Promises and Punishment*

Martin Dufwenberg[†] Flora Li[‡] Alec Smith[§]

August 30, 2021

Abstract

We study the effect of communication on beliefs, behavior, and efficiency in the context of hold-up problems with a punishment option. 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. 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 that costly punishment results from belief-dependent anger and frustration. Promises drive the effect of communication on beliefs and broken promises lead to higher rates of costly punishment.

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

[*] We are particularly grateful to Pierpaolo Battigalli, Gary Charness, Uri Gneezy, and many seminar and conference participants for their helpful comments. The Center for Peace Study and Violence Prevention at Virginia Tech provided financial support.

[†] University of Arizona; University of Gothenburg; CESifo. martind@eller.arizona.edu.

[‡] Corresponding author. Economics Experimental Lab, Nanjing Audit University. florali@nau.edu.cn

[§] Department of Economics, Virginia Tech. alecsmith@vt.edu.

1 Introduction

Communication can foster trust and cooperation. A recent literature explores why people keep their promises, focusing on the motivation of the promisor.[1] We explore a new and complementary explanation, whereby it is the promisee that is affected. If a promise is broken this induces dashed hopes and frustration, which triggers anger and aggression.[2] If anticipated, this creates an incentive for promisors not to renege.

We develop this idea for environments which augment a simple trust game with a punishment stage. The resulting structures 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.[3] We explore such settings both theoretically and experimentally:

First, we apply the model of frustration and anger from Battigalli et al. (2015, 2019) (BDS) to a three-stage hold-up problem, allowing us to examine the impact of communication in general and promises in particular. The basic ideas are: 1) decision-makers experience anger when they are frustrated; 2) frustration results when material payoffs are less than expected; and 3) anger leads to aggression and the urge to retaliate; 4) promises may enhance the just-mentioned effects, by shaping expectations such that promise-keeping is expected; 5) if these effects are anticipated, trust and cooperation ensues. 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.[4] 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. 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. Ellingsen and Johannesson (2004a,b), who also study communication and hold-up in an experiment, are important precursors to our

[1] Charness and Dufwenberg (2006) develop a theoretical argument based on "guilt aversion" and Vanberg (2008) similarly explores "a preference for keeping one's word." Some of the large subsequent literature is surveyed by Cartwright (2019).

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

[3] 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.

[4] The approach involves belief-dependent utilities and draws on the framework of psychological game theory (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009, 2020).

study.[5] 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]

We limit our attention to messages from the second-mover to the first-mover in this paper. 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, 2021).

Section 2 presents theory. We describe the games we study, apply BDS' model of belief-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 results. Section 5 concludes.

2 Theory

2.1 A hold-up game with costly punishment

We study a class of 2-player, 3-stage games, as shown in Figure 1, where the numbers and variables at end nodes represent monetary payoffs. The game may be interpreted as a mini-trust game with an added punishment option, an ultimatum mini-game with an added 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 monetarily

[5] See Yang (2021) for a recent review of the experimental literature on hold-up problems.

[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.

[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 of the effect of broken promises we focus on a hold-up environment that allows for costly punishment after opportunistic behavior.

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 $(In, Accept), Take)$. 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 $((Out, Accept); Take)$; when $b = 5$ and players care only for monetary payoffs, there are two SPEs: $((Out, Accept); Take)$ and $((In, Accept); Take)$.

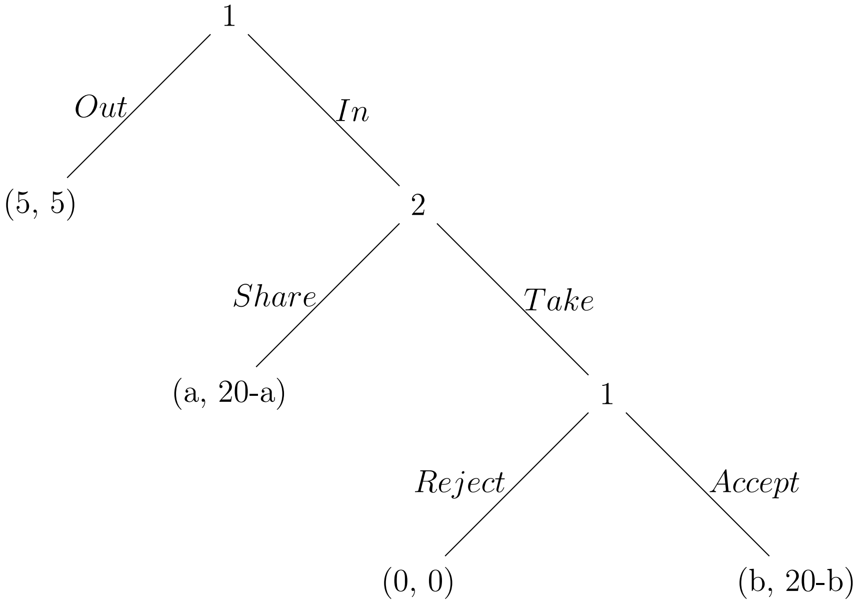


Figure 1. A hold-up game with punishment.

2.2 Frustration and anger

We apply the frustration and anger model of Battigalli et al. (2019).[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

[8] Battigalli et al. (2019) 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). Battigalli et al. (2019) focus on two-stage "leader-follower" games; an earlier working paper (Battigalli et al., 2015, Section 6) develops an extension to general multi-stage games.

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

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.

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

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

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

In the game forms defined in Figure 1, if Player 2 gets the move her maximal payoff is still available. According to Equation 1, she cannot be frustrated ($F_2(In;\alpha_2) = 0$), and so her behavior after *In* is indistinguishable from material payoff maximization. We therefore focus our analysis on the beliefs and behavior of Player 1.

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 higher order beliefs are correct in the sense that the marginals of the higher order beliefs are equal to the lower order beliefs. In equilibrium, the belief systems of both players coincide, so we may drop the subscripts and generically refer to beliefs p, q, and r.

