Threats*

Martin Dufwenberg[1], Flora Li[2], and Alec Smith[3]

[1] *University of Arizona, University of Gothenburg, and CESifo. Email: martind@eller.arizona.edu.*

[2] *Fralin Biomedical Research Institute at Virginia Tech Carilion, and Experimental Economic Laboratory, Nanjing Audit University. Email: florali@vtc.vt.edu.*

[3] *Department of Economics, Virginia Tech. Email: alecsmith@vt.edu.*

June 4, 2019

Preliminary Draft

Abstract

We study the effect of communication on deterrence and costly punishment. We show that a theoretical model of belief-dependent anger captures the relationship between messages, beliefs, and behavior and implies that threats can generate credible commitments. We test our model in a between-subjects experiment with belief elicitation where one-sided communication is available as a treatment. The evidence supports the theory, demonstrating that communicated threats change beliefs and payoff expectations and lead to greater rates of costly punishment. Threats successfully deter co-players from exploiting the strategic environment to their advantage.

Keywords: Communication, Belief-Dependent Motivation, Threats, Bargaining
JEL: C78, C92, D91,

*Corresponding Author: We appreciate comments received from Southern Economic Association 88th Annual Meetings, 2018 North-American ESA Conference, Virginia Association of Economists 2018 Annual Meeting, and 2018 Spring School in Behavioural Economics.

1 Introduction

Threats are communicated conditional plans to cause harm or loss to another person. In game theoretic analyses, threats that are too costly to carry out are typically judged to be non-credible according to behavioral concepts such as sequential rationality. In addition, communication is ancillary to traditional strategic analyses of one-shot games with unique equilibria. In these environments behavior is determined by the costs and benefits of actions, so communicated threats are judged to be "cheap talk" that cannot influence behavior.

However, explicit threats are common in everyday life. Psychological studies have shown that expressing threats is essential to human bargaining situations, and there is a psychological tendency to use threats when available (Deutsch and Krauss, 1960). Threats are a commonplace aspect of politics and international diplomacy (e.g. Huth and Russett, 1984; Guzzini, 2013). For example, President Trump threatened to shut down the federal government twice during his first Presidential term, and he followed through his threats both times. Many of the work in early game theory on bargaining and negociation centered on the analysis of threats and the role of deterrence (e.g. Schelling, 1956, 1958; Smith and Price, 1973). In addition, animals often settle disputes through threat displays rather than resorting to violence as well (Manning and Dawkins, 1998; Bradbury and Vehrencamp, 1998). The prevalence of threats in social, psychological, economic, and political life suggests that they are central to the analysis of strategic interaction, yet the mechanism through which explicit, communicated threats might work is not well understood.

In this paper we argue that explicit threats can shape strategic outcomes when decision-makers are prone to anger. Anger is one of the five basic emotions (Ekman, 1992), and all healthy humans experience anger (Averill, 1983, 2012). We build upon the model of frustration and anger of Battigalli, Dufwenberg, and Smith (2018) (BDS), which formalizes the idea that frustration builds up from goal blockage and diminished payoff expectations, and motivates aggression (Dollard et al., 1939; Berkowitz, 1989). Because the behavior of anger-prone players is belief-dependent, communication can affect strategic outcomes to the extent that it changes expectations about behavior. In contrast to the predictions of models that focus solely on material payoffs, explicit threats now change beliefs about outcomes, and anger-prone players are more willing to engage in costly punishment when behavior deviates from expectations and leads to frustration cumulation. With belief-dependent motivations, threats gain strong commitment power, and therefore, they can serve to deter opportunistic behavior (e.g. entry into a market, renegotiating a contract, developing nuclear weapons) in situations where via traditional analyses such messages would be deemed non-credible. The belief-dependent frustration and anger model provides a plausible explanation to every-day phenomenon involving threats.

We design an experiment using a two-person, two-stage deterrence game to examine the relationship between communicated threats and deterrence. This game shares the same strategic structure as the chain-store game (Selten, 1978) and the ultimatum minigame (Gale et al., 1995).[1] In stage one of the deterrence game, the first mover (P1) proposes either a fair split (which is automatically accepted) or a greedy one. If P1 grabs the larger share, then in stage two, the second mover (P2) has the option to punish the opponent, so that the initial endowment vanishes. As a treatment, we allow free-form messages from P2 to P1. To address our concern that players might not feel it appropriate to send threats if they are not provoked, we also study a three-stage variation of this game (staggered entry game) where P1 must choose the greedy offer twice, and in the communication treatment of this staggered entry game, P2 sends messages only after P1's first greedy offer. In traditional analyses of both of these games, messages from P2 should have no impact on behavior, since self-interested players will treat any communicated threats as cheap talk. To test the belief-dependent motivation, both players' 1st order beliefs about themselves and their opponents are elicited. In the communication treatment, all beliefs are elicited once before receiving messages and once after receiving messages; therefore, we can observe directly the influence of communication on reported beliefs.

A few studies have tested BDS' frustration-anger model with experiments. Persson (2018) finds that individuals react to unexpected material losses emotionally, but not behaviorally. Instead, his results are consistent with versions of the theory that modulate anger with blame. Aina et al. (2018) test the frustration-anger model in an ultimatum minigame via both the direct response (emotion relevant) and the strategy method (emotion irrelevant). Consistent with the theory they find that individuals punish with high initial expectations in the direct response condition but not using the strategy method. They also find gender differences that females are more consistent with belief-dependent motivations than males. In a companion paper to this one (Dufwenberg et al., 2018), we study the relationship between promises and costly punishment. The results in that paper are consistent with the belief-dependent notion that promises lead to cooperation and broken promises lead to costly punishment.

A large literature in economics studies communication in strategic environments (e.g.

[1] For a thorough literature review covering experiments using ultimatum games, see Güth and Kocher (2014).

Crawford and Sobel, 1982; Crawford, 1998; Charness and Dufwenberg, 2006; Balliet, 2010), but only a few experiments study communicated threats and deterrence. Rankin (2003) studied communication in ultimatum games, where responders could make a non-binding and non-freeform request to the proposer. Rankin (2003) found that not only did proposers increase the amount of offers when responders requested higher amount (analogous to threats), but also responders rejected more often when they were allowed to request. Croson et al. (2003) examined both deception and threats in ultimatum games. Croson et al.'s results showed that responders who threatened to reject low offers received higher offers, and they were more likely to reject the low offers. Masclet et al. (2013) examined threats and punishment in public goods game where in one treatment, non-binding and non-freeform threats were allowed. They found that threats significantly increased contributions though their effectiveness diminished with repetition. García et al. (2015) studied threats in a sequential hawk-dove game experiment. They found that when the game is played repeatedly, players learned that threats not only can work in their benefit, but the success of deterrence is also related to the threat's credibility.

One closely related work, Ellingsen and Johannesson (2004) studies promises and threats in a hold-up experiment. They find that individuals tend to keep their promises, but that they tend not follow through on threats. Ellingsen and Johannesson test the effectiveness of both promises and threats (separately); however, they did not elicit beliefs, and they have only a few data points. They observe a total of 9 threats, of which 5 were actionable. Of the 5 actionable threats in their experiment, only a single one was actually followed through by the participant. To explain their data they propose a behavioral model that combines distributional preferences (Fehr and Schmidt, 1999) and preferences for consistency.

We describe the game structure used for the experiment and we briefly discuss the theoretical model of belief-dependent anger incorporate with explicit threats in Section 2. We present the experiment design details, experiment procedure, and derived hypotheses in Section 3. Section 4 presents results, and Section 5 concludes.

2 Deterrence, Anger, and Threats

2.1 Deterrence Game

We focus on the deterrence game depicted in Figure 1, where the numbers and variables at the end nodes denote monetary payoffs. The variables a and b take the following values: $0 < a < b < 20$ and $a + 10 = b$. Messages from P2 to P1 can be used to examine the role of threats in a strategic environment. In stage 1, P1 can choose either *Share* to give a larger share to P2 and end the game, or *Grab* to take a larger share for herself and let P2 make the next decision. If the game continues to stage 2, P2 can either *Accept* the proposed offer, or *Punish* the proposer and both players receive 0. The amount $20 - b$ represents the cost of punishment: it is the monetary amount that P2 must forgo to reduce P1's payoff to 0 after *Grab*.

