Promises and Punishment*

Martin Dufwenberg† Flora Li‡ Alec Smith§

June 26, 2026

Abstract

We study the effect of communication on beliefs, behavior, and efficiency in the context of trust games with a punishment option, which have the structure of hold-up problems. We apply a novel behavioral motivation, frustration-dependent anger, that links unmet payoff expectations with the willingness to forgo material payoffs to punish others, and we conjecture that communication works through this mechanism to raise expectations about the likelihood of belief-dependent costly punishment and to increase trust, cooperation, and efficiency. Promises drive the effect of communication on beliefs and broken promises lead to costly punishment. In an experiment we allow communication in the form of a single pre-play message. We measure beliefs and our design permits the observation of promises and deception. The results are consistent with the theory.

Keywords: Communication, Hold-up, Frustration and anger, Promises, Punishment, Psychological game theory

*We thank Pierpaolo Battigalli, Gary Charness, Uri Gneezy, two referees, the editor David Cooper, and many seminar and conference participants for their helpful comments. The Center for Peace Study and Violence Prevention at Virginia Tech provided financial support.

†Purdue University and University of Gothenburg. Email: madufwenb@purdue.edu.

‡Corresponding author. Guangdong Institute of Intelligence Science and Technology. Email: lizhuncheng@gdiist.cn.

§Department of Economics, Virginia Tech. Email: alecsmith@vt.edu.

1 Introduction

Communication can foster trust & cooperation. A large recent literature explores why people often keep their promises, focusing on the motivation of the person who issued a promise. Charness and Dufwenberg (2006) develop a theoretical argument based on "guilt aversion" and Vanberg (2008) similarly explores "a preference for keeping one's word."[1]

We explore a new and complementary explanation, whereby it is the person who received the promise who is emotionally affected. If a promise is broken this induces dashed hopes and frustration, which triggers anger and aggression.[2] When anticipated, this creates an incentive for those who issue promises not to renege. We develop this idea for environments which augment trust games with a punishment stage, e.g., as in Figure 1:

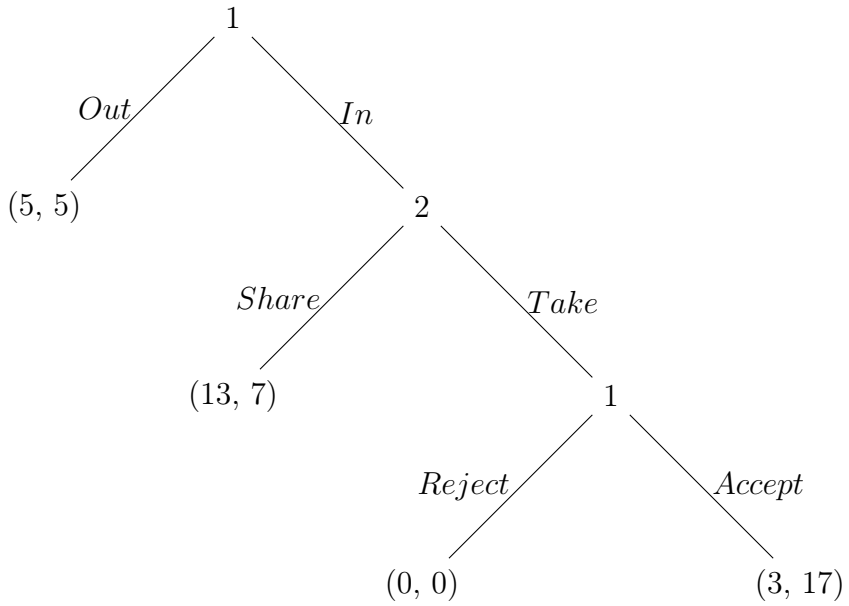


Figure 1. The game form.

We study such settings both theoretically and experimentally: First, we apply the model of frustration and anger from Battigalli, Dufwenberg, and Smith (2015; 2019) (BDS), adding pre-play communication from Player 2 to Player 1. The basic ideas are: (1) Players experience anger when frustrated. (2) If Player 2 chooses *Take*, then 1 is frustrated to the extent that his material payoff is less than expected; (3) this translates into anger and an urge for 1 to punish 2 by choosing *Reject*. (4) A promise from 2 to 1 may enhance these effects, by

[1] Much of the subsequent literature is surveyed by Cartwright (2019) (see especially Section 4.2) and Di Bartolomeo et al. (2023) (Section 2).

[2] Psychologists associate frustration with aggression; see, e.g., Dollard et al. (1939); Berkowitz (1989).

making Player 1 expect promise-keeping. (5) If all of this is anticipated, the path-of-play will be (*In*, *Share*), but 1 would have chosen *Reject* had 2 chosen *Take*.

The approach requires a formulation of utility where a player's preferences depend both on material payoffs and on beliefs about his own and others' behavior.[3] Messages become relevant to the extent that they influence expectations about payoffs, thus linking communication, beliefs, and the willingness to forgo material payoffs to punish others.

Second, we design an experiment to test the predictions of the theory.[4] We allow pre-play communication as a treatment in order to study whether promises are sent and whether their effect on beliefs and behavior is as we predicted. A key contribution of our paper is that, in addition to recording messages and behavior, we elicit the beliefs of both players before messages are sent and after they are received. We measure beliefs about co-player choices, and also, in a novel contribution, about players' own behavior at subsequent stages of the game. These measures allow us to carefully examine the relationship between communication, beliefs, and behavior. In particular, we measure how promises change beliefs and how expectations about behavior influence the decision to engage in costly punishment.

The game forms that we explore may be viewed as particular forms of hold-up problems, where relationship-specific investments and incomplete contracts expose one party to opportunistic renegotiation, potentially resulting in underinvestment.[5] Ellingsen and Johannesson (2004a,b), who also study communication and hold-up in an experiment, are important precursors to our study. However, since they did not conduct their exercise with BDS' theory in mind, they did not measure the beliefs which are central to our tests.[6] Less closely related are several experimental studies of hold-up games that do not focus on the impact of communication. See Yang (2021) for a recent review.

Section 2 presents theory. We describe the games we study, apply BDS' model of belief-

[3]The approach involves belief-dependent utilities and draws on the framework of psychological game theory (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009, 2022).

[4]BDS (2019, pp. 17, 29, 31) discuss previous attempts by economists to address frustration and anger, either theoretically or experimentally. Two of the experimental studies – Persson (2018) and Aina et al. (2020) – relate directly to BDS, although unlike us these authors do not explore issues of communication.

[5]See Williamson (1971); Klein et al. (1978); Grout (1984); Grossman and Hart (1986); Tirole (1986); North and Weingast (1989); and Hart and Moore (1990); compare, e.g., Ellingsen and Johannesson (2004a), Ellingsen and Johannesson (2004b), Che and Sákovics (2008), and Dufwenberg et al. (2013) who explain how the setup we consider involves the sub-class of hold-up problems with a punishment option.

[6]Ellingsen and Johannesson suggest that their data is consistent with Fehr and Schmidt's (1999) model of inequality aversion combined with a preference for consistency, and that communication serves to change beliefs about co-player types. This interpretation is quite different from the theory that we test. Later on, we address how models of inequity aversion relate to our data.

dependent anger, and discuss the extension needed to incorporate the ideas we have regarding the effect of promises on trust, credibility, and costly punishment. Section 3 presents details of the experimental design and implementation, and states hypotheses to be tested. Section 4 reports summary statistics, main results as regards hypotheses, and additional observations. Other motivations (e.g., distributional preferences, guilt aversion, or reciprocity) may also have impact on beliefs and behavior in the settings we explore. In Section 5 we discuss and distinguish their effects from those of frustration and anger. Section 6 concludes.

2 Theory

2.1 A hold-up game with costly punishment

We study a class of 2-player, 3-stage game forms, as shown in Figure 2, where the numbers and variables at end nodes represent monetary payoffs. The game may be interpreted as a mini-trust game with an added (subsequent) punishment option, a mini-ultimatum game with an added (preceding) entry decision, or as a hold-up game where sellers can destroy the proceeds of a relationship-specific investment.[7] In the first stage, Player 1 can choose *In* to make an investment of her entire endowment of \$5, or *Out* to not invest and walk away with her initial endowment. If Player 1 invests, the endowments of both players double, and Player 2 can then propose how to divide the proceeds. To make the problem simple, Player 2 can propose one of two possible splits. One option is to choose *Share*, which is monetarily favorable (or at least as good as the other option) for Player 1. The other is to choose *Take*, which is (potentially) monetarily favorable for Player 2. If Player 2 *Take*s, Player 1 can then *Reject*, in which case both players receive 0, or *Accept* to settle with a less favorable offer in the third stage. The parameters a and b reflect the payoffs to Player 1 after, respectively, $(In, Share)$ and $(In, Take, Accept)$. We impose the following parameter restrictions: $20 > a \geq 5 \geq b > 0$, and $a \neq b$. When players care only for monetary payoffs and $b < 5$, the unique subgame perfect equilibrium (SPE) is $((Out, Accept); Take)$, which is inefficient. When $b = 5$ and players care only for monetary payoffs, there is an additional SPE $((In, Accept); Take)$.

[7] In general, hold-up may occur in environments with or without the opportunity for punishment or "vengeance" (Dufwenberg et al., 2013). In order to study the effect of broken promises we focus on a hold-up environment that allows for costly punishment after opportunistic behavior.

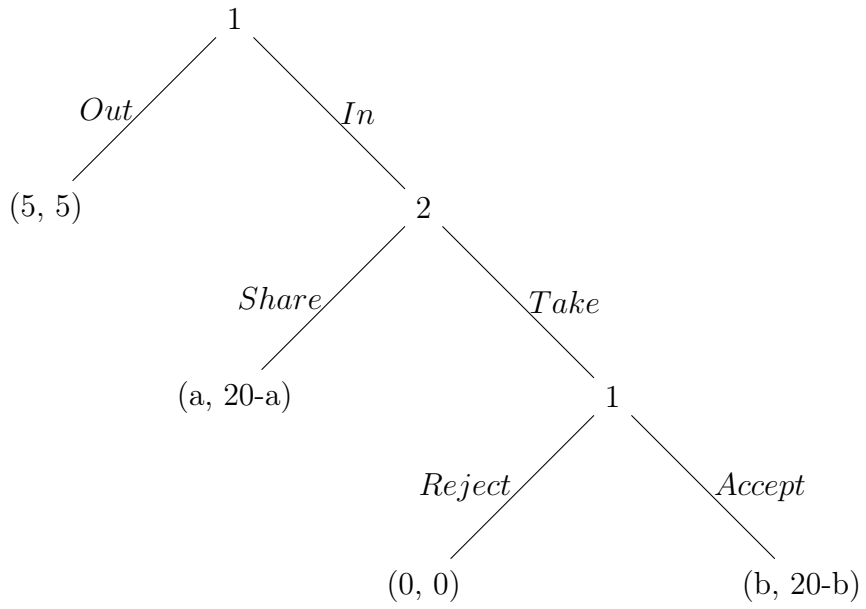


Figure 2. A hold-up game with punishment.

2.2 Frustration and anger

We apply BDS's frustration and anger model.[8] In this model, anger is motivated by frustration, and the tendency to hurt others is proportional to frustration, following the frustration-aggression hypothesis from psychology (Dollard et al., 1939; Berkowitz, 1989). In general, one feels frustrated if one's initial expectation is not met. Frustration is modeled as the gap (if positive) between one's initial expected payoff and the current best possible outcome. At any history h, Player i's frustration is

$$F_i(h;\alpha_i) = \max\left\{\bar{\pi}_i(h_0) - \max_{a_i \in A_i(h)} \mathbb{E}[\pi_i|h;\alpha_i], 0\right\}, \tag{1}$$

where $\bar{\pi}_i(h_0) = \mathbb{E}[\pi_i|h_0;\alpha_i]$ denotes Player i's expected payoff at the initial history h_0 given his first-order belief α_i about Player j's behavior, $a_i \in A_i(h)$ denotes Player i's action choice at the history h, so $\max_{a_i \in A_i(h)} \mathbb{E}[\pi_i|h;\alpha_i]$ gives the maximum possible expected payoff available to Player i at the history h.

[8]BDS model three versions of belief-dependent frustration and anger: 1) Simple anger (SA), 2) Anger from blaming behavior (ABB), and 3) Anger from blaming intentions (ABI). In the hold-up environment studied here, the predictions of all three models coincide (although the math below reflects the SA-formulation). BDS (2019) focus on two-stage "leader-follower" games; an earlier working paper BDS (2015, Section 6) develops the extension to general multi-stage games.

Player i's (expected) utility from action a_i at history h is

$$u_i(h, a_i; \alpha_i) = \mathbb{E}[\pi_i|(h, a_i); \alpha_i] - \theta_i F_i(h; \alpha_i)\mathbb{E}[\pi_j|(h, a_i); \alpha_i], \tag{2}$$

where $\theta_i \geq 0$ is Player i's sensitivity to anger. A frustrated individual is motivated to hurt the other player, if the cost is low enough. Frustration increases the negative weight placed on Player j's material payoff, and motivates aggression.

In the game forms defined in Figure 2, Player 1 is the party who might get frustrated, so we apply Equations (1) and (2) with $i = 1, j = 2$. Let the probability that Player 1 assigns to choosing *Out* be $p_1 = \alpha_1(Out|h^0) \in [0, 1]$. Let $q_1 \in [0, 1]$ denote the probability that Player 1 assigns to Player 2 choosing *Share* if stage 2 is realized, i.e., $q_1 = \alpha_1(Share|In)$ and let $r_1 = \alpha_1(Reject|In, Take) \in [0, 1]$ denote the probability that Player 1 assigns to choosing *Reject* conditional on the 3rd stage being reached. We can also define analogously a similar belief system (p_2, q_2, r_2) for Player 2. We further assume that beliefs are coherent, so the marginals of the higher order beliefs are equal to the lower order beliefs.

Next, we derive equilibrium predictions, applying the sequential equilibrium (SE) concept from Battigalli et al. (2019).[9] With utility as defined in Equation 2 there are two pure-strategy SEs of this game: an efficient one, where Player 1 chooses *In*, Player 2 *Shares*, and Player 1 *Accepts*; and an inefficient one, which coincides with the subgame perfect equilibrium for material-payoff maximizing players.

In any SE, beliefs must be correct in the sense that they are consistent with behavior. This means that the first order belief of Player i about what Player j will do matches Player j's behavior strategy. In addition, in equilibrium the belief systems of both players coincide, so for expedience, in the rest of this subsection we drop the subscripts and generically refer to beliefs p, q, and r. We focus here on SE's involving pure strategies when $b < 5$.

If Player 1's sensitivity to anger θ_1 is small, then the unique SE coincides with the SPE for players who only care for material payoffs: $((Out, Accept); Take)$, with beliefs $p = 1, q = 0, r = 0$ for both players. Player 1's initial expected material payoff is $5p + a(1-p)q + b(1-p)(1-q)(1-r) = 5$. With these beliefs, frustration equals $5 - b$ after *Take*. In that case Player 1 compares 0 (the payoff from *Reject*) to $b - \theta_1(5-b)(20-b)$ (the payoff from *Accept*), and chooses *Accept* if $\theta_1 < \frac{b}{(5-b)(20-b)}$. We refer to this SE as the "inefficient equilibrium."

