Perceived Motives and Reciprocity

A. Yeşim Orhun*†

Abstract

In reciprocal interactions, both genuine kindness and self-interested material gain may motivate socially beneficial actions. The paper presents results from two experiments that distinguish the role of perceived motives in reciprocal decision making from the role of outcomes or perceived intentions. The results indicate that positive reciprocity triggered by the same beneficial action is lower when the first-mover is more likely to be motivated by strategic incentives. Therefore, stronger incentives for beneficial behavior may not increase total welfare..

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

JEL: C91, C92, D63, D64, D84.

*Corresponding author can be contacted at aorhun@umich.edu, Ross School of Business, University of Michigan, 701 Tappan St. Ann Arbor, MI 48109.

†I am indebted to the Associate Editor and two anonymous referees for their guidance. I also thank Gary Bolton, Jonathan Carmel, Bogachan Celen, Yan Chen, James Cox, Seda Ertac, Emel Filiz-Ozbay, Aradhna Krishna, Steve Leider, Yusufcan Masatlioglu, Axel Ockenfels, Steve Salant, Andrew Schotter, Katharina Schüssler, Severine Toussaert, Neslihan Uler, Peter Werner, and seminar participants at Erasmus University Rotterdam, George Mason University, New York University, University of Cologne, University of Michigan, and University of Texas at Dallas for their comments and suggestions. Lillian Chen, Michael Payne, Arun Varghese, Roshni Kalbavi, Hannah Lee, Valerie Laird and Catherine Dolan provided excellent support in conducting experimental sessions.

1 Introduction

An action that yields a beneficial outcome for others can be altruistically or strategically motivated.[1] Consider a man who lends money to his nephew with the intention of helping him with his college payments. On the one extreme, this act may be entirely motivated by altruism, with no expectation of reciprocity. On the other extreme, the man may be motivated by his intention to elicit a larger material favor from his nephew in the future and would not have helped him without this strategic motive. Does the nephew's inference about his uncle's motives influence how he reciprocates?

Economists have conjectured that perceived motives influence kindness perceptions and reciprocal decision making. Bellemare and Shearer (2011, p. 861) speculate that gifts "clearly in the short-term interests of the firm" may not be perceived as kind. Rabin (1998, p. 22) notes that "a crucial feature of the psychology of reciprocity is that people determine their dispositions toward others according to motives attributed to these others.... If you think somebody has been generous to you solely to get a bigger favor from you in the future, then you do not view his generosity to be as pure as if he had expected no reciprocity from you." As discussed in more detail in the next section, the impact of perceived *intentions* on reciprocal behavior has been a topic of great interest. Although perceived *motives* are likely central to kindness attributions, the experimental literature is silent on whether these perceptions matter for reciprocal behavior, possibly due to challenges in experimentally identifying the impact of perceived motives from that of perceived intentions.

Intentions and motives, while related, are distinct constructs. Intention refers to *what* an individual meant his or her action to yield as a consequence. The theoretical work on sequential reciprocal interactions has defined the kindness of a first-mover's intention as depending on (i) the voluntariness of the action and (ii) how he thought his action would affect the utility of the second-mover in equilibrium (Dufwenberg and

[1]The distinction between strategic versus altruistic motives in reciprocal interactions—and, relatedly, intrinsic and instrumental reciprocity—has been recognized in prior literature (Cabral, Ozbay and Schotter, 2014; Dreber, Fudenberg, and Rand, 2014; Gneezy, Güth, and Verboven, 2000; Reuben and Suetens, 2012; Segal and Sobel, 2008; Sobel, 2005).

Kirchsteiger, 2004; Falk and Fischbacher, 2006).[2] In contrast, motive refers to *why* the individual wanted to achieve the intended consequence. As the opening example demonstrates, different motives may drive the same intent. The distinction between intent and motive is not merely a semantic one, as demonstrated by its role in psychological attributions (Heider, 1958; Kelley, 1973; Ross and Fletcher, 1985) and its relevance in criminal law.[3]

In this paper, I examine whether perceived motives influence reciprocal behavior above and beyond the actualized consequences and perceived intent of an action. I present treatments that manipulate beliefs about the first-mover's strategic motive without generating confounding movements in perceived intentions and without the need to deceive subjects about the nature of the interaction. The experiments also elicit second-order beliefs and inferences about a first-mover's altruism that support the hypothesized impact of perceived motives by revealing players' mental models and providing evidence for their strategic thinking.

In related work, Stanca, Bruni, and Corazzini (2009) and Johnsen and Kvaloy (2016) show that the degree of positive reciprocity triggered by the first-mover's helpfulness is higher in treatments in which the first-mover mistakenly believed that he could not materially benefit from being helpful compared with treatments in which the first-mover knew that second-movers could positively reciprocate. However, it is unclear whether the differences in reciprocity are driven by a reaction to perceived motives or to perceived intentions. In treatments in which the first-mover may harbor expectations of rewards for being helpful, he is more likely to be strategically motivated rather than intrinsically motivated by altruism. However, because rewards come at a cost, his helpful action is associated with worse expected material consequences for the second-

[2]Henceforth, for simplicity, the male pronoun is used to refer to the first-mover and the female pronoun is used to refer to the second-mover.

[3]Criminal courts must determine a defendant's intent behind an action that caused harm to another. Did the defendant expect his or her conduct to cause harm to the victim and desire this outcome? The court also must establish the defendant's motive. Was the intentional harm inflicted in self-defense, or was it fueled by revenge? The intent element of a crime may exist without any malicious motive—or may even exist with a perceived benevolent motive, as in the case of mercy killing.

mover, and therefore the helpful action is perceived to be driven by less kind intentions (Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). Therefore, varying perceptions about whether the first-mover expected rewards for his helpful action, while intuitively in line with the example of the uncle helping his nephew with college tuition payments, produces confounding movements in both perceived intent and perceived motive. Instead, the identification strategy used in this paper rests on manipulating what the first-mover expects to lose by not being helpful. When the helpful action is strategically motivated to avoid punishment, it also protects the second-mover from a worse expected material payoff that she would have obtained due to having levied a costly punishment, and thus the intention of the first-mover cannot be perceived as less kind. Therefore, lower degrees of positive reciprocity when strategic motives are perceived as being stronger cannot simultaneously be explained by an account of intentions.

To demonstrate, consider two firms, one in California and one in Texas, that each install special filters in their factories. Both firms expect the filters to decrease employees' exposure to air pollutants, and both desire this outcome. Therefore, both firms intend to improve working conditions. However, the California firm is motivated mainly by the fear that its employees will strike if it does not install the filter, while the Texas firm is motivated mainly by a concern for its employees' well-being. All else being equal, how will the employees view and react to their respective firm's actions? A consideration of motives would clearly identify the Texas firm's action as kinder than that of the California firm, and one could conjecture that employees of the Texas firm may be more likely to oblige a request to work over the weekend if needed. Intent alone, however, cannot capture this intuition, as both firms expect and desire to bring about the same consequence.[4]

I present data from two main experiments that rely on the same identification strategy in the context of a two-player reciprocal interaction. The results show that reciprocation hinges on whether the beneficiary believes that the benefactor made a sac-

[4]The predictions of intention-based reciprocity models are formally derived in Experiment 1, which closely resembles this example.

rifice for strategic reasons or out of altruism. Fixing the first-movers' intentions about the impact of his helpful action on the second-mover, in cases where second-movers can punish, first-movers are more likely to be helpful, but second-movers are less likely to positively reciprocate. Moreover, the within-person data on beliefs provide direct support for the proposed mechanism of perceived motives. Second-movers become more likely to reward the first-mover's helpful action in the strategic interaction the more they expect the first-mover to be helpful even in the absence of a strategic incentive. Several robustness treatments prove that these findings are not driven by menu-effects and are robust to many variations in experimental design, including variations in the type of decisions, whether beliefs are elicited, whether actions are elicited directly or via the strategy method, and whether the treatments are within- or between-subjects.

The findings have important implications for contract design. The welfare gains achieved in reciprocal interactions depend not only on the initial action but also on the degree of reciprocity it triggers from the other party. Prior studies have shown that people tend to reward helpful actions and punish hurtful ones and that these responses can lead to large welfare gains by encouraging socially beneficial actions that would otherwise not be incentive-compatible.[5] Although punishments and rewards increase helpfulness behaviors for those who are subject to them, if perceived motives matter, these incentives may have a hidden cost. The existence of strong extrinsic incentives may taint the beneficiaries' perceptions of kindness as the motive behind the helpful actions and thus damage, or at least diminish, their willingness to reciprocate, thereby decreasing the overall welfare gains the incentives aimed to achieve. Indeed, the results of the experiments herein show that providing stronger extrinsic incentives does not necessarily lead to an increase in total welfare, even if it is effective in increasing transfers to the second-mover. Englmaier and Leider (2012) show that profit-maximizing firms can make the most of workers' feelings of reciprocity and effectively reach outcomes that benefit both parties if offers of higher wages are interpreted as earnest

[5]For example, see Fehr, Gächter, and Kirchsteiger (1997) and Andreoni, Harbaugh, and Vesterlund (2003). A recent review of experimental results is provided by Cooper and Kagel (2016).

acts of kindness. However, if perceived motives matter in reciprocal decision making, the degree to which self-interested parties can take advantage of and influence others' reciprocal behaviors may be limited.

Section 2 discusses related literature, Section 3 presents the experiments and analyzes their results, and Section 4 discusses the implications of these results for the existing and future work in this area.

2 Related Literature

The outcome-based models of altruism and reciprocity (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) explain the positive relationship between a helpful action and a reaction only in terms of preferences for payoff allocations. These models capture the foremost determinants of other-regarding behavior. A class of reciprocity models also considers the impact of the perceived kindness of one's intentions on reciprocity feelings (Rabin, 1993; Dufwenberg and Kirchsteiger, 2004; Falk and Fischbacher, 2006). As discussed previously, the notion of intent has two dimensions: expectations of consequences and voluntariness. Therefore, intention-based reciprocity models define intent in terms of the consequence the first-mover wanted the action to have on the second-mover. In particular, Dufwenberg and Kirchsteiger (2004) define the kindness of the first-mover's action based on the difference between the second-mover's expected payoff and what the second-mover could have obtained had the first-mover behaved differently. Falk and Fischbacher (2006) define the kindness of the first-mover's action based on the comparison between the first-mover's and second-mover's expected payoffs. Moreover, Falk and Fischbacher (2006) explicitly allow voluntariness of the first-mover's actions to impact kindness perceptions.

