Reciprocating to Strategic Kindness[*]

A. Yeşim Orhun

Ross School of Business, University of Michigan

September 16, 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 degree of strategic incentives the first-mover has for taking that action are stronger. 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. These results imply that perceived motives, in addition to outcomes and perceived intentions, play an important role in shaping reciprocal responses.

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

[*] I thank Gary Bolton, Yan Chen, Emel Filiz Ozbay, Aradhna Krishna, Erin Krupka, Steve Leider, Yusufcan Masatlioglu, Neslihan Uler, Axel Ockenfels, Erkut Ozbay and Tanya Rosenblat 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

When a man aims the gun and shoots an animal in the head, his intention is to kill. His motive could be to save it from the pain of a terminal injury, or to boast to his friends about his hunting skills. Both motives and intentions are important, yet distinct elements of decision-making. Intentions refer to the outcome the person meant to bring about with an action, whereas motives refer to the reasons why he meant to bring about that outcome. Attribution theory (Heider 1958, Kelley 1967; 1973, Ross and Fletcher 1985) posits that causal inference, the process of taking an actor's motives, constraints, and intentions into account, is essential for determining the appropriate response to an action.

The very nature of a reciprocal interaction features potential rewards for helpful behavior and/or potential punishment for selfish or hurtful behavior. Therefore, 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 (Gneezy et al., 2000; Sobel, 2005; Segal and Sobel, 2007; 2008; Cabral et al. 2014). Outcome-based models of reciprocity (Fehr and Schmidt, 1999; Bolton and Ockenfels, 1998, 2000; and the basic model in Charness and Rabin, 2002) focus only on how people reciprocate to the consequences of an action, and do not incorporate any role of perceptions regarding the first-mover's choice process in bringing about that consequence. Attribution theory, however, would suggest that the beneficiary may ponder what the benefactor intended to for her to obtain as a result of his choice (his intent), and the reason why the benefactor preferred this outcome (his motive) in order to assess the kindness of his gesture and respond appropriately. Intention-based reciprocity theories (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006) have highlighted the role of perceived intent in reciprocity. This paper builds on this literature by considering the additional role of perceived motives behind the intended outcome, when the perceived intentions and motives may lead to different judgments of kindness.

We present evidence from two experiments designed to isolate the role of perceived motives on reciprocal behavior. The experiments manipulate perceived motives of the person taking a helpful action in the first-stage by varying the access of the second-mover to a costly punishment option conditional on the first-mover being selfish. Importantly, the designs allow for a direct comparison of positive reciprocity across treatments by keeping constant the response options of the second-mover conditional on the first-mover being helpful. In order to isolate the role of perceived motives from the role of perceived intentions, the experiments place certain restrictions on the payoffs such that the two accounts would make different qualitative predictions based on second-order expectations. Finally, both experiments elicit first- and second-order beliefs in order to expore the mental models of players and to provide a direct test of the predictions of the motive-based account of reciprocity.

The first experiment features a modified binary trust game where the first-mover decides between an unequal distribution of payoffs that gives himself three dollars more than it gives the second-mover, and a more equitable option that transfers a dollar from his pot to that of the second-mover. Conditional on the first-mover giving a dollar to the second-mover, the second-mover decides whether to send fifty cents to the first-mover, which is tripled by the experimenter before being added to his pot. The experiment varies whether the second-mover has the opportunity to punish the first-mover for not transferring a dollar to her pot. In the treatment where the second-movers can punish, the first-movers suspect substantial punishment in the event of non-transfer, and second-movers expect the first-movers to suspect this punishment. As a result, when punishment potential is present, first-movers are much more likely to transfer money, and second-movers understand that a larger fraction of those who transferred money were motivated by strategic considerations, and a smaller fraction helped due to altruistic motives. If perceived motives matter, we would expect positive reciprocity to the same helpful action (the dollar transfer) to be weaker when second-movers have the option to punish.