[9] The SE concept was extended to psychological games by Battigalli and Dufwenberg (2009). Battigalli et al. (2019) focus on leader-follower games. The game form in the present study is not a leader-follower game, but the definitions in that paper extend naturally. For a full development, see Battigalli et al. (2015, Section 6).

If θ_1 is sufficiently large, this (psychological) game has a unique SE involving the strategy profile $((In, Reject); Share)$ where Player 1 chooses *In*, Player 2 chooses *Share*, and if Player 2 instead chooses *Take* then Player 1 chooses *Reject*. For $((In, Reject); Share)$ to be an SE, the correct belief system is $p = 0, q = 1, r = 1$ for both players. Player 1's initial expected material payoff is $5p + a(1-p)q + b(1-p)(1-q)(1-r) = a$, and at the history $(In, Take)$ Player 1's frustration equals $a - b$. If he gets the move after *Take*, Player 1 then compares the payoff of 0 from choosing *Reject* to the payoff $u_1 = b - \theta_1(a-b)(20-b)$ from *Accept*. Given equilibrium beliefs, Player 1 will *Reject* if $\theta_1 > \frac{b}{(a-b)(20-b)}$, demonstrating the uniqueness of the equilibrium for large θ_1. We refer to this as the "efficient equilibrium."

Notice that the threshold value of θ_1 that sustains the efficient equilibrium is (i) increasing in b, the cost of punishment; (ii) decreasing in $a - b$, the "*Take* amount"; and (iii) decreasing in $20 - b$, Player 2's payoff from *Accept*.

For intermediate values of θ_1, both the inefficient and efficient equilibrium exist. To see why, recall that Player 1's frustration following *Take* is $5 - b$ in the inefficient SE and $a - b$ in the efficient SE. Since $a \geq 5$, then $5 - b \leq a - b$ (with strict inequality if $a > 5$). Hence, the lowest value of θ_1 that makes it a best response for Player 1 to *Reject* in an efficient SE is at least as low as the highest value of θ_1 that makes it a best response for Player 1 to *Accept* in an inefficient SE.

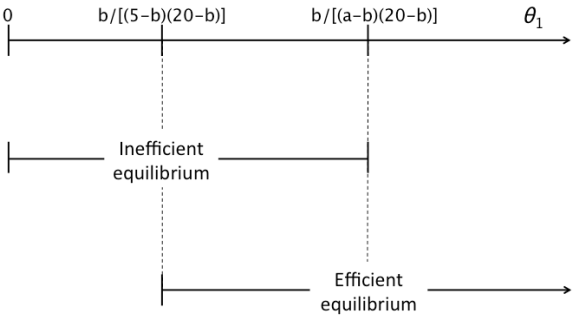


Figure 3. Sequential Equilibria, as a Function of the Anger Sensitivity θ_1 of Player 1, when $b < 5$.

When $b = 5$, there is an additional SE $((In, Accept); Take)$, with appropriate beliefs. Player 1 initially expects material payoff 5. At the history $(In, Take)$, Player 1 is not frustrated, since 5 is still available. Knowing this, Player 2 chooses *Take* after *In*, and at the

root, Player 1 is indifferent between *Out* and *In*, and in this case, Player 1 selects *In*.

Example. *Consider the game form in Figure 1, which in our parametrized setting corresponds to* $a = 13, b = 3$.

Consider first the inefficient SE $((Out, Accept); Take)$*, with beliefs* $p = 1, q = 0, r = 0$ *for both players. Player 1 initially expects the material payoff of 5, from* Out*. We must check that the off-path play is sequentially rational. After* $Take$ *Player 1's payoff from* Accept *is* $3 - \theta_1 \cdot F(\cdot) \cdot 17$*, and Player 1's payoff from* Reject *is 0. Player 1's frustration after* Take *is* $F(\cdot) = 5 - 3 = 2$*. Therefore, Player 1's equilibrium action choice of* Accept *is sequentially rational if* $3 - \theta_1 \cdot 2 \cdot 17 \geq 0$*. The expression holds if* $\theta_1 \leq \frac{3}{34}$.

Next consider the efficient SE $((In, Reject); Share)$*, with beliefs* $p = 0, q = 1, r = 1$*. Player 1 initially expects the material payoff* $a = 13$*. After* Take*, Player 1 is frustrated in the amount* $F(\cdot) = a - b = 13 - 3 = 10$*. We must check that* Reject *is sequentially rational. Given the assumed beliefs, Player 1's payoff from* Accept *is* $3 - \theta_1 \cdot 10 \cdot 17$*, and the payoff from* Reject *is again 0. Player 1 then prefers to* Reject *if* $0 \geq 3 - \theta_1 \cdot 10 \cdot 17$*, or* $\theta_1 \geq \frac{3}{170}$.

When $\theta_1 \in \left[\frac{3}{170}, \frac{3}{34}\right]$*, corresponding to the middle region in Figure 3, both the inefficient and efficient SE's are plausible, given appropriate supporting beliefs.*

The anger sensitivity θ_2 of Player 2 plays no role in the analysis. If Player 2 gets the move, then her maximal payoff is still available and according to the model, she cannot be frustrated ($F_2(\cdot) = 0$); her behavior is indistinguishable from material payoff maximization. Therefore we (mainly) focus our analysis on the beliefs and behavior of Player 1.

To summarize: For small values of θ_1, the unique SE coincides with the inefficient subgame perfect equilibrium for material-payoff maximizers. For large values of θ_1, the unique SE is efficient: Player 2 *Shares* to avoid being punished, and so Player 1 chooses *In*. For intermediate values of θ_1 there exist two SE in pure strategies, the inefficient one and the efficient one. Figure 3 summarizes how these SE map to the anger sensitivity of Player 1.

2.3 Communication and Promises

If the players are selfish (in particular, if $\theta_1 = 0$, since as noted θ_2 is not relevant) and $b < 5$, the game has a unique backward induction solution: $((Out, Accept); Take)$. The logic of that prediction is not affected by whether or not there is an opportunity for pre-play

communication and promises.

By contrast, if θ_1 is large enough, promises may plausibly affect behavior and beliefs. We predict that a promise-to-*Share* will (i) increase the likelihood of choices *In*, *Share*, and *Reject*, and (ii) that p_i will decrease while q_i and r_i will increase, for $i = 1, 2$.

A special case is an instance of equilibrium selection, if $\frac{b}{(5-b)(20-b)} < \theta_1 < \frac{b}{(a-b)(20-b)}$ (compare Figure 3). Recall that there are two SE, $((Out, Accept); Take)$ and $((In, Reject); Share)$. Assume that absent communication players play $((Out, Accept); Take)$. Assume that following a promise-to-*Share* players play $((In, Reject); Share)$. In this case a promise-to-*Share* allows the players to coordinate on the Pareto-efficient equilibrium.[10]

Predictions (i) and (ii) are not limited to equilibria though. Rather, the idea is that messages feed self-fulfilling chains of beliefs that better choices will be made more frequently. In particular, suppose Player 2 issues a promise-to-*Share*:

- If Player 1 attaches credibility to Player 2's promise, then q_1 rises.
- With a higher q_1, Player 1's frustration following $(In, Take)$ increases (see (1)).
- Player 1's increased frustration makes *Reject* a better choice for Player 1, suggesting an increased frequency of *Reject* as well as higher r_i's.
- The increased r_2 makes *Share* a better choice for Player 2, suggesting an increased frequency of *Share* as well as higher q_2.
- When q_1 and r_1 increase, as long as the increase in q_1 is large enough, this makes *In* a better choice for Player 1, suggesting an increase in p_i.

To sum it up, under the assumption that certain messages influence beliefs, the belief-dependent frustration and anger model implies that, with promises, Player 1 is more likely to trust Player 2 (choose *In*), Player 2 is more likely to keep her promises (to *Share*), and Player 1 is more likely to punish broken promises (by *Reject*ing).

Are our BDS-based predictions unique or do they also arise under other motivations such as inequity aversion, guilt aversion, reciprocity, or a preference to honor promises? We defer a full discussion, including proper introductions to such other motivations, to

[10]See Crawford (2016) for a broad discussion of how there is some empirical support for communication to have such efficiency-enhancing effects, as well as a nuanced critical discussion of the difficulty in establishing clear game-theoretic underpinnings.

Section 5. We nonetheless flag three points now that will be central there. First, standard consequentialist models such as selfishness and inequity aversion (e.g., Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), unlike our theory, do not predict that *Reject* varies with ex-ante beliefs about whether the co-player will *Share*: a purely selfish player never engages in costly rejection, and an inequity-averse player's decision to reject depends only on the final payoff allocation, not on beliefs. Second, neither guilt aversion (e.g., Charness and Dufwenberg, 2006; Battigalli and Dufwenberg, 2007) nor a preference to honor promises (Vanberg, 2008) can explain costly punishment by the promise-receiver, as both place the behavioral force on the promise-maker. A preference to honor promises is, moreover, belief-independent by construction, whereas our prediction turns on belief-dependent frustration. Third, while sequential reciprocity (Dufwenberg and Kirchsteiger, 2004) predicts a positive relationship between *Reject* and the *Take* amount, it does not predict a direct relationship between ex-ante beliefs about *Share* and *Reject*. Our predictions that the plan to *Reject* is driven positively by both the *Take* amount and the belief about *Share* therefore isolate the promise-receiver implication of our theoretical model.

3 Experiment

To study the effect of promises on trust and punishment we implemented a laboratory experiment with the class of game forms depicted in Figure 2. We employed a within-subject design where subjects played variations of the game over multiple rounds, with fixed roles, paired with anonymous partners with random rematching each round. Each session included a communication and a no-communication block, with the order counterbalanced across sessions.

3.1 Procedures

The experiment was programmed using z-Tree (Fischbacher, 2007) and conducted at the Virginia Tech Economics Laboratory. A sample of the experiment instructions is reproduced in the Appendix. We conducted a total of 11 sessions, with 200 total participants.[11] Each session included 14-20 participants with an average of 18.4 per session. Sessions took about 1.75 hours to complete.

[11] We dropped the data from one additional session that was interrupted by a software malfunction.

At the beginning of each session, participants were randomly assigned to the role of either Player 1 or Player 2, which remained fixed throughout the experiment. Before each round, participants were randomly and anonymously matched with a partner of the opposite role (i.e., we used stranger matching). After the experiment, participants were paid according to the outcome of one randomly selected round. Excluding the show-up fee, participants earned an average of $12.24.[12]

Each session consisted of 20 rounds separated into two blocks: 10 rounds of communication, and 10 rounds where no communication was allowed. After each round both players were informed of the outcome. We counterbalanced the order of the communication block across sessions, so that in 5 of the 11 sessions the first 10 rounds involved pre-play messages from Player 2 to Player 1, and the no-message block followed; the other 6 sessions experienced the no-message block first. The only restrictions on message sending were that the message had to be less than 140 characters long, and to retain confidentiality, individuals were not allowed to reveal their identity in the message.

In each block, participants played 10 different variations on the game in Figure 2, in random order. The game variations are shown in Table 1, where all the numbers are in US dollars. A change of the parameter b (Cost of *Reject*) indicates changing the cost of *Reject*, and we vary the cost of *Reject* (b) from 1 to 5. The difference $a - b$ indicates the "*Take* amount", which takes one of two values in our design: either $a - b = 4$ to indicate a low *Take* amount, or $a - b = 10$ to indicate a high *Take* amount. The payoff splits in Stage 2 and Stage 3 are asymmetric, such that $a \neq 10$, to reduce the salience of an equal split.

Table 1. Experiment Design – Game Structure.

Game	Value of *Share* (a)	Cost of *Reject* (b)	*Take* Amount ($a - b$)
LT1	5	1	4
LT2	6	2	4
LT3	7	3	4
LT4	8	4	4
LT5	9	5	4
HT1	11	1	10
HT2	12	2	10
HT3	13	3	10
HT4	14	4	10
HT5	15	5	10

It is common in experimental designs to make use of the "strategy method", where players

[12]After Session 4, we increased the show-up fee from $5 to $10 to improve turnout.

make conditional choices before the game is played. However, the BDS model implies that players experience zero frustration in this set-up, as frustration can only arise in the course of play. Consistent with our motivation, Aina et al. (2020) demonstrate that frustration and anger are more relevant with the direct response method than with the strategy method, and that costly punishment exhibits greater belief-dependence in sequential decisions. In addition, Brandts and Charness (2011) show that costly punishment is more frequently observed with the direct response method than with the strategy method. Accordingly, we used a direct response design, such that players move sequentially.

3.2 Belief Elicitation

During the experiment we elicited each participant's probabilistic beliefs about their co-player's actions, conditional upon future play. We also asked participants to report the probability with which they expect to take an action conditional upon future play in the game. In each round, we measured the first-order beliefs that participants held about their own (in the case of first movers) and their co-players' behavior. We elicited Player 1's beliefs regarding the likelihood of choosing *Out* (p_1), Player 1's conditional first order beliefs of Player 2's probability of choosing *Share* (q_1), and Player 1's beliefs regarding the likelihood of choosing *Reject* (r_1) conditional on entering the 3rd stage. We interpret players' beliefs about their own choices as revealing their *plans*. Our data on plans allow us to gauge players' would-have-been behavior at nodes that are not actually reached when the game is played out (e.g., if 1 chose *Out*), even though we did not use the strategy method.

To examine how messages influence beliefs, in the communication treatment we measure Player 1's beliefs both before and after messages are received. If Player 1 chose *In*, we elicited Player 2's second order belief about *Share* (q_2) and first order belief about the conditional probability that Player 1 will choose *Reject* (r_2) after Player 2 made a decision on the 2nd stage.

When combining the direct response method with belief elicitation, as we do, there is a potential conflict between incentives for behavior and for reporting truthful beliefs (e.g., Rutström and Wilcox, 2009; Blanco et al., 2010).[13] In sequential play designs, incentivizing truthful belief reports by, e.g., a scoring method can create a spillover effect where players have incentives to continue the game in order to receive payment for a reported belief in a future stage, or to choose actions that are consistent with a previously reported belief.

[13]See also Schotter and Trevino (2014) for a review of the methodology of eliciting beliefs.

The problem is exacerbated when eliciting beliefs about a player's own future behavior.[14] Trautmann and Kuilen (2015) find that flat fee incentives perform almost as well as more complicated methods for eliciting beliefs such as proper scoring rules. We therefore eschew the use of a scoring rule for payment, instead incentivizing belief reports with a flat fee payment of \$5. In addition, we asked participants to pledge to answer these questions "to the best of my knowledge," in an attempt to trigger a desire for honest response (see the instructions in the Appendix).

3.3 Hypotheses

Our hypotheses are based upon two main assumptions. First, participants in the experiment are motivated by belief-dependent anger. Second, messages are informative.

With regard to belief-dependent anger, the model implies that unmet expectations regarding material payoffs will increase the disutility from a co-player's payoff. Hypothesis 1 is motivated by the theoretical assumption that diminished payoff expectations make aggression and costly punishment more attractive (Section 2.2).