Experimental literature provides two types of tests of the intention-based reciprocity account. One group of experiments tests whether varying the voluntariness of an action affects the reciprocity it generates. These experiments compare a control group, in which the first-mover can voluntarily choose what action to take among a set of alternatives, with a treatment group, in which the first-mover cannot choose either because

there is no alternative (McCabe et al., 2003) or because the choice is determined by an external process, such as chance (e.g., Blount, 1995; Charness and Haruvy, 2002; Charness, 2004; Charness and Levine, 2007; Falk et al., 2008; Klempt, 2012; Offerman, 2002). Overall, this literature finds positive reciprocity to be higher in treatments in which the action is perceived to be voluntary and interprets the difference to be consistent with an account of intentions. However, with regards to the experiments that compare voluntary actions to actions determined by a-priori fair chance, Bolton et al. (2005) present data that reject this interpretation in favor of the procedural fairness account. Another group of experiments tests whether reciprocity is influenced by what the second-mover could have obtained if the first-mover had behaved differently. Brandts and Solà (2001) and Nelson (2002) find positive reciprocity with respect to the same action to be higher in treatments in which the first-mover could not have chosen a better outcome for the second-mover compared with treatments in which the first-mover could have been more helpful.

Importantly, as I show in the current paper, intention-based reciprocity models do not account for kindness perceptions based on motive attributions. Similarly, previous experimental work on the role of intentions manipulates perceptions about *what* the first-mover expected his action to yield, but it does not independently vary perceptions about *why* he wanted it.

To provide evidence for the role of motives, the experimental design must create variation in beliefs regarding the first-mover's motives for an action without generating confounding shifts in beliefs regarding intended consequences or voluntariness. The experiments in this paper shift the proportion of helpful first-movers who are strategically versus altruistically motivated and achieve this objective by varying the strategy space of the second-mover. Stanca, Bruni, and Corazzini (2009), Straissmair (2009), and Johnsen and Kvaloy (2016) also present designs that shift this proportion by manipulating beliefs about what first-movers expect to gain from being helpful. Straissmair (2009) does not find any difference in the reciprocity of second-movers as a result of a shift in the expected gains of the first-mover from being helpful. In contrast, Stanca,

Bruni, and Corazzini (2009) and Johnsen and Kvaloy (2016) find that second-movers reciprocate more in treatments in which the first-mover mistakenly believed that he could not gain from being helpful. However, these manipulations fail to separate the impact of perceived motives from that of perceived intentions, because in treatments in which the action of the first-mover is more likely to be strategically motivated, the action is also associated with worse expected material consequences for the second-mover. The main contribution of the experiments described in this paper is the ability both to isolate the role of motives from the role of intentions and to shift beliefs about motives without having to mislead subjects about the nature of the interaction.

3 Experimental Investigation

3.1 Experiment 1

3.1.1 The reciprocal interaction

Experiment 1 involves a two-stage reciprocity game (Game Γ_1), depicted in Figure 1. The first-mover (player A) chose between (H) and (S), where (H) was associated with higher material payoffs for the second-mover. Note that player A chose between the same number of choice options and had full control over his choice in both treatments. The second-mover (player B) chose between a reciprocal (r) or material payoff-maximizing (l) option in response to either choice. Conditional on the first-mover choosing (H), the second-mover chose between options that induced the same material payoffs at the end nodes in both treatments. However, conditional on the first-mover choosing (S), the treatments varied in the material payoffs for player A.

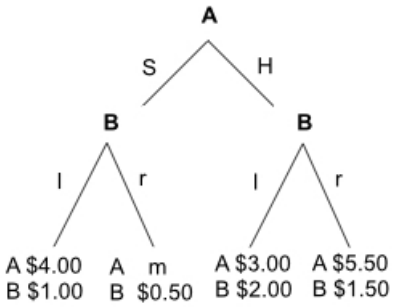

Figure 1: Game Γ_1

All participants in a given session made decisions in either treatment 1 or treatment 2 versions of Game Γ_1. Game Γ_1 took on different natures depending on the value of m, which was set to \$1.50 in treatment 1 and \$6.50 in treatment 2. Therefore, treatment 1 gave player B a costly punishment option if player A chose (S), whereupon player B could decide to sacrifice \$0.50 to decrease player A's payoff by \$2.50. Treatment 2 gave player B a costly reward option if player A chose (S), whereupon player B could decide to sacrifice \$0.50 to increase player A's payoff by \$2.50. In short, in treatment 1, the second-mover had the option to negatively reciprocate to the choice of (S), but she did not have this option in treatment 2.[6] In both treatments, the game gave player B the same costly reward option if player A chose (H). The interaction was framed as player As choosing between (\$4, \$1) and (\$3, \$2) in the first stage and player Bs choosing between keeping these payoffs unchanged and paying \$.50 to alter them in the second stage. Player B responses were elicited directly and conditional on observing player A's choice. More subjects were recruited for the treatment 2 sessions to achieve a comparable number of instances in which player Bs made a decision in response to (H) across the two treatments.

In treatment 1, the first-movers who choose (H) could be motivated by altruism, the hope of receiving a reward, and/or the fear of punishment, whereas in treatment 2, the

[6]I also conducted a treatment in which player B had the choice of three options if player A chose (S): (\$4, \$1), (\$6.50, \$0.50), and (\$1.50, \$0.50). The online appendix presents a comparison of behavior in this alternative treatment (referred to as treatment 1d) with the behavior in a treatment that parallels treatment 2 (referred to as treatment 2d). The results mirror the results from Experiment 1.

fear of punishment was not present. Given that previous research shows that sanctions in combination with rewards are more motivating than rewards alone (Andreoni et al. 2003), treatment 1 should motivate more player As to choose (H) for strategic reasons. If player Bs expect a higher proportion of player As who chose (H) to have done so for strategic motives, they should be less likely to reward player As. Therefore, the central hypothesis of Experiment 1 is that *in response to player A choosing (H), a lower proportion of player Bs will choose (r) in treatment 1 than in treatment 2.* This prediction is not captured in the Dufwenberg and Kirchsteiger (2004) or the Falk and Fischbacher (2006) intention-based reciprocity models.[7]

3.1.2 Additional decisions

The experiment included four parts. The two-stage reciprocity game (Game Γ_1) was presented in part 3 of the four-part experiment. The focal aim of the experiment is to document how the reciprocal behavior in Game Γ_1 varies as the strength of strategic incentives to help in the first stage is manipulated. Here, I briefly explain the other parts and the motivation for including them in the experimental design. The experiment is reproduced in the online appendix.

Part 1 presented binary choices in modified dictator games that reflected the same trade-offs the player would make later in the context of the reciprocal interaction in part 3. Part 1 also included other binary modified dictator games and trade-offs, partly motivated by not wanting to draw too much attention to the repetition of the choices of interest between part 1 and part 3 to avoid stickiness in choices.[8] In part 1, only player As made decisions in dictator games, while player Bs waited. Asking player As

[7]The Appendix provides a formal analysis of these models' predictions in Game Γ_1. Because the Falk and Fischbacher (2006) model does not take into account what the first-mover could have obtained if he chose differently, because the material payoffs at the end nodes following (H) are held constant across the two treatments, and because player B earns less than player A, this model always predicts that player B will choose l after player A chooses H in both treatments. The Dufwenberg and Kirchsteiger (2004) model predicts positive reciprocity to be more prevalent in treatment 1 because player B's expected material payoffs, conditional on player A choosing (S), are lower in equilibrium.

[8]To the extent that stickiness is a concern, it applies equally across treatments and does not impact the main results. However, it would lead to an underestimation of reciprocity, rendering the results conservative.

to make choices in modified dictator games that present the same choice options as in the first-stage of the reciprocal game provides information about their other-regarding preferences in the absence of reciprocation. This baseline behavior helps identify the degree to which the helpful behavior in the first stage of the reciprocal interaction results from strategic considerations and the degree of reciprocity versus altruism.

Part 2 elicited subjects' predictions about the percentage of player As in that session who had chosen each option in a subset of games featured in part 1. The subset included predictions that were relevant for comparisons with the predictions in part 4. The predictions elicited in part 2 serve as baseline beliefs about the degree of altruism in the population of participants in a given session. This information is useful in determining whether the beliefs elicited in part 4 reflect an understanding of the strategic considerations of the first-movers and makes it possible to conduct within-person analyses of beliefs. Subjects earned \$4 if their predictions were exact, and their earnings declined quadratically as a function of their inaccuracy. Summaries of the results from parts 1 and 2 appear in Table 3 in the Appendix.

Finally, part 4 elicited subjects' predictions about behavior in the sequential reciprocity game using the same accuracy incentives as in part 2. In particular, part 4 elicited player A's first-order beliefs about player B's responses and player B's first-order beliefs about player A's choices. For example, player As were asked "What percentage of Person Bs chose each option (r or l) in response to S?" and were instructed that the percentage of choices should add up to 100%. In addition to eliciting first-order beliefs, part 4 also elicited player Bs' second-order beliefs (i.e., their expectations of player As' first-order beliefs) with the following question: "We asked Person As, 'What percentage of Person Bs chose each option (r or l) in response to S?' What do you think was the average of their predictions?" In response, Player Bs completed statements such as "On average, Person As expected ___% of Person Bs to choose r (or l) in response to S." These beliefs are useful in establishing internal validity of the experiment, providing evidence for the mental models of the players, and addressing the concern that differences in behaviors may be confounded with other strategic issues (Bolton,

Brandts and Ockenfels, 1998).