If Player 1's sensitivity to anger θ_1 is sufficiently large, this (psychological) game has a unique sequential equilibrium (SE) $((In, Reject); Share)$ 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 beliefs 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

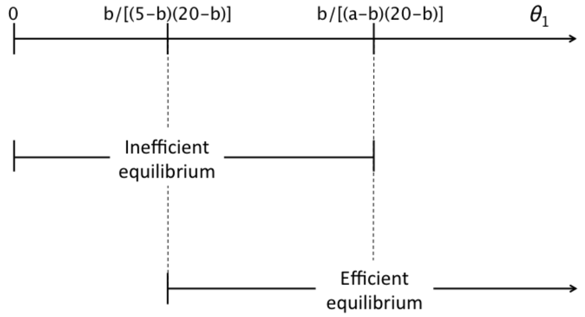


Figure 2. Sequential Equilibria as a Function of the Anger Sensitivity θ_1 of Player 1.

$u_1 = b - \theta_1(a-b)(20-b)$ from *Accept*. Given equilibrium beliefs, Player 1 will *Reject* if $\theta_1 > \frac{b}{(a-b)(20-b)}$, demonstrating the uniqueness of the efficient equilibrium for large θ_1.

If θ_1 is small, then the unique SE coincides with the subgame perfect Nash equilibrium for self-interested players $((Out, Accept); Take)$, with beliefs $p = 1, q = 0, r = 0$ for both players. Player 1's initial expected material payoff is $5p + a(1-p)q + b(1-p)(1-q)(1-r) = 5$. Experienced frustration equals to $5-b$ if stage 3 is realized. Player 1 compares 0 to $u_1 = b - \theta_1(5-b)(20-b)$, and chooses *Accept* if $\theta < \frac{b}{(5-b)(20-b)}$.

For intermediate values of θ_1 there are two SE in pure strategies, the efficient one and the inefficient one. Figure 2 shows the SE as a function of the anger sensitivity of Player 1.

2.3 Communication and Promises

With belief-independent preferences, communication and promises affect neither expectations nor behavior. The belief-dependent frustration-anger model, by contrast, suggests a channel whereby particular messages can play a role. Under the assumption that certain message types influence beliefs, the model implies that, with promises, Player 1 is more likely to trust Player 2 and to invest, Player 2 is more likely to keep her promises, and Player 1 is more likely to punish broken promises.

3 Experiment

To study the effect of promises on trust and punishment we implemented a laboratory experiment with the class of games depicted in Figure 1. 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 programed 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.[9] Each session included 14-20 participants with an average of 18.4 per session. Sessions took about 1.75 hours to complete.

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.[10]

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 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 1, in

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

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

random order. The game variations are shown in Table 1, where all the numbers are in US dollars. A change of the parameter b (Payoff from *Accept*) indicates changing the cost of punishment, and we vary the cost of punishment from 1 to 5. The difference $a - b$ indicates the "*Take* amount": 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 saliency of an equal split.

Table 1. Experiment Design – Game Structure.

Game	a	Cost of Punishment (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 experiment designs to make use of the "strategy method", where players 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 experimentally that frustration and anger are more relevant with direct response method than with strategy method, and 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, compared to strategy method. Therefore, we used a direct response design, such that players move sequentially.

3.2 Belief Elicitation

During the experiment we elicited 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 own plan of choosing *Reject* (r_1) conditional on entering the 3rd stage. 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).[11] 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. The problem is exacerbated when eliciting beliefs about a player's own future behavior.[12] 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 assumption that diminished payoff expectations make aggression and costly punishment more attractive.

Hypothesis 1. *Player 1's higher 1st order belief about Player 2's probability of cooperation*

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

[12] 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 also do not employ methods that involve proxies such as similar others or the average belief in the room (as in Charness and Dufwenberg, 2006).

leads to higher rates of Reject choices.

Besides beliefs, game structure can influence Player 1's behavior as predicted by the belief-dependent anger theory. A unique efficient sequential equilibrium $((In, Reject); Share)$ can be achieved when anger sensitivity parameter $\theta_1 > \frac{b}{(a-b)(20-b)}$. Therefore, as the *Take* amount $(a-b)$ increases, it is easier for Player 1 to reach to the threshold for θ_1 to *Reject*. Similarly, as the cost of punishment (b) increases, it is harder for Player 1 to reach to the threshold for θ_1 to *Reject*.

Hypothesis 2. *Player 1 is more likely to Reject when the Take amount (a-b) is high; Player 1 is less likely to Reject as the cost of punishment (b) increases.*

We next turn to the effect of communication. We expected that communication would increase the frequency of cooperative outcomes and improves efficiency, consistent with a number of studies of communication and cooperation (Charness and Dufwenberg, 2006; Balliet, 2010; Battigalli et al., 2013), and studies of communication and efficiency (Blume and Ortmann, 2007; Avoyan and Ramos, 2020; Fehr and Sutter, 2019).