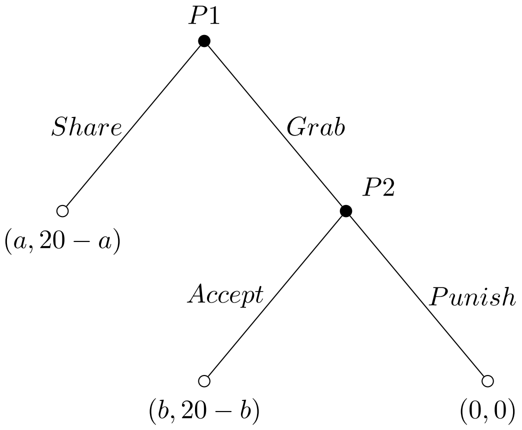


Figure 1. Deterrence Game

Outcome (*Grab*; *Accept*) is monetarily advantageous for P1, and outcome (*Share*) is monetarily advantageous for P2. Both players equally dislike outcome (*Grab*; *Punish*) monetarily. When players care only for monetary payoffs, there is a unique subgame perfect equilibrium (SPE): (*Grab*; *Accept*).

2.2 Frustration and Anger

With either self-interested or distributional preferences (Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000), communication cannot affect behavior in games with unique SPEs. However, if costly punishment is belief-dependent, then messages can influence behavior by changing expectations. BDS propose 3 different versions of a belief-dependent frustration-anger model: 1) Simple anger (SA), 2) Anger from blaming behavior (ABB), and 3) Anger from blaming intentions (ABI). SA models anger where the tendency to hurt others is proportional to frustration, formalizing the frustration-aggression hypothesis from psychology (Dollard et al., 1939; Berkowitz, 1989). ABB adds a simple notion of blame to SA, where players can only be blamed if their actions cause frustration. In two-player games, SA and ABB's predictions coincide. With ABI, players only blame others who intend to frustrate them, and so this approach relies upon higher-order beliefs. In this paper, we focus on SA/ABB.

In the belief-dependent frustration-anger model, anger is motivated by frustration. A player is frustrated if her initial payoff expectation is not met (goal blockage). Frustration is expressed as the positive difference between the initial expected material payoff and the current best possible outcome, given beliefs. For example, in the game depicted in Figure 1, if P2 assigns positive probability to P1 choosing *Share*, but the game reaches Stage 2, then P2 will experience frustration. At any history h, P2's frustration is

$$F_2(h;\alpha_2) = \left[\bar{\pi}_2(h_0) - \max_{a_2\in A_2(h)} \mathbb{E}[\pi_2|h;\alpha_2]\right]^+, \tag{1}$$

where $\bar{\pi}_2(h_0) = \mathbb{E}[\pi_2|h_0;\alpha_2]$ denotes P2's initial expectation (at h_0) given her initial set of beliefs α_2. The expression $\max_{a_2\in A_2(h)} \mathbb{E}[\pi_2|h;\alpha_2]$ denotes the maximum possible expected payoff available to P2 at the history h, where $a_2 \in A_2(h)$ represents P2's action choice at the history h.

The SA version of the frustration-anger model assumes that P2's utility from action a_2 at history h is

$$u_2^{SA}(h,a_2;\alpha_2) = \mathbb{E}[\pi_2|(h,a_2);\alpha_2] - \theta_2 F_2(h;\alpha_2)\mathbb{E}[\pi_1|(h,a_2);\alpha_2], \tag{2}$$

where $\theta_2 > 0$ denotes P2's anger sensitivity parameter. If one is frustrated, her utility consists of both material payoff and a disutility from being frustrated. Frustration increases the negative weight put on the other player's material payoff. Therefore, a frustrated individual tends to hurt the other player if the cost is low enough.

In the deterrence game defined in Figure 1, let the probability that P1 assigns to choosing *Grab* be $p_1 = \alpha_1(Grab|h^0) \in [0, 1]$. Let $q_1 \in [0, 1]$ denotes the probability that P1 assigns to P2 choosing *Punish* if stage 2 is realized, i.e. $q_1 = \alpha_1(Punish|Grab)$. We can also define analogously a similar belief system (p_2, q_2) for P2. 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, and q.

The deterrence game has multiple psychological sequential equilibria (SE) depending on P2's anger sensitivity parameter θ_2. For $(Share; Punish)$ to be a SE, the correct beliefs system is $p = 0, q = 1$. P2 initially expects $20 - a$, and experienced frustration equals $b - a$ if stage 2 is realized. Therefore, P2 will *Punish* the offer if $\theta_2 > \frac{20-b}{(b-a)b}$. The unique SPE $(Grab; Accept)$ consists another SE. When P2 expects $(Grab; Accept)$, her initial monetary payoff is $20 - b$. If P1 chooses *Grab*, P2 experiences 0 frustration. P2 chooses *Accept* with all possible θ_2.

2.3 Threats

To study threats, we allow communication as a treatment. In the experiment, P2 can send a free-form message to P1 in communication treatment. There should be no difference in behavior across treatments if agents are indeed self-interested as assumed in classic economics with complete information, as message contents should be irrelevant to players' decisions. However, if players are motivated by expectations, communication could potentially influence behavior.

With a message from P2 to P1 at the beginning of the game, we are able to observe how P1 reacts to P2's threats about *Punish*. An explicit threat looks like "if you choose *Grab*, I will *Punish*." If P2 fails to deter, and P1 chooses to *Grab*, we can then observe whether the threats are credible, or alternatively, are bluffs.

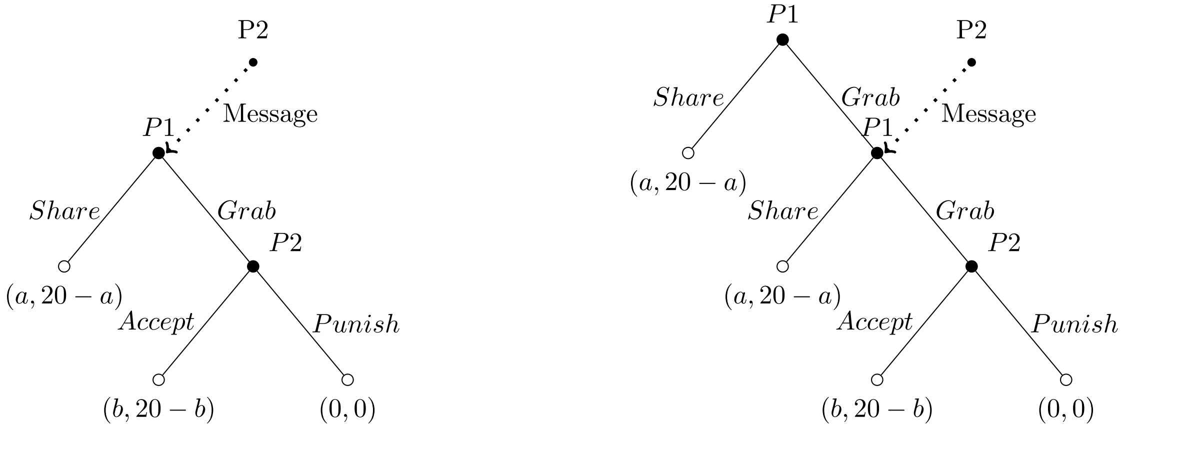
Belief-dependent frustration and anger provides a plausible explanation of how communication might influence behavior. In particular, if messages contain threats and affect expectations, P1 is more likely to *Share*, and anger-prone threateners are more likely to

Punish when deterrence fails.

3 Experiment

3.1 Design

We use a between-subject design where the treatment variable is pre-play communication.[2] In the communication treatment, P2 is allowed to send a free-form message to P1, while no message is allowed in the no-message treatment. Along with the benchmark deterrence game described in the previous section, we also study a three-stage *staggered entry* game, shown in Figure 2(b). The only difference between the two games is that in the staggered entry game P1 has to choose *Grab* and advance twice before P2 can make a decision. In the message treatment, in contrast of the pre-play message in the deterrence game, P2 is able to send a message only if P1 chooses *Grab* in the first stage of the staggered entry game. In the staggered entry games, P1's *Grab* action in stage 1 can be seen as a negative signal to challenge P2, and therefore,P2 is more likely to threat. In addition, the staggered entry design allows us to observe P1's response to a threat when comparing her choice in stage 1 and 2.



(a) Deterrence Game with Message **(b)** Staggered Entry Game with Message

Figure 2. Game Structure

In the staggered entry game, we elicit beliefs using the variables m, p, and q, where subscripts indicate the player holding the beliefs. Thus $m_1 = \alpha_1(Grab|h^0)$ is the probability

[2] Dufwenberg et al. (2018) showed that communication effect is persistent throughout the whole session. Therefore, we employ a between-subject design for communication treatment in this paper.