Hypothesis 1. *Reject choice frequencies and plans to Reject (r_1) are increasing in Player 1's belief about the probability of Share (q_1).*

Our model predicts that beliefs and payoffs interact to influence Player 1's behavior (see Section 2.2). The efficient SE $((In, Reject); Share)$ is unique when the anger sensitivity parameter is sufficiently large, such that $\theta_1 > \frac{b}{(a-b)(20-b)}$. This expression shows that when the *Take* amount $(a-b)$ is high, the minimum value of θ_1 that supports the choice of *Reject* in the efficient equilibrium is lower. Similarly, as the cost of *Reject* (b) increases, higher values of θ_1 are necessary to support the choice to *Reject*. These comparative statics motivate the following:

Hypothesis 2. *(i) Player 1s are less likely to Reject (and plan to Reject, r_1) as the cost of Reject (b) increases; (ii) Player 1s are more likely to Reject (and plan to Reject) when the Take amount (a-b) is high.*

We next turn to the effect of communication. With reference to the arguments made in Section 2.3, we expect that communication will increase the frequency of cooperative outcomes and lead to greater total material payoffs (which we refer to as efficiency).

[14] See also the discussion of incentivizing own beliefs in Toussaert (2018), who addresses this issue by eliciting beliefs about a "similar other." Because we are interested in *own* beliefs as the relevant variable for anger and costly punishment, we do not employ methods that involve proxies such as similar others.

Hypothesis 3. *Communication increases the frequency of the cooperative outcome* $(In, Share)$ *and improves efficiency.*

Motivated by the theoretical description in Section 2.3 and the idea that promises may change beliefs (analogous to Charness and Dufwenberg, 2006), Hypothesis 4 connects communication and costly punishment through beliefs:

Hypothesis 4. *Communication influences beliefs via promises, such that promises shift Player 1's reported beliefs and plans in the direction of increased likelihood of In* (p_1)*, Share* (q_1)*, and Reject* (r_1)*; non-promises have no impact on beliefs.*

If promises are believed and then broken, the higher initial expectation of cooperation generates greater frustration and leads to a higher likelihood of rejection in the 3rd stage:

Hypothesis 5. *Broken promises lead to a higher Reject rate, and promises lead to a higher cooperation (Share) rate relative to non-promises.*

4 Results

We begin our examination of the results by reporting summary statistics on behavior in Section 4.1, and then present our main results following the order of our preconceived hypotheses (Section 4.2). We report additional observations in Section 4.3.

4.1 Data

Our dataset includes choices, elicited beliefs and plans, and messages (when communication was available) from 11 experiment sessions involving 200 total participants, each of whom made choices in 20 rounds of game play. Figure 4 shows aggregate results for each stage, pooling participants and games. Participants chose *In* 1,542/2,000 times (77.1%). This is clearly much greater than the upper bound of 20% implied by the SPE for selfish players.[15] Of games that advanced to the 2nd stage, participants chose *Share* 1,068/1,542 times (69.3%). In the third stage, participants selected *Reject* 179/474 times (37%). In both stages 2 and 3, the SPE prediction for selfish players is unique for all 10 games, involving *Take* and *Accept*. Thus, the majority of our observations involve departures from purely selfish behavior.

[15] This upper bound of 20% is due to Games LT5 and HT5, where Player 1s give up a payoff of 5 to select *Reject*, the same amount that could have been earned by selecting *Out*.

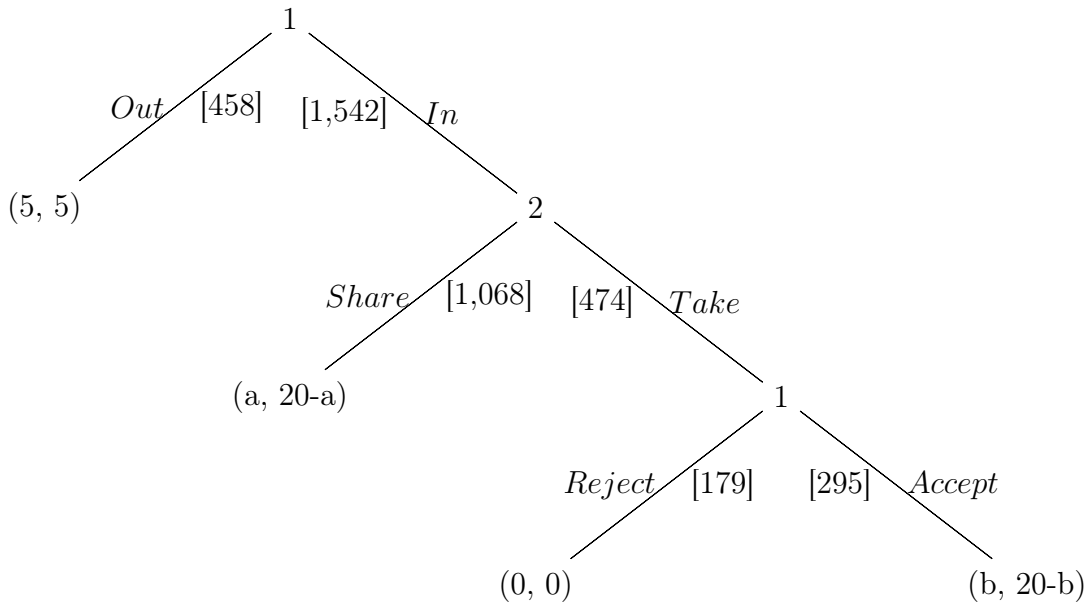


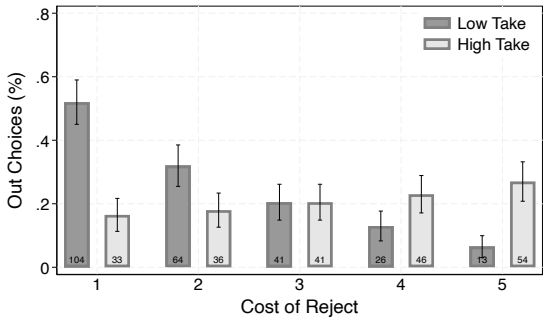
Figure 4. Summary of experimental results

Behavior. Figure 5 shows the relative frequencies of *Out*, *Share*, and *Reject* choices, arranged by the cost of *Reject*(b) and the *Take* amount.

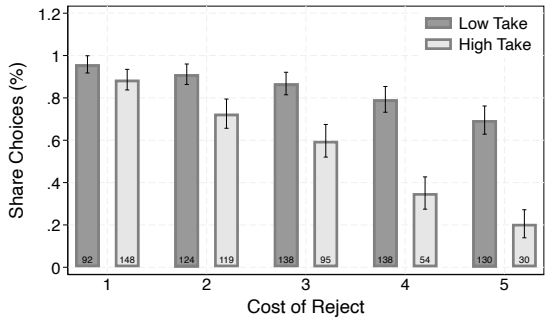
In Low Take games (LT1-LT5), the relative frequency of *Out* choices declines from 52% (when the cost of *Reject* (b) is 1) to 6.5% (when the cost of *Reject* is 5). In High Take games (HT1-HT5), the relative frequency of *Out* choices is 16.5% when $b = 1$, rising to 27% when $b = 5$.

In all games we observe a decreasing tendency for Player 2s to select *Share* as b increases. Share rates are close to 1 when $b = 1$. When $b = 5$, *Share* is chosen nearly 70% of the time in the Low Take game (LT5) but about 20% of the time in the High Take game (HT5). These differences reflect the differing incentives for Player 2 in the two types of games: the payoff to Player 2 from selecting *Take* after *In* is 11 in game LT5 but only 5 in the High Take game.

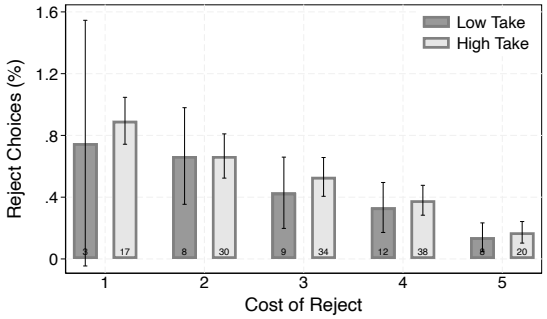
We also observe a decreasing tendency for Player 1s to select *Reject* as b increases. When $b = 1$, the relative frequency of *Reject* choices is 75% for the Low Take game LT1 and 89% for the High Take game HT1. However, our standard errors are especially large in the LT1 game, where in only 4/200 instances did the game reach the third stage. The proportion of *Reject* choices was 14% in game LT5 and 17% in game HT5. In general, there appears to be a pattern of greater proportions of *Reject* choices in the High Take games, though again standard errors are large.



(a) Stage 1: *Out* choices by Player 1.



(b) Stage 2: *Share* choices by Player 2.

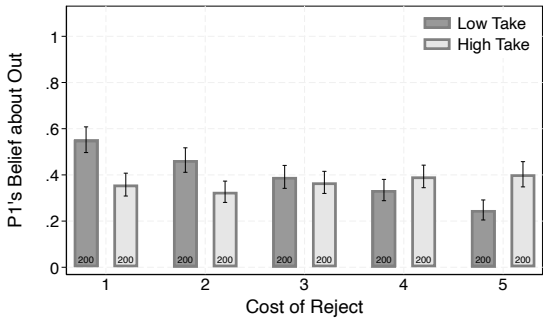


(c) Stage 3: *Reject* choices by Player 1.

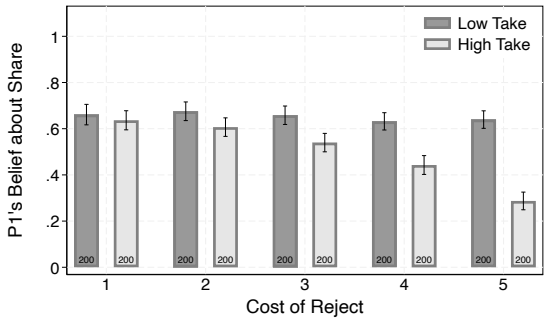
Figure 5. Relative frequency of player actions, by the cost of *Reject* and *Take* amount.

Beliefs and Plans. Figure 6 shows means of Player 1's self-reported plans and beliefs by game. In general, the patterns across two *Take* amounts and costs of *Reject* are similar to the observed choices. The data here is suggestive of bias: the mean self-reported *Out* plan is typically greater than the empirical probability of *Out* choices and the mean *Share* belief is mostly below the empirical probability of *Share* choices. In addition, the mean reported likelihood of choosing *Reject* is lower than the empirical probability of *Reject* in all but the games where $b = 5$. Next, we investigate the relation between beliefs and behavior more thoroughly.

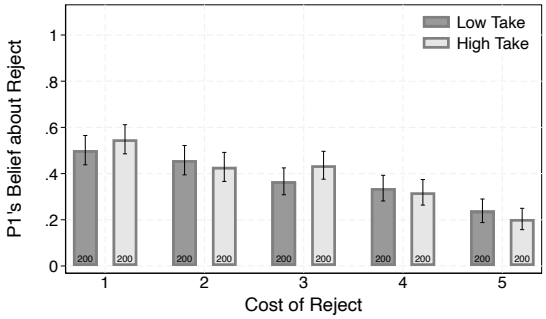
Relating behavior, beliefs, and plans. Probability models are well-calibrated when probabilities match observed relative frequencies. For example, a weather forecasting model is well calibrated if, when the forecast indicates an 80% chance of rain, it rains 80% of the time. In addition, probability models should also have good resolution, meaning that when the observed outcome is very likely (unlikely), the predictor variable is close to 1 (0). To measure the calibration and resolution of a prediction, one can estimate a simple linear probability model that regresses the observed outcome on the elicited probabilities. The fitted model will have slope 1 and intercept 0 if it is perfectly calibrated; the higher the R^2 of the regression, the greater the resolution of the prediction.



(a) Player 1's *Out* plans.



(b) Player 1's *Share* beliefs.



(c) Player 1's *Reject* plans.

Figure 6. Player 1's beliefs and plans, by the cost of *Reject* and *Take* amount.

In the machine learning literature, it is commonplace to evaluate classifier performance by plotting Receiver Operating Characteristic (ROC) curves. These plot the true positive rate (sensitivity) of a predictor versus the false positive rate (1 minus the specificity) as a function of the threshold that determines the prediction. One can then compute the Area Under the Curve (AUC) summary statistic, with values closer to 1 indicating better classification.[16]

In the Appendix, we report the results from both types of analyses, for each of Player 1's 3 self-reported probabilities of *Out*, *Share*, and *Reject* (Supplementary Figures 1, 2, and 3). Elicited plans are strong predictors of behavior. For *Out* and *Reject* plans, the linear probability models yield slopes of 0.77 and 0.74, and intercepts of -0.07 and 0.13 (Supplementary Figures 1(a) and 3(a)). Though these estimates exhibit moderate bias, the high R^2 values of 0.45 and 0.38 confirm excellent resolution.

Furthermore, the ROC curve analysis yields AUC values of 0.9167 (for *Out*, Supplementary Figure 1(b)) and 0.8370 (for *Reject*, Supplementary Figure 3(b)) representing "outstanding" and "excellent" discrimination, respectively.[17] Elicited beliefs about co-player behavior

[16]See, e.g., Murphy (2012, Chapter 5).

[17]Hosmer Jr et al. (2013, Section 5.2.4) suggest that AUC values between 0.8 and 0.9 indicate "excellent"

are less accurate predictors, but remain informative: a simple regression of *Share* choices by Player 2 on Player 1's beliefs has slope of 0.52, intercept 0.36, and R^2 of 0.10, with an AUC statistic of 0.6828 (Supplementary Figures 2(a) and 2(b)).

Taken together, these calibration and ROC metrics demonstrate that despite the use of a flat-fee incentive for belief elicitation, participants' reported beliefs and plans carry substantial informational content and maintain high predictive validity. Furthermore, because we use these elicited beliefs as regressors in our subsequent hypothesis testing, it is important to note the econometric implications of measurement error. Classical measurement error in a predictor leads to attenuation bias, pushing coefficient estimates toward zero. Therefore, any noise introduced by our elicitation method would theoretically bias subsequent hypothesis tests toward the null, rendering any detected effects as conservative lower bounds of the true relationships.

4.2 Main Results

In this section we revisit our preconceived hypotheses. We start by evaluating the relationship between Player 1's self-reported beliefs, plans, and choices (Hypothesis 1), followed by an examination of the relationship between game structure and behavior (Hypothesis 2). Then, we test for the effect of communication on behavior and payoffs (Hypothesis 3). Finally, we investigate how message content in the form of promises influences beliefs and behavior (Hypothesis 4 and 5).

4.2.1 Testing H1: The effect of beliefs about *Share* on *Reject* choices and plans

To study the role of beliefs in driving costly punishment, in Table 2 we show regression analyses that study the link between beliefs, *Reject* choices, and *Reject* plans. These analyses account for the cost of choosing *Reject*, the *Take* amount, and whether communication was available. In this subsection, we focus on the role of beliefs; in subsections 4.2.2 and 4.2.3 we discuss the effects of the *Take* amount and of communication.

Columns A-B in Table 2 report the results of logistic regressions where the dependent variable, P1's *Reject* Choice, is equal to 1 if Player 1 *Reject*s the offer in stage 3, and equal to 0 if Player 1 *Accepts* the offer in stage 3. We find, in the model in Column B, a significant

discrimination, while AUC values greater than 0.9 are termed "outstanding".

relationship between Player 1's first order belief about *Share* and decision to *Reject* the offer after *Take*. A 10% increase in the elicited probability of *Share* increases Player 1's chance of rejecting by 2.488%, which is consistent with Hypothesis 1.