Because this experiment manipulates perception of the first-mover's motives by changing the second-mover's strategy space outside of the sub-game of interest, any reciprocity demand difference across the two treatments cannot be explained by outcome-based theories of reciprocity (Fehr and Schmidt, 1999; Bolton and Ockenfels, 1998, 2000; and the basic model in Charness and Rabin, 2002). Since the design keeps the agency and the choice set of the first mover constant, intention-based theories (Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006) predict positive reciprocity to differ across the two treatments only to the degree that the second-movers differ in what they thought about how the first-mover expected them to respond (second-order expectations). However, given the equilibrium play and second-order expectations in the experiment, intention-based theories predict higher, rather than lower levels of positive reciprocity to the helpful action in the treatment where punishment option is available[1]. Therefore, the first experiment allows for testing the role of perceived motives above and beyond the role of distributional concerns or perceived intentions in reciprocal decision making.

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. Fear of punishment is usually a stronger strategic motivator than hope of rewards (Andreoni et al. 2003). Therefore, we expect that a suspicion of reward-seeking motives will lead to a lesser degree of reciprocation deterioration compared to the impact brought by a suspicion of punishment-avoidance motives. Similar to the first experiment, hypothesized differences in positive reciprocity to the same helpful action across treatments cannot be reconciled with outcome- or intention-based theories alone.

Second, this experiment 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. Presumably, if you think somebody has been helpful to you mainly to avoid punishment or to seek future rewards, then you do not perceive him to be as altruistic as if he had not expected any form of reciprocity from you. Therefore, the second experiment relies on a within-subjects design to test whether an increase in the strength of strategic incentives to be helpful leads to a decline in perceived altruism of a helpful second-mover, and whether these perceptions in turn shape reciprocal responses.

By providing evidence from both a between-subjects and a within-subject design, the two experiments show that perceptions of *why* a helpful action was taken is important for the reciprocal response the helpful action triggers: the stronger strategic incentives the first-mover has for choosing the helpful action, the lower the degree of positive reciprocity is to that action. In addition,

[1] We discuss the predictions of the models proposed by Dufwenberg and Kirchsteiger, 2004 and Falk and Fischbacher, 2006 after we present each of the experimental results. We also provide a detailed analysis in the Appendix.

the decline in the level of reciprocity is shown to be associated with the deterioration of altruism inferences regarding the person who took the helpful action. These results suggest that people are quite sophisticated about others' mental models and contemplate their motives when deciding on the appropriate reciprocal response.

Notwithstanding its central role in understanding the fundamentals of reciprocal behavior, the influence of perceived motives on reciprocal decision-making remains largely unexplored, plausibly due to the difficulties in isolating its impact. Because incentives are inherent to all reciprocal interactions, and motives behind helpful behavior are therefore ambiguous, the findings in this paper have broad implications across a variety of contexts where reciprocal preferences influence decision making, such as labor contracts and socially responsible businesses. Our hope is that the impact of perceived motives gets more attention in reciprocity research, and that the experimental design and results presented in this article are useful in this regard.

2 Related Literature

The literature has proposed three broad sources of reciprocal behavior. Outcome-based models of altruism and reciprocity (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000; Charness and Rabin, 2002; Andreoni and Miller, 2002) model the positive relationship between a helpful action and a reaction based on preferences over payoff allocations. These models are able to explain a large body of reciprocal behavior by focusing on reactions to consequences, but do not allow for any influence of perceptions regarding the process with which the outcome was realized.

A second class of models promote the idea that people's reciprocal behavior is influenced 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.

A third class of models have been proposed to 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[2] that the first-mover intended the second-stage player's material payoff to be larger than some benchmark[3] as a result of his action. In addition, Falk and Fischbacher (2006) propose the degree of agency (whether the first-stage player had full control over his action and had an alternative to choose differently) to amplify kindness perceptions of helpful outcomes.

A large body of experimental work investigates whether reciprocity is sensitive to the perceptions of intentionality as proposed by Dufwenberg and Kirchsteiger (2004) (DK) and Falk and Fischbacher (2006) (FF). A group of experiments highlight the significant role of agency by comparing reciprocal responses to a decision made by the first-mover to a 'decision' made by an external process such as a random draw from a distribution or by a third party (Blount, 1995; Offerman, 2002; Charness and Haruvy, 2002; Charness, 2004; Falk et al., 2008). For example, Charness (2004) finds positive reciprocity to be 35% higher towards a volitional choice than the same choice being made without the first-mover's volition.

