b^{''}_{RO}(r|S)[(y-k)-(x+5k)] + (1-b^{''}_{RO}(r|S))[y-x]$, and the relative outcome kindness of (H) is $\kappa_B^{RO}(H) = b^{''}_{RO}(r|H)[(y+t-k)-(x-t+5k)] + (1-b^{''}_{RO}(r|H))[(y+t)-(x-t)]$. In treatment RP where $m = x - 5k$, the relative outcome kindness of (S) is $\kappa_B^{RP}(S) = b^{''}_{RP}(r|S)[(y-k)-(x-5k)] + (1-b^{''}_{RP}(r|S))[y-x]$, and the relative outcome kindness of (H) is $\kappa_B^{RP}(H) = b^{''}_{RP}(r|H)[(y+t-k)-(x-t+5k)] + (1-b^{''}_{RP}(r|H))[(y+t)-(x-t)]$.

If $b^{''}_{RO}(r|H) = b^{''}_{RP}(r|H)$, action (H) looks equally unkind in both treatments, as the payoffs are the same for this sub-game across the treatments and player B earns less than player A. If second order expectations are in line with predicted equlibrium play (and the SOE we elicit in the data), then $b^{''}_{RO}(r|H) > b^{''}_{RP}(r|H)$. Interestingly, according to the Falk and Fischbacher (2006) model this would imply that choosing (H) in RP is less unkind ($\kappa_B^{RP}(H) > \kappa_B^{RO}(H)$). This prediction would not be able to rationalize a higher degree of positive reciprocity in response to (H) in treatment RO.

Experiment 2. In treatment N, since player B has no choice, this considerations is reduced to calculating the difference between player B's payoffs and player A's payoffs as a result of player A's choices. If player A chooses H, the distance between player B's payoff from player A's payoff is $\kappa_B^{N}(H) = y - x + 4t$.

In the other two treatments, the second order beliefs of player B matter. Let's denote player B's beliefs regarding player A's expectations of player B choosing R in response to H in treatment NR as $b^{''}(R|H)$. Similarly, let's denote player B's beliefs regarding player A's expectations of player B choosing P in response to S in treatment NP as $b^{''}(P|S)$. Then, in treatment NR, the relative outcome kindness of (H) is $\kappa_B^{NR}(H) = b^{''}(R|H)[(y+3t-k)-(x-t+3k)] + (1-b^{''}(R|H))[(y+3t)-(x-t)]$. If $b^{''}(R|H) = 0$, action (H) looks equally unkind in treatment NR as it does in treatment N. However, if $b^{''}(R|H) > 0$, then action (H) looks more unkind in treatment NR since it leads to a larger disadvantaged payoff for player B compared to the action (H) in treatment N (by an amount of $4k \cdot b^{''}(R|H)$). However, action (H) looks equally unkind in treatment NP compared to the action (H) in treatment N, $\kappa_B^{NP}(H) = y - x + 4t$. For completeness, this model would produce the following perceived kindness of (S) across treatments: $\kappa_B^{N}(S) = (y-x)$, $\kappa_B^{NR}(S) = (y-x)$, and $\kappa_B^{NP}(S) = b^{''}(P|S)(y-x+2k) + (1-b^{''}(P|S))(y-x)$, leading to the prediction that punishment should be higher (and equal to each other) in treatments N and NR and lowest in treatment in NP. This prediction is also contradicted by the Experiment 2 data.