3.1.3 Protocol

Each session had an even number of subjects, ranging from 10 to 20. Half the subjects in a session were randomly and anonymously assigned the role of player A, and the rest were assigned the role of player B. Players kept these roles throughout the four-part experiment. A total of 258 subjects, all over the age of 18 years, from the University of Michigan student and staff population were recruited through ORSEE to participate in 18 60-minute lab sessions for this experiment.[9]

Subjects earned a fixed participation fee of $5. They also earned additional payments from each of four parts of the experiment. If the parts included more than one task, subjects were informed that one task was selected at random from each part to determine additional payments. Each player A was randomly and anonymously matched with one player B for each task, and all communications about decisions of the matched players were anonymous. Participants interacted using the software z-Tree (Fischbacher, 2007).

Subjects were informed that their payments from each part were independent of their choices in both future and previous parts of the experiment. Each part was introduced with its own set of instructions to all subjects at the same time. Subjects completed all four parts without receiving feedback on their performance or others' behaviors until the end of the experiment.[10] At the end of the study, subjects learned the randomly selected tasks for each part and received payments in a double-blind payoff protocol.

[9]Subjects who participated in one experiment were not allowed to participate in another.

[10]The exception is in part 3 of Experiment 1, in which each matched pair learned the decision of their partner in the reciprocal interaction because a direct elicitation method was employed. In Experiment 2, responses were elicited using a strategy method in part 3; therefore, no information was shared until the end of the experiment.

3.1.4 Results

Table 1 summarizes first-movers' actions and first-order expectations (A FOE), second-movers' responses, second-movers' first-order expectations (B FOE), and second-movers' second-order expectations (B SOE) in Game Γ_1, in which a total of 129 pairs of player As and Bs interacted.

Table 1: Beliefs and Actions in Game Γ_1 across Treatments 1 and 2

	N	A choice	B FOE	B choice		A FOE		B SOE	
		H	H	r|S	r|H	r|S	r|H	r|S	r|H
		(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
Treatment 1	59	55	72%	1	19	41%	30%	54%	27%
Treatment 2	70	46	41%	1	26	19%	41%	16%	43%

A choice (column 1) and B choice (columns 3-4) report the number of subjects making the indicated choice. B FOE and A FOE columns report average first-order expectations of players B and A. The B SOE column reports player Bs' second-order expectations.

Player B responses The central hypothesis of Experiment 1 is that *in response to player A choosing (H), a lower proportion of player Bs will choose (r) in treatment 1 than in treatment 2.* In support of this hypothesis, only 19 of the 55 (34.5%) player Bs rewarded (H) in treatment 1, whereas 26 of the 46 (56.5%) player Bs rewarded (H) in treatment 2 (chi-square test: $\chi^2(1) = 4.90$, $p = 0.027$).

This result is in line with the conjecture that second-movers are less likely to reciprocate positively to the same helpful action when the reciprocal interaction provides stronger strategic incentives for the first-movers to be helpful. In further support of this conjecture, the findings also show that (i) among the helpful player As, treatment 1 included a higher proportion of strategically motivated players, and (ii) players' mental models with respect to one another were congruent. I discuss these results next.

Actions of player A Because player Bs cannot respond to player As in dictator games, player As' choices in part 1 of the experiment reflect their intrinsic preferences. To judge whether the action of a player A in Game Γ_1 is strategically motivated, it is possible to compare it with his choice in the dictator game that involved the same trade-off. The first stage of Game Γ_1 presents a choice between (\$4, \$1) and (\$3, \$2), in which the first amount denotes the payoff to player A and the second denotes the payoff to player B. When player As chose between these options in part 1, only 37 of the 129 (29%) player As chose to transfer \$1 from their payment to the other player, and the rest chose to keep all \$4 for themselves (see Table 3, Appendix). Due to strategic considerations, player As were more willing to sacrifice \$1 to help player B in the first stage of Game Γ_1 than in part 1, in both treatment 1, which included the possibility of rewards and punishments (93% in Game Γ_1 vs. 29% in the dictator game; McNemar test: $\chi^2(1) = 39$, $p = 0.000$), and treatment 2, which included the possibility of rewards (66% in Game Γ_1 vs. 29% in the dictator game; McNemar test: $\chi^2(1) = 23.15$, $p < 0.001$).[11] Furthermore, more player As chose (H) in treatment 1 than in treatment 2 (93% vs. 66%; chi-square test: $\chi^2(1) = 14.25$, $p < 0.001$), due to the additional incentives that treatment 1 presented by allowing player Bs to punish unhelpful actions.[12] Thus, the manipulation of incentives across treatments achieved its objective: the proportion of player As who were strategically motivated within the set of player As who choose (H) was higher in treatment 1.

Beliefs Identifying the impact of perceived motives relies on exogenously shifting second-movers' beliefs about the reason a first-mover helped her. Experiment 1 provides evidence for such a shift by showing that (i) player A's beliefs about layer B's response to his actions and (ii) player B's beliefs about these beliefs are shifted.

Player A's first-order expectations are summarized in columns 5-6 of Table 1. In treatment 1, on average, player As expected 30% of player Bs to reward a helpful

[11] The McNemar test accounts for the paired nature of the responses across part 1 and part 3.

[12] The estimations that player Bs made about the proportion of player As who would choose (H) were conservative (biased toward the uniform), but their beliefs correctly reflected the ordering across treatments.

action and 41% of player Bs to punish an unhelpful action. In treatment 2, on average, player As expected 41% of player Bs to reward a helpful action and expected 19% of player Bs to reward an unhelpful action.[13] Player As' first-order expectations reveal two important features of their mental models. First, they expected a considerable degree of punishment (41%) in response to an unhelpful behavior in treatment 1. With such expectations, choosing to be unhelpful does not maximize expected returns from Game Γ_1, which may explain why 93% of player As chose to be helpful in this treatment. Second, player As expected positive reciprocity in response to their helpfulness, but only in treatment 2. Expectations of positive reciprocity can be identified by comparing how likely player As believed participants were to choose (r) over (l) in a modified dictator game that presented the same options (elicited in part 2) with how likely they believed this choice was when the person was responding to (H) in Game Γ_1 (elicited in part 4). Player As reported an average expectation of only 24% of people making the same choice in part 1 (see Table 3, Appendix). Within-subject differences in expectations reveal that player As expected positive reciprocity (over and above an expectation of altruism) in treatment 2 (Wilcoxon sign-ranked test: $z = -4.25$, $p < 0.001$) but not in treatment 1 (Wilcoxon sign-ranked test: $z = -1.01$, $p = 0.311$). The fact that player As did not expect positive reciprocity for being helpful in treatment 1 reflects their understanding of the role of perceived motives with respect to player Bs' reciprocal feelings.

Player Bs also understood player As' considerations. Player B's first-order expectations (shown in column 2 of Table 1) reflect a significant difference in the extent to which player Bs thought player As were willing to choose (H) in treatment 1 (72%) versus in treatment 2 (41%) (two-sample Wilcoxon rank-sum test: $z = 5.78$, $p < 0.001$). This result suggests that second-movers understood the incentives of each treatment

[13]This expectation may seem optimistic given that only 1 out of 24 player Bs who faced a choice in this subgame chose ($0.50, $6.50) over ($1, $4); however, the number of observations in the subgame is too low to conclude that the belief is biased. In fact, player Bs also expected a similar (16%) fraction of other player Bs to do so. For reference, 31% of subjects who faced a choice between the same options picked ($0.50, $6.50) over ($1, $4) when these options were presented in a binary modified dictator game context in part 1.

for first-movers, which is a prerequisite for contemplating the first-mover's motives.[14]

Most importantly, player B's second-order expectations (shown in columns 7-8 of Table 1) were closely aligned with player As' first-order expectations. In treatment 1, on average, player Bs believed that player As expected 27% of player Bs to reward (H) and 54% of player Bs to punish (S). In treatment 2, on average, player Bs believed that player As expected 43% of player Bs to reward (H) and 16% of player Bs to reward (S). Congruent with the main result, as well as player As' expectations, player Bs intuited that the expectations of rewards are higher in treatment 2 than in treatment 1 (43% vs. 27%; two-sample Wilcoxon rank-sum test: $p = 0.006$).

Overall, these data show that actions, first-order and second-order expectations were all closely aligned, suggesting that players had a clear understanding of each others' preferences and strategic considerations. Most importantly, the experiment manipulated player B's second-order expectations as intended, shifting the composition of the potential motive they attributed to helpful player As.

Welfare Experiment 1 demonstrates that a fear of punishment increases first-movers' helpfulness and generates larger transfers to second-movers. If the second-movers continued to reward helpfulness at the same rate as in the situation in which punishment was not possible, the 27% increase in first-stage helpfulness would lead to a 15% increase in the number of player A-B pairs achieving the highest total welfare option (\$5.50, \$1.50). However, the welfare gains achieved in a reciprocal interaction depend not only on the initial action but also on the degree of reciprocity it triggers from the other party. Experiment 1 shows that the existence of strong extrinsic incentives taints the perception of the motives behind helpful actions and decreases positive reciprocity. This decrease in positive reciprocity more than offsets the welfare gains that could have been achieved as a result of an increase in helpfulness in the first stage. The average earnings in Game Γ_1 are \$5.59 in treatment 1 and \$5.77 in treatment 2. This

[14]Note that because responses in Game Γ_1 are directly elicited, player Bs' beliefs about the percentage of player As in the session who chose (H) or (S) are elicited after player Bs observe the choices of the player As matched with them for part 3. This information may increase the accuracy of player Bs' first-order beliefs. However, other beliefs are not influenced by this information.

result serves as a stylized example of how increasing the strength of incentives may not lead to increases in welfare due to its deleterious effect on perceived motives and, as a consequence, reciprocity.

3.1.5 Additional checks and evidence

I conducted three additional paired treatments to check the robustness of the results. For these additional treatments, subjects who had not participated in treatments 1 and 2 of Experiment 1 were recruited. The protocol and detailed results of these treatments appear in the online appendix.