The models in Columns A-B do not include subject or session level controls. When either of these controls are included, the coefficient on Belief about *Share* is not significant (see Supplementary Table 1), implying that the relation in Column B is driven by variability between individuals or sessions. This is not inconsistent with our theory, which simply posits a relation between beliefs and behavior. We also have limited data: the 474 *Take* choices imply that on average we observe 4.74 choices to *Accept* or *Reject* at the end of the game (min 1, max 10). Thus, our data may not be sufficient to establish a within-participant relation between Beliefs about *Share* and *Reject* choices.

Table 2. The Effect of Belief about *Share* on P1's *Reject* Choice and Plan.

	P1's *Reject* Choice		P1's *Reject* Plan	
	A mfx / se	B mfx / se	C mfx / se	D mfx / se
Cost of *Reject*	-0.1985***	-0.1814***	-0.0736***	-0.0677***
	(0.0224)	(0.0236)	(0.0085)	(0.0070)
High *Take*	0.0442	0.0769	0.0087	0.0272***
	(0.0674)	(0.0581)	(0.0073)	(0.0077)
Communication	0.0174	0.0111	0.0552***	0.0449**
	(0.0501)	(0.0597)	(0.0176)	(0.0189)
Period	0.0137***	0.0129***	0.0122***	0.0118***
	(0.0042)	(0.0048)	(0.0014)	(0.0017)
Belief about *Share*		0.2488**		0.1230***
		(0.1100)		(0.0340)
Observations	474	474	2000	2000
BIC	560.4	558.9	589.4	571.6
Subject controls	No	No	Yes	Yes

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Note: mfx: marginal effect. Marginal effects for continuous variables are evaluated at means, and for binary variables are evaluated as the discrete change from 0 to 1. se: standard error. Standard errors are bootstrapped at the session level. Logistic regressions are employed for P1's *Reject* Choice, and fixed effect linear regressions are employed for P1's *Reject* Plan. See Supplementary Table 1 and 2 for additional regressions of *Reject* Choice and *Reject* Plan.

In Table 2, Columns C-D, we employ fixed effects linear regressions to study the determinants of Player 1's reported *Reject* plan (divided by 100, to scale between 0 and 1). Here,

we have data for each game played, for a total of 2,000 observations (20 per participant in the role of Player 1), and so these models include participant level controls. Focusing on the role of beliefs, the model in Column D shows that Player 1's first order Belief about *Share* has a positive and statistically significant association with Player 1's *Reject* plan, consistent with the frustration and anger model and with Hypothesis 1. Thus, within-participant variation in the probabilistic belief that Player 2's would choose *Share* was linked to changes in participants' probabilistic plan to *Reject*: an increase of 10% in the Belief about *Share* is associated with about a 1.2% increase in the reported Plan to *Reject*.

4.2.2 Testing H2: The effect of the cost of *Reject* and the *Take* amount on choices and plans

Our model posits that participants make economic tradeoffs between monetary rewards and anger, and that larger *Take* amounts should lead to increased frustration after *Take*. Thus in our experimental design (see Table 1), we varied both the cost of *Reject* (b) and the *Take* amount ($a - b$, either 4 or 10).

The results in Figures 5(c) and 6(c) show declining rates of *Reject* choices and plans to *Reject* as b increases and are consistent with the idea that participants are sensitive to the cost (foregone reward) of choosing *Reject*. After pooling High and Low *Take* games, a nonparametric test for trend (Cuzick, 1985) rejects the null hypothesis that *Reject* rates are unrelated to the cost of *Reject* (b) ($p < 0.001$). A nonparametric test for trend also rejects the null hypothesis that *Reject* plans are unrelated to the cost of *Reject* ($p < 0.001$). Furthermore, the regression analyses in Table 2 show that the coefficient estimate on the variable "Cost of *Reject*" (b) is significant and negative for both *Reject* choices (Models A-B) and plans (Models C-D). The magnitude of the effect varies from an almost 20% increase per unit (Models A-B, choices) to about a 7% increase per unit (Models C-D, plans). Participants were clearly sensitive to the cost of choosing *Reject*, consistent with Hypothesis 2.[18]

Our theory also implies a relationship between the *Take* amount ($a-b$) and *Reject* choices and plans, so next we return to the regression analyses in Table 2. In neither of the regression analyses for Player 1's *Reject* choice is the coefficient on High *Take* significant (Models A-B), though the sign is positive. In the Appendix, Supplementary Table 1 reports the results from

[18]As noted in Section 2.2, games where the "Cost of *Reject*" (b) is equal to 5 (LT5 and HT5) are edge cases that admit an additional sequential equilibrium in which Player 1 is indifferent between *In* and *Out*. Including an indicator variable for $b = 5$ in our regressions does not meaningfully affect any of the reported estimates.

additional specifications; in only one of the models is the coefficient for High *Take* significant (Model D). Turning to our fixed-effects linear regression analyses for *Reject* plans (r_1), we again find that the coefficient on High *Take* is positive but not statistically significant in Model C. However, after including Player 1's Belief about *Share*, Model D finds that the High *Take* condition adds about 2.7% to the reported likelihood that Player 1 will assign to *Reject*. While consistent with Hypothesis 2, the effect of High *Take* is conditional on beliefs. Additional model specifications give similar results; see Supplementary Table 2.

We interpret these results through the lens of the preceding analyses indicating the importance of player's beliefs about whether their co-players will cooperate. Clearly, Player 1s' *Reject* choices and plans are closely linked to their beliefs about whether Player 2 will *Share*, and given beliefs, on the *Take* amount $(a - b)$.

4.2.3 Testing H3: The effect of communication

Each experimental session included both no-communication and communication blocks. In the latter, Player 2 was given the opportunity to send a pre-play free-form message to Player 1. We first investigate how the communication treatment affects game outcomes in aggregate, and the result is shown in Table 3. The effect of communication across the full spectrum of 10 game structures is shown in Supplementary Table 8.

Table 3. The Effect of Communication.

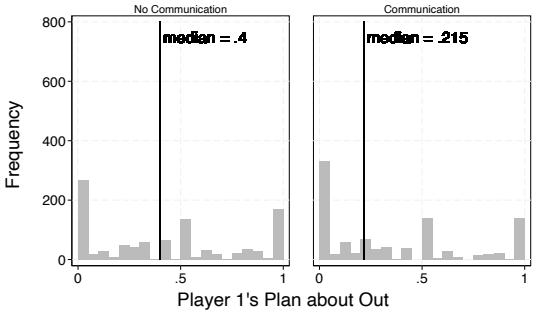
	Out	*Share*	*Reject*	*Accept*	Total
	263	467	97	173	1000
No Communication	26.30%	46.70%	9.70%	17.30%	100.00%
			35.93%	64.07%	100.00%
	195	601	82	122	1000
Communication	19.50%	60.10%	8.20%	12.20%	100.00%
			40.20%	59.80%	100.00%
	458	1068	179	295	2000
Total	22.90%	53.40%	8.95%	14.75%	100.00%
			37.76%	62.24%	100.00%

Note: Row 1: number of observations; row 2: percent of total observations; row 3: percent of observations that reach the third stage.

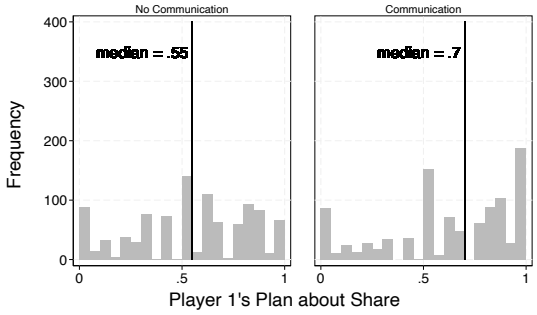
The cooperative outcome *(In,Share)* is more prevalent in the communication treatment

(60.10% vs. 46.70%). A 1-sided Fisher's exact test confirms that the cooperation rate is higher in the communication treatment (p-value < 0.001). This result is consistent with the belief-dependent model of frustration and anger and with Hypothesis 3, that communication will increase cooperation. A chi-squared test shows that allowing communication has a significant effect on the distribution of outcomes (terminal histories) (p-value < 0.001). The conditional *Reject* rate is also higher in the communication treatment (40.20% vs. 35.93%), but this difference is not significant (1-sided Fisher's exact test: p-value $= 0.197$).

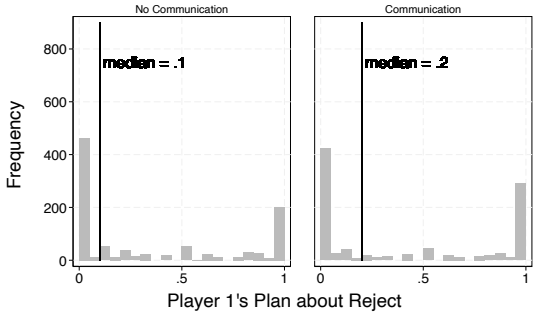
Communication also affects reported beliefs. Figure 7 presents histograms of Player 1's self-reported plans to choose *Out* and *Reject*, and beliefs that Player 2 will choose *Share*, in the communication and the no-communication treatments. In the communication treatment we measured beliefs both before and after messages were received; unless otherwise noted the belief data we report for the communication treatment were recorded after messages were received. Epps-Singleton tests (Epps and Singleton, 1986) confirm that the distributions of reported beliefs are significantly different in the communication vs. the no-communication treatment (plan to choose *Out*: p-value < 0.001; belief about *Share*: p-value < 0.001; plan to choose *Reject*: p-value < 0.001).



(a) P1's *Out* Plan.



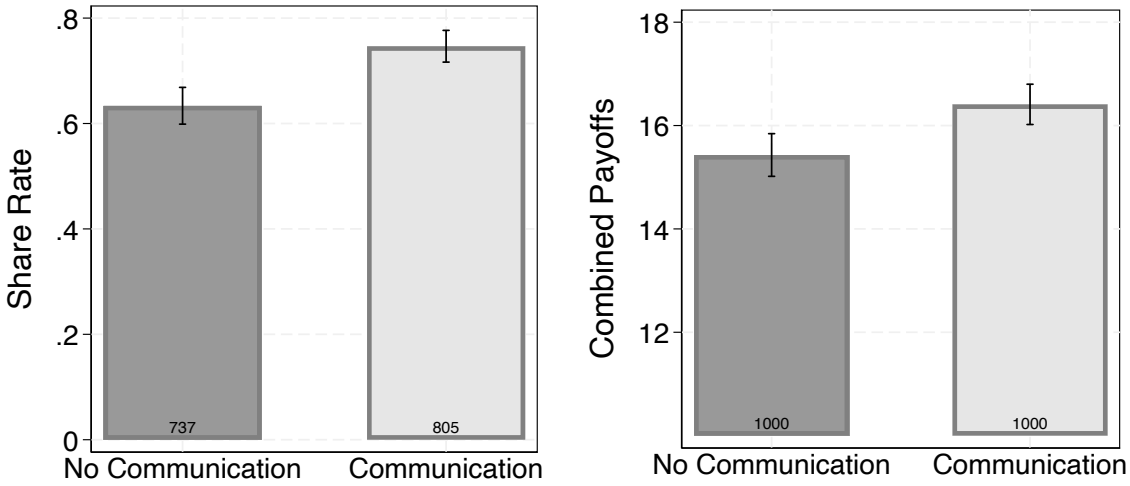
(b) P1's *Share* Belief.



(c) P1's *Reject* Plan.

Figure 7. Histograms of P1's Plans (beliefs about own actions) and Beliefs (about P2), by Communication.

As predicted by Hypothesis 3, communication has a strong effect on efficiency and cooperation. Figure 8(a) compares *Share* outcomes from the no-message and message blocks, pooling the data from all sessions, with number of observations labeled on the bars. We observe a significantly higher cooperation rate with communication on a subject level (1-sided t-test: p-value < 0.001). Additionally, we show that in Supplementary Figure 4(a), relatively higher cooperation rates are observed across 10 different game variations in the communication treatment.



(a) P2's *Share* Rate with Communication. **(b)** Combined Payoffs per Pair of Players.

Figure 8. Communication Improves Cooperation and Earnings.

To test whether communication improves efficiency, we first compare participants' combined payoffs (Figure 8(b)). On average, combined payoffs are significantly higher in the communication treatment (\$16.41 vs. \$15.43, rank sum test: p-value = 0.014). Next, we look into Player 1 and Player 2's payoffs separately.

Figure 9 shows that on average, the payoffs of Player 2s are insignificantly greater in the communication treatment (\$9.34 vs. \$9.03, rank sum test: p-value = 0.127); whereas Player 1's average earnings are significantly larger with communication (\$7.07 vs. \$6.40, rank sum test: p-value < 0.001). This suggests that social welfare or efficiency increases if communication is allowed. This result is consistent with our Hypothesis 3, that communication improves efficiency.

The effect of communication varies across game parameterizations in a pattern broadly consistent with the theory. As illustrated in Figure 3, the range of anger sensitivities sup-

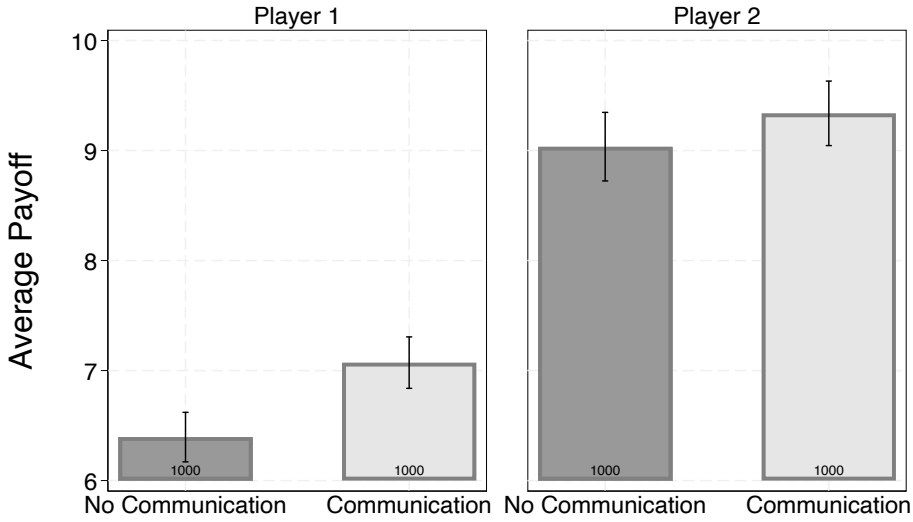


Figure 9. The Effect of Communication on Payoffs by Player Role.

porting multiple equilibria depends on the cost of Reject (b) and the *Take* amount (a-b). While the population distribution of anger sensitivities is unobservable, we designed our ten game variations to span a wide range of thresholds. Supplementary Table 8 shows that the increase in cooperative outcomes due to communication tends to be largest at intermediate parameterizations (e.g., 28% in LT2, 24% in LT3, and 17% in HT3). At the extremes the effect is more muted: where the cost of *Reject* is low (LT1, HT1), high baseline cooperation leaves little room for communication to add (+14% and +8%), while the +10% in HT5 reflects the $b = 5$ edge case noted in Section 2.2, where indifference at the entry node makes cooperation especially sensitive to communication.