Hypothesis 3. *Communication increases the frequency of cooperative outcomes and improves efficiency.*

Motivated by the results of Charness and Dufwenberg (2006) and the subsequent literature, we also hypothesized that the content of the free-form messages would play an important role in connecting communication with behavior via beliefs. In particular, we predicted that promises would change beliefs and plans in the direction of increased investment, cooperation, and punishment. Hypotheses 1 and 4 connect communication and costly punishment via the effect of communication on beliefs. With regard to message content, we hypothesize:

Hypothesis 4. *Communication influences beliefs via promises, such that promises shift Player 1's reported beliefs in the direction of increased likelihood of investment, cooperation, and costly punishment; but cheap talk (non-promise messages) has no impact over beliefs.*

We predicted that the effect of promises on beliefs would carry through to behavior, through the mechanism of belief-dependent anger as described in Hypothesis 1. In particular, an implication of the frustration-anger model is that 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:

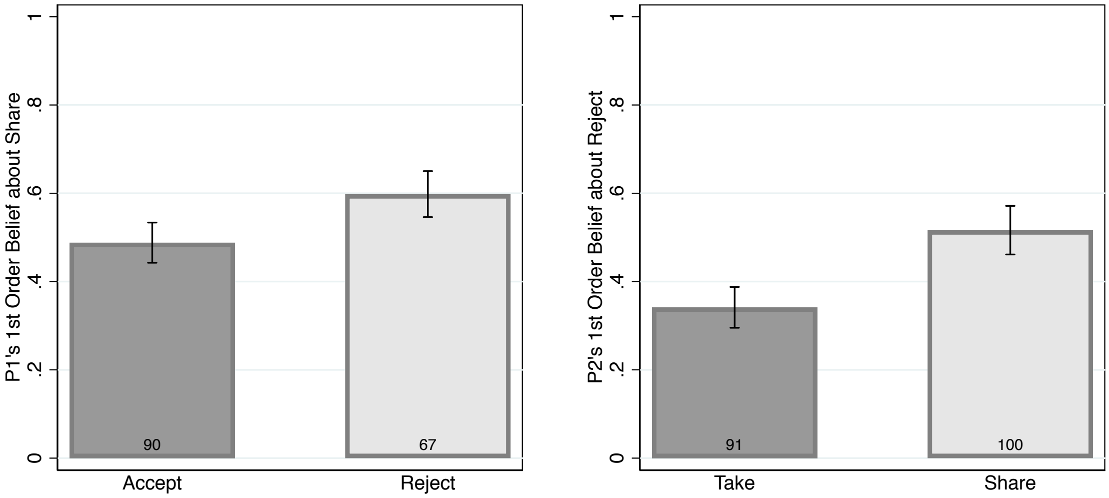
Hypothesis 5. *Broken promises lead to a higher rejection rate, and promises lead to a higher cooperation rate relative to cheap talk*

4 Results

We begin our examination of the results by evaluating the relationship between self-reported beliefs and behavior and the relationship between game structure and behavior. Next, we measure the effect of communication on behavior and welfare. We then investigate how message content in the form of promises influences beliefs and behavior.

4.1 Belief-Dependent Anger

We first evaluate the correlations between self-reported beliefs and subjects' choices (Hypothesis 1). Spearman's rho shows a positive correlation between Player 1's *Reject* choice and belief about Player 2's *Share* (rho $=$ 0.222 p-value $<$ 0.001) and a positive correlation between Player 2's *Share* decision and belief about Player 1's *Reject* (rho $=$ 0.212 p-value $<$ 0.001). Figure 3 further demonstrates that players' behavior is consistent with belief-dependent anger. Figure 3(a) shows that Player 1 who chooses *Reject* over *Accept* reports a higher 1st order belief about Player 2's *Share* (rank sum test p-value $=$ 0.002). Figure 3(b) shows that Player 2 who believes that Player 1 is more likely to *Reject* is more likely to choose *Share* (rank sum test p-value $<$ 0.001). The above results are conforming with Hypotheses 1.



(a) P1's *Share* Belief vs. Reject Choice. **(b)** P2's *Reject* Belief vs. *Share* Choice.

Figure 3. Relationship Between Beliefs and Behavior.

To study the role of beliefs in driving costly punishment, in Table 2 Columns A-B, we run fixed effect logistic regressions with the dependent variable reject = 1 if Player 1 *Reject*s the offer in stage 3, and reject = 0 if Player 1 *Accepts* the offer in stage 3. The key to evaluate the belief-dependent anger is to investigate how beliefs about *Share* influence *Reject* behavior (Hypothesis 1), since models of self interest and of distributional preferences imply that these beliefs should have no impact on behavior in the 3rd stage of the game, after controlling for the cost of punishment (b), high *Take* condition (*Take* amount $a - b = 10$), communication treatment, and round number. We find, in Column B, a significant relationship between Player 1's first order belief about *Share* and decision to *Reject* the offer after *Take.* A 10% increase in Belief about *Share* increases Player 1's chance of rejecting by 2.488%, which is consistent with Hypothesis 1.

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

	P1's *Reject* Choice		P1's *Reject* Plan	
	A	B	C	D
	mfx / se	mfx / se	mfx / se	mfx / se
Cost of Punishment	-0.1985***	-0.1814***	-0.0736***	-0.0677***
	(0.0219)	(0.0245)	(0.0077)	(0.0069)
High *Take*	0.0442	0.0769	0.0087	0.0272***
	(0.0584)	(0.0606)	(0.0078)	(0.0082)
Communication	0.0174	0.0111	0.0552***	0.0449***
	(0.0494)	(0.0483)	(0.0170)	(0.0167)
Period	0.0137***	0.0129***	0.0122***	0.0118***
	(0.0044)	(0.0045)	(0.0014)	(0.0017)
Belief about *Share*		0.2488**		0.1230***
		(0.1173)		(0.0360)
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. Fixed effect logistic regressions are employed for P1's *Reject* Choice, and fixed effect linear regressions are employed for P1's *Reject* Plan.