Pursuing a similar question, Charness and Levine (2007) and Klempt (2012) allow luck to alter first-mover's choices and compare reciprocity towards outcomes the first-mover intended to generate versus outcomes that are actually realized. They show that reciprocity is higher towards intentionally beneficial actions. In these experiments, the alteration by luck is common information, and therefore what the first-mover intended to choose is clear. In contrast, Rand et al. (2013) and Toussaert (2014) vary the degree to which the first-mover's intended choices are observed

[2] 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) captured similar drivers of intention-based reciprocity without relying on beliefs. They modeled 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, given the first-mover's choice compared to some reference point.

[3] This benchmark in the Falk and Fischbacher (2006) depends on the second-mover's second-order expectations regarding the payoffs of the first-mover model. The second-mover thinks that the first-mover has acted kindly if she thinks that he meant the second-mover to obtain a higher payoff than himself. The benchmark in the Dufwenberg and Kirchsteiger (2004) model depends on second-mover's second-order expectations regarding what she could have obtained had the first-mover behaved differently. Therefore, while the Falk and Fischbacher (2006) model defined kindness of an action based on comparisons of outcomes across players, the Dufwenberg and Kirchsteiger (2004) model defines it based on comparisons between potential outcomes for the second-mover.

in situations where first-movers' choices are implemented with noise. These articles show that the second-mover is sensitive to the degree of ambiguity regarding the agency behind a realized outcome. Toussaert (2014) also shows that the first-movers are sophisticated about this sensitivity and prefer to send clearer signals of their choices if they choose to help.

Another set of experiments vary the first-stage player's choice alternatives off the actual path of play in order to investigate how reciprocity responds to perceptions of whether the first-mover could have chosen an action more beneficial for the second-mover (for example, Bolton et al., 1998; Brandts and Sola, 2001; Nelson, 2002; McCabe et al., 2003). With the exception of Bolton et al. (1998), these experiments show that positive reciprocity towards a helpful action is higher when the first-mover could not have chosen better, given the choice alternatives he had.

The body of experimental work designed to test for the role of perceived intentions does not provide insights about the role of perceived motives. Notably, Stanca et al. (2009) provide a manipulation that is aimed to vary the perception of motives behind a helpful action. 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. In this design, the first-movers can only be motivated by altruism in treatment 1, but potentially also by reward-seeking motives in treatment 2. Therefore, the account of perceived motives would predict more reciprocity to the same helpful action in treatment 1, because second-movers know that first-movers expected to earn more by making transfers in treatment 2 than they did in treatment 1.

The results show that the slope of the second-movers' rewards (increasing with the corresponding generosity of first-movers) is steeper when first-stage choices are made in absence of knowledge of the second stage. The authors present this evidence as supportive evidence for the role of perceived motives[4]. However, the experimental design confounds the role of intentions and motives, because

[4] The authors do not find significant differences in the average rewards across the two treatments, but only find

it studies a context where their predictions overlap. To see how the results can be expleind by the role of perceived intentions, note that any positive second-order expectation regarding positive reciprocity in treatment 2, the same helpful action leads to lower expected payoffs for the second-mover than it would in treatment 1, as rewards are costly and not giving any rewards is the de facto outcome in treatment 1. Therefore, the FF model would predict lower degree of reciprocity to transfers in treatment 2. Consequently, the authors rely on the FF intention-based reciprocity model to explain their results.

The current paper extends this work in several important ways. Both Experiment 1 and Experiment 2 are designed i) to provide evidence of differences in the average rewards in response to the same helpful action as a function of perceived motives, ii) to isolate the role of motives from the role of perceived intentions, and, iii) to do so without misleading subjects about the nature of the interaction.

3 Experimental Investigation of the Role of Perceived Motives

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, preferably without relying on any surprises about the true nature of the interaction. This article makes use of manipulations of motives that involve a costly punishment option conditional on the first-mover being selfish. It also restricts the payoffs such that the second-mover's payoffs are never higher than that of the first-mover. As a result, punishing the first-mover decreases the inequality of material payoffs. While this article hypothesizes that the rewards for a helpful action in such contexts are higher when it could not be motivated by a fear of punishment, intention-based models of reciprocity either predict the opposite result or do not predict any differences.