Our counterbalanced design also reveals that the effect of communication on beliefs persists into subsequent no-communication periods, a finding we explore in detail in Section 4.3.1. When we restrict the *Reject* plan regressions in Table 2 to the first 10 periods to isolate the pure treatment contrast, the evidence for our mechanism strengthens: the key theoretical variables increase in magnitude (Supplementary Table 7).

4.2.4 H4 and H5: Communication Shifts Beliefs through Promises

We now consider the effect of communication on beliefs (measured after the message was received in the communication treatment). Figure 10 shows that communication affects Player 1's reported beliefs, and this result is consistent with our Hypothesis 4. Player 1s

report a higher likelihood that Player 2 will cooperate (1st order belief about *Share*, q_1) when communication is allowed. Communication affects Player 1's own plans as well. With communication, Player 1 believes that she is less likely to play *Out* but more likely to *Reject* if 3rd stage is reached. 1-sided t-tests confirm that Player 1's beliefs are significantly different in the communication treatment and the no-communication treatment (plan to choose *Out*: p-value = 0.009; 1st order belief about *Share*: p-value < 0.001; plan to choose *Reject*: p-value = 0.069). In addition, the direction of how communication influences beliefs is consistent with belief-dependent anger.

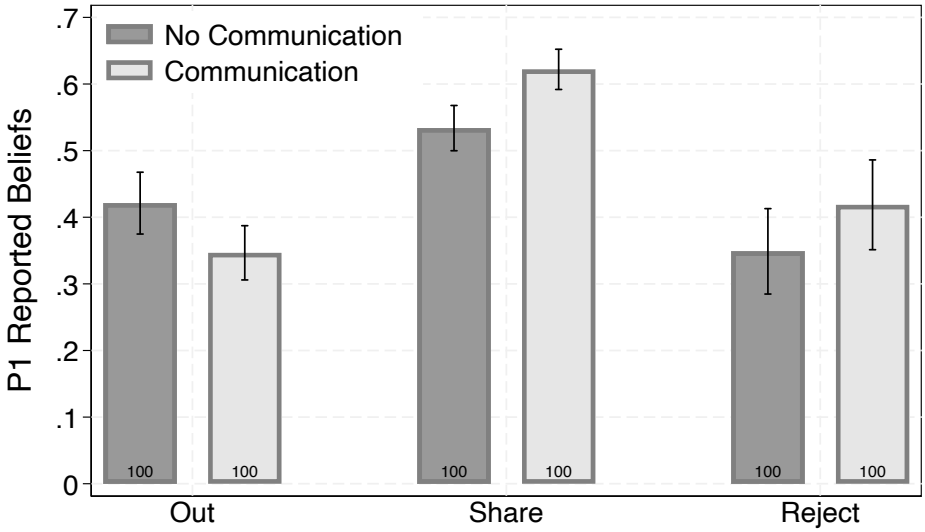


Figure 10. Communication Influences P1's Reported Beliefs.

To examine the links between message content, beliefs, and behavior, the authors coded the messages, then resolved minor discrepancies through discussion. A message is identified as a promise if it followed a clear conditional commitment to a specific action, such as "If you choose *In*, I will choose *Share*." Following this protocol, we identify 32% of messages as promises, and the median percentage of promises per session was 32.2%. A selection of representative sample messages and their categories are presented in Supplementary Table 6, while the full dataset of 1,000 messages and their associated coding is available upon publication.

As noted above, in the communication treatment, we measured beliefs both before and after receiving a message. Figure 11 shows that promises have a strong effect on Player 1's reported beliefs. Promises increase Player 1's belief about Player 2's cooperative behavior (1st order belief about *Share*). Promises also influence Player 1's beliefs about their own actions (plans to choose *Out* and *Reject*). After receiving a promise message, Player 1s

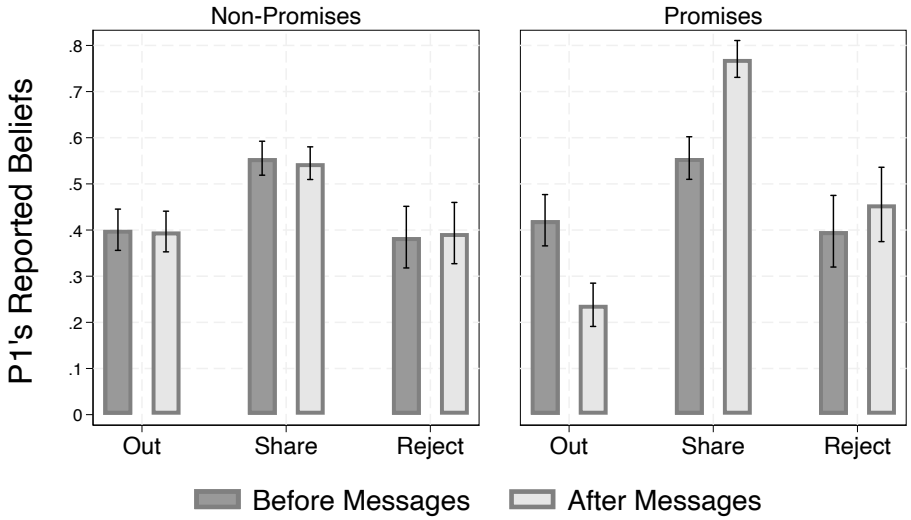


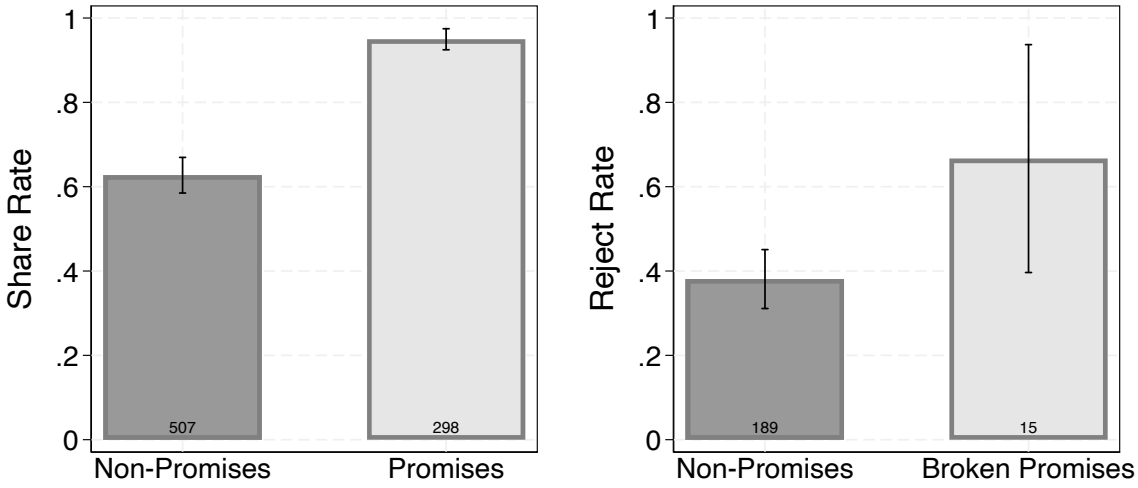
Figure 11. Belief Change After Receiving a Message.

report that they will be less likely to choose *Out*, but will be more likely to punish Player 2. 1-sided t-tests show a significant difference in the change in Player 1's reported beliefs after receiving a promise, compared to receiving a message that did not involve a promise (plan to choose *Out*: p-value < 0.001; 1st order belief about *Share*: p-value < 0.001; plan to choose *Reject*: p-value $= 0.011$). In addition, two-sided t-tests confirm that the change in Player 1's reported beliefs after non-promise messages is not significantly different from 0 (plan to choose *Out*: p-value $= 0.779$; 1st order belief about *Share*: p-value $= 0.343$; plan to choose *Reject*: p-value $= 0.390$). Promises have a significant effect upon beliefs, while non-promises have an insignificant effect, consistent with Hypothesis 4.

4.2.5 Promises Influence Behavior

To further demonstrate the effect of promises on behavior as predicted in Hypothesis 5, we look at behavior differences under promises and non-promises. Supplementary Table 3 shows the outcome distribution with respect to promises and non-promises is consistent with belief-dependent anger. A chi-squared test shows that the distribution of outcomes is significantly different with and without promises (p-value < 0.001). Figure 12 shows that the proportion of both *Share* and *Reject* choices is higher when a promise is made. This result is consistent with Hypothesis 5: promises foster cooperation, but broken promises lead to higher rates of punishment. The effect of promises is greater than that of communication alone: comparing Supplementary Figures 4(a) and 4(b), separating promise from non-promise messages yields

a greater improvement in cooperation across the 10 game variations than the communication treatment alone.



(a) P2's *Share* Rate with Promises. **(b)** P1's *Reject* Rate with Broken Promises.

Figure 12. Kept and Broken Promises.

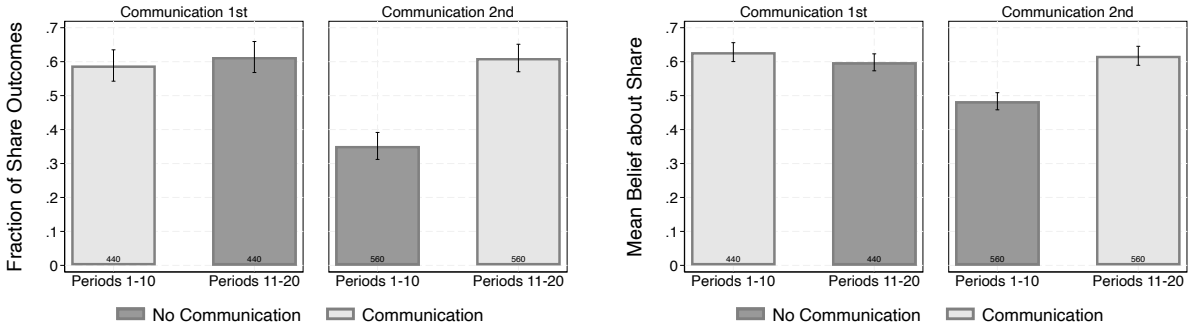
The result shown in Figure 12(a) is consistent with the frustration and anger model, in that if Player 2 anticipates the change in Player 1's beliefs following a promise, Player 2 will have increased motivation to choose *Share* in order to avoid punishment from a *Reject* choice after *Take*. When we compare Player 2's behavior after non-promises vs. after promises, the *Share* rate is significantly higher following promises (1-sided Fisher's exact test: p-value < 0.001). A rank sum test confirms that subject level *Share* rate is also higher with promises (p-value < 0.001). This result holds on the level of each of the 10 different game variations as well. Supplementary Figure 4(b) shows that the *Share* rate for promises is higher across all 10 game variations. Promises affect not only cooperative behavior but also rejection. As predicted by the frustration and anger model, Player 1's beliefs change following promises, and Player 1 is more likely to punish with broken promises. Figure 12(b) shows that the *Reject* rate is higher with broken promises compared to non-promises, consistent with Hypothesis 5 (1-sided Fisher's exact test: p-value $= 0.030$). A rank sum test confirms that subject level *Reject* rate is also higher with broken promises (p-value $= 0.068$).

4.3 Additional Observations

This section reports additional observations regarding our data. We first discuss the observed persistent effect of communication, then we present evidence for gender differences. We then present a post-experimental survey of self-reported anger ratings.

4.3.1 The Persistent Effect of Communication

Our counterbalanced design reveals that the effect of communication persists even after it is removed. Figure 13 shows that in the communication-first sessions, there is a persistent effect of communication: the higher rate of cooperation (*Share* outcomes) is sustained in the subsequent no-communication periods. Restricting attention to the first 10 rounds of each treatment, there is a significantly higher cooperation rate in the communication-first sessions than in the communication-second treatment (58.86% vs. 35.18%, 1-sided Fisher's exact test: p-value < 0.001). The difference disappears in rounds 11-20 (61.36% vs. 61.07%, 2-sided Fisher's exact test: p-value $= 0.948$). This suggests that communication durably shifts beliefs about cooperation, such that the cooperative equilibrium is sustained even without continued communication. This durable effect of communication arises not just in outcomes, but also in Player 1's reported beliefs. Player 1s have a significantly higher first order belief about *Share* in the first 10 periods of the communication-first sessions relative to the no-communication sessions (two-sided t-test for difference in session means: $t = 3.628$, $df = 9$, p-value $= 0.006$), but there is no difference in beliefs about *Share* when comparing rounds 11-20 of the communication-first and communication-second sessions ($t = -0.416$, $df = 9$, p-value $= 0.6874$).



(a) Player 2's *Share* Choices. **(b)** Player 1's Belief about *Share*.

Figure 13. Persistent Communication Effect.

Because of this persistent effect of communication, we examine the distribution of outcomes after restricting the sample to include only the first 10 rounds. Figure 14 compares the effects of communication on the distribution of outcomes with all 20 rounds and with the first 10 rounds only, when the no-communication group has no experience with messages. The contrast of communication vs. no communication is stronger when we look at the first 10 rounds only. The mean fraction of *Share* outcomes in the communication treatment in the first 10 rounds is 58.86%, which is close to the overall mean for 20 rounds (60.10%, see also in Table 3), but the cooperation rate without communication in the first 10 rounds decreases to 35.18%. A chi-squared test shows that the communication treatment has a significant effect on the distribution of outcomes for the first 10 rounds of the experiment (p-value < 0.001). These results demonstrate that communication has a strongly positive and persistent effect on cooperation.

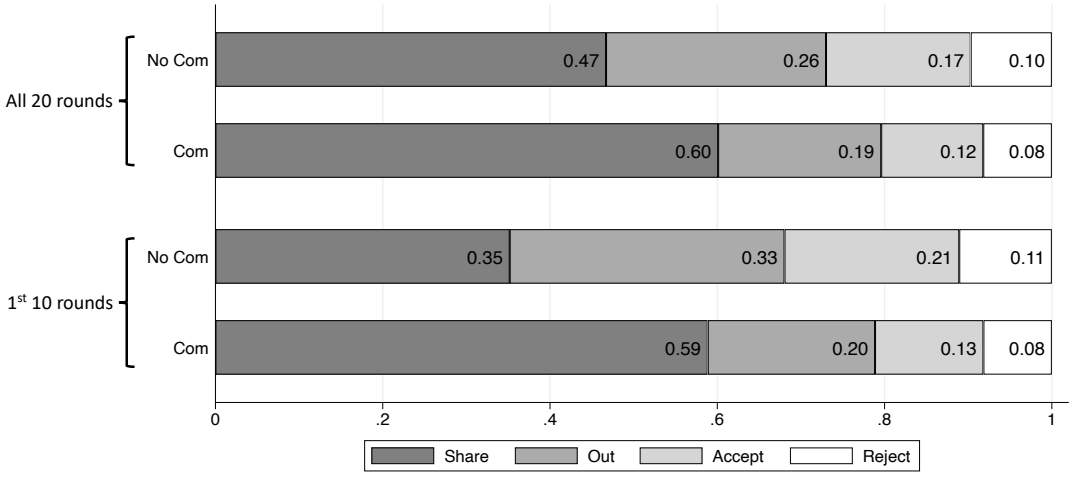


Figure 14. Outcomes Compare All 20 vs. 1st 10 Rounds.