P1 assigns to choosing *Grab* herself in stage 1, $p_1 = \alpha_1(Grab|Grab)$ is the probability P1 *Grab*s again in stage 2, and $q_1 = \alpha_1(Punish|Grab, Grab)$ is P1's 1st order belief on P2's *Punish* choice. A similar belief system (m_2, p_2, q_2) for P2 is defined analogously.

We vary the decision problem with different payoff structures in different periods, while holding the strategic aspect of the game fixed so that $b - a = 10$, as in section 2. The payoff structures are described in Table 1, where all the values are denoted in dollars. DG stands for deterrence games, and SE represents staggered entry games. As the belief-dependent frustration-anger model specifies the significance of timing issue, we implement a standard direct-response method.[3]

Table 1. Game Variations

Game	a	20-b
DG1 & SE1	9	1
DG2 & SE2	8	2
DG3 & SE3	7	3
DG4 & SE4	6	4
DG5 & SE5	5	5

3.2 Procedures

The experiment was programmed with z-Tree (Fischbacher, 2007) and conducted at the Virginia Tech Economics Laboratory. We invited 7 to 10 pairs of participants per session. Upon entering the laboratory and signing consent forms, participants were randomly assigned to seats based on randomly drawing numbers. The experiment instructions are reproduced in the Appendix. Instructions were presented to participants on their computer monitors, and participants were also given paper copies of the instructions. At the start of the experiment the experimenters read the instructions aloud. Player roles were assigned randomly and were fixed throughout the session. Participants received feedback on both players' choices after each round.

Each session consisted of 20 rounds with stranger matching. Each session was divided in to two blocks of 10 rounds. In each block, participants played all 10 variations of the games (DG1-5 and SE1-5) in a random order. Individual level beliefs were elicited and

[3] See Brandts and Charness (2011) for evidence that results from strategy method are significantly different from that of sequential play if the game involves costly punishment.

were incentivized via a flat fee.[4] Participants received \$5 for reporting their beliefs. In the deterrence games with no message, we elicited P1's plan of choosing *Grab* (p_1), P1's 1st order belief of P2 choosing *Punish* (q_1) conditional on reaching 2nd stage, P2's 1st order belief about P1 choosing *Grab* (p_2), and P2's conditional plan of *Punish* (q_2). All beliefs were elicited at the beginning of the game. In message treatment, the same beliefs were elicited twice, before and after P1 receiving the messages.

In the staggered entry games, P1 reported her own plan about choosing *Grab* (m_1) in stage 1, her own plan about choosing *Grab* (p_1) in stage 2 conditional on reaching the stage, and 1st order belief about P2's conditional probability of choosing *Punish* (q_1). P2 reported 1st order beliefs on 1st and 2nd stage conditionally (m_2, p_2), and her own plan of choosing *Punish* (q_2) conditional on reaching to the 3rd stage. In both the message and the no message treatments, beliefs were measured twice, once at the beginning of the game, and once before stage 2 if stage 2 was reached. The detailed experiment timeline is presented in Figure 3.

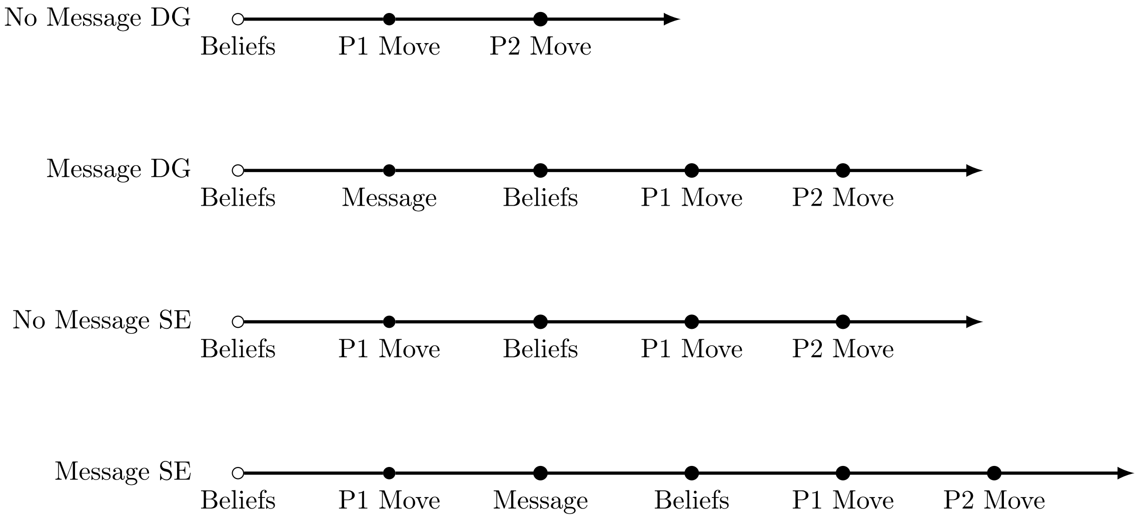


Figure 3. Experiment Timeline

At the end of the experiment, one randomly selected round is realized for actual payment. The final payment included \$10 for showing up, \$5 for belief elicitation, and amount of money earned in the randomly selected round. Participants earned \$23.68 total on average. At the end of the decision task, the participants were asked to fill out a survey on their self-reported anger ratings (second movers only), socioeconomic status, and selective questions about risk preference and social preferences based upon the survey questions in the Global Preference Survey of Falk et al. (2015). The data comprise 16 sessions of a total of 294 participants

[4] Other works employing this method include Toussaert (2018); Ameriks et al. (2007) and Dufwenberg et al. (2018).

(average of 18 participants per session). Half of the sessions were message treatment sessions, with the remaining sessions being no message treatment sessions.

3.3 Hypotheses

We test several hypotheses derived from the frustration-anger model, regarding behavioral outcomes and elicited beliefs.

Knowing that P2 is prone to anger, BDS implies that P1 believes that P2 will *Punish* more often with threats. Therefore, we expect that P1 will *Share* more frequently when receiving threats, compared to when receiving cheap talk.

Hypothesis 1. *Threats lead to a higher rate of deterrence.*

With P2 prone to anger, the frustration-anger model predicts that sending a threat should increase the probability that P1 selects *Share.* When P2's raised expectation is not met, P2 is more likely to *Punish.* We expect to observe more *Punish* outcomes with threats when reaching to stage 2, relative to messages involving no threats (cheap talk).

Hypothesis 2. *Threats lead to a higher rate of costly punishment.*

We expect that P1 will report a lower probability to *Grab* (m_1, p_1), and a higher 1st order belief about *Punish* (q_1) after receiving a threat. P2 also reports a lower 1st order belief about *Grab* (m_2, p_2), and a higher probability to *Punish* (q_2) when sending a threat.

Hypothesis 3. *Communication in the form of threats drives the effect of messages on beliefs.*

As predicted by the frustration-anger model, we not only see that threats affect behavioral outcomes, and threats drive changes in beliefs, but also we expect to detect a relationship between threats, beliefs, and behavior.

Hypothesis 4. *The effect of threats on behavior is belief-dependent.*

BDS suggests that since threats impact expectations, threats can serve as a tool for equilibrium selection. With threats, we hypothesize that we will observe a tendency for more deterrent outcomes.

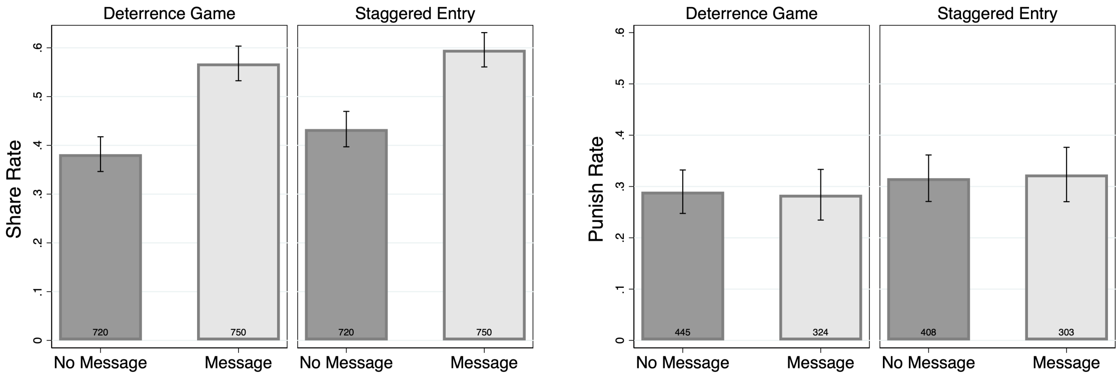
Hypothesis 5. *Players eventually reach to one of the two Sequential Equilibrium* $(\{Share, Punish\}$ *and* $\{Grab; Accept\})$. *Threats select* $\{Share; Punish\}$ *to be reached more often.*

4 Results

This section is organized as follows: Section 4.1 summarizes the overall behavioral results on treatment effect. We focus on analysis of communication treatment effect on cooperation and costly punishment. Section 4.2 presents results on threats vs. cheap talk. In Section 4.2, we conduct non-parametric and regression analyses to test Hypothesis 1 and 2. Section 4.3 tests Hypothesis 3 and 4 regarding participants belief-dependent motivations.