As reported in Table 2, only 474 out of 2000 observations survive in third stage. Therefore, we use Player 1's *Reject* plan (Player 1's reported belief about *Reject* at the start of the game) as a proxy for Player 1's actual behavior in 3rd stage. Spearman's rho shows significant correlations between plans and choices (Player 1's *Out*: rho = 0.617 p-value < 0.001; Player

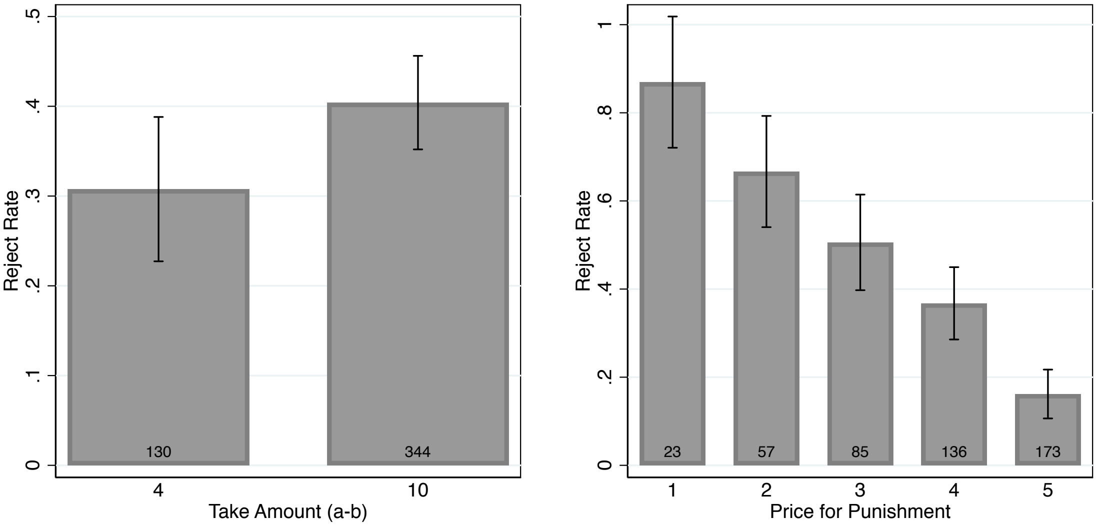
2's *Share*: rho = 0.430 p-value < 0.001; Player 1's *Reject*: rho = 0.598 p-value < 0.001). Supplementary Figure 3 shows Receiver Operating Characteristic (ROC) curves relating plans with behavior. The reported plans are good predictors of subsequent decisions, with Area Under the Curve (AUC) greater than 0.8 for each measure.

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). We observe that in Column D, with subject level control, Player 1's first order belief about *Share* is significantly infleuncing Player 1's *Reject* plan. Since *Reject* plan serves as a high quality proxy of *Reject* choice, we find strong evidence that Player 1's behavior is consistant with belief-dependent anger (Hypothesis 1).

4.2 The Effect of the *Take* Amount

Our experimental design varies both the *Take* amount $(a - b)$ and the cost of punishment (b). Figure 4(a) shows that Player 1 is more likely to *Reject* in the third stage when the *Take* amount is high (1-sided Fisher's exact p-value = 0.033; rank sum test p-value = 0.054; 1-sided t test p-value = 0.027), consistent with Hypothesis 2. This suggests that participants are motivated by more than distributional preferences, as the *Take* amount is not instrumental to Player 1's decision after *Take*. Figure 4(b) shows that as the cost of punishment increases, Player 1 is less likely to *Reject* (Fisher's exact p-value < 0.001; Spearman's rho = -0.413 p-value < 0.001), confirming that participants are also sensitive to the cost of punishment (Hypothesis 2).

After including controls for communication, round number, and Player 1's belief about *Share*, with subject level control, fixed effect linear regression shows that both the cost of punishment and the high *Take* condition predict Player 1's *Reject* plan in Table 2 Column D. In the regression, as predicted by Hypothesis 2, an increase in the cost of punishment is associated with a significant decrease in reported plans to choose *Reject*. Consistent with Hypothesis 2, after controlling for the cost of punishment, participants indicate a greater likelihood of choosing *Reject* in the High *Take* condition.



(a) *Reject* Rate vs. *Take* Rate. **(b)** *Reject* Rate vs. Cost of Punishment.

Figure 4. *Reject* Rate by Game Structure.

4.3 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, and then measure the effect of allowing communication affects overall welfare and efficiency. We then consider how communication influences participants' probabilistic beliefs about their own and their co-player's future actions.

As predicted by Hypothesis 3, communication has a strong effect on efficiency and cooperation. Figure 5 compares outcomes from the no-message and message blocks, pooling the data from all sessions. The cooperation outcomes are more prevalent in the communication treatment (60.10% vs. 46.70%), see Table 3. 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 models of frustrated anger and guilt aversion and with Hypothesis 3, that communication increases cooperative outcomes. A chi-squared test shows that the communication treatment 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 p-value $= 0.197$).

Table 3. The Effect of Communication.

	Out	Cooperation	Rejection	Acceptance	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: fraction of total observations; row 3: fraction of observations reaches to the third stage.

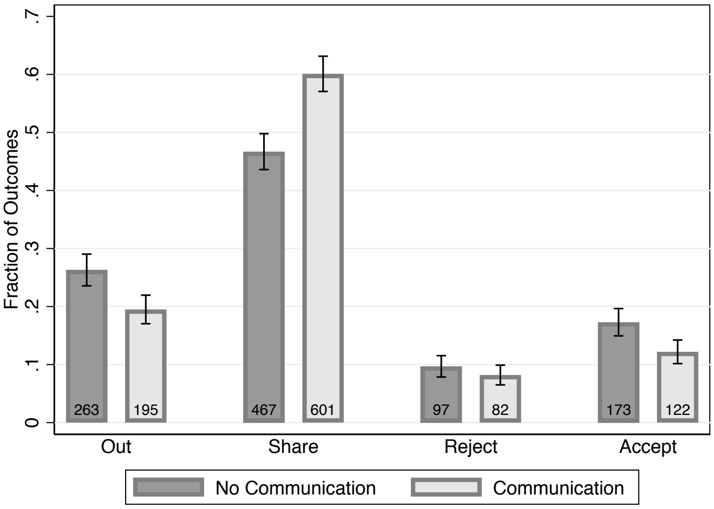


Figure 5. Outcomes and Communication Summary.