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, in addition to these design features regarding the sequential reciprocity games, the experiments in this article also elicit several belief measures (first- and second-order beliefs,

differences in the slope of the reaction. Therefore, a more nuanced account of motives is required to fully explain the results.

altruism inferences). These data are useful in establishing internal validity of the experiments, 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.

3.1 Experiment 1

Experiment 1 presented severl decision tasks across four parts. Part 1 elicited other-regarding preferences from first-movers, part 2 elicited expectations of other-regarding preferences among the first-movers in the given session from all the subjects, part 3 elicited first- and second-movers' choices in a sequential reciprocity game, and part 4 elicited first- and second- order expectations of these choices in the sequential reciprocity game. The reciprocity game is the focal investigation of Experiment 1. Therefore, we present this game before we detail the entire protocol and explain the role of the remaining parts.

The reciprocal interaction

Experiment 1 features 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 delivers a highly unequal payoff distribution where 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 the same amount, and delivers a more equitable distribution (\$3, \$2).

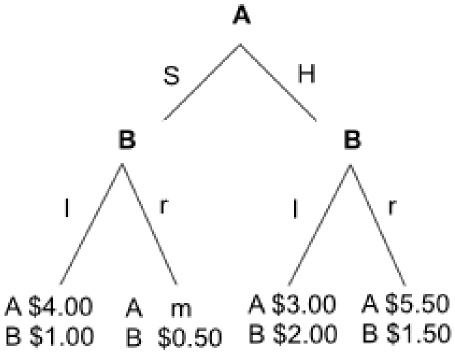

Figure 1: Game Γ_1

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), which leaves the distribution player A delivered unchanged, and a costly reward option (r), whereby player B 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), which leaves the distribution player A delivered unchanged, and a costly option whereby Player B 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.

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 altruistic and/or reward-seeking motives. In treatment RP, Game Γ_1 is a judgment game that also offers the potential of being punished for not choosing (H)[5]. Clearly, treatment RP gives stronger strategic incentives for the first-mover to choose (H): the first-movers who choose (H) could be motivated by altruistic, reward-seeking and/or punishment-avoidance motives. Given the stronger incentives to choose (H), the proportion of people motivated by altruistic motives among the helpful first-movers is expected to be lower.

There are several design features of Game Γ_1 that allow for the identification of the role of perceived motives. The game keeps the action space of the first-mover constant. The action space of the second-mover is also constant conditional on the first-mover choosing (H), but varies 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 and thus the preferences that govern their concern for others are kept constant, and 2) punishing player A always decreases the magnitude of the inequality of material payoffs between player A and player B, and rewarding him

[5] 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.

always increases it. In addition to these restrictions, the material payoffs of player B are the same across the two treatments.

The Protocol

A total of 258 participants (recruited through ORSEE) participated in eighteen 60-minute sessions at the University of Michigan's School of Information 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[6] (Charness and Rabin, 2002). 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

[6] 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.

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 provide support for internal validity, allos us to explore the mental models of players regarding the game and the other player, and help rule out alternative explanations.

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.*

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 the sequential reciprocity game 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[7]	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 subgame 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 incentives in the game, and each others' decision processes.

Discussion of Experiment 1 Results

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. Second-movers' expectations regarding the difference in the degree of helpfulness across the two treatments point to an understanding of the first-mover's motives.

Relation to intention-based reciprocity models Here we provide an inuitive discussion of how the results relate to the models of intention-based reciprocity proposed by Dufwenberg and Kirchsteiger (2004) and Falk and Fischbacher (2006), and refer the technicalities to the Appendix. 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 subgame reached after (H) to the average of her expected payoffs across both subgames. Given that all of player Bs payoffs are held constant across the two treatments, the difference in the perceived kindness of (H) arises from the differences in player B's second-order beliefs across the two treatments. Contrary to the account of motives, 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 more player Bs sacrifice 50 cents to punish (S) in treatment RP than they reward (S) in treatment RO (and believe that player As expect them to). 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. Therefore, the DK model predicts the opposite of the main result of Experiment 1: a higher degree of positive reciprocity to (H) in condition RP.