4.1 The Effect of Communication on Cooperation & Costly Punishment

Overall, we find that communication has a strong deterrence effect. Table 2 summarizes the outcomes of each game using session-level averages. First, when communication is not allowed, P2 chooses *Punish* 30.25% of the time. Second, there is an obvious difference in behavior between the communication and no communication treatment, indicating that messages are not just "cheap talk." Comparing the two treatments, we observe a substantial increase in the aggregated *Share* outcomes (58.20% vs. 40.76%, 1-sided Fisher's exact, p < .001) when messages are allowed. The effect of communication treatment is also apparent when looking at individual games. For both the deterrence and the staggered entry games, the *Share* rate is significantly higher with communication, confirmed with the Wilcoxon ranksum tests reported in Table 2. This result is also illustrated in Figure 4(a), with the vertical bar representing the 95% confidence interval.



(a) P1's *Share* Rate

(b) P2's *Punish* Rate

Figure 4. Outcome Summary with Communication Treatment Effect

Table 2. Communication Treatment Effect on Behavior

	P1's *Share* Rate			P2's *Punish* Rate		
DG	No Com	Com	p-value	No Com	Com	p-value
DG1	68.06%	85.33%	0.010	65.22%	50.00%	0.634
DG2	65.28%	74.67%	0.091	48.00%	50.00%	0.627
DG3	35.42%	63.33%	0.006	37.63%	30.91%	0.226
DG4	13.89%	35.33%	0.004	23.39%	23.71%	0.833
DG5	8.33%	25.33%	0.002	8.33%	19.64%	0.109
SE	No Com	Com	p-value	No Com	Com	p-value
SE1	77.78%	90.00%	0.005	68.75%	46.67%	0.663
SE2	61.11%	81.33%	0.010	53.57%	39.29%	0.268
SE3	43.06%	62.67%	0.031	36.59%	41.07%	0.833
SE4	22.92%	40.00%	0.013	22.52%	37.78%	0.156
SE5	11.81%	24.00%	0.004	17.32%	20.18%	0.207
All	40.76%	58.20%	0.001	30.25%	30.30%	0.466

Note: p-values are obtained from session level averages using Wilcoxon ranksum (Mann-Whitney) tests. Games are defined by the "Payoff from *Accept*", so that *e.g.* DG1 represents a deterrence game where the Payoff from *Accept* equals 1 for P2.

At first glance the communication treatment does not seem to have an effect on P2's *Accept* vs. *Punish* choices, as shown in Table 2. When focusing only on P2's behavior in the last stage, we notice a slightly higher but non-significant *Punish* rate in communication treatment (30.30% vs. 30.25%, 1-sided Fisher's exact, p = .513). When looking at each of the 10 games separately, we see no significant difference from Wilcoxon ranksum tests comparing individual games. The results are also graphically represented in Figure 4(b). We see roughly the same *Punish* rate in both treatments in the deterrence and the staggered entry games. Although we do not see a clear difference in P2's *Punish* behavior comparing the different treatments, we cannot simply conclude that communication impacts only P1 and not P2. Dufwenberg et al. (2018) show that there can be some selection bias when individuals play sequential games involving costly punishment using the direct response method. In order to draw conclusions about the factors determining the decision to choose *Punish*, we investigate the communication treatment effect further using players' self-reported plans as an indicator/proxy for their actual behavior, allowing us to examine what P2 plans to do in the last stage of every game played.

We perform linear probability regressions for players' choices and linear regressions for players' plans. Since the communication treatment is implemented at the session level (between subjects) we report the results from linear regressions that pool the data for a given

Table 3. Regression Results – The Effect of Communication on P1's *Share* Choice and Plan

	P1's *Share* Choice		P1's *Share* Plan	
	A	B	C	D
	coef / se	coef / se	coef / se	coef / se
Payoff from *Accept*	-0.169***	-0.169***	-0.089***	-0.089***
	(0.007)	(0.005)	(0.007)	(0.005)
Staggered Entry	0.041*	0.041**	-0.050**	-0.050***
	(0.022)	(0.017)	(0.019)	(0.013)
Communication		0.171***		0.180***
		(0.017)		(0.013)
Constant	0.984***	0.899***	0.739***	0.649***
	(0.027)	(0.026)	(0.025)	(0.019)
Observations	160	160	160	160
AIC	-172.304	-246.877	-217.097	-345.702
BIC	-163.079	-234.576	-207.872	-333.401

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

Note: We ran linear probability regressions for P1's *Share* Choice and linear regressions for P1's *Share* Plan. Data for each game are aggregated at the session level.

Table 4. Regression results – The Effect of Communication on P2's *Punish* Choice and Plan

	P2's *Punish* Choice		P2's *Punish* Plan	
	A	B	C	D
	coef / se	coef / se	coef / se	coef / se
Payoff from *Accept*	-0.105***	-0.105***	-0.086***	-0.086***
	(0.014)	(0.014)	(0.007)	(0.007)
Staggered Entry	0.034	0.034	0.034*	0.034*
	(0.035)	(0.035)	(0.020)	(0.020)
Communication		-0.016		0.056***
		(0.035)		(0.020)
Constant	0.674***	0.683***	0.676***	0.648***
	(0.058)	(0.055)	(0.029)	(0.031)
Observations	160	160	160	160
AIC	-22.534	-20.756	-197.084	-202.852
BIC	-13.308	-8.455	-187.858	-190.551

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

Note: We ran linear probability regressions for P2's *Punish* Choice and linear regressions for P2's *Punish* Plan. Data for each game are aggregated at the session level.

game at the session level. In the regressions, we use "Payoff from *Accept*" (20-b) and the indicator variable "Staggered Entry" to control for each individual games. Indicator variable "Communication" tests for communication treatment effect. Consistent with previous nonparametric results, when regressing P1's *Share* choice (Table 3), communication increases *Share* rate significantly, and when regressing P2's *Punish* choice (Table 4) , communication does not seem to affect *Punish* rate.

In practice plans are good predictors of their subsequent choices. The correlation between P1's plan and choice is 0.6851 ($p < .001$), and the correlation between P2's plan and choice is 0.7332 ($p < .001$). In addition, the quality of the reported beliefs is demonstrated in Figures 16 & 17 (in Appendix), where we plot nonparametric estimates of Receiver Operating Characteristic (ROC) curves that measure how well players' reported beliefs predict their behaviors. We find that players' reported beliefs and plans are very accurate predictors of behavior, and that the areas under the ROC curves are all well above 0.80 (probability that Players' reported beliefs represents their final choices). Since players' plans are elicited once at the beginning of the game, there is no selection bias for plans.

When we look at linear regressions where the dependent variable is the players' plan, we detect a stronger effect of communication. Communication significantly affects both P1's *Share* and P2's *Punish* decisions. In addition, the coefficient on "Staggered Entry" becomes marginally significant. P2 reports that she is more likely to choose *Punish* in the staggered entry games.

Another notable observation is that in terms of material payoffs, communication helps P2 (the message sender) to increase payoffs, but hurts P1 (the message receiver) as demonstrated in Figure 5. In total, communication helps to increase welfare (P1's and P2's payoffs combined) by \$1.05 (1-sided Fisher's exact, $p < .001$).