The persistent effect of communication provides additional evidence for our belief-based mechanism. To demonstrate this, we reran our main regression models (focusing on plans, Table 2 and Appendix Table 2), restricted to the first 10 periods of the experiment (see Supplementary Table 7.). Restricting our analysis to observations before any spillover effects of communication can emerge substantially strengthens our findings. The magnitude of the coefficients for the key theoretical variables increases noticeably: the marginal effect of the High Take condition roughly doubles in the first 10 periods compared to the full sample, and the impact of Belief about *Share* is similarly amplified. Furthermore, in this restricted specification, the effect of Communication is limited. This provides additional evidence for the frustration and anger mechanism: communication matters precisely because it shifts beliefs, which in turn drive punishment.

4.3.2 Gender Differences

We started this project with BDS' theory in mind and the intention to test hypotheses 1-5. We recorded subjects' genders without any preconceived conjectures as regards whether results would differ between women and men. Aina et al. (2020), however, report that men are more affected by anger than women. As it turns out, we have comparable findings. Female and male behavior is relatively similar, except that when promises are made, males tend to *Reject* more often (*Reject* rate 70% vs. 30%, rank sum test: p-value = 0.077). We ran the fixed effects linear regressions for *Reject* plans separately for females and males and report the results in Table 4. The effect of communication survives with females, but disappears with males. In contrast, beliefs about *Share* are significantly positively associated with plans to *Reject* for males, but not for females. Consistent with Aina et al. (2020), the coefficient estimates for "High *Take*" and "Belief about *Share*" are positive and statistically significant predictors of Player 1's *Reject* plan in the regression for males, but not in the female-only analysis. In our data as well, men's beliefs and choices are more consistent with the frustration and anger model.

Table 4. Linear Regressions – Gender Effect of P1's *Reject* Plan.

	Females		Males	
	A	B	C	D
	mfx / se	mfx / se	mfx / se	mfx / se
Cost of *Reject*	-0.0744***	-0.0703***	-0.0732***	-0.0668***
	(0.0133)	(0.0143)	(0.0094)	(0.0086)
High *Take*	-0.0027	0.0104	0.0165	0.0365**
	(0.0129)	(0.0157)	(0.0164)	(0.0169)
Period	0.0110***	0.0107***	0.0130***	0.0126***
	(0.0031)	(0.0030)	(0.0030)	(0.0033)
Communication	0.0592	0.0539	0.0543*	0.0397
	(0.0367)	(0.0361)	(0.0318)	(0.0355)
Belief about *Share*		0.0734		0.1539***
		(0.0732)		(0.0324)
Observations	880	880	1100	1100
BIC	154.8	157.5	451.8	437.6
Subject controls	Yes	Yes	Yes	Yes

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Note: mfx: marginal effect. Marginal effects for continuous variables are evaluated at means, and for binary variables are evaluated as the discrete change from 0 to 1. se: standard error. Standard errors are bootstrapped at the session level. Fixed effect linear regressions are employed for P1's *Reject* Plan. See Supplementary Tables 4 and 5 for additional regressions of *Reject* Plan separating for females and males.

4.3.3 Emotion Self-Reports

In a post-experiment survey, we collected self-reported data to examine Player 1's emotional response to the opponent's choice. The survey consisted of two questions: first, a rating of anger intensity on a 5-point scale (ranging from 1 "not angry at all" to 5 "very angry") following an outcome where Player 1 chose *In* and Player 2 chose to *Take*. Second, Player 1 performed a cross-treatment comparison to identify whether they would feel more intense anger if this outcome occurred in a round with messaging versus a round without messaging.

The results show that 62% of Player 1s reported feeling angry following a *Take* choice. The mean anger rating (2.28) is significantly larger than not feeling angry at all ($H_0 = 1$, p-value < 0.001). Furthermore, self-reported anger is significantly higher in the communication

treatment compared to the no-communication treatment (One-sample t-test, mean = 0.59 vs. $H_0 = 0.5$, p-value = 0.036). The survey results provide some evidence for the frustration and anger mechanism.

5 Alternative Motivations

Although our focus in this paper is the BDS model and its application to our setting, people are, of course, motivated by factors other than frustration & anger. We now discuss the extent to which these other motivations make similar predictions to ours, and whether participant behavior matches these predictions. We consider, in turn, inequity aversion, guilt aversion, reciprocity, fixed costs of lying, and social norms.

Inequity aversion These models (e.g., Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), applied to our setting, predict that players will choose to *Reject* because they dislike disadvantageous inequality. This is consistent with our Hypothesis 2(a). However, inequity aversion is a consequentialist theory and predicts that decision-makers will make the same choice every time from a given choice set. Thus it cannot explain the effects of changes in the *Take* amount, of communication, or of variation in beliefs about *Share*. However, it is possible that our sample involves heterogeneity and that some of our participants may be motivated by inequity aversion to various degrees.

One implication of inequity aversion is that if Player 1 ever *Reject*s a high offer in the 3rd stage, then she/he would never *Accept* a lower offer, regardless of communication or beliefs. Using this idea we can classify subjects into three categories, shown in Table 5. The analysis is conservative in that some participants may have had few opportunities to make choices in the third stage of the game, if they primarily chose *Out* or their co-players chose *Share*. In the table, "Selfish" refers to players who always choose *Accept*. "Inequity Averse" subjects' behavior is consistent with inequity aversion in that they never *Reject* a higher and *Accept* a lower offer, nor *Reject* and *Accept* the same offer (e.g., rejecting an offer of 3 in one round and accepting 3 in another). "Other" represents everyone else.

Table 5 shows that 39% of subjects demonstrate behavior that is inconsistent with either self-interest or inequity aversion, 28% of subjects demonstrate behavior consistent with inequity aversion, while 33% of subjects behave as if they care only for material payoffs. Moreover, the number of subjects whose behavior is inconsistent with inequity aversion or selfishness increases when subjects have more decisions in the 3rd stage. Table 6 repeats

Table 5. Classification of Player 1 behavior.

	Selfish	Inequity Averse	Other
# of Subjects	33	28	39
# of 3rd Stage Decisions	4.27	4.79	5.10

the analyses of Tables 2 and 4 separately for each group of participants from Table 5. Only the Other group shows significant coefficients for High *Take* and Belief about *Share*. We conclude that at minimum, greater than one third of our sample shows behavior consistent with the BDS model.

Table 6. 3 Types of Player 1's *Reject* Plan Regressions

	Selfish	Inequality Averse	Other
	A	B	C
	mfx / se	mfx / se	mfx / se
Cost of *Reject*	-0.0333***	-0.0924***	-0.0799***
	(0.0111)	(0.0155)	(0.0096)
High *Take*	0.0120	-0.0010	0.0576***
	(0.0119)	(0.0177)	(0.0154)
Period	0.0007	0.0154***	0.0199***
	(0.0012)	(0.0021)	(0.0035)
Communication	-0.0000	0.1412***	-0.0007
	(0.0162)	(0.0189)	(0.0415)
Belief about *Share*	0.1061	0.0606	0.1857***
	(0.0739)	(0.0684)	(0.0460)
Observations	660	560	780
BIC	-87.5	183.6	359.3
Subject controls	Yes	Yes	Yes

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Note: mfx: marginal effect. Marginal effects for continuous variables are evaluated at means, and for binary variables are evaluated as the discrete change from 0 to 1. se: standard error. Standard errors are bootstrapped at the session level. Fixed effect linear regressions are employed for P1's *Reject* Plan.

Guilt aversion is a well-studied belief-dependent motive for why people repay trust.[19]

[19]See Dufwenberg (2002) and Dufwenberg and Gneezy (2000) for early papers, Battigalli and Dufwenberg (2007) for general theory, and Cartwright (2019) and Attanasi et al. (2023, Table A1; 2025) for surveys of (and in the latter two cases contributions to) the experimental literature.

While our focus for why Player 2 may *Share* concerns 2's dread that Player 1 may *Reject*, guilt aversion provides an alternative reason. We cannot separately identify these two potential motives behind the choice to *Share*. However, guilt aversion cannot explain Player 1's decision to engage in costly punishment, as players solely motivated by guilt aversion will never choose *Reject*.

Cost of lying Another strand of models addresses subjects' direct preference to honor promises (e.g., Vanberg, 2008). Guilt aversion and/or fixed costs of lying may help explain why communication increases the frequency of *Share* choices. However, as with guilt aversion, a fixed cost of lying can explain only Player 2's decision to share, but not Player 1's decision to engage in costly punishment. Our results indicate that the motivation to avoid frustration, anger, and costly punishment in our game has additional effects. Finally, after promises, participants in our study choose to *Share* a striking 95% of the time. This amount of promise-keeping is much higher than in comparable studies without a punishment stage (e.g., Charness and Dufwenberg, 2006).

Reciprocity An alternative motivation for punishing broken promises might involve reciprocity, in which individuals are motivated to reward kindness and punish hostility through beliefs. However, sequential reciprocity (Dufwenberg and Kirchsteiger, 2004) allows for the possibility that players engage in mutual unkindness on the equilibrium path. In our setting, this would imply that a Player 1 whose assigned a low probability to Player 2 choosing *Share* could also report plans (beliefs about her own subsequent actions) that involve a high probability of choosing out *Out*, and a low probability of choosing *Reject*. This pattern of beliefs is ruled out by the frustration and anger model, and in fact, we do not observe it in our data. Player 1s' reported beliefs about the likelihood of *Share* are negatively associated with self-reported plans to choose *Out*, and positively associated with plans to choose *Reject* (Figure 15).[20] This observation does not constitute a refutation of reciprocity theory, as it may be possible to observe such mutual unkindness in other settings (e.g., feuds), but it does suggest that the beliefs and plans we elicit are consistent with the frustration and anger model.

Social Norms Another potential motivation is the desire to conform to social norms. Many individuals are willing to pay a cost to punish norm violators (Fehr and Gächter, 2000; Xiao and Houser, 2011). In our setting, Player 1 might regard Player 2's decision to *Take* as a violation of fairness norms. Player 1 can then choose to punish this norm violation. The frustration and anger explanation and the punishing-norm-violators story are

[20]Bosman and Van Winden (2002) observed a similar pattern in the context of a power-to-take game.

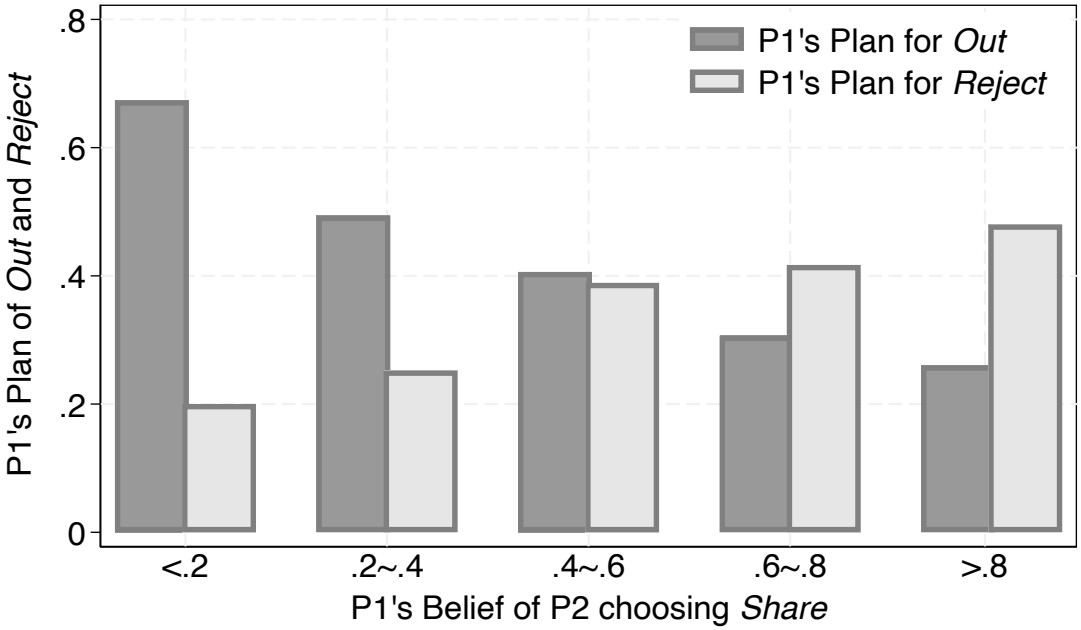


Figure 15. Player 1's self reported plans vs. beliefs about *Share*.

not mutually exclusive, as frustration is likely to result from the violation of the fairness norm. Krupka et al. (2017) show that social (in)appropriateness ratings (norms) and second-order beliefs (guilt aversion) are both good predictors of behavior involving informal agreements. This suggests that further work is necessary to separately identify the motivations of belief-dependent anger from the desire to punish norm violators in the context of broken promises.

6 Conclusion

We develop a new model whereby broken promises induce belief-dependent frustration, triggering anger and aggression. This deters those who issue promises from reneging in environments that augment trust games with a punishment stage. We test our mechanism in an experiment and find considerable support. While many behavioral motivations may influence play, alternative models cannot explain the full pattern of beliefs and behavior. Our framework uniquely predicts how frustration over unmet expectations drives punishment of broken promises.

The literature on hold-up explores how relationship-specific investments and unenforceable contracts may result in underinvestment and inefficiency. The settings we explore have

the structure of hold-up problems. Our analysis shows that anger, costly punishment, and communication combine to facilitate cooperation and efficiency. Other approaches have, by contrast, tended to focus on incomplete information and aspects of the bargaining process. It follows from our results that even if it is not possible to fully commit to uphold a contract, the hold-up problem can be resolved at least in part if players are prone to anger.

Our solution has some similarities with the approach of Hart and Moore (2008, H&M). They model agents for whom contracts serve as a reference point by which outcomes are evaluated. Parties who are shortchanged relative to the reference contract feel "aggrieved" and shade on performance. Our focus is different; the starting point for H&M is the contract as a reference point, while ours is the belief-dependent model of BDS in which the "reference point" against which outcomes are judged is expected payoffs. We study how anger and communication can help to resolve the hold-up problem, while H&M focus on how behavioral considerations generate tradeoffs between flexible and rigid contracts.

In studying the role of communication, we have limited our attention to messages from the second-mover to the first-mover. This is restrictive. In particular, messages from the first-mover to the second-mover could involve threats that gain credibility with anger and frustration in the picture. The topic is so interesting that it warrants its own research exercise, which, in fact, we run as a companion project (Dufwenberg, Li, and Smith, 2026).

Appendices

A Instructions

Below is an example of the instructions for sessions with the communication treatment before the no communication treatment. The instructions for the second part of the experiment were given to all the subjects after the communication block was completed.

Part I Instructions

Welcome to the experiment. The purpose is to study how people make decisions in a particular situation. Please do not speak to other participants during the experiment. Feel free to ask a question at any time by raising your hand.

You will receive $5 for participating. You have the potential to earn additional money based on your own and others' decisions, as described below. Your decisions and payoffs will remain confidential. You will be paid individually and privately, in cash, at the end of the experiment.