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). 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. The only way (H) can be perceived as a kinder action in treatment RO 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 that would produce a misalignment between equilibrium beliefs and the exact behavior the model is trying to explain. Indeed, the elicited second-order beliefs also fail to support this condition. Therefore, the FF model also fails to capture the main finding of Experiment 1.

Relation to reciprocity models based on revealed altruism A different class of reciprocity models promote the idea that people's reciprocal behavior is driven by their inferences regarding the altruism of the person. The model proposed by Gül and Pesendorfer (forthcoming) (GP), and the earlier model of Levine (1998) suggest reciprocation towards altruistic people to be higher. As we noted in the introduction, there is a close relationship between perceptions of motives and perceptions of altruism of the person taking the helpful action. Therefore, 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.

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[8]. 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

[8] 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).

player A can obtain, it can be particularly suitable for thinking about the role of motives. Let us consider a simple extention 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 this paper.

3.2 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. Informed by revealed-altruism based reciprocity theories, 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. Third, Experiment 2 elicits other-regarding preferences from both first- and second-movers in order to control for heterogeneity in altruism as an alternative explanation for differences in reciprocity (Cox, 2004).

This experiment also presented decision tasks across four parts. Part 1 elicited other-regarding preferences from all participants, part 2 elicited expectations of other-regarding preferences, part 3 elicited first- and second-movers' choices in a sequential reciprocity game, and part 4 elicited first-order expectations of behavior in the sequential reciprocity game and altruism inferences regarding player As who chose (H). Again, we present the reciprocal interaction before we detail the entire protocol.

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).

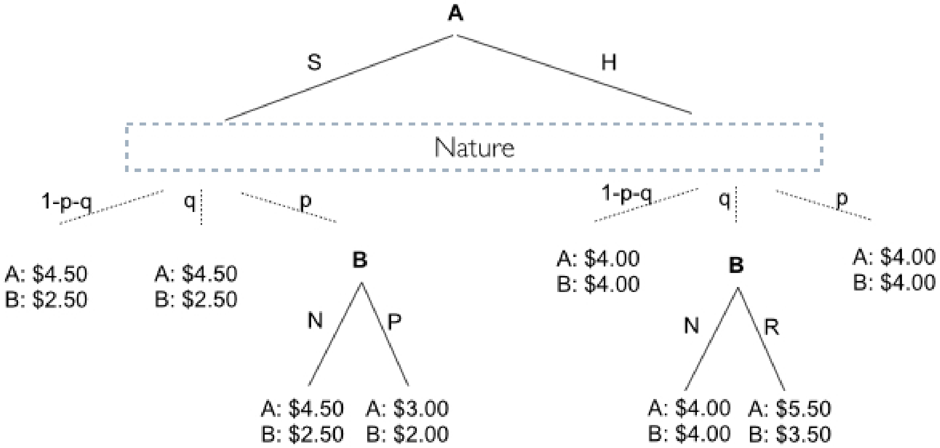

Figure 2: Game $\mathit{\Gamma}_2$

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) acros 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.[9] In particular, player Bs are asked to

[9] 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.

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 in this treatment, even if they are not likely to be able to exercise this choice.[10]

Player B's choice between (R) and (N) across the three treatments is the main comparison of interest. Is player B more likely to want to reward (H) when player A expected her not to be able to respond, when he expected her to be able to reward, or when he expected her to be able to punish? Note that unlike Experiment 1, Experiment 2 focuses on one strategic motivation in each treatment, minimizing any expectations of the alternative strategic motivation. As a result, it permits comparing the implications of a punishment-avoidance motive or a reward-seeking motive to the case where there are no strategic motivations.[11]

The Protocol

A total of 176 participants participated in eleven 45-minute sessions conducted at the University of Michigan's Ross School of Business Behavioral 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).

[10] A concern the reader may have with eliciting strategies with a low probability of implementation is that the participants may not be revealing their true preferences. 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. Moreover, note that Nature is equally unlikely to choose 1 in treatments N and NP. Therefore, any bias generated by the low probability of the event is common to both treatment N and NP, allowing us to interpret the difference between them causally.