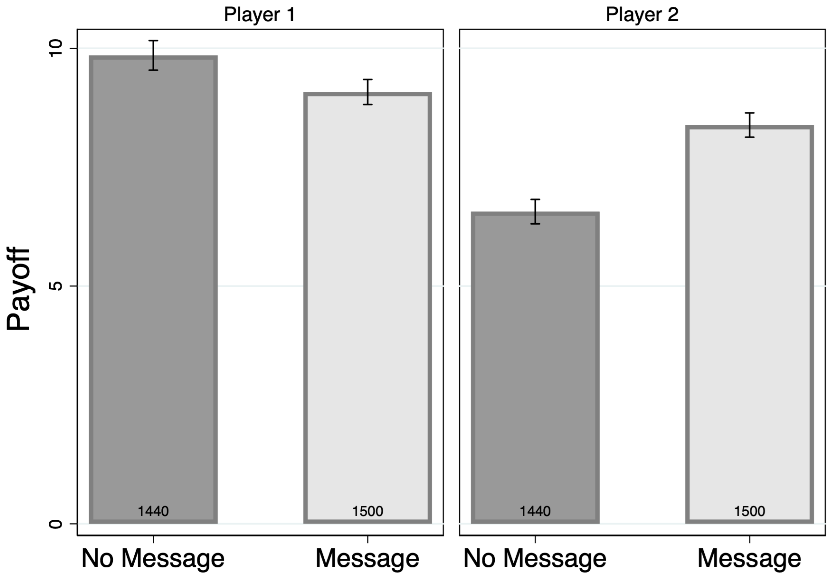


Figure 5. Payoff Distribution

4.2 The Credibility of Threats

To examine the effect of message contents on behavior (Hypotheses 1 and 2), we manually categorize the messages as either threats, or cheap talk. We define threats as messages that convey the intention to punish the opponents. For example, threats share the similar pattern of "If you choose *Grab*, I will *Punish*." We define cheap talk as messages that are not threats. Those messages are not necessarily meaningless in our strategic environment, but we categorize them as cheap talk since they are not relevant to the study of threats.

Figure 6 shows that the use of threats increases over rounds before leveling off around the middle of the experiment. There is a surprisingly high frequency of threats in the communication sessions: When P2 is allowed to send a message to P1, 54.24% of the messages include threats. P2 sends fractionally more threats in the staggered entry games than in the deterrence games (55.29% vs. 53.47%). However, the difference is not statistically significant (1-sided Fisher's exact, p = 0.274).

For the analysis of threats we focus on the data from the communication treatment. As presented in Table 5, in the deterrence games, when P1 receives a message, P1 *Share*s with a higher probability when she receives a threat compared to when she receives cheap talk (65.84% vs. 46.42%, 1-sided Fisher's exact, p < .001). We note a similar result for the staggered entry games. There is a higher *Share* rate with threats, and a lower *Share* rate with

Table 5. The Effect of Threats on Behavior

Deterrence Game	Share	Accept	Punish	Total
	162	154	33	349
Cheap Talk	46.42%	44.13%	9.46%	100%
		82.35%	17.65%	100%
	264	78	59	401
Threats	65.84%	19.45%	14.71%	100%
		56.93%	43.07%	100%
	426	232	92	750
Total	56.80%	30.93%	12.27%	100%
		71.60%	28.40%	100%

Staggered Entry	Share	Share (2nd)	Accept	Punish	Total
	193	85	126	38	442
Cheap Talk	43.67%	19.23%	28.51%	8.60%	100%
			76.83%	23.17%	100%
	0	169	79	60	308
Threats	0%	54.87%	25.65%	19.48%	100%
			56.83%	43.17%	100%
	193	254	205	98	750
Total	25.73%	33.87%	27.33%	13.07%	100%
			67.66%	32.34%	100%

Note: Each data entry consists three values: 1) Frequency of the outcome, 2) Proportion of the outcome, and 3) Outcome distribution in the last stage.

cheap talk (54.87% vs. 34.14%, 1-sided Fisher's exact, $p < .001$). We are especially careful when analyzing the staggered entry games data, since 25.73% of the games end at stage 1, before P2 has a chance to send a message. In Table 5 we conservatively categorize these games as involving cheap talk; however, we do not actually know the potential messages. Therefore, when analyzing *Share* rate for threats and cheap talk, we treat those games as missing values.

The above results are consistent with Hypothesis 1, that threats result in a higher *Share* rate in both games. These results are graphically presented in Figure 7(a), with the vertical bars representing the 95% confidence intervals.

To test Hypothesis 2, we examine P2's behavior with both threats and cheap talk. Table 5 demonstrates that for the deterrence games, the conditional *Punish* rate is significantly higher with threats (43.07% vs. 17.65%, 1-sided Fisher's exact, $p < .001$). The same result holds for the staggered entry games (43.17% vs. 23.17%, 1-sided Fisher's exact, $p < .001$).

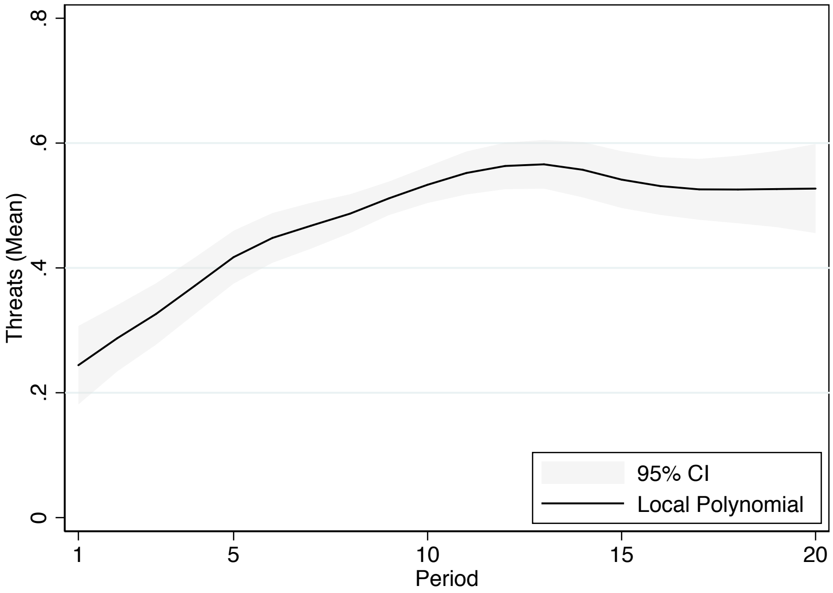
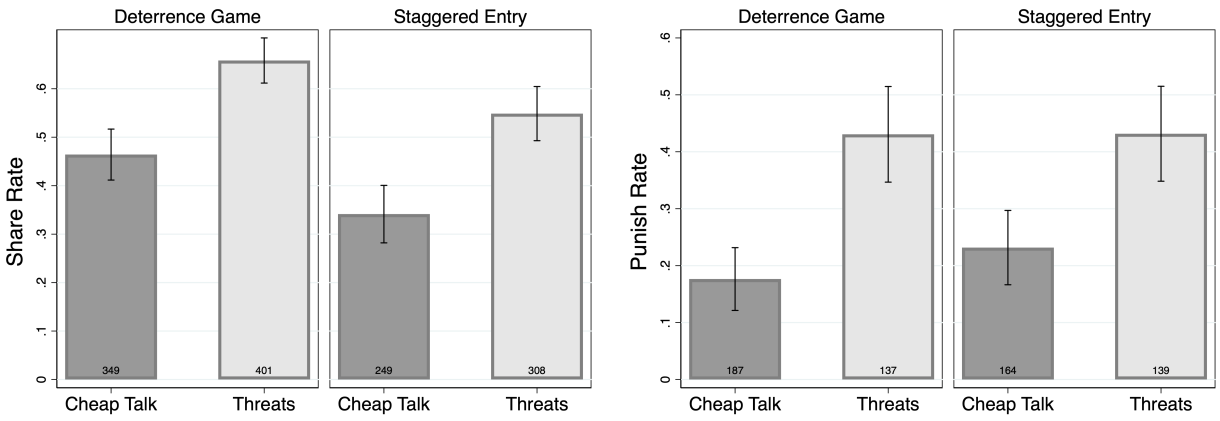


Figure 6. Number of Threats in Each Period

Figure 7(b) demonstrates that P2 *Punish*s more often when sending a threat instead of sending cheap talk. This is consistent with our Hypothesis 2 and the frustration-anger model, that P2 is more likely to engage in costly punishment when threats are made.

Using only the communication data, we examine the effect of threats on behavior with subject level fixed effect logistic regressions in Table 6. "Payoff from *Accept*" and the indicator variable "Staggered Entry" are used to control for individual games, and "Period" is used to control for extent of time. Regression models B and D show that threats are associated with an increase in the rate of both *Share* and *Punish* choices. In addition, we observe in these regression analyses that our staggered entry procedure produces higher rates of both *Share* and *Punish* choices.



(a) P1's *Share* Rate

(b) P2's *Punish* Rate

Figure 7. Outcome Summary Comparing Threats vs. Cheap Talk

Table 6. Logistic Regressions – Effect of Threats on Players' Behavior

	P1's *Share* Choice		P2's *Punish* Choice	
	A coef / se	B coef / se	C coef / se	D coef / se
Payoff from *Accept*	-0.859***	-0.872***	-0.475***	-0.523***
	(0.052)	(0.054)	(0.054)	(0.067)
Staggered Entry	0.134*	0.179**	0.229**	0.214*
	(0.077)	(0.082)	(0.112)	(0.124)
Period	0.050***	0.044***	0.036	0.013
	(0.007)	(0.006)	(0.024)	(0.023)
Threats		0.418***		1.230***
		(0.161)		(0.237)
Observations	1500	1500	627	627
AIC	1546.424	1537.484	640.517	607.180
BIC	1562.363	1558.737	653.840	624.944
Subject controls	Yes	Yes	Yes	Yes

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

Note: Coef.: Coefficient. SE: standard error. Standard errors are clustered at the session level.