First, I checked whether the differences in positive reciprocity across treatments 1 and 2 were driven by the mere existence of different options in the subgame that followed (S). Is the second-mover's choice mainly a response to the different choice options or to differences in the strategic considerations of the first-mover? To answer this question, treatments 1a and 2a replicate treatments 1 and 2, but with one important difference: player A has no choice to make, and (H) is selected by the computer. Subjects were informed that the computer picked (H) and asked player Bs to choose between ($3, $2) and ($5.50, $1.50) either when their choice would have been between ($4, $1) and ($1.50, $0.50) had the computer picked (S) (treatment 1a) or when their choice would have been between ($4, $1) and ($6.50, $0.50) had the computer picked (S) (treatment 2a). There were no differences in player Bs' choices across the two treatments. Of 63 player Bs, 38% chose ($5.50, $1.50) over ($3, $2) in treatment 1a, and 39% made the same choice in treatment 2a. Therefore, the difference in second-movers' choices in the subgame reached after (H) in Experiment 1 is not driven by the mere existence of different alternatives in the subgame reached after (S), but require the attribution of the first-stage decision to the first-mover.

Second, I checked for the robustness of beliefs. The beliefs elicited in part 4 are closely aligned with actual choices and other players' beliefs, which can be interpreted as a clear understanding of the strategic considerations of Game Γ_1. However, player Bs' first-order beliefs may be accurate not due to an understanding of player As' strategic

considerations but because player B is exposed to the decision of one player A in the course of Game Γ_1. In addition, participants may report beliefs that make themselves feel better about their actions, such as reporting a low second-order expectation of rewards when choosing not to reward. To check whether the reported beliefs reflect subjects' understanding of the game, treatments 1b and 2b present the belief questions from part 4 to 121 third parties who did not participate in Game Γ_1. In a within-subject design, third parties predicted that more player As would choose (H) in treatment 1 (60%) than in treatment 2 (37%), and this difference was highly significant (Wilcoxon sign-ranked test: $z = 8.05$, $p < 0.001$). Conditional on player As choosing (H), third parties expected 30% of player Bs to reward player As in treatment 1 and 34% of them to do so in treatment 2 (Wilcoxon signed-rank test: $z = 2.93$, $p = 0.003$). Conditional on player A choosing (S), third parties expected 47% of player Bs to punish player As in treatment 1 and 18% of player Bs to reward player As in treatment 2. In addition, these third parties were asked to predict the proportion of player As who would have helped in the absence of any strategic considerations among those who chose to be helpful in each of the reciprocal interactions. First, they were asked to predict the proportion of player As who would choose (\$3, \$2) over (\$4, \$1) in part 1, in which player Bs could not respond. They predicted that 25% of the player As would make this choice. Second, they were asked to predict the proportion of player As who made the same choice among the player As who chose (H) over (S) in part 3. On average, the third parties predicted that only 43% of helpful player As in treatment 1 would also choose (\$3, \$2) over (\$4, \$1) when selecting between these options in part 1. In other words, they predicted that the remaining 57% of the helpful player As were motivated by strategic considerations in treatment 1. In contrast, they predicted that 49% of the helpful player As in treatment 2 were motivated by strategic considerations (Wilcoxon signed-rank test: $z = 3.79$, $p < 0.001$). These beliefs indicate that third parties recognize the difference in the mix of motives between treatments 1 and 2. Overall, the results suggest that the beliefs reported by player Bs in Experiment 1 reflect a clear understanding of the differences in player As' potential motives across

the two treatments.

Finally, I checked whether the results of Experiment 1 are robust to changing the material payoffs in the second stage such that (i) the maximum payoff player A could obtain by choosing (S) is constant across treatments and (ii) the inequality between material payoffs that follow (S) is not smaller following (S). In particular, in treatment 1c, (S, l) paid (\$5, \$5), (S, r) paid (\$0, 4.50), (H, l) paid (\$3, \$8), and (H, r) paid (\$9, \$7.50). In treatment 2c, the choice options and material payoffs that followed (H) were the same as in treatment 1, but the second-mover did not have a choice after (S), which paid (\$5, \$5). Therefore, treatment 1c presented a costly punishment option to player B if player A chose (S), but treatment 2c did not. The protocol differed from Experiment 1 in two ways. First, treatments 1c and 2c presented only the reciprocal interaction and omitted the decision tasks in parts 1, 2, and 4. This simplification makes it possible to assess whether having to make decisions in multiple parts, and in particular not having the results of these decisions realized until the end of Experiment 1, could have driven behavioral differences across treatments 1 and 2 (Cox et al., 2015). Second, the reciprocal interaction was framed as player A deciding between two sets of choice options that player B would choose from. Due to stronger strategic incentives, more player As chose (H) in treatment 1 (85% vs. 66%; chi-square test: $\chi^2(1) = 4.23$, $p = 0.04$). Because stronger strategic incentives to choose (H) lead to a lower perception of genuineness motives, lower degrees of positive reciprocity were expected in treatment 1c. Indeed, a lower proportion of the second-movers rewarded (H) in treatment 1c than in treatment 2c (44% vs. 77%; chi-square test: $\chi^2(1) = 9.49$, $p = 0.002$), regardless of the design and protocol differences these treatments presented.

3.1.6 Summary of results

Experiment 1 compares the levels of positive reciprocity of second-movers toward helpful first-movers when first-movers could have been motivated to help by the fear of punishment (treatment 1) with the level of positive reciprocity to the same helpful action when there was no punishment option in the second stage (treatment 2). In

treatment 1, the choice of (H) triggered lower reciprocity from player Bs. The robustness treatments show that this result (i) requires attribution of the first-stage decision to the first-mover, (ii) is not driven by the fact that the material payoff inequality is lower after (S) in treatment 1, or (iii) is not explained by the maximum payoff being higher after (S) in treatment 2. Overall, the results provide strong support for the distinct role of perceived motives in reciprocal decision making.

3.2 Experiment 2

Experiment 2 extends Experiment 1 in several ways. First, Experiment 2 makes it possible to compare the role of perceived punishment-avoidance motive separately with the no-incentive benchmark (where the first-movers are motivated by altruism), whereas first-movers could have been motivated by reward seeking in both treatments of Experiment 1. Second, and more importantly, Experiment 2 explores the mechanism by which the existence of strategic incentives may influence reciprocity. Given the close relationship between perceptions of the first-mover's motives and his altruism, it is likely that the second-movers' altruism inferences about helpful first-movers are more positive in treatment 2 than in treatment 1 of Experiment 1. Such a difference in altruism inferences seems plausible based on two data patterns from Experiment 1. First, second-movers expected a higher fraction of strategically motivated first-movers among those who were helpful in treatment 1. Second, third parties expected lower degrees of altruism among the helpful first-movers in treatment 1 than among the helpful first-movers in treatment 2. Therefore, Experiment 2 elicits beliefs about the altruism of helpful first-movers in a within-subjects design.

3.2.1 The reciprocal interaction

Experiment 2 investigates positive reciprocity in the context of a probabilistic sequential game in which 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) and (H). Then, nature chooses 1, 2, or 3. If nature

chooses 2, 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 (H); however, if player A chose (S), player B decides whether to pay \$0.50 to *decrease* player A's earnings by \$1.50 (P). If nature chooses 3, the game ends if player A chose (S); however, if player A chose (H), player B decides whether to pay \$0.50 to *increase* player A's earnings by \$1.50.

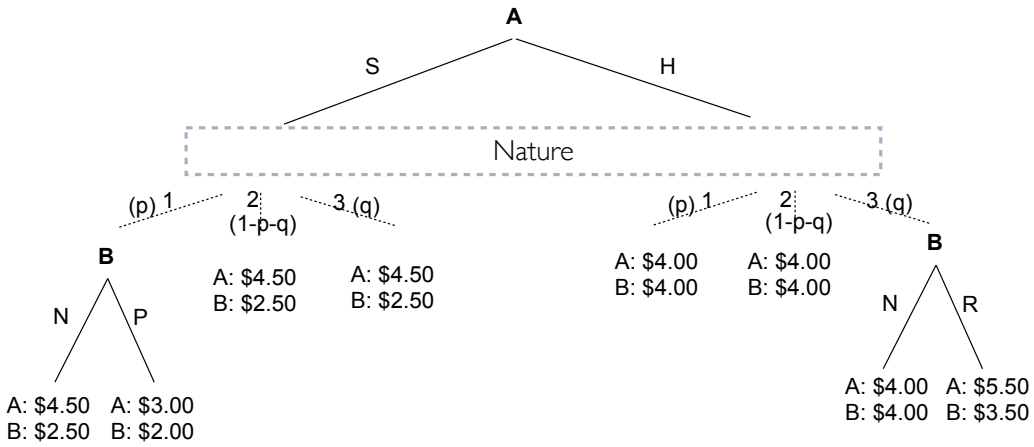

Figure 2: Game Γ_2

Let p be the probability that nature chooses 1, and let q be the probability that nature chooses 3. Consider how changing p and q may affect player A's and player B's behavior. First, consider $p + q$ approaching zero. Then, player A would choose (H) only if he genuinely prefers the more equitable allocation (\$4, \$4) to the more profitable allocation (\$4.50, \$2.50). In contrast, consider p approaching 1. Then, if player A chooses (H), he will earn \$4. If player A chooses (S), however, player B may choose (P), giving player A only \$3. Therefore, player A may be inclined to choose (H) to avoid potential punishment. Thus, the larger p becomes, the greater is the proportion of player As who are motivated by punishment avoidance among player As who choose (H). Experiment 2 features three treatments. In treatment 1, $p = .98$, and $q = .01$; therefore, player A expects player B (almost always) to have the option

to punish a selfish action. In treatment 2, $p = .01$, and $q = .01$; therefore, player A expects player B to be a passive recipient most of the time. In treatment 3, $p = .01$, and $q = .98$; therefore, player A expects player B (almost always) to have the option to reward a helpful action. In treatment 2, a helpful player A is most likely to be motivated by intrinsic preferences. However, in treatments 1 and 3, a helpful player A could also be motivated by punishment avoidance and reward seeking, respectively. In each of the three treatments, player Bs designate their response in each contingency using the strategy method. The probabilistic design makes it possible to elicit reward demand from player Bs even when they are not likely to be able to reward player As, without misleading participants about the nature of the interaction.