[11] Note that the comparison of positive reciprocity across the two treatments in Stanca et al. (2009) and treatments N and NR in Experiment 2 of this article 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?" However, Experiment 2 does not rely on an element of surprise to manipulate the expectations and thus the motives of player A.

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 (Cox, 2004) and the role of strategic incentives in Game Γ_2. Among these dictator games, player As were presented 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. 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.[12] 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[13] 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

[12] 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.

[13] 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 subgame that was implemented with 98% chance.

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.[14] 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.

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).*

[14] 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.

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[15]. 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 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.

[15] 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.

<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. 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 difference in the altruism of helpful player As , compared to those who continue to reward action (H) (Wilcoxon rank-sum (Mann-Whitney) test: $z = -2.12$, $p = 0.034$).[16] 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).[17] In sum, the results show that a within-person increase

[16] 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).

[17] 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

(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 forhypothesis *H4*.

Discussion of Experiment 2 Results

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. They show that the 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. Also, 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.

Relation to intention-based reciprocity models While the intention-based reciprocity theories also predict the reciprocity difference between treatments N and NR, they differ in their predictions from the hypotheses in this article regarding the comparisons between treatment N and NP. Note that Experiment 2 holds fixed the intentionality of player A across treatments and restricts the payoffs such that punishing decreases the inequality of players' payoffs. 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 as a result of choosing (H) versus player A's intended outcome for himself. In treatments N and NP[18], player A expects player B not to have a choice if player A chooses H, therefore second-order expectations do not differ across

inference deterioration of these twenty-six player Bs is significantly lower (Wilcoxon rank-sum (Mann-Whitney) test: z=3.39, p=0.0001).

[18] We simplify the discussion by considering slightly modified versions of treatment N (p=q=0) and NP (p=1), both here and in the Appendix. This simplification greatly aids discussion without impacting the differences in the predictions of different theories.

these treatments. Moreover, the payoffs of both players are the same across the two treatments when player A chooses (H). As a result, the FF model predicts equal reciprocity towards (H) in treatments N and NP. Given the large difference in reciprocal responses across the two treatments, the main evidence presented by Experiment 2 cannot be captured by the FF model.

In both treatment N and treatment NP, if player A chooses (H), player B has no choice. In treatment NP, player B can pay to punish player A for choosing (S). If player A believes that a positive proportion of player Bs would pay to punish (S), it means that he expects player B to receive a lower payoff in treatment NP than in treatment N, should he choose (S). Therefore, the DK model predicts a higher perceived kindness of (H) in treatment NP than in treatment N, because the reference point of payoffs associated with choosing (S) is lower in treatment NP. This prediction is also contradicted by Experiment 2 results.

Relation to reciprocity models based on revealed altruism Overall, the results of Experiment 2 can be explained by the GP model. 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 (extended) CFS model that we discussed in Section 3.2. can partially capture these results. Across all the treatments of Experiment 2, the maximum payoff player A can get if he chooses (S) is $4.50. 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. On the other hand, comparing treatments N and NR, the extended CFS model could predict higher levels of positive reciprocity in treatment N than in treatment NR, as we document in Experiment 2.

4 Conclusion and directions for future research

Both genuine kindness and self-interested material gain may drive decision-making in reciprocal interactions. 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. How do people reciprocate to helpful actions when the benefactor may be simply people throwing a sprat to catch a mackerel? This paper presents data from two experiments designed to isolate the role of perceived motives on reciprocal behavior. By providing evidence from both a between-subjects and a within-subject design, the results show that the stronger strategic incentives the first-mover has for choosing the helpful action, the lower the degree of positive reciprocity is to that action. Also, the decline in the level of reciprocity is shown to be associated with the deterioration of altruism inferences regarding the person who took the helpful action. 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. The current research notes the ambiguity of motives in reciprocal interactions. Future research studying how people make inferences about these ambiguous motives needs to complement research on how these inferences impact behavior. Also, the current research does not compare the importance of perceived motives in relation to other important determinants of reciprocal behavior, such as distributional preferences and perceived intentions. Future experiments design to assess the relative importance of each of these factors are needed.