There are two parts to the experiment. Both parts consist of multiple rounds of simple games that will be described below. The order in which choices are made in the games will remain the same in each round, but the payoff to different actions may change, so please pay careful attention to the payoffs in each round. At the end of the experiment, you will be privately paid for one randomly selected round from the entire experiment.

At the beginning of the experiment you will be randomly assigned to the role of either Player 1 or Player 2, and your role will not change throughout the experiment. In each round you will be randomly matched with another person in the room to play the game.

Prior to the start of each round, Player 2 will have the option to send messages to Player 1 (maximum 140 characters). Player 2 may say anything that he or she wishes in these messages, with one exception: no one is allowed to identify him or herself by name or number or gender or appearance. Violations of this rule may result in the loss of Player 2's payment for that part of the experiment (experimenter discretion). In that case the paired Player 1 will receive the average amount received by other Player 1s in this session.

Please raise your hand now if you have any questions. Select Continue when you are ready.

The game consists of three stages. The picture below may help and will be shown in each round. Payoffs will change in each round, so please familiarize yourself with the picture. Player 1's payoffs are listed above Player 2's payoffs. The game proceeds as follows:

- Player 1 goes first and must decide between A and B.
  - If A is chosen, the game ends and both players receive $5.
  - If B is chosen, the game proceeds to stage 2.
- If Player 1 chooses B, Player 2 must decide between C and D.
  - If C is chosen, the game ends with payoffs specified for that round.
  - If D is chosen, Player 1 will make another decision.

- If Player 2 chooses D, Player 1 will decide between E and F.
  - If E is chosen, the game ends and both players receive $0.
  - If F is chosen, the game ends with payoffs specified for that round.

Please raise your hand now if you have any questions. Select Continue when you are ready.

In each game you will be asked to guess how likely it is that certain events (decisions made by you or the other player) will happen. Your response is very important to our research. You will be asked to state the percent chance that each event will happen. You may select any number between 0 and 100, with the number you select indicating the likelihood of the event occurring (100 = certain the event will happen, 0 = certain the event will not happen). You will be rewarded with $5 for answering these questions. You have an option to choose to pledge to answer the guessing questions to the best of your knowledge by checking the box below:

□ By checking this box, I pledge that I will answer all guessing questions to the best of my knowledge.

Please raise your hand now if you have any questions. Select Continue when you are ready.

Part II Instructions

Thank you for completing the first part of the experiment. In the second part of the experiment your assigned role will not change. The second part of the experiment is like the first part, with one change: no messages will be exchanged. As before, this part consists of multiple rounds. In each round you will be randomly matched with another person in the room to play the game.

The only difference from the first part is that no messages will be exchanged for the second part of the experiment.

Please raise your hand now if you have any questions. Select Continue when you are ready.

As before, the game consists of three stages. The picture below may help and will be shown in each round. Payoffs will change in each round, so please familiarize yourself with

the picture. Player 1's payoffs are listed above Player 2's payoffs. The game proceeds as follows:

- Player 1 goes first and must decide between A and B.
  - If A is chosen, the game ends and both players receive $5.
  - If B is chosen, the game proceeds to stage 2.
- If Player 1 chooses B, Player 2 must decide between C and D.
  - If C is chosen, the game ends with payoffs specified for that round.
  - If D is chosen, Player 1 will make another decision.
- If Player 2 chooses D, Player 1 will decide between E and F.
  - If E is chosen, the game ends and both players receive $0.
  - If F is chosen, the game ends with payoffs specified for that round.

Please raise your hand now if you have any questions. Select Continue when you are ready.

B Supplementary Tables and Figures

Supplementary Table 1. Determinants of P1's *Reject* Choice (Robustness Checks).

	A	B	C	D	E
	mfx / se	mfx / se	mfx / se	mfx / se	mfx / se
Cost of *Reject*	-0.1985***	-0.1814***	-0.0806*	-0.1680***	-0.1604*
	(0.0219)	(0.0245)	(0.0416)	(0.0283)	(0.0915)
High *Take*	0.0442	0.0769	0.0170	0.1407**	0.0728
	(0.0584)	(0.0606)	(0.0368)	(0.0717)	(0.0862)
Communication	0.0174	0.0111	0.0053		
	(0.0494)	(0.0483)	(0.0260)		
Period	0.0137***	0.0129***	0.0058	0.0116	0.0155
	(0.0044)	(0.0045)	(0.0043)	(0.0071)	(0.0277)
Belief about *Share*		0.2488**	0.0684	0.4949***	0.3771
		(0.1173)	(0.0906)	(0.1397)	(0.3374)
Promise				0.0692	0.0564
				(0.2282)	(0.6753)
Observations	474	474	474	204	204
AIC	539.6	533.9	462.9	230.2	186.1
BIC	560.4	558.9	483.7	250.1	202.7
Session controls	No	No	Yes	No	Yes
Subject controls	No	No	No	No	No

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Note: Logistic regressions with P1's *Reject* choice as the dependent variable. mfx: marginal effect. Marginal effects for continuous variables are evaluated at means, and for binary variables are evaluated as the discrete change from 0 to 1. se: standard errors. Standard errors are clustered at the session level in the pooled logit models and bootstrapped at the session level (200 replications) in the session fixed-effects models.

Supplementary Table 2. Determinants of P1's *Reject* Plan (Robustness Checks).

	A	B	C	D	E
	coef / se	coef / se	coef / se	coef / se	coef / se
Cost of *Reject*	-0.0723***	-0.0736***	-0.0677***	-0.0741***	-0.0690***
	(0.0083)	(0.0081)	(0.0071)	(0.0073)	(0.0087)
High *Take*	0.0068	0.0087	0.0272***	0.0208	0.0377***
	(0.0102)	(0.0081)	(0.0073)	(0.0157)	(0.0112)
Communication	0.0698*	0.0552***	0.0449**		
	(0.0374)	(0.0182)	(0.0175)		
Period		0.0122***	0.0118***	0.0196***	0.0196***
		(0.0015)	(0.0014)	(0.0038)	(0.0038)
Belief about *Share*			0.1230***	0.1122**	0.2232***
			(0.0336)	(0.0506)	(0.0576)
Promise				0.0064	-0.0089
				(0.0191)	(0.0333)
Constant	0.5623***	0.4445***	0.3561***	0.3408***	0.2527***
	(0.0545)	(0.0527)	(0.0413)	(0.0820)	(0.0866)
Observations	2000	2000	2000	1000	1000
AIC	682.198	561.399	538.041	121.620	1062.800
BIC	704.601	589.403	571.646	151.067	1092.247
Session controls	No	No	No	No	Yes
Subject controls	Yes	Yes	Yes	Yes	No

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Note: Linear regressions with *Reject* plan as dependent variable. mfx: marginal effect. Marginal effects for continuous variables are evaluated at means, and for binary variables are evaluated as the discrete change from 0 to 1. se: standard errors. Standard errors are bootstrapped at the session level.

Supplementary Table 3. The Effect of Promises on Outcomes (Communication Treatment Only).

	Out	*Share*	*Reject*	*Accept*	Total
	22	283	10	5	320
Promises	6.88%	88.44%	3.12%	1.56%	100.00%
			66.67%	33.33%	100.00%
	173	318	72	117	680
Non-Promises	25.44%	46.76%	10.59%	17.21%	100.00%
			38.10%	61.90%	100.00%
	195	601	82	122	1000
Total	19.50%	60.10%	8.20%	12.20%	100.00%
			40.20%	59.80%	100.00%

Note: Row 1: number of observations; row 2: percent of total observations; row 3: percent of observations that reach the third stage.

Supplementary Table 4. Determinants of P1's *Reject* Plan (Females)

	A	B	C	D	E	F
	coef / se	coef / se	coef / se	coef / se	coef / se	coef / se
Cost of *Reject*	-0.0726***	-0.0744***	-0.0703***	-0.0751***	-0.0717***	-0.0717***
	(0.0149)	(0.0160)	(0.0133)	(0.0179)	(0.0190)	(0.0180)
High *Take*	-0.0027	-0.0027	0.0104	0.0006	0.0099	0.0106
	(0.0136)	(0.0134)	(0.0163)	(0.0179)	(0.0242)	(0.0211)
Communication	0.0742**	0.0592*	0.0539			
	(0.0377)	(0.0352)	(0.0379)			
Period		0.0110***	0.0107***	0.0253***	0.0253***	0.0253***
		(0.0030)	(0.0032)	(0.0036)	(0.0040)	(0.0040)
Belief about *Share*			0.0734		0.0609	0.0668
			(0.0769)		(0.0929)	(0.0881)
Promise				0.0241	0.0098	0.0031
				(0.0268)	(0.0163)	(0.0235)
Constant	0.5172***	0.4143***	0.3572***	0.3058***	0.2575**	0.2557**
	(0.0948)	(0.0861)	(0.0734)	(0.0940)	(0.1036)	(0.1146)
Observations	880	880	880	440	440	440
AIC	179.545	130.935	128.772	13.019	13.474	438.206
BIC	198.665	154.835	157.451	33.453	37.995	462.727
Session controls	No	No	No	No	No	Yes
Subject controls	Yes	Yes	Yes	Yes	Yes	No

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Note: Linear regressions with *Reject* plan as dependent variable for females subjects only. mfx: marginal effect. Marginal effects for continuous variables are evaluated at means, and for binary variables are evaluated as the discrete change from 0 to 1. se: standard errors. Standard errors are bootstrapped at the session level.

Supplementary Table 5. Determinants of P1's *Reject* Plan (Males)

	A	B	C	D	E	F
	coef / se	coef / se	coef / se	coef / se	coef / se	coef / se
Cost of *Reject*	-0.0724***	-0.0732***	-0.0668***	-0.0823***	-0.0773***	-0.0740***
	(0.0090)	(0.0094)	(0.0078)	(0.0082)	(0.0082)	(0.0075)
High *Take*	0.0128	0.0165	0.0365**	0.0065	0.0289	0.0428**
	(0.0182)	(0.0171)	(0.0155)	(0.0225)	(0.0220)	(0.0212)
Communication	0.0708	0.0543	0.0397			
	(0.0494)	(0.0363)	(0.0348)			
Period		0.0130***	0.0126***	0.0148**	0.0150***	0.0150***
		(0.0031)	(0.0030)	(0.0063)	(0.0053)	(0.0054)
Belief about *Share*			0.1539***		0.1527***	0.2427***
			(0.0320)		(0.0477)	(0.0435)
Promise				0.0335	0.0069	-0.0010
				(0.0369)	(0.0336)	(0.0378)
Constant	0.6018***	0.4742***	0.3703***	0.5308***	0.4174***	0.3476***
	(0.0742)	(0.0664)	(0.0608)	(0.1114)	(0.1050)	(0.0954)
Observations	1100	1100	1100	550	550	550
AIC	493.885	426.762	407.562	126.671	117.458	524.642
BIC	513.897	451.777	437.580	148.220	143.318	550.501
Session controls	No	No	No	No	No	Yes
Subject controls	Yes	Yes	Yes	Yes	Yes	No

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Note: Linear regressions with *Reject* plan as dependent variable for male subjects only. mfx: marginal effect. Marginal effects for continuous variables are evaluated at means, and for binary variables are evaluated as the discrete change from 0 to 1. se: standard errors. Standard errors are bootstrapped at the session level.

Supplementary Table 6. Sample Messages. The full list of messages will be made available after publication.

	Message Content	Promise?	Outcome
1	Considering some irrational guys, I will stop at C for get money not for angry	Yes	Share
2	If I were you I will stop at A. If you choose E, I would consider you as an irrational guys. You can be nice or not	No	Share
3	If you choose B, then I will choose C. Cross my heart and swear to die. :)	Yes	Reject
4	Have yourself a merry little Christmas	No	Accept
5	Hi, I hope you have a good rest of the day. Thanks for participating in this research.	No	Accept
6	hello love. Choose B and lets both make more money!!! i promise i'lll pick C	Yes	Reject
7	if you choose B, i'll choose C (you'll be making the same amount but also helping me–ut prosim amiright?)	Yes	Share
8	Hello, I am a poor, broke, college student, plz be reasonable and considerate and generous	No	Out
9	In my opinion, pineapple is a pretty good topping on pizza.	No	Accept
10	Thanks for choosing F	No	Out
11	Hokies play UVA in baseball today at 5:30 at home. Pick B and I'll choose C. Go Hokies!!	Yes	Share
12	Hello! I hope you're having a wonderful day :)	No	Reject
13	Don't YAWN if you yawn I am gonna pick C. If you do not YAWN I am probably gonna pick C... End of story, you may YAWN and I am gonna pick C	Yes	Share
14	If there was a meme or gif that could convey to you that I was picking C, I would send it to you 13 times not 17 because I want 13	Yes	Share
15	i know there is no way in hell we're making it to F so if you hook me up by picking B I will choose C	Yes	Share

Supplementary Table 7. The Effect of Communication on P1's *Reject* Plan (1st 10 Periods vs. Full Sample)

	First 10 Periods		Full Sample	
	A	B	C	D
	mfx / se	mfx / se	mfx / se	mfx / se
Cost of *Reject*	-0.0633***	-0.0544***	-0.0736***	-0.0677***
	(0.0118)	(0.0104)	(0.0080)	(0.0070)
High *Take*	0.0211*	0.0590***	0.0087	0.0272***
	(0.0120)	(0.0152)	(0.0079)	(0.0082)
Communication	0.1027	0.0666	0.0552***	0.0449**
	(0.0690)	(0.0664)	(0.0173)	(0.0176)
Period	0.0264***	0.0268***	0.0122***	0.0118***
	(0.0035)	(0.0027)	(0.0015)	(0.0014)
Belief about *Share*		0.2499***		0.1230***
		(0.0391)		(0.0400)
Observations	1000	1000	2000	2000
BIC	977.6	949.9	589.4	571.6
Subject controls	No	No	Yes	Yes

* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

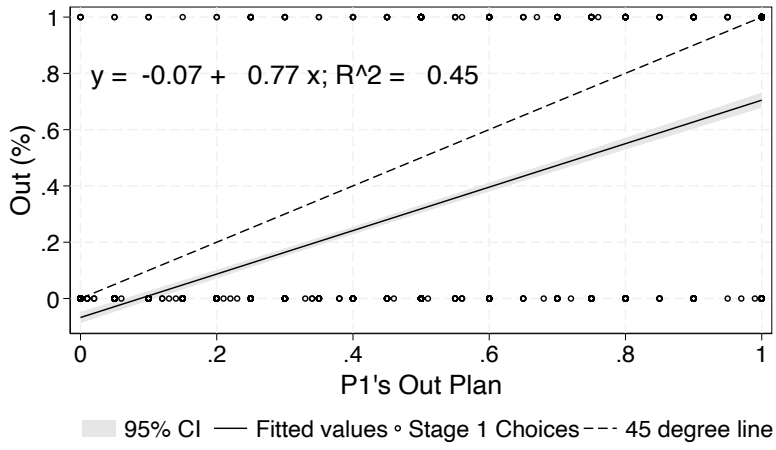
Note: mfx: marginal effect. Marginal effects for continuous variables are evaluated at means, and for binary variables are evaluated as the discrete change from 0 to 1. SE: standard error. Standard errors are bootstrapped at the session level. Linear regressions are employed for the first 10 periods, and fixed effect linear regressions are employed for the full sample.