In contrast with Experiment 1, Experiment 2 focuses on one strategic motivation in each treatment, minimizing any expectations of the alternative strategic motivation. This makes it possible to compare the punishment-avoidance motive and the reward-seeking motive with the case in which there are no strategic motivations. Although comparing reciprocity differentials between treatments 2 and 3 is interesting, as highlighted in the discussion of Stanca et al. (2009) in Section 2, such differences can also be driven by intention-based reciprocity.[15] Therefore, the main goal of Experiment 2 is to test for reciprocity differences across treatments 1 and 2.

3.2.2 The protocol

In total, 176 participants, all over the age of 18 years, were recruited from the undergraduate and graduate student population at the University of Michigan to participate in 11 45-minute lab sessions. Experiment 2 reflected the same central design and protocol elements of Experiment 1 and is also reproduced in the online appendix. Experiment 2 also had a few unique features. Payoffs were denoted in tokens, where 200 tokens = \$1. In part 1, all subjects made decisions in modified dictator games.[16] Part

[15]Moreover, nature is equally unlikely to choose 3 in treatments 1 and 2 but is highly likely to choose 3 in treatment 3. There may be concern about comparing responses elicited across subgames that have drastically different probabilities of being carried out.

[16]At the end of the experiment, pairs of subjects and their roles were randomly assigned. The main reason for asking player Bs to make choices in part 1 was to elicit their altruism, which helps

2 elicited beliefs about behavior in the dictator games (part 1) from all participants. The participants were incentivized based on the accuracy of their reported expectations using a linear scoring rule for simplicity. A summary of results from parts 1 and 2 appears in Table 4 in the Appendix. Part 3 presented three within-subject treatments of Game Γ_2.[17] 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.

In Experiment 1, shifts in player Bs' second-order expectations were assumed to correspond to shifts in their perceptions regarding the proportion of helpful player As who are strategic (i.e., those who would not have helped if it were not for the reciprocal nature of Game Γ_1). In Experiment 2, part 4 directly elicited player Bs' altruism attributions regarding player As who were helpful in 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? (500, 0); (400, 300)." 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 to increase the payoff of player Bs by 300 tokens. Therefore, player Bs' beliefs about helpful player As' choices in this modified dictator game provide an indication of their beliefs about how player As would choose in the first stage of Game Γ_2 if it were not for strategic considerations. The elicitation of choices in the dictator games (part 1) makes it possible to incentivize truthful reporting about expectations of genuine kindness (part 4).

3.2.3 Results

Table 2 summarizes first-movers' actions and first-order expectations (A FOE), second-movers' responses, and second-movers' first-order expectations (B FOE) across treat-

assess whether their predictions about the altruism of helpful player As (elicited in part 4) suffer from projection bias, and pick the correct econometric approach for making inferences based on those beliefs.

[17]The order of the three treatments was counterbalanced among the following three treatment sequences (1-2-3, 2-3-1, 3-2-1). The ordering of the treatments did not affect any of the results.

ments 1 and 2 in Game Γ_2.[18]

Table 2: Beliefs and Actions in Game Γ_2 across Treatments 1 and 2

Treatment	N	A choice	B choice		B FOE	A FOE
		H	R | H	P | S	of H	of P | S
		(1)	(2)	(3)	(4)	(5)
Treatment 1	88	71	37	44	75%	43%
Treatment 2	88	42	55	38	51%	-

Column (4) reports the average predictions of player Bs about the percentage of player As who chose (H). Column (5) reports the average predictions of player As about the percentage of player Bs who chose would punish (S) in treatment 1.

Player B responses The central hypothesis is that player Bs are more likely to demonstrate positive reciprocity in treatment 2 than in treatment 1. The percentage of player Bs choosing R in response to (H) in treatment 2 is 63% (55 out of 88). As hypothesized, player Bs reciprocated less in treatment 1. Only 42% of player Bs (37 out of 88) indicated that they would choose R if player A chose (H) in treatment 1 (matched-pairs sign test, $p < 0.001$).

In this experiment, we can also compare player Bs' responses in Game Γ_2 to their choices in part 1. To reward player As for being helpful, player Bs must go from an equal distribution of 800 tokens for each player to 700 tokens for themselves and 1,100 tokens for player A. Only 20% of player Bs do so in part 1 in the context of a modified dictator game (see Table 4 in the Appendix). Therefore, in both treatments, player Bs demonstrate positive reciprocity (above and beyond distributional preferences), but do so more strongly in treatment 2, as hypothesized.

Actions of player A In part 1, 51% of player As chose the option (800, 800) over the option (900, 500) (Table 4, Appendix). This choice probability should be roughly the same in treatment 2, because it gives no additional strategic incentives to be helpful, but significantly higher in treatment 1 due to the possibility of punishment in the

[18]Table 5 in the Appendix reports results from treatment 3.

second stage. When player As faced the same binary choice in the first stage of Game Γ_2, 71 out of 88 (81%) of them in treatment 1 and 42 out of 88 (48%) of them in treatment 2 chose (H) over (S) (matched-pairs sign test, $p < 0.001$).

Beliefs Player Bs correctly expected a larger proportion of player As to choose (H) in treatment 1 than in treatment 2 (75% vs. 51%; matched-pairs sign test, $p < 0.001$). Player As would not have been motivated by the mere existence of the punishment option in treatment 1 if they did not believe that player Bs were likely to use it. Indeed, player As expected, on average, 43% of player Bs to choose P in treatment 1 if they chose (S), and 44 out of 88 (50% of) player Bs did so.

Altruism attributions The design of Experiment 2 makes it possible to test the notion that reciprocity is driven by beliefs about the genuine altruism of the first-mover. Recall that part 4 of Experiment 2 elicited attributions of altruism. First, I checked if attributions of altruism were different across treatments. If player Bs are sophisticated about how strategic motivations induce different selections of player As to be helpful, they should report higher expectations of altruistic (H)-choosing player As in treatments in which strategic incentives are weaker. On average, player Bs predicted that 73% of (H)-choosing player As in treatment 2 would choose (400, 300) over (500, 0) in part 1. However, they predicted that only 54% in treatment 1 would make the same choice (matched-pairs sign test, $p < 0.001$). These results suggest that player Bs believed that punishment avoidance led to a lower proportion of truly generous people among those who chose the helpful action. Player Bs also correctly inferred that the (H) choosers in treatment 1 were not kinder than the population of player As in general, as they reported an expectation (elicited in part 2) of 56% of player As choosing (400, 300) over (500, 0). These results show very different levels of inferred altruism from the same helpful action.

Next, I evaluated whether reciprocation responds to differences in altruism attributions, hypothesizing that an increase in player Bs' altruism attributions would increase her willingness to reward (H). This hypothesis specifically refers to within-

subject changes in beliefs rather than asserting a relationship between beliefs and reciprocity, because a within-subject differencing approach eliminates the potential confounds arising from possible correlation of preferences and beliefs.[19] To establish that a within-subject increase (decrease) of kindness inferences about helpful player As is associated with an increase (decrease) in player Bs' propensity to reward player As for being helpful, I tested whether those who withdrew rewards in treatment 1 differ in terms of the change in their altruism inferences from those who continued to reward (H) choosers in treatment 1.[20]

Among the 55 player Bs who rewarded (H) choosers in treatment 2, 21 did not reward (H) choosers in treatment 1. On average, these player Bs reported a 31.9% decline in the percentage of genuinely kind player As among (H) choosers. In contrast, player Bs who rewarded (H) choosers in both treatments reported, on average, a 16.6% decline in the percentage of genuinely kind player As among (H) choosers. Therefore, player Bs who rewarded action (H) in treatment 2 but stopped rewarding it in treatment 1 perceived a larger difference in the altruism of helpful player As than those who continued to reward action (H) in treatment 1 (31.9% vs. 16.6%; Wilcoxon rank-sum test: $z = 2.31$, $p = 0.017$).

I also compare changes in the kindness inferences of player Bs who did not reward

[19]Player Bs who are more altruistic have higher expectations of altruism, given helpful behavior. If I were to simply test whether subjects with high kindness inferences are more likely to reciprocate helpful behavior, this would confound the causal impact of kindness inferences with their baseline willingness to help in the given subgame. For example, there is a significant correlation between choosing ($3.50, $5.50) over ($4, $4) in part 1 and beliefs about the percentage of altruistic (H) choosers in treatment 1 ($p = 0.033$) and in treatment 2 ($p = 0.087$). Similarly, there is a significant correlation between choosing ($4, $4) over ($4.50, $2.50) in part 1 and beliefs about the percentage of altruistic (H) choosers in treatment 1 ($p = 0.037$) and treatment 2 ($p = 0.000$). Using a within-subject differencing approach eliminates the potential confounds arising from possible correlation of preferences and beliefs. Other researchers have taken different approaches to deal with endogeneity concerns arising from projection bias (Bellemare, Kröger and van Soest, 2008, 2011b; Bellemare, Sebald and Strobel, 2011a; Costa-Gomes, Huck and Weizsäcker, 2014).

[20]The identifying assumption is that player Bs who are more altruistic do not have lower degrees of deterioration in kindness inference across treatments. The data verify this assumption. There is no significant correlation between choosing ($3.50, $5.50) over ($4, $4) in part 1 and the degree to which beliefs about the percentage of altruistic (H) choosers decline between treatment 2 and treatment 1 ($p = 0.539$). Similarly, there is no significant correlation between choosing ($4, $4) over ($4.50, $2.50) in part 1 and the degree to which beliefs about the percentage of altruistic (H) choosers decline between treatment 2 and treatment 1 ($p = 0.138$).