Such research can also help clarify seemingly contradictory results in the literature. For example, Bolton and Ockenfels (1998) showed that 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 in the Güth van Damme's three person bargaining game. 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 Guth 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 study the relative importance of outcomes versus perceived motives in a design that allows us to vary 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.

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. 2004. “How to identify trust and reciprocity.” Games and Economic Behavior, 46, 260-281.

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.

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.

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.

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.

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. In this discussion, 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 ($y > k$, $x > 5k$, $5k > t$).

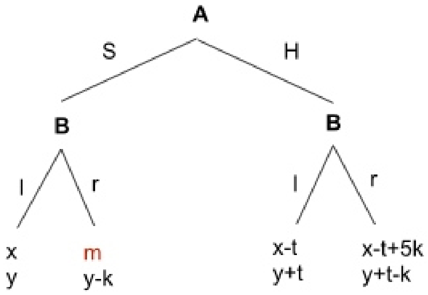

Figure 3: Generalized Game Γ_1

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.

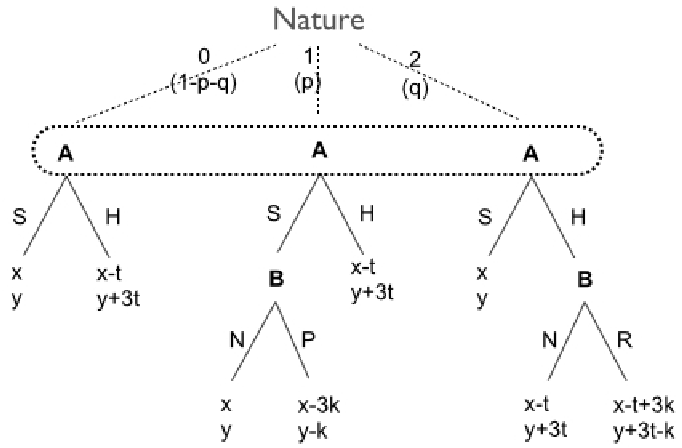

Figure 4: Generalized Game Γ_2

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).[19] 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 $\kappa^{RP}_B(H) = \frac{1}{2}\{b^{''}_{RP}(r|H)(y+t-k) + (1-b^{''}_{RP}(r|H))(y+$

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

$t)\} - \{b''_{RP}(r|S)(y-k) + (1 - b''_{RP}(r|S))(y)\}$. 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 $b''_{RO}(r|S) > b''_{RP}(r|S)$ 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 $\kappa^N_B(H) = 1.5t$.

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 $\kappa^{NR}_B(H) = 1.5t - 0.5b''(R|H)k$. 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 $\kappa^{NP}_B(H) = 1.5t + 0.5b''(P|S)k$. 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.[20]

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 kind in both treatments, as the payoffs are the same for this subgame across the treatments. 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 kinder ($\kappa_B^{RP}(H) > \kappa_B^{RO}(H)$), leading to less positive reciprocity in treatment RO in response to (H), producing the exact opposite prediction of the behavior we are trying to explain. Therefore,

[20] 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.

the model cannot explain the main hypothesis of this paper, namely the higher degree of positive reciprocity in response to (H) in treatment RO.[21]

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 kind in treatment NR as it does in treatment N . However, if $b^{''}(R|H) > 0$, then action (H) looks less kind 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 kind in treatment NP compared to the action (H) in treatment N, $\kappa_B^{NP}(H) = y - x + 4t$. Therefore, the Falk and Fishbacher (2006) model would predict that the positive reciprocity towards (H) would be (at least weakly) higher in treatment N than in treatment NR, and equal in treatments N and NP, clearly contradicting the predictions and findings in this paper.[22]

[21] As long as $b^{''}_{RO}(r|S)$ is not much larger than $b^{''}_{RP}(r|S))$, the model also predicts that $\kappa_B^{RP}(S) > \kappa_B^{RO}(S)$, since the punishment option yields payoffs that are more equal than the reward option. The data shows that more player B's choose to punish (S) in treatment RP than they choose to reward (S) in treatment RO and player B's second order expectations are also aligned with this ordering ($b^{''}_{RO}(r|S) < b^{''}_{RP}(r|S)$). The experiment is not designed to test the implications of this prediction.

[22] 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.