Supplementary Table 8. The Effect of Communication on Outcomes by Game Form

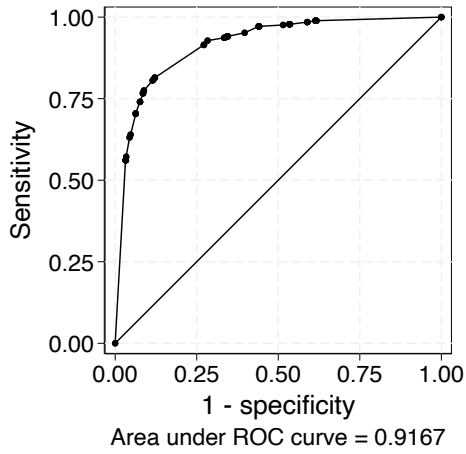
High Take										
Cost of *Reject*	No Communication					Communication				
	Out	*Share*	*Reject*	*Accept*	Total	*Out*	*Share*	*Reject*	*Accept*	Total
1	21	70	8	1	100	12	78	9	1	100
2	23	55	13	9	100	13	64	17	6	100
3	21	39	19	21	100	20	56	15	9	100
4	21	23	22	34	100	25	31	16	28	100
5	26	10	10	54	100	28	20	10	42	100
Total	112	197	72	119	500	98	249	67	86	500
Low Take										
Cost of *Reject*	No Communication					Communication				
	Out	*Share*	*Reject*	*Accept*	Total	*Out*	*Share*	*Reject*	*Accept*	Total
1	58	39	3	0	100	46	53	0	1	100
2	44	48	6	2	100	20	76	2	2	100
3	28	57	7	8	100	13	81	2	4	100
4	14	68	5	13	100	12	70	7	11	100
5	7	58	4	31	100	6	72	4	18	100
Total	151	270	25	54	500	97	352	15	36	500
Grand Total	263	467	97	173	1000	195	601	82	122	1000

Note: Each cell reports the number of observations of a given outcome, out of 100 total, for the indicated game form and communication condition; values can thus be read as percentages. Game forms are ordered by the cost of *Reject* (b) within the Low *Take* (LT1–LT5) and High *Take* (HT1–HT5) blocks; see Table 1 for parameters.

C Beliefs, Plans, and Behavior

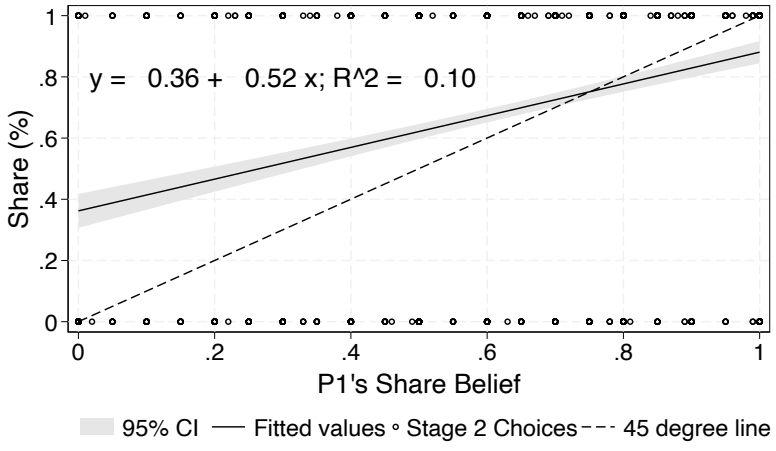


(a) Linear fit

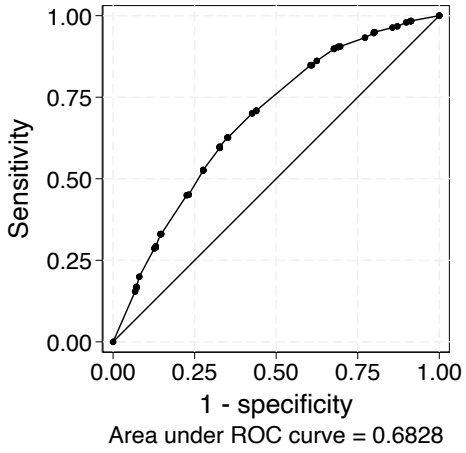


(b) ROC curve

Supplementary Figure 1. P1's *Out* choices vs. P1's Plans

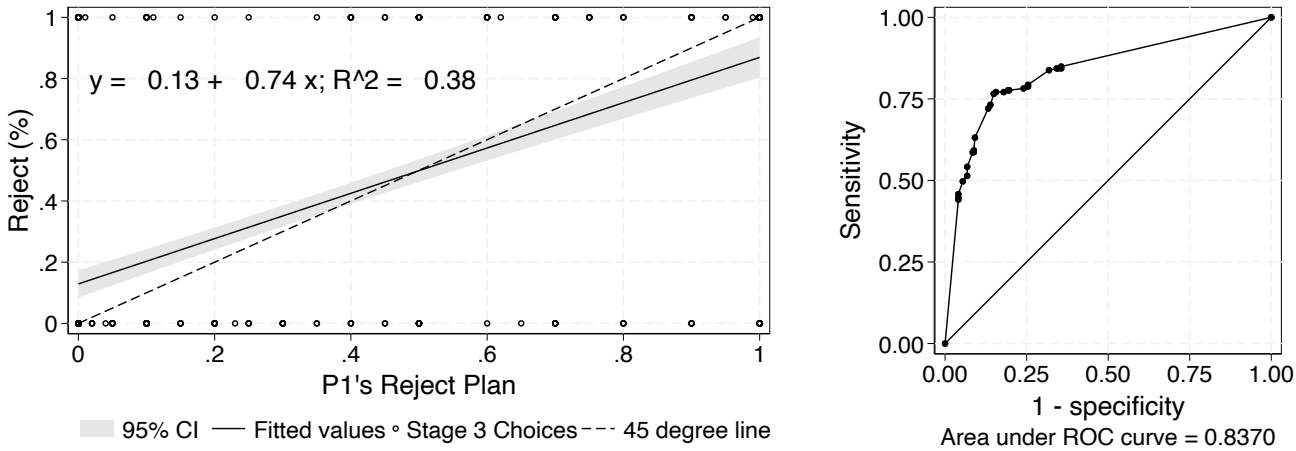


(a) Linear fit



(b) ROC curve

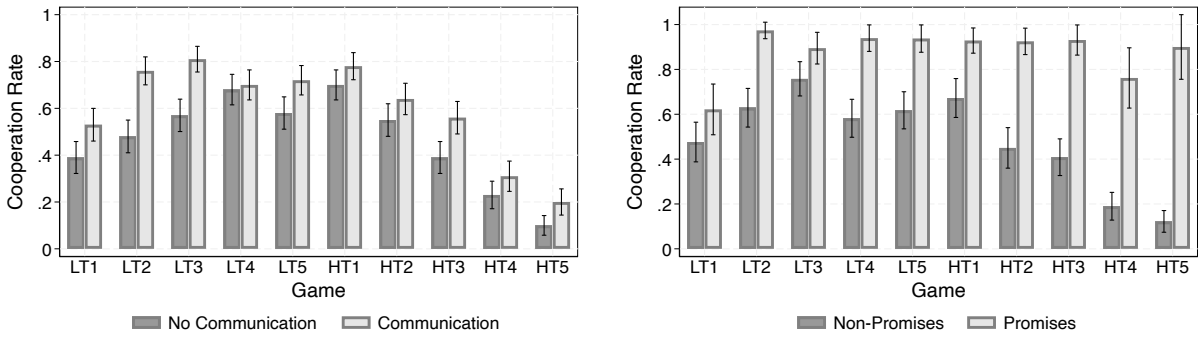
Supplementary Figure 2. P2's *Share* choices vs. P1's Beliefs



(a) Linear fit **(b)** ROC curve

Supplementary Figure 3. P1's *Reject* choices vs. P1's Plans

D Communication, Promises, and Cooperation



(a) High cooperation with communication. **(b)** High cooperation with promises.

Supplementary Figure 4. Cooperation Rate by Communication and Promises.

References

Aina, C., Battigalli, P., and Gamba, A. (2020). Frustration and anger in the ultimatum game: An experiment. *Games and Economic Behavior*, 122:150–167.

Attanasi, G., Battigalli, P., Manzoni, E., and Nagel, R. (2025). Disclosure of belief–dependent preferences in a trust game. *Economic Theory*, pages 1–38.

Attanasi, G., Rimbaud, C., and Villeval, M. C. (2023). Guilt aversion in (new) games: Does partners' payoff vulnerability matter? *Games and Economic Behavior*, 142:690–717.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *American Economic Review*, 97(2):170–176.

Battigalli, P. and Dufwenberg, M. (2009). Dynamic psychological games. *Journal of Economic Theory*, 144(1):1–35.

Battigalli, P. and Dufwenberg, M. (2022). Belief-dependent motivations and psychological game theory. *Journal of Economic Literature*, Forthcoming.

Battigalli, P., Dufwenberg, M., and Smith, A. (2015). Frustration and anger in games. *CESifo Working Paper Series No. 5258*.

Battigalli, P., Dufwenberg, M., and Smith, A. (2019). Frustration, aggression, and anger in leader-follower games. *Games and Economic Behavior*, 117:15–39.

Berkowitz, L. (1989). Frustration-aggression hypothesis: Examination and reformulation. *Psychological Bulletin*, 106(1):59.

Blanco, M., Engelmann, D., Koch, A. K., and Normann, H.-T. (2010). Belief elicitation in experiments: is there a hedging problem? *Experimental Economics*, 13(4):412–438.

Bolton, G. E. and Ockenfels, A. (2000). ERC: A theory of equity, reciprocity, and competition. *American Economic Review*, pages 166–193.

Bosman, R. and Van Winden, F. (2002). Emotional hazard in a power-to-take experiment. *The Economic Journal*, 112(476):147–169.

Brandts, J. and Charness, G. (2011). The strategy versus the direct-response method: a first survey of experimental comparisons. *Experimental Economics*, 14(3):375–398.

Cartwright, E. (2019). A survey of belief-based guilt aversion in trust and dictator games. *Journal of Economic Behavior & Organization*, 167:430–444.

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601.

Che, Y.-K. and Sákovics, J. (2008). *Hold-Up Problem*. In: Palgrave Macmillan (eds) The New Palgrave Dictionary of Economics. Palgrave Macmillan, London.

Crawford, V. P. (2016). New directions for modelling strategic behavior: Game-theoretic models of communication, coordination, and cooperation in economic relationships. *Journal of Economic Perspectives*, 30(4):131–50.

Cuzick, J. (1985). A wilcoxon-type test for trend. *Statistics in medicine*, 4(1):87–90.

Di Bartolomeo, G., Dufwenberg, M., and Papa, S. (2023). Promises and partner-switch. *Journal of the Economic Science Association*, 9(1):77–89.

Dollard, J., Miller, N. E., Doob, L. W., Mowrer, O. H., and Sears, R. R. (1939). *Frustration and aggression*. Yale University Press.

Dufwenberg, M. (2002). Marital investments, time consistency and emotions. *Journal of Economic Behavior & Organization*, 48(1):57–69.

Dufwenberg, M. and Gneezy, U. (2000). Measuring beliefs in an experimental lost wallet game. *Games and economic Behavior*, 30(2):163–182.

Dufwenberg, M. and Kirchsteiger, G. (2004). A theory of sequential reciprocity. *Games and Economic Behavior*, 47(2):268–298.

Dufwenberg, M., Li, F., and Smith, A. (2026). Credible threats. Manuscript in preparation.

Dufwenberg, M., Smith, A., and Van Essen, M. (2013). Hold-up: With a vengeance. *Economic Inquiry*, 51(1):896–908.

Ellingsen, T. and Johannesson, M. (2004a). Is there a hold-up problem? *Scandinavian Journal of Economics*, 106(3):475–494.

Ellingsen, T. and Johannesson, M. (2004b). Promises, threats and fairness. *The Economic Journal*, 114(495):397–420.

Epps, T. W. and Singleton, K. J. (1986). An omnibus test for the two-sample problem using the empirical characteristic function. *Journal of Statistical Computation and Simulation*, 26(3-4):177–203.

Fehr, E. and Gächter, S. (2000). Fairness and retaliation: The economics of reciprocity. *Journal of Economic Perspectives*, 14(3):159–181.

Fehr, E. and Schmidt, K. M. (1999). A theory of fairness, competition, and cooperation. *Quarterly Journal of Economics*, 114(3):817–868.

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178.

Geanakoplos, J., Pearce, D., and Stacchetti, E. (1989). Psychological games and sequential rationality. *Games and Economic Behavior*, 1(1):60–79.

Grossman, S. J. and Hart, O. D. (1986). The costs and benefits of ownership: A theory of vertical and lateral integration. *Journal of Political Economy*, 94(4):691–719.

Grout, P. A. (1984). Investment and wages in the absence of binding contracts: a Nash bargaining approach. *Econometrica*, pages 449–460.

Hart, O. and Moore, J. (1990). Property rights and the nature of the firm. *Journal of Political Economy*, 98(6):1119–1158.

Hart, O. and Moore, J. (2008). Contracts as reference points. *Quarterly Journal of Economics*, 123(1):1–48.

Hosmer Jr, D. W., Lemeshow, S., and Sturdivant, R. X. (2013). *Applied logistic regression*. John Wiley & Sons.

Klein, B., Crawford, R. G., and Alchian, A. A. (1978). Vertical integration, appropriable rents, and the competitive contracting process. *Journal of Law and Economics*, 21(2):297–326.

Krupka, E. L., Leider, S., and Jiang, M. (2017). A meeting of the minds: Informal agreements and social norms. *Management Science*, 63(6):1708–1729.

Murphy, K. P. (2012). *Machine learning: a probabilistic perspective*. MIT Press.

North, D. C. and Weingast, B. R. (1989). Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century england. *Journal of Economic History*, 49(4):803–832.

Persson, E. (2018). Testing the impact of frustration and anger when responsibility is low. *Journal of Economic Behavior & Organization*, 145:435–448.

Rutström, E. E. and Wilcox, N. T. (2009). Stated beliefs versus inferred beliefs: A methodological inquiry and experimental test. *Games and Economic Behavior*, 67(2):616–632.

Schotter, A. and Trevino, I. (2014). Belief elicitation in the laboratory. *Annual Review of Economics*, 6(1):103–128.

Tirole, J. (1986). Procurement and renegotiation. *Journal of Political Economy*, 94(2):235–259.

Toussaert, S. (2018). Eliciting temptation and self-control through menu choices: a lab experiment. *Econometrica*, 86(3):859–889.

Trautmann, S. T. and Kuilen, G. (2015). Belief elicitation: A horse race among truth serums. *The Economic Journal*, 125(589):2116–2135.

Vanberg, C. (2008). Why do people keep their promises? An experimental test of two explanations. *Econometrica*, 76(6):1467–1480.

Williamson, O. E. (1971). The vertical integration of production: market failure considerations. *American Economic Review*, 61(2):112–123.

Xiao, E. and Houser, D. (2011). Punish in public. *Journal of Public Economics*, 95(7-8):1006–1017.

Yang, Y. (2021). A survey of the hold-up problem in the experimental economics literature. *Journal of Economic Surveys*, 35(1):227–249.