To test whether communication improves efficiency, we compare average payoffs in both treatments. Figure 6 shows that Player 2's average payoffs are insignificantly higher in the communication treatment; whereas, Player 1's average payoffs significantly increase \$0.68 from no-communication to communication treatment (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.

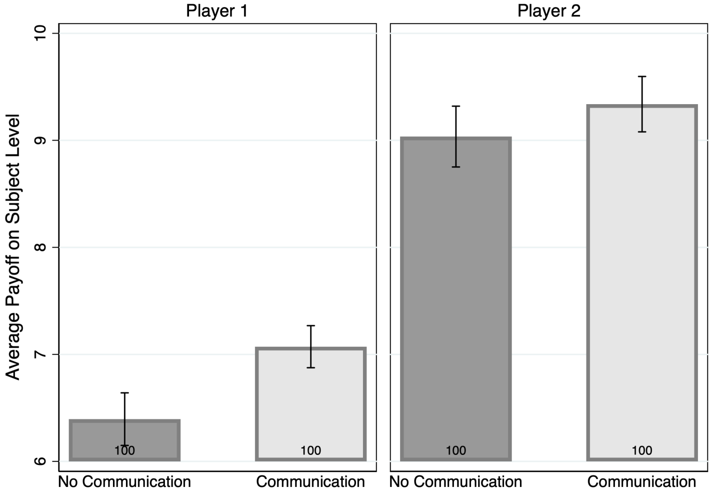


Figure 6. The Effect of Communication on Payoffs by Player Role.

We now consider the effect of communication on beliefs. Figure 7 shows that communication affects Player 1's reported beliefs, and this result is consistent with our Hypothesis 4. Player 1 believes that Player 2 will cooperate with higher probability (1st order belief about *Share*) 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 for *Out* p-value = 0.009, 1st order belief about *Share* p-value < 0.001, plan for *Reject* p-value = 0.069). In addition, we observe that the direction of how communication influences expectations is consistent with belief-dependent anger.

There are some difference between sessions with communication first and sessions with communication second. We discuss this further in Appendix A.

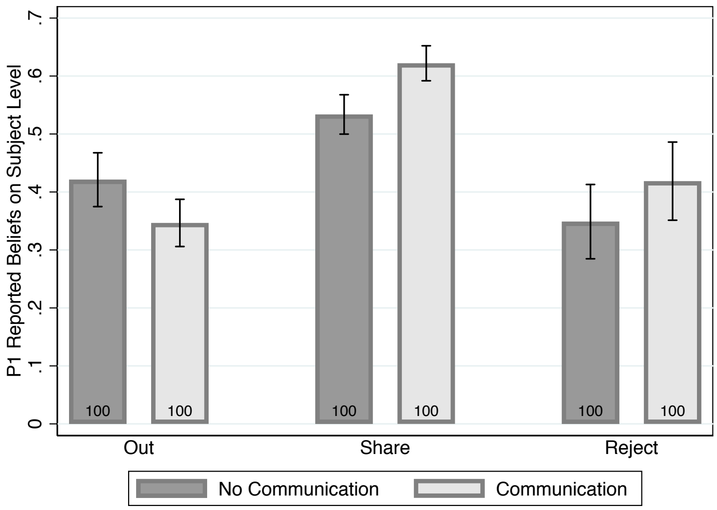


Figure 7. Communication Influences P1's Reported Beliefs.

4.4 Promises

In this section, we restrict attention to data from the Communication block and examine the role of message content. We categorize messages as either Promises or Cheap Talk, and then examine the differential effect of promises on beliefs and behavior.

The Effect of Promises on Beliefs

To examine the relationship among message content, beliefs, and behavior, we manually coded messages as promises if they follow the pattern of "If you choose *In*, I will *Share*." Using this approach, we identify 32% of messages as promises, and the median number of promises per session was 32.2%.

Figure 8 shows that promises produce 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 (plan for *Out* and *Reject*). Player 1s report that they will be less likely to choose *Out*, but will be more likely to punish Player 2 after receiving a promise. 1-sided t-tests show a significant difference in Player 1's reported beliefs change between promises and cheap talk (plan for *Out* p-value < 0.001, 1st order belief about *Share* p-value < 0.001, plan for *Reject* p-value $= 0.011$). In addition, two-sided t-tests confirm that Player 1's reported beliefs change with cheap talk is not different from 0 (plan for *Out* p-value $= 0.779$, 1st order belief about