4.3 Threats and Belief-Dependent Anger

Motivated by the theoretical modeling of BDS, we hypothesized that messages containing threats would drive changes in beliefs and expectations (Hypothesis 3) and that threats would work through the mechanism of belief-dependent frustration and anger to generate a self-fulfilling effect on behavior (Hypothesis 4). To test Hypothesis 3, we investigate the relationship between players' reported beliefs and the content of the messages. In addition, we examine the relationship between players' reported beliefs and their actual behavior to test Hypothesis 4.

During the experiment we elicited a rich set of beliefs and plans for both players. Before the game is played, we measured probabilistic first-order beliefs about players' own actions (their plans) and about their co-player's behavior at each history. In the communication treatment, we also measured beliefs both before and after messages were received. In this section we exploit this data to study the relationship between messages and player's belief-dependent motivations.

Table 7 presents summary statistics for self-reported beliefs (both players' beliefs about *Share* and *Punish*) recorded after messages are received, and Figures 18 and 19 (in the Appendix) present the histograms of these beliefs. These data are most likely to capture the beliefs participants held when choosing actions, and as discussed in Section 4.1, self-reported beliefs and plans are good predictors of participant behavior (see Figures 16 & 17 in the Appendix for ROC analyses).

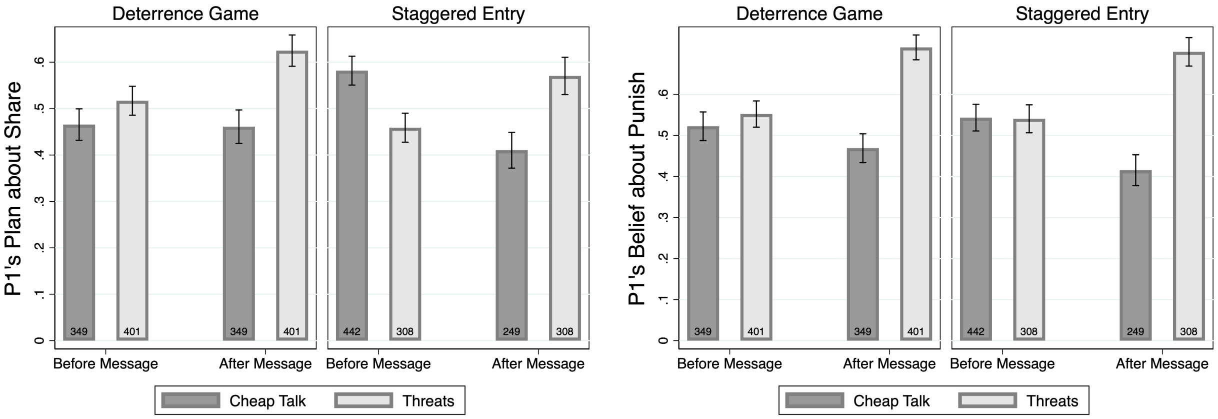
For both players and for both types of games, the effect of communication on reported beliefs is driven by the messages containing threats (Figures 8 & 9), consistent with Hypothesis 3.

We first examine the effect of threats on P1's beliefs and plans. Because we elicit beliefs both before and after P1 receives messages, we can directly detect the change in reported beliefs caused by receiving the messages. In the deterrence games, we see a significant increase in P1's reported probability of choosing *Share* when receiving a threat, but we observe no such change with cheap talk (Figure 8(a)). In the staggered entry games we notice a similar result. In addition, when P1 receives a cheap-talk message, we detect a statistically significant decrease in the self-reported probability of choosing *Share*, suggesting that P1 anticipates receiving threats and that she is more likely to engage in opportunistic behavior if she does not receive a threat.

Table 7. Summary Statistics – Reported Beliefs

	No Communication		Communication		Total
	DG	SE	DG	SE	
P1's Plan re: *Share*	720	513	750	557	2540
	0.396	0.293	0.549	0.499	0.443
	(0.342)	(0.278)	(0.353)	(0.346)	(0.347)
P1's Belief re: *Punish*	720	513	750	557	2540
	0.408	0.407	0.601	0.575	0.501
	(0.329)	(0.315)	(0.343)	(0.338)	(0.344)
P2's Belief re: *Share*	720	513	750	557	2540
	0.308	0.190	0.445	0.385	0.342
	(0.245)	(0.237)	(0.278)	(0.295)	(0.281)
P2's Plan re: *Punish*	720	513	750	557	2540
	0.381	0.394	0.453	0.450	0.420
	(0.400)	(0.418)	(0.443)	(0.447)	(0.428)

Note: Each data entry contains 1) number of observation, 2) mean, and 3) standard deviation in parentheses. Only beliefs of interests are presented. All beliefs presented in communication treatment are elicited after sending/receiving the message. Beliefs on *Share* in the staggered entry games present only second stage beliefs.



(a) P1's Plan about *Share*

(b) P1's Belief about *Punish*

Figure 8. P1's Reported Beliefs

We note a similar pattern in P1's reported 1st order beliefs about P2's *Punish* choices. Figure 8(b) shows that P1s' reported 1st order belief about *Punish* increases with threats but stays roughly the same with cheap talk in the deterrence game. But in the staggered entry games, P1 believes that P2's *Punish* rate is increasing with threats, but is decreasing with cheap talk. Therefore, when receiving threats, P1 is more likely to *Share*, and she believes that P2 is more likely to follow through on the threats.

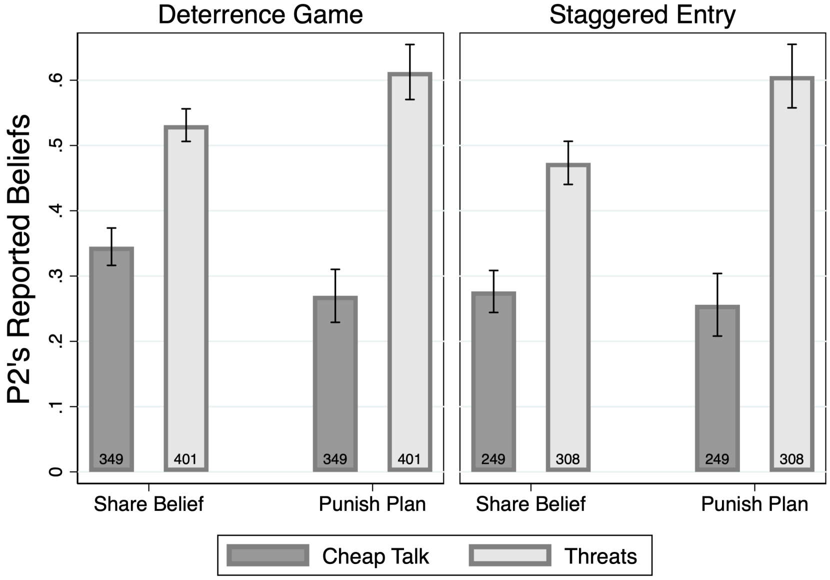


Figure 9. P2's Reported Beliefs

Figure 9 demonstrates that on average, P2 reports a higher 1st order belief about *Share*, and a higher probability to choose *Punish* when messages include threats, in both deterrence and staggered entry games. This indicates that with threats, P2 believes that P1 is more likely to *Share* (successful deterrence), and P2 is more likely to punish and follow through on her own threats when game reaches the last stage. The above results are supportive of our Hypothesis 3.

We also run logistic regressions to test Hypothesis 4, focusing on whether participants' 1st order beliefs are associated with P1's choice between *Share* and *Grab* and P2's choice between *Punish* and *Accept*. In Table 8, we run separate logistic regressions on the full sample, the no communication treatment sample, and the communication treatment sample with subject level control to illustrate the relationship between P1's reported beliefs and P1's choice of *Share*. In all three samples, when controlling for individual games ("Payoff from *Accept*" and "Staggered Entry") and experience ("Period"), we see that both P1's belief about *Punish* and plan to *Share* is positively associated with P1's *Share* choice. For the communication treatment sample, comparing Table 6 regression model B to Table 8 regression model H, the effect of threats diminishes after adding P1's 1st order belief about *Punish*. These results imply that although we observe behavioral differences between threats and cheap talk, the behavioral results are driven by beliefs. The result is even stronger when looking at Table

8 model I. After controlling for both P1's belief and plan, the effect of threats is no longer statistically significant. This result is consistent with Hypothesis 4.