(H) choosers in treatment 1. Among the 51 player Bs who did not reward (H) choosers in treatment 1, 30 of them also did not reward (H) choosers in treatment 2. On average, these player Bs reported a 12.1% decline in the percentage of genuinely kind player As among (H) choosers. Compared with the 21 player Bs who rewarded (H) choosers in treatment 2 even though they did not reward them in treatment 1, the average inference deterioration of these 26 player Bs is significantly lower (Wilcoxon rank-sum test: $z = 3.39$, $p < 0.001$).

3.2.4 Robustness checks and evidence

Treatments 1a and 2a were run with 146 subjects who did not participate in Experiment 2 to check the sensitivity of the results to the within-subject and multitask design of Experiment 2 and the choice of p and q. Treatment 1a (2a) presented part 3 of treatment 1 (treatment 2) to pairs of participants in the role of player A and player B in a between-subjects design. Subjects did not complete the tasks in parts 1, 2, and 4. By setting $p = .8$ and $q = .1$ in treatment 1a and $p = .1$ and $q = .1$ in treatment 2a, this also increased the chances that second-movers' response options elicited by the strategy method would be implemented. All other features of Game Γ_2 were kept the same as in Experiment 2.

The results replicate Experiment 2's findings. A greater chance of punishment is motivating: 96% of player As chose (H) in treatment 1a compared with 67% in treatment 2a (two-sided Pearson chi-square test: $\chi^2 = 8.97$, $p = 0.003$). In support of the main hypothesis that positive reciprocity with respect to the same helpful action is lower when the interaction strongly incentivizes that helpful action, 15% of player Bs rewarded (H) in treatment 1a compared with 46% in treatment 2a (two-sided Pearson chi-square test: $\chi^2 = 5.51$, $p = 0.019$).[21] Further details of the study are available in the online appendix.

[21]The degree of positive reciprocity is lower than that in Experiment 2, possibly because of the higher probability of being motivated by strategic considerations due to nonnegligible chances of punishment or rewards in treatment 2a.

4 Discussion

The experimental evidence in this paper shows that positive reciprocity declines based on the degree to which a helpful action is perceived to be strategically motivated. The findings suggest the need for further research into when reciprocal incentives can be useful to reach otherwise nonenforceable outcomes that benefit both parties. Clearly, rewards and punishments are inherent to many professional and personal reciprocal relationships. These incentives can help motivate socially desirable actions (Andreoni, Harbaugh, and Vesterlund, 2003) and lead to large efficiency gains (Fehr, Gächter, and Kirchsteiger, 1997). Because of their success, however, incentives can obscure the motives of people who act generously in these interactions. The results indicate that strong incentives may negate some of the social welfare gains they were designed to induce because people are less inclined toward positive reciprocity in response to helpful actions motivated by strategic considerations. This finding primes the question of optimal reciprocal incentives when perceived motives matter and underscores the need for reciprocity theories that consider the role of perceived motives.

Relatedly, the results shed light on the type of reciprocity model that can incorporate the role of perceived motives. The data patterns suggest at least two possibilities. One is to allow perceptions of what the first-mover expects to gain or lose as a result of his behavior to influence the perceived kindness of an action. Cox et al. (2008a) capture this intuition by considering the maximum payoff the first-mover could have achieved if he had acted differently. They define a first-mover's action to be more generous than another if it induces an opportunity set for the second-mover with a higher maximum payoff and if it does not increase the first-mover's own opportunity more than that of the second-mover. In turn, more generous actions induce higher positive reciprocity. To capture the evidence produced in this paper, however, their model would need to be extended (i) to define a more continuous definition of generosity based on the difference between how much choosing one action versus the other increases the second-mover's potential income minus how much it increases that of the first-mover and (ii) to consider expected, rather than maximum, payoffs in the set of

opportunities induced by the first-mover's action.

Another possibility is to model reciprocity as a response to the revealed altruism of the first-mover. Experiment 2 provides direct support for the idea that people respond to the altruism of the person performing the action rather than merely responding to the action itself (Levine, 1998). This finding suggests that reciprocity models aiming to capture the role of motives can benefit from considering how a person's altruism can be gleaned from his or her actions in a strategic context. Recent work has proposed models of reciprocal behavior in games in which (i) individuals care about others to the extent that others are altruistic and (ii) altruism is private information (Arbak and Kranich, 2005; Dur, 2009; Gül and Pesendorfer, 2016; Non, 2012). In particular, Gül and Pesendorfer's (2016) model considers inferences about the underlying altruism of the first-mover directly as a decision-making input for the second-mover. Their model predicts higher rewards for the same helpful action when the person performing the action is perceived to have a higher degree of altruism. Therefore, the model can incorporate the role of perceived motives, as documented in the experiments presented herein.

Whether perceived motives matter for reciprocity is related to a broader question that has been pivotal in recent research on reciprocity: how to evaluate kindness. This is an important question to answer across many domains that involve reciprocal considerations. In recent work, Celen, Blanco and Schotter (2014) offer a definition of kindness based on a notion of blame, similar to the notion of relative kindness of players in the Gül and Pesendorfer (2016) model. Similarly, the current set of results suggest that kindness judgments are not only about the action but also about the person himself. Future experimental research could test different notions of kindness and the relative importance of consequences, intentions, and motives on perceptions of kindness.

Finally, a consideration of perceived motives may also have implications for the so-called positive reciprocity puzzle. There is an emerging consensus that the propensity to punish harmful behavior is stronger than the propensity to reward friendly behavior

(e.g., Fehr and Gächter, 2000; Offerman, 2002; Charness and Rabin, 2002; Cox and Deck, 2005). Does the role of perceived motives also contribute to this asymmetry? In the case of intentionally hurtful actions in a reciprocal context, the motives of the first-mover are unambiguously unkind and therefore deserve retribution. However, the motives behind intentionally helpful actions in a reciprocal context can be ambiguous, as the examples in this paper demonstrate. A positive reciprocal response may not be as strong as it would have been had the helpful action been unambiguously driven by kindness. A closer assessment of the asymmetry between positive and negative reciprocity could help disentangle the potential role of perceived motives.

References

Andreoni, J., Harbaugh, W., Vesterlund, L., 2003. The carrot or the stick: Rewards, punishments, and cooperation. The American Economic Review 93 (3), 893–902.

Arbak, E., Kranich, L., 2005. Can wages signal kindness? Working Paper du GATE 2005-11.

Bellemare, C., Kröger, S., Van Soest, A., 2008. Measuring inequity aversion in a heterogeneous population using experimental decisions and subjective probabilities. Econometrica 76 (4), 815–839.

Bellemare, C., Kröger, S., van Soest, A., 2011a. Preferences, intentions, and expectation violations: A large-scale experiment with a representative subject pool. Journal of Economic Behavior & Organization 78 (3), 349–365.

Bellemare, C., Sebald, A., Strobel, M., 2011b. Measuring the willingness to pay to avoid guilt: estimation using equilibrium and stated belief models. Journal of Applied Econometrics 26 (3), 437–453.

Bellemare, C., Shearer, B., 2011. On the relevance and composition of gifts within the firm: Evidence from field experiments. International Economic Review 52 (3), 855–882.

Blount, S., 1995. When social outcomes aren't fair: The effect of causal attributions on preferences. Organizational behavior and human decision processes 63 (2), 131–144.

Bolton, G. E., Brandts, J., Ockenfels, A., 1998. Measuring motivations for the reciprocal responses observed in a simple dilemma game. Experimental Economics 1 (3), 207–219.

Bolton, G. E., Brandts, J., Ockenfels, A., 2005. Fair procedures: Evidence from games involving lotteries. The Economic Journal 115 (506), 1054–1076.

Bolton, G. E., Ockenfels, A., 2000. Erc: A theory of equity, reciprocity, and competition. American economic review, 166–193.

Brandts, J., Solà, C., 2001. Reference points and negative reciprocity in simple sequential games. Games and Economic Behavior 36 (2), 138–157.

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

Celen, B., Blanco, M., Schotter, A., 2014. On blame and reciprocity: An experimental study. New York University, mimeo.

Charness, G., 2004. Attribution and reciprocity in an experimental labor market. Journal of labor economics 22, 665–688.

Charness, G., Haruvy, E., 2002. Altruism, equity, and reciprocity in a gift-exchange experiment: an encompassing approach. Games and Economic Behavior 40 (2), 203–231.

Charness, G., Levine, D. I., 2007. Intention and stochastic outcomes: An experimental study. The Economic Journal 117 (522), 1051–1072.

Charness, G., Rabin, M., 2002. Understanding social preferences with simple tests. The Quarterly Journal of Economics 117 (3), 817–869.

Cox, J. C., 2004. How to identify trust and reciprocity. Games and economic behavior 46 (2), 260–281.

Cox, J. C., Deck, C. A., 2005. On the nature of reciprocal motives. Economic Inquiry 43 (3), 623–635.

Cox, J. C., Friedman, D., Sadiraj, V., 2008a. Revealed altruism. Econometrica 76, 31–69.

Cox, J. C., Sadiraj, K., Sadiraj, V., 2008b. Implications of trust, fear, and reciprocity for modeling economic behavior. Experimental Economics 11 (1), 1–24.

Cox, J. C., Sadiraj, V., Schmidt, U., 2015. Paradoxes and mechanisms for choice under risk. Experimental Economics 18 (2), 215–250.

Dreber, A., Fudenberg, D., Rand, D. G., 2014. Who cooperates in repeated games: The role of altruism, inequity aversion, and demographics. Journal of Economic Behavior & Organization 98, 41–55.

Dufwenberg, M., Kirchsteiger, G., 2004. A theory of sequential reciprocity. Games and economic behavior 47 (2), 268–298.

Dur, R., 2009. Gift exchange in the workplace: Money or attention? Journal of the European Economic Association 7 (2-3), 550–560.

Falk, A., Fehr, E., Fischbacher, U., 2008. Testing theories of fairness - intentions matter. Games and economic behavior 62, 287–303.

Falk, A., Fischbacher, U., 2006. A theory of reciprocity. Games and economic behavior 54, 293–315.