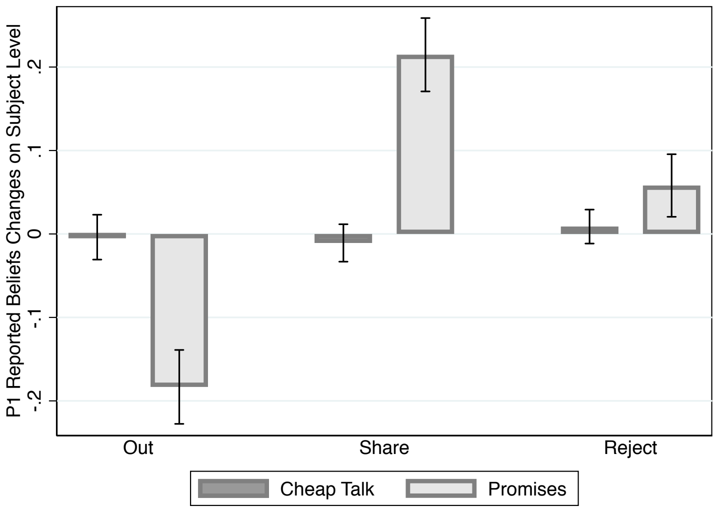


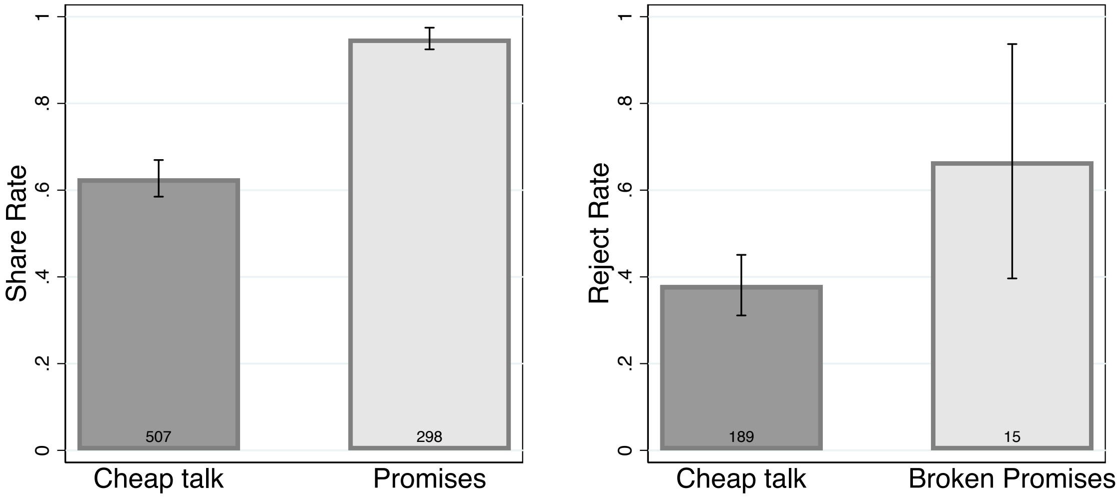
Figure 8. Belief Change After Receiving a Message.

Share p-value = 0.343, plan for *Reject* p-value = 0.390). The result indicates that promises have a significant effect upon beliefs, but that non-promise messages (cheap talk) have an insignificant effect, consistent with Hypothesis 4.

The Effect of Promises on Behavior

To further demonstrate the effect of promises on behavior as prediced in Hypothesis 5, we look at beahvior differences under promises and cheap talk. Supplementary Table 1 shows the outcome distribution respect to promises and cheap talk 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). In Figure 9, we observe higher conditional *Share* and *Reject* rates when a promise is made, consistent with Hypothesis 5 that promises foster cooperation, but broken promises leads to higher level of punishments. The effect of promises is greater than the effect of communication, and messages other than promises have no effect on behavior.

The result shown in Figure 9(a) is consistent with the frustration-anger model that if Player 2 anticipates Player 1's beliefs change following promises, Player 2 will be motivated to choose *Share* to avoid punishment. When we compare Player 2's behavior after cheap talk 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 for individual games as



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

Figure 9. Kept and Broken Promises.

well. Supplementary Figure 4(b) shows that the *Share* rate for promises is higher across all 10 games compared to games with messages categorized as cheap talk. Promises affect not only cooperative behavior but also rejection. As predicted by the frustration-anger model, Player 1's beliefs change following promises, and Player 1 is more likely to punish with broken promises. Figure 9(b) shows that the *Reject* rate is higher with broken promises compared to cheap talk, 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).

4.5 Gender Differences

We started this project with BDS' theory in mind and the intention to test hypotheses 1-5. Our design records subjects' genders, but we didn't have any preconceived conjectures as regards whether results would differ between women and men. Aina et al. (2020), however, report that men are more effected by anger than women. As it turns out, we have comparable findings. Females and males' behavior are relatively close except for when promises are made, males tend to *Reject* more often (*Reject* rate 70% vs. 30%, rank sum test p-value = 0.077). We run the fixed effects linear regressions for *Reject* plan separately for females and males (Table 4). The communication effect survives with females, but disappears with males. Whereas, beliefs about *Share* is significant across models for males, but not for females.

Consistent with Aina et al. (2020), "High *Take*" and "Belief about *Share*" are significantly affecting Player 1's *Reject* plan for males but not for females. Therefore, we conclude that males' behavior is more inline with the frustration-anger model.

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

	Females		Males	
	A mfx / se	B mfx / se	C mfx / se	D mfx / se
Cost of Punishment	-0.0744***	-0.0703***	-0.0732***	-0.0668***
	(0.0144)	(0.0135)	(0.0083)	(0.0087)
High *Take*	-0.0027	0.0104	0.0165	0.0365**
	(0.0138)	(0.0148)	(0.0161)	(0.0157)
Period	0.0110***	0.0107***	0.0130***	0.0126***
	(0.0029)	(0.0031)	(0.0030)	(0.0030)
Communication	0.0592*	0.0539	0.0543*	0.0397
	(0.0333)	(0.0351)	(0.0319)	(0.0318)
Belief about *Share*		0.0734		0.1539***
		(0.0714)		(0.0308)
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.

5 Discussion

We study the effect of communication on strategic behavior in environments that allow for trust, promises, deception, and punishment. Communication increases cooperation and impacts beliefs. Beliefs are shaped by promises, and punishment increases with broken promises. The results support the idea that communication, beliefs, and costly punishment are linked through the mechanism of belief-dependent frustration and anger.

Experimental and behavioral economists have convincingly argued that models of social preferences are needed to explain human behavior, but little of such work factors in anger and frustration.[13] One may wonder if doing so is necessary. For example, can models of inequity aversion (e.g. Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) explain our results? 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 four categories, shown in Table 5. "IA Violation" represents subjects whose behavior is inconsistent with inequity aversion: they either *Reject* a higher and *Accept* a lower offer, or they both *Reject* and *Accept* the same offer (e.g. rejecting an offer of 3 in one round and accepting 3 in another). "Inequity Averse" subjects' behavior is always consistent with inequity aversion, "Self-interest" refers to players who always *Accept* any offer, and "Unclassified" are subjects that faced fewer than two different offers.

Table 5. Classification of Player 1 behavior.