Table 9 presents logistic regressions with subject level controls in order to illustrate the relationship between P2's reported beliefs and P2's choice of *Punish*. We study this relationship again on three samples: the full sample, the no communication treatment sample, and the communication treatment sample. As in Table 8, we control for individual games and experience. In regression models B, E, and G, we note that P2's 1st order belief about *Share* is positively associated with P2's probability of choosing *Punish*. Even after controlling for "Threats" (model H) in the communication treatment sample, P2's 1st order belief about *Share* shows a strong association with *Punish* decisions. We note that, at the time of choice, this belief is not consequential with either self-interested or distributional preferences. Therefore, both beliefs and the contents of the messages affect P2's decisions. Finally, if we include P2's plan about *Punish* (models C, F, and I), we find that P2's plan is significant and the effect of P2's 1st order beliefs and threats disappeared. This provides further evidence that P2's plan about *Punish* predicts P2's actual *Punish* choice well, and that it is reasonable to treat P2's plan as a close proxy for P2's choice.

Table 8. Logistic Regressions – Effect of Beliefs on P1's *Share* Choice

	Full			No Com			Com		
	A	B	C	D	E	F	G	H	I
	coef / se	coef / se	coef / se	coef / se	coef / se	coef / se	coef / se	coef / se	coef / se
Payoff from *Accept*	-0.804***	-0.672***	-0.577***	-0.887***	-0.817***	-0.709***	-0.729***	-0.739***	-0.608***
	(0.051)	(0.063)	(0.057)	(0.087)	(0.121)	(0.153)	(0.048)	(0.059)	(0.067)
Staggered Entry	0.197***	-0.575***	-0.522***	0.283***	-0.878***	-0.701***	-0.349***	-0.354***	-0.356**
	(0.048)	(0.080)	(0.099)	(0.067)	(0.142)	(0.177)	(0.119)	(0.123)	(0.147)
Period	0.045***	0.013	-0.012*	0.051***	0.029*	-0.001	0.019*	0.015	-0.010
	(0.004)	(0.008)	(0.007)	(0.008)	(0.015)	(0.014)	(0.010)	(0.011)	(0.013)
P1's Belief re: *Punish*		2.497***	1.520***		1.471***	0.929**	2.658***	2.475***	1.416***
		(0.286)	(0.198)		(0.430)	(0.418)	(0.477)	(0.497)	(0.385)
P1's Plan re: *Share*			5.261***			5.096***			5.042***
			(0.302)			(0.719)			(0.367)
Threats								0.350*	0.255
								(0.189)	(0.166)
Observations	2940	2540	2540	1440	1233	1233	1307	1307	1307
AIC	3210.241	2464.330	1693.729	1414.725	1040.062	745.313	1239.062	1235.631	869.921
BIC	3228.200	2487.690	1722.928	1430.542	1060.530	770.899	1259.764	1261.509	900.974
Subject controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

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

Note: Coef.: Coefficient. SE: standard error. Standard errors are clustered at the session level.

Table 9. Logistic Regressions – Effect of Beliefs on P2's *Punish* Choice

	Full			No Com			Com		
	A	B	C	D	E	F	G	H	I
	coef / se	coef / se	coef / se	coef / se	coef / se	coef / se	coef / se	coef / se	coef / se
Payoff from *Accept*	-0.573***	-0.532***	-0.577***	-0.679***	-0.639***	-0.651***	-0.425***	-0.481***	-0.443***
	(0.057)	(0.059)	(0.047)	(0.068)	(0.074)	(0.070)	(0.050)	(0.057)	(0.086)
Staggered Entry	0.206***	0.311***	0.120	0.171	0.273**	0.078	0.291**	0.269**	0.234
	(0.059)	(0.062)	(0.124)	(0.122)	(0.121)	(0.171)	(0.135)	(0.133)	(0.273)
Period	0.032***	0.032***	-0.037	0.024**	0.024**	-0.045	0.035	0.015	-0.034
	(0.011)	(0.011)	(0.023)	(0.011)	(0.010)	(0.033)	(0.024)	(0.025)	(0.028)
P2's Belief re: *Share*		1.385***	0.184		0.924**	0.495	1.794***	1.309***	-0.107
		(0.271)	(0.505)		(0.438)	(0.658)	(0.335)	(0.364)	(0.666)
P2's Plan re: *Punish*			5.230***			5.740***			4.831***
			(0.397)			(0.413)			(0.449)
Threats								1.039***	-0.013
								(0.245)	(0.295)
Observations	1480	1480	1480	853	853	853	627	627	627
AIC	1550.847	1520.562	821.977	830.822	826.736	426.300	618.956	598.097	355.493
BIC	1566.747	1541.762	848.476	845.069	845.731	450.044	636.719	620.302	382.139
Subject controls	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes	Yes

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

Note: Coef.: Coefficient. SE: standard error. Standard errors are clustered at the session level.

4.4 Sequential Equilibrium Selection

In Figure 10, we see a pattern for convergence in either of the Sequencial Equilibrium ($\{Share; Punish\}$ and $\{Grab; Accept\}$). In addition, we see a equilibiurm selection over $\{Share; Punish\}$ as well. The non-equilibrium outcome *Punish* happens least frequent; 15.24% of the games end with *Punish*, with no notable change in rate throughout the experiment. Figure 10 shows that roughly same amount of games starts with either *Share* or *Accept* outcomes, but towards the end of the experiment, there are more games end with *Share* compared to *Accept* (last 5 periods: 1-sided Fisher's exact p $<$ .001; last period: 1-sided Fisher's exact p $<$ .001). In addition, *Share* outcomes increase throughout the experiment (first vs. last 5 periods: 41.09% vs. 59.59%, ranksum p $<$ .001; first vs. last period: 47.625% vs. 61.22%, ranksum p $=$ .019). Whereas, *Accept* outcomes decrease thoughout the experiment (first vs. last 5 periods: 44.35% vs. 26.67%, ranksum p $<$ .001; first vs. last period: 43.54% vs. 25.17%, ranksum p $<$ .001).

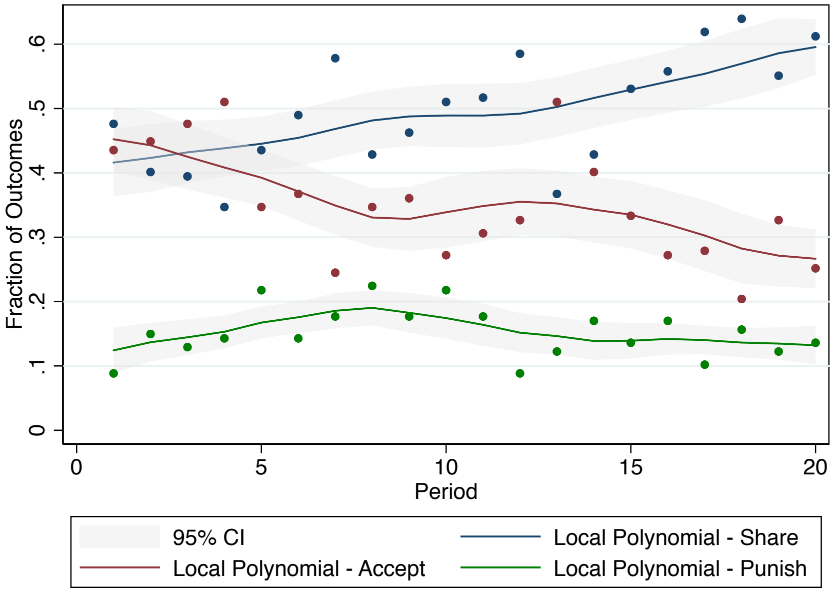


Figure 10. Equilibrium Convergence

pool every 5 rounds, to see change in outcome distributions and beliefs, last 5 rounds with low Punish outcome (beliefs). Want to check on threats vs. cheap talk as well.