Fehr, E., Gächter, S., 2000. Fairness and retaliation: The economics of reciprocity. The journal of economic perspectives 14 (3), 159–181.

Fehr, E., Gächter, S., Kirchsteiger, G., 1997. Reciprocity as a contract enforcement device: experimental evidence. Econometrica 65 (4), 833–860.

Fehr, E., Schmidt, K. M., 1999. A theory of fairness, competition and cooperation. The quarterly journal of economics 114 (3), 817–868.

Fischbacher, U., 2007. z-tree: Zurich toolbox for ready-made economic experiments. Experimental economics 10, 171–178.

Gneezy, U., Güth, W., Verboven, F., 2000. Presents or investments? Journal of Economic Psychology 21 (5), 481–493.

Gül, F., Pesendorfer, W., 2016. Interdependent preference models as a theory of intentions. Journal of Economic Theory 165, 179–208.

Johnsen, A., Kvaloy, O., 2016. Does strategic kindness crowd out prosocial behavior? Journal of Economic Behavior & Organization 132, 1–11.

Kelley, H., 1973. The processes of causal attribution. American psychologist 28 (2), 107–128.

Klempt, C., 2012. Fairness, spite, and intentions. Economics letters 116 (3), 429–431.

Levine, D. K., 1998. Modeling altruism and spitefulness in experiments. Review of economic dynamics 1, 593–622.

McCabe, K. A., Rigdon, M. L., Smith, V. L., 2003. Positive reciprocity and intentions in trust games. Journal of economic behavior & organization 52, 267–275.

Nelson, W. R., 2002. Equity or intention: it is the thought that counts. Journal of economic behavior & organization 48, 423–430.

Non, A., 2012. Gift-exchange, incentives, and heterogeneous workers. Games and economic behavior 75, 319–336.

Offerman, T., 2002. Hurting hurts more than helping helps. European economic review 46 (8), 1423–1437.

Rabin, M., 1993. Incorporating fairness into game theory and economics. The American Economic Review 5, 1281–1302.

Rabin, M., 1998. Psychology and economics. Journal of economic literature 36, 11–46.

Reuben, E., Suetens, S., 2012. Revisiting strategic versus non-strategic cooperation. Experimental Economics 15, 24–43.

Ross, M., Fletcher, G., 1985. Attribution and social perception. The handbook of social psychology 2, 73–114.

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 2, 392–436.

Stanca, L., Bruni, L., Corazzini, L., 2009. Testing theories of reciprocity. Journal of economic behavior & organization 71 (2), 233–245.

Strassmair, C., 2009. Can intentions spoil the kindness of a gift? University of Munich, mimeo.

Appendix

Table 3: Behavior and Beliefs about Behavior in Modified Dictator Games in Experiment 1

Choice Question	N	Option 1 Choice	Option 1 Beliefs	
(Option 1) vs. (Option 2)		Player As	Player As	Player Bs
		(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%

* Column (1) reports the frequency and the percentage of player As who chose Option 1. Columns (2) and (3) report the average predictions of player As and Bs, respectively, about the percentage of player As who chose Option 1 in a given session.

Table 4: Behavior and Beliefs about Behavior in Modified Dictator Games in Experiment 2

Choice Question	N	Option 1 Choice		Option 1 Beliefs	
(Option 1) vs. (Option 2)		Player As	Player Bs	Player As	Player Bs
		(1)	(2)	(3)	(4)
(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	49%	41%		46%
(500, 900) vs. (400, 1200)	88	69%	73%	78%	
(500, 0) vs. (400, 300)	88	25%	20%	45%	44%
(900, 0) vs. (800, 200)	88	27%	27%		
(400, 600) vs. (300, 1100)	88	69%	72%	66%	
(500, 900) vs. (400, 600)	88	81%	79%		

* Columns (1) and (2) report the frequency and the percentage of player As and Bs who chose Option 1. Columns (3) and (4) report the average predictions of player As and Bs about the percentage of subjects who chose Option 1 in a given session.

Table 5: Observed Behavior and First-Order Beliefs in Treatment 3 of Experiment 2

Treatment	N	% choice	B FOE	% choice	A FOE	% choice
		H	of H	R | H	of R | H	P | S
Treatment 3	88	72%	67%	52%	40%	42%

Predictions of intentions-based reciprocity models

I detail the predictions of the intention-based reciprocity models proposed by Dufwenberg and Kirchsteiger (2004; DK model) and Falk and Fischbacher (2006; FF model) regarding player B's behavior in the experiments. I focus only on the equilibrium behavior of player B because the experiments are designed to rule out intentions-based reciprocity theories based on the differences in player B behaviors across treatments. The DK model predicts (at least weakly) a higher propensity to choose (r|H) in treatment 1 than in treatment 2, and the FF model predicts no difference in player B's actions after player A chooses (H) across the two treatments.

Preliminaries

Let player i's set of behavior strategies be A_i, let $B_{ij} = A_j$ be the set of possible player i's beliefs about the strategy of player j, and let $C_{jij} = B_{ij} = A_j$ be the set of possible beliefs that player j holds about the beliefs of player i about the strategy of player j. Define $A = \prod_{n \in i,j} A_n$ and let $\pi_i : A \to \mathbb{R}$ denote player i's material payoff function, which maps the strategy profile played to payoffs assigned at the end nodes. Because intentions are determined by beliefs, the reciprocity payoff depends on beliefs about

beliefs. Profile $a^* \in A$ is a sequential reciprocity equilibrium (SRE) if, for all players i, it holds that (i) $a_i^* \in argmax_{a_i \in A_i} U_i(a_i, b_{ij}, c_{iji})$, (ii) $b_{ij} = a_j^*$, and (iii) $c_{iji} = a_i^*$.

Dufwenberg and Kirchsteiger (2004)

Player i's utility in the DK model is $U_i(a_i, b_{ij}, c_{iji}) = \pi_i(a_i, b_{ij}) + r_i \cdot \kappa_{ij}(a_i, b_{ij}) \cdot \lambda_{iji}(b_{ij}, c_{iji})$, which includes the material payoffs of player i and a reciprocity payoff composed of three terms: r_i (the reciprocity parameter), which reflects the weight of the reciprocity payoff compared with the material payoff for player i and is assumed to be positive; κ_{ij}, which measures how kind player i is being to player j by choosing a_i; and λ_{iji}, which captures how kind player i thinks player j is being to player i. The kindness of player i to player j is the difference between the material payoff player i expects player j to obtain due to his action a_i and an equitable payoff for player j: $\kappa_{ij}(a_i, b_{ij}) = \pi_j(a_i, b_{ij}) - \frac{1}{2}\{max(\pi_j(a_i, b_{ij})) + min(\pi_j(a_i, b_{ij}))\}$, where the equitable payoff is defined as the midpoint between the expected minimum and maximum payoff player j could obtain as a result of actions available to player i.[22] The perception of how kind player i thinks player j is being to player i is $\lambda_{iji}(b_{ij}, c_{iji}) = \pi_i(b_{ij}, c_{iji}) - \frac{1}{2}\{max(\pi_i(b_{ij}, c_{iji})) + min(\pi_i(b_{ij}, c_{iji}))\}$.

Experiment 1

In the context of Experiment 1, I denote player B's beliefs about player A's beliefs that player B will choose r after player A chooses (H) as $c_{BAB}(r|H)$ and her second-order beliefs about choosing r after player A chooses (S) as $c_{BAB}(r|S)$. Note that there are only two choices available to each player. Therefore, the equitable payoff is defined as the midpoint between the expected material payoffs generated by each action available to a player. Given the material payoffs specified in the game, the utility player B derives from choosing r in response to H is $U_B(r|H) = \underbrace{1.5}_{\pi_B} + r_B \cdot \underbrace{\frac{1}{2}(5.5 - 3)}_{\kappa_{BA}} \cdot \underbrace{\frac{1}{2}([1.5c_{BAB}(r|H) + 2 \cdot (1 - c_{BAB}(r|H))] - [0.5c_{BAB}(r|S) + 1 \cdot (1 - c_{BAB}(r|S))])}_{\lambda_{BAB}}$. Note that the perceived kindness of (H) from the perspective of player B is positive because it results in strictly higher material payoffs for B than choosing (S) does. By the same logic, the perceived kindness of (S) is negative. Simplifying this expression and repeating the same for all utilities that player B derives from her possible choices, I obtain the following: $U_B(r|H) = 1.5 + \frac{5}{16} r_B(2 - c_{BAB}(r|H) + c_{BAB}(r|S))$, $U_B(l|H) = 2 - \frac{5}{16} r_B(2 - c_{BAB}(r|H) + c_{BAB}(r|S))$, $U_B(r|S) = 0.5 + \frac{1}{4}(2 - \frac{m}{2}) r_B(2 + c_{BAB}(r|S) -$

[22] DK also requires actions considered in $min(\pi_j(a_i, b_{ij}))$ to belong to the set of efficient strategies (defined on p. 276) to avoid pathological cases in which a dominated strategy makes everything else look kind by comparison. Both (H) and (S) are in the efficient set of actions for player A; therefore, this detail is omitted here.

$c_{BAB}(r|H)$), and $U_B(l|S) = 1 - \frac{1}{4}(2 - \frac{m}{2})r_B(2 + c_{BAB}(r|S) - c_{BAB}(r|H))$, where $m = 1.5$ in treatment 1 and $m = 6.5$ in treatment 2.

Treatment 2: $m = 6.5$

Observation 1: If player A chooses (S), choosing (l) is player B's unique equilibrium behavior.

Note that for any possible strategy of player B, player B gets less when player A chooses (S) than when he chooses (H). It follows that whatever player A believes about player B's strategy, player A's choice of (S) is unkind, and thus player B must believe that it is unkind. When $m = 6.5$, choosing (r) would reward player A ($\kappa_{BA} > 0$), and thus the reciprocity payoff is negative. Therefore, the lower material payoff, as well as the lower reciprocity payoff, makes player B choose (l).