	IA violation	Inequality averse (IA)	Self-interest	Unclassified
# of Subjects	36	28	33	3
# of 3rd Stage Decisions	5.42	4.79	4.27	1.33

Table 5 indicates that 36% of subjects are 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 self-interest. Moreover, the number of subjects whose behavior is inconsistent with inequity aversion or self-interest increases when subjects have more decisions in the 3rd stage. This suggest that inequity aversion cannot explain the behavior of at least one-third of our participants, and models that allow for non-consequential behavior such as BDS may be needed to fully capture the range of behavior demonstrated.

Another strand of models addresses subjects' tendency to honor promises, e.g. work on guilt aversion (see Charness and Dufwenberg (2006); compare Battigalli and Dufwenberg (2007)) or a direct preference to honor a promise (e.g. Vanberg, 2008). Related models help explain why communication increases the frequency of *Share* choices, but our results indicate

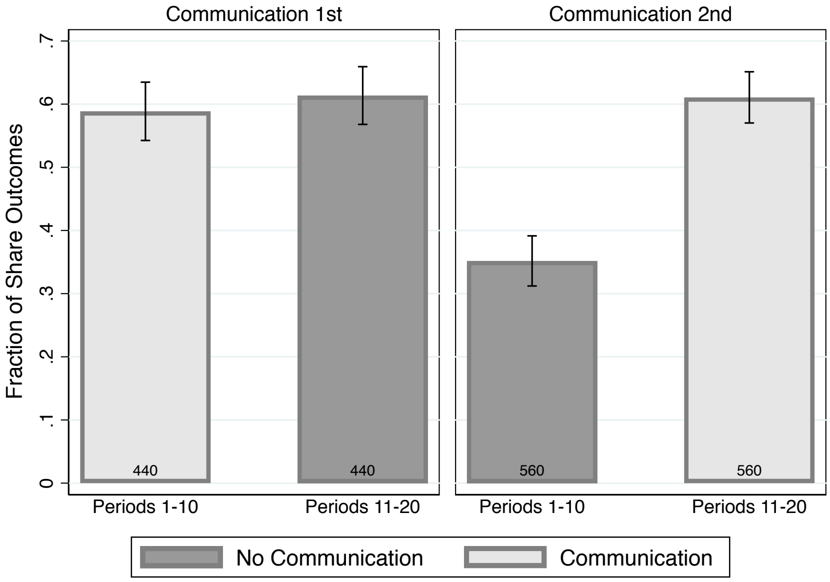
[13] BDS (2019, pp. 17, 29, 31) discuss previous attempts by economists to address frustration and anger, either theoretically or experimentally. Two of the experimental studies - Persson (2018) and Aina et al. (2020) - relate directly to BDS, although unlike us these authors do not explore issues of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

that frustration and anger in our game has additional effects. First, models of a tendency to honor promises first cannot explain the behavioral results we observe in the third stage of the games, regarding increased rates of punishment when promises are breached. Second, the overall rates of *Share* choices that we observe are much higher than in comparable studies that do not allow for a punishment stage (e.g. Charness and Dufwenberg, 2006).

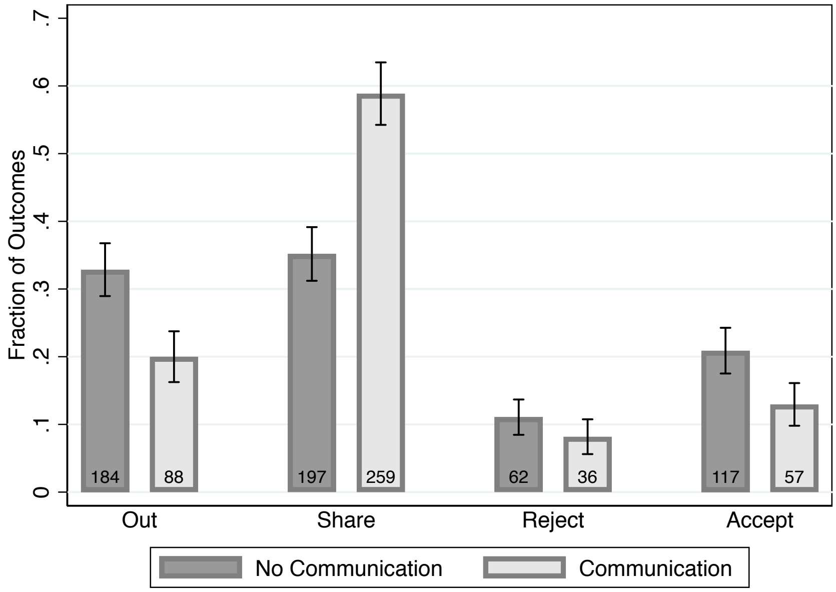
Appendices

A Persistent Communication Effect

There is a significant difference between sessions with communication first and sessions with communication second. Supplementary Figure 1 shows that there is a persistent effect of communication on outcomes. In the first 10 rounds, there is a significant higher cooperation rate in the communication-first sessions (58.86% vs. 35.18%, 1-sided Fisher's exact p-value < 0.001). This difference disappears in rounds 11-20 (61.36% vs. 61.07%, 2-sided Fisher's exact p-value $= 0.948$). This suggests that the communication effect is so strong that after being exposed to the communication environment, participants behave as if they are still sending and receiving messages, even in the no-communication treatment.



Supplementary Figure 1. Persistent Communication Effect.



Supplementary Figure 2. First 10 Round Outcome Summary.