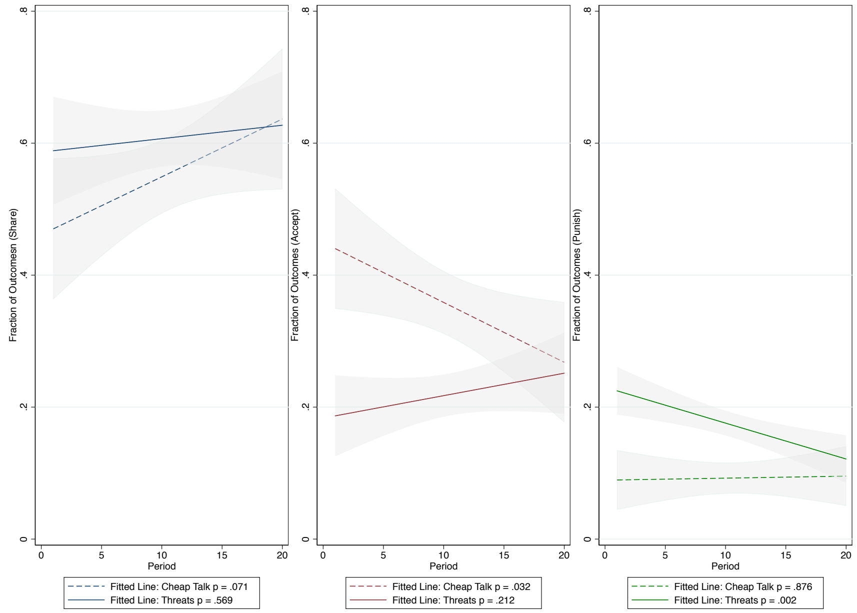


Figure 11. Equilibrium Convergence

5 Conclusion

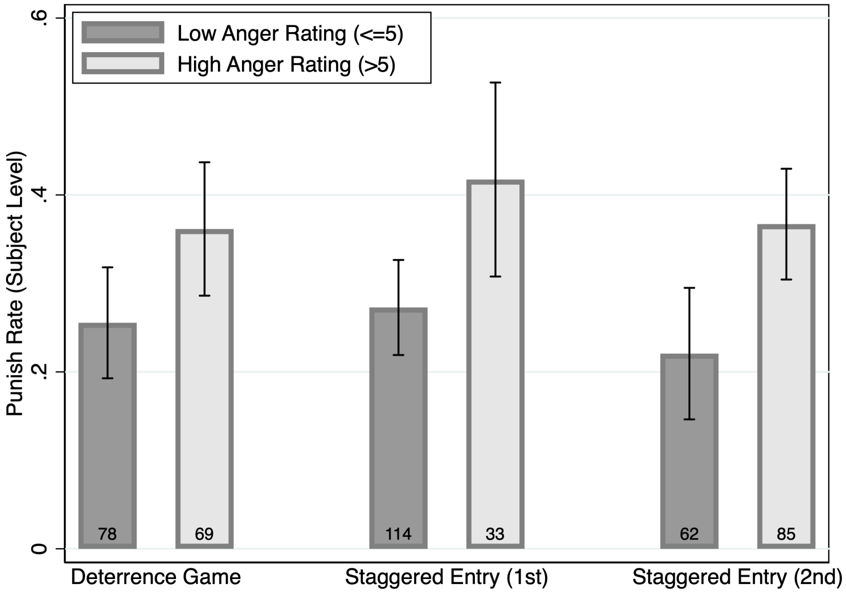
In this paper, we study the relationship between threats, credibility, and costly punishment, deriving theoretical predictions from the model of belief-dependent anger of Battigalli et al. (2017). When combined with the notion that communicated messages influence beliefs, our model implies that threats will be self-fulfilling. When threats are disregarded, frustration and the propensity to engage in costly punishment (aggression) increases. Knowing this, message recipients deem threats credible.

In our deterrence experiments the content of messages drives the effect of communication. Threats successfully deter first movers, and second movers tend to follow through on their threats when they are disregarded. We also find that belief changes mediate the effect of communication on behavior. Threats change beliefs, while other messages have no effect. These results are consistent with the idea that threats, beliefs, and behavioral outcomes are linked through the mechanism of belief-dependent frustration and anger.

Appendices

A Self-Reported Anger

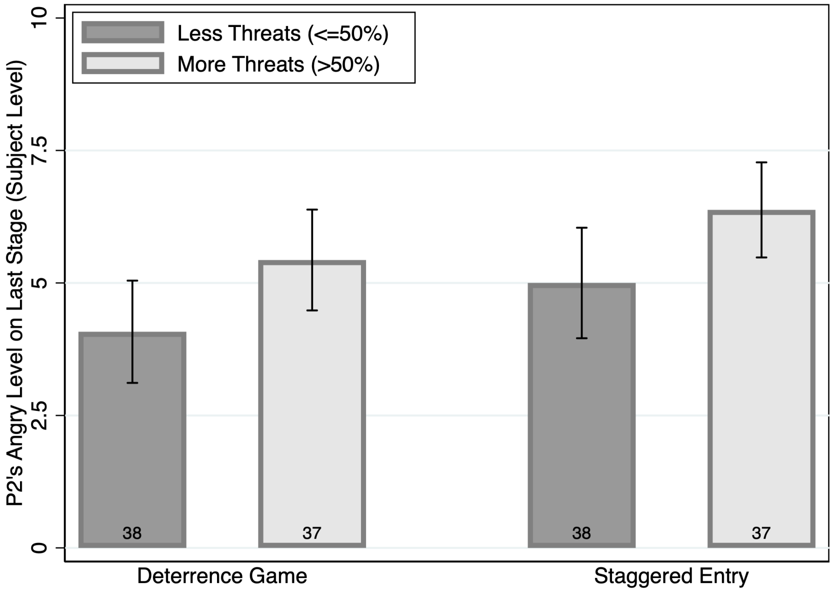
After the experiment concludes, we elicited self-reported measures of anger from participants assigned the role of Player 2. We are able to examine whether an individuals' level of anger is correlated with their behavior. Various studies have shown that the ultimatum game induces negative emotions especially anger (e.g. Xiao and Houser, 2005; Grecucci et al., 2013; Güth and Kocher, 2014). In the survey, P2 reports anger on a scale from 0 (not angry at all) to 10 (very angry) in 3 different strategic scenarios: 1) If P1 chose *Grab* in the deterrence games, 2) If P1 chose *Grab* in the 1st stage of the staggered entry games, and 3) If P1 chose *Grab* in the 2nd stage of the staggered entry games. Questions 1-3 in Supplementary Table 10 include the working of these questions. On average P2 reports some degree of anger in all three scenarios (DG: mean 4.60 sd 2.92, SE 1st: mean 3.19 sd 2.80, SE 2nd: mean 5.39 sd 3.20).



Supplementary Figure 12. Greater Anger with Higher Punish Rate

In Supplementary Figure 12, We compare participants who report anger ratings above 5 to those who report ratings below or equal to 5. We find that P2s who report high anger *Punish* more often in all three scenarios (Wilcoxon ranksum: DG p-value = .039, SE 1st p-value = .012, SE 2nd p-value = 0.001). We also note that when opponents choose *Grab* on the 2nd stage, individuals report higher anger ratings, compared to when opponents choose

Grab on the 1st stage in the staggered entry games (1 sided t-test p-value $< .001$). P2's anger builds up with opponent's *Grab* actions, and this might be the reason why P2 is more likely to *Punish* in the staggered entry games than in the deterrence games.



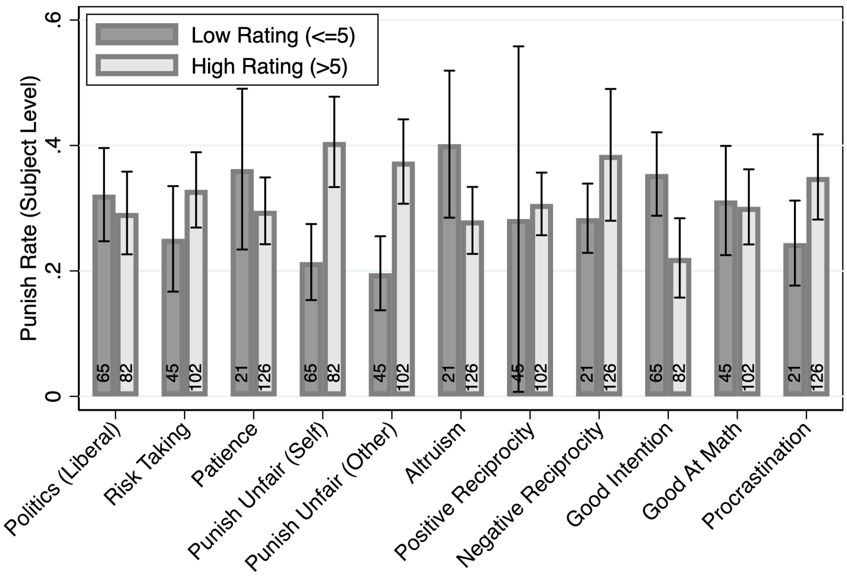