Observation 2: If player A chooses (H), the following holds in all SRE:
(1) if $r_B > 4/5$, player B chooses r;
(2) if $r_B < 2/5$, player B chooses l;
(3) if $4/5 > r_B > 2/5$, player B chooses r with a probability of $p = 2 - \frac{4}{5r_B}$.

Proof. Note that $U_B(r|H) = U_B(l|H)$ when $r_B = 4/5(2+c_{BAB}(r|S)-c_{BAB}(r|H))$. When r_B is larger than this threshold, $U_B(r|H) > U_B(l|H)$. In equilibrium, the second-order beliefs must be correct. Therefore, when $U_B(r|H) > U_B(l|H)$, $r_B > 2/5$ because $c_{BAB}(r|S) = 1$ (by observation 1) and $c_{BAB}(r|H) = 1$ in equilibrium. Similarly, $U_B(l|H) > U_B(r|H)$ if $r_B < 4/5(2+c_{BAB}(r|S)-c_{BAB}(r|H))$. Substituting $c_{BAB}(r|S) = 1$ (by observation 1) and $c_{BAB}(r|H) = 0$, a threshold of $2/5$ is obtained. For intermediate values ($4/5 > r_B > 2/5$), neither a choice of (r) or a choice of (l) can be a part of an equilibrium. To have an equilibrium that involves random choice, it must be that $U_B(r|H) = U_B(l|H)$. Because in equilibrium second-order beliefs must be correct, the actual probability that player B choose (r) should be $2 - \frac{4}{5r_B}$. □

Treatment 1: $m = 1.5$

Observations 3 and 4 characterize equilibrium responses of player B's in treatment 1 of Experiment 1.

Observation 3: If player A chooses (S), player B's equilibrium behavior is characterized by one of the following possibilities:
(1) if $r_B > 2/5$, player B chooses r;
(2) if $r_B < 2/5$, player B chooses l.

Observation 4: If player A chooses (H), player B's equilibrium behavior is characterized by one of the following possibilities:
(1) if $r_B > 2/5$, player B chooses r;
(2) if $r_B < 2/5$, player B chooses l.

Proof. Note that $U_B(r|H) = U_B(l|H)$ when $r_B = 4/5(2+c_{BAB}(r|S)-c_{BAB}(r|H))$. When r_B is larger than this threshold, player B chooses (r|H); when it is smaller, player B

chooses (l|H), and player B randomizes when this equality is strict. When $m = 1.5$, $U_B(r|S) = U_B(l|S)$ when $r_B = 4/5(2+c_{BAB}(r|S)-c_{BAB}(r|H))$ as well. Note that the utility from an action in a given subgame depends on second-order beliefs regarding behavior in the other subgame. These beliefs need to be correct in equilibrium. Importantly, optimal behavior in each subgame changes with the reciprocity parameter, because unlike in treatment 2, neither subgame has a dominant strategy for player B. Therefore, the reciprocity parameters for which an action is optimal following (H) need to satisfy the optimality of the believed behavior following (S), and vice versa.

Consider the case in which player B chooses (l|H) and believes that $c_{BAB}(r|S) = 0$. Based on the thresholds defined above, player B finds (l) to be her best response to (H) when $r_B < 2/5$ and when $c_{BAB}(r|S) = 0$. Player B finds (l) to be her best response to (S) when $r_B < 2/5$ and when $c_{BAB}(r|H) = 0$. Consider the case in which player B chooses (r|S) and believes that $c_{BAB}(r|H) = 1$. Based on the thresholds noted above, $U_B(r|S) > U_B(l|S)$ when $r_B > 2/5$. The belief $c_{BAB}(r|H) = 1$ holds in equilibrium in this parameter range because $U_B(r|H) > U_B(l|H)$ also when $r_B > 2/5$ and $c_{BAB}(r|S) = 1$.[23]□

Summary

There are no reciprocity parameters such that positive reciprocity in response to (H) is part of an equilibrium in treatment 2 but not in treatment 1. In particular, the predictions of the DK model are the same for both treatments when reciprocity parameter is either less than 2/5 or more than 4/5. However, in the intermediate range of parameters, the model predicts that second-movers always positively reciprocate to (H) in treatment 1, but it predicts a mixed strategy in treatment 2. The reason lies in how kind player B thinks the choice of (H) is given her second-order beliefs. Player B is more likely to choose (r|S) in treatment 1 and expects player A to expect this to be the case. Note that (r|S) gives her a lower material payoff than (l|S). The more she thinks player A expected her to choose (r|S), the kinder the choice of (H) seems compared with the choice of (S).

Experiment 2

I want to compare how likely player B is to choose R in response to H across treatments 1 and 2. The DK model applies to multistage games without nature. To derive the predictions of the DK model for Experiment 2, it is necessary to make a natural modification to reflect the fact that player B evaluates the kindness of player

[23]In general, there can be multiple equilibria depending on player B's beliefs, which in equilibrium can be self-fulfilling. To find all equilibria, it is necessary to check every action and belief combination, of which there are nine in this case (Player B choosing (r) with probability one, choosing (l) with probability one, and randomizing after observing (H), crossed with beliefs about player B choosing (r) with probability one, choosing (l) with probability one, and randomizing after observing (S)). In the context of Experiment 1, these comparisons are greatly aided by the fact that the thresholds for both subgames are the same. For brevity, the details of nonequilibrium cases are omitted.

A based on the beliefs player A held at the time he made his decision, rather than based on player A's updated beliefs after nature moves. Recall that the probability of nature choosing 1 (p) or 3 (q) differs across treatments. In treatment 1, $p = .98$, $q = 0.01$, but in treatment 2, $p = q = 0.01$. Therefore, in treatment 1, player B thinks that player A expects the average material payoff consequences of choosing (S) to be $2.5 - .5c_{BAB}(P|S)$ and the consequence of choosing (H) to be $4 - .005c_{BAB}(R|H)$ for player B. In treatment 2, player B thinks that A expects average material payoff consequences of choosing (S) to be $2.5 - .005c_{BAB}(P|S)$ and the consequence of choosing (H) to be $4 - .005c_{BAB}(R|H)$ for player B. I define the r_B value for which $U_B(R|H) = U_B(N|H)$ as $r_B^*(H)$ and the r_B value for which $U_B(P|S) = U_B(N|S)$ as $r_B^*(S)$. In treatment 1, $r_B^*(H) = r_B^*(S) = {}^{2}/_{3(1.5+.5c_{BAB(P|S)}-.005c_{BAB(R|H)})}$. In treatment 2, $r_B^*(H) = r_B^*(S) = {}^{2}/_{3(1.5+.005c_{BAB(P|S)}-.005c_{BAB(R|H)}}$. Therefore, in treatment 2, player B's unique pure strategy is reciprocation if $r_B > 4/9$, and it is nonreciprocation if $r_B < 4/9$. In treatment 1, player B's unique pure strategy is reciprocation if $r_B > 4/9$, and it is nonreciprocation if $r_B < 1/3$. When $4/9 > r_B > 1/3$, both a reciprocation equilibrium and a nonreciprocation equilibrium are possible. If $c_{BAB}(R|H) = c_{BAB}(P|S) = 1$, these beliefs are fulfilled in equilibrium, and player B reciprocates. If $c_{BAB}(R|H) = c_{BAB}(P|S) = 0$, these beliefs are also fulfilled in equilibrium, and player B does not reciprocate. The derivation closely follows the derivation of equilibrium responses in Experiment 1 and is omitted for brevity. In summary, the DK model predicts a (weakly) larger set of reciprocity parameters for which player B chooses (R|H) in equilibrium in treatment 1 than in treatment 2.

Falk and Fischbacher (2006)

Player i's utility in the FF model is $U_i(a_i, b_{ij}, c_{iji}) = \pi_i(a_i, b_{ij}) + r_i \cdot \sigma_{ij}(a_i, b_{ij}) \cdot \Delta_{iji}(b_{ij}, c_{iji}) \cdot \epsilon_j$, which includes the material payoffs of player i and a reciprocity payoff composed of four terms: r_i (the reciprocity parameter), which reflects the weight of the reciprocity payoff compared with the material payoff for player i; σ_{ij}, which measures how much player i alters the payoff of player j by choosing a_i; Δ_{iji}, which only depends on i's belief about whether j intended i to receive more than j wants for himself; and ϵ_j, which reflects player j's intentionality and lies in the unit interval $[0, 1]$. Note that intentionality is kept constant across the treatments in this paper. In the context of comparing $U_B(r, c_{BAB})$ and $U_B(l, c_{BAB})$, the FF model would define $\Delta_{BAB}(b_{BA}, c_{BAB}) = \pi_B(c_{BAB}) - \pi_A(c_{BAB})$ and $\sigma_{BA} = \pi_A(a_B) - \pi_A(c_{BAB})$, where $\pi_A(a_B)$ is player A's final material payoff as a result of player B's action and $\pi_i(c_{BAB})$ is player B's beliefs about what player A expected player i to receive at the time he chose between (H) and (S). While $\pi_A(c_{BAB})$ and $\pi_B(c_{BAB})$ differ across treatments, within each treatment of either experiment, $U_B(r|H) > U_B(l|H)$ if and only if $\frac{\pi_B(l|H) - \pi_B(r|H)}{\pi_A(r|H) - \pi_A(l|H)} > r_B \cdot (\pi_B(c_{BAB}) - \pi_A(c_{BAB})) \cdot \epsilon_A$. Note that $r_B \geq 0$, $\epsilon_A \in [0, 1]$ by definition. Given that $\frac{\pi_B(l|H) - \pi_B(r|H)}{\pi_A(r|H) - \pi_A(l|H)} > 0$ and $(\pi_B(c_{BAB}) - \pi_A(c_{BAB})) < 0$ for all $c_{BAB} \in [0, 1]$ in both experiments, the above inequality is always violated. Therefore,

the FF model that predicts player B will always choose l after player A chooses H in treatments 1 and 2 (of both experiments). The reason lies in the fact that the FF model does not take into account the payoffs the first-mover could have obtained if he chose differently.