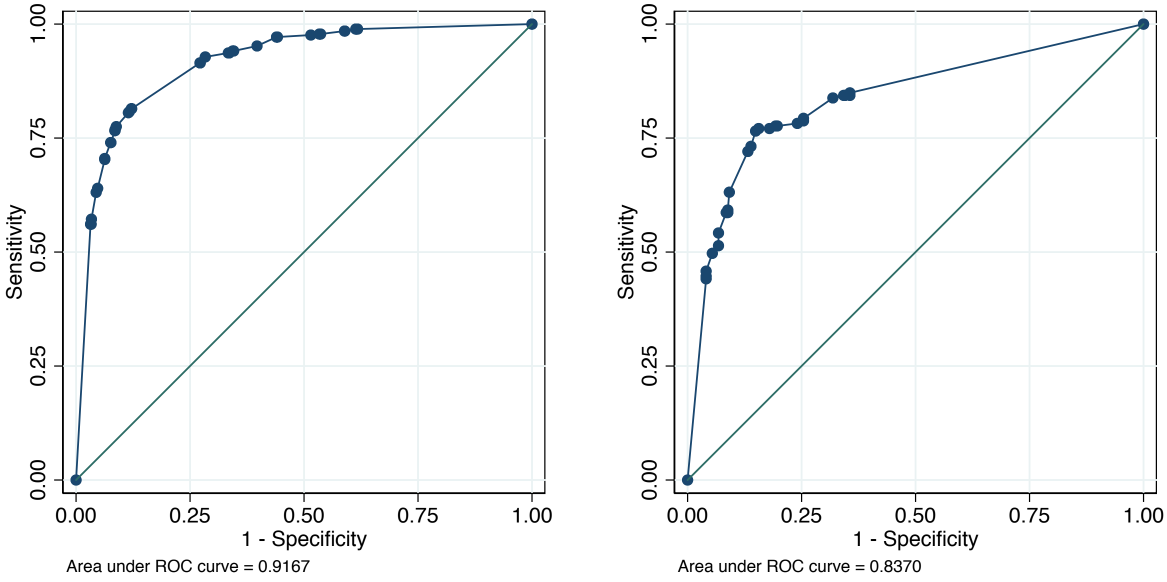
Because of the persistent effect of communication, we examine the distribution of outcomes after restricting the sample to include only the first 10 rounds. Supplementary Figure 2 shows the effects of communication on the distribution of outcomes in the first 10 rounds only, when the no-communication group has no experience with messages. The figure demonstrates a much stronger effect of communication.

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%), 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). The difference demonstrates that communication has a strong and persistent effect.

B Supplementary Graphs and Tables

Supplementary Table 1. The Effect of Promises on Outcomes.

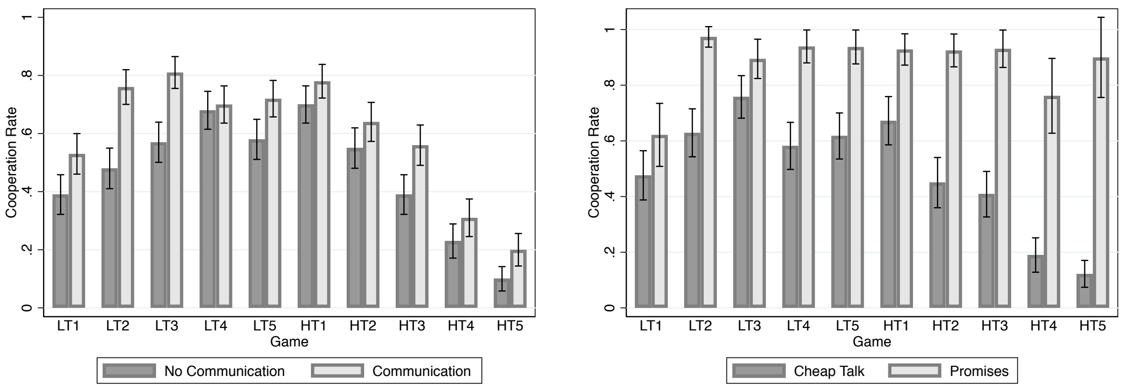
	Out	Cooperation	Rejection	Acceptance	Total
Promises	22	283	10	5	320
	6.88%	88.44%	3.12%	1.56%	100.00%
			66.67%	33.33%	100.00%
Cheap Talk	173	318	72	117	680
	25.44%	46.76%	10.59%	17.21%	100.00%
			38.10%	61.90%	100.00%
Total	195	601	82	122	1000
	19.50%	60.10%	8.20%	12.20%	100.00%
			40.20%	59.80%	100.00%



(a) P1's Plan about *Out.*

(b) P1's Plan about *Reject.*

Supplementary Figure 3. Reported Plan Predicts Own Behaviors.



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

Supplementary Figure 4. Cooperation Rate by Communication and Promises.

C 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.

Your 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 this 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 1's 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 Player2's payoff. 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.

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.

Avoyan, A. and Ramos, J. (2020). A road to efficiency through communication and commitment. *SSRN 2777644*.

Balliet, D. (2010). Communication and cooperation in social dilemmas: A meta-analytic review. *Journal of Conflict Resolution*, 54(1):39–57.

Battigalli, P., Charness, G., and Dufwenberg, M. (2013). Deception: The role of guilt. *Journal of Economic Behavior & Organization*, 93:227–232.

Battigalli, P. and Dufwenberg, M. (2007). Guilt in games. *The 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. (2020). 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. *IGIER*, (539).

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.

Blume, A. and Ortmann, A. (2007). The effects of costless pre-play communication: Experimental evidence from games with pareto-ranked equilibria. *Journal of Economic Theory*, 132(1):274–290.

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

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. *The New Palgrave Dictionary of Economics: Volume 1–8*, pages 2778–2782.

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

Dufwenberg, M., Li, F., and Smith, A. (2021). Threats.

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.

Fehr, D. and Sutter, M. (2019). Gossip and the efficiency of interactions. *Games and Economic Behavior*, 113:448–460.

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

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(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: Journal of the Econometric Society*, 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.

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

North, D. C. and Weingast, B. R. (1989). Constitutions and commitment: the evolution of institutions governing public choice in seventeenth-century england. *The 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. *The American Economic Review*, 61(2):112–123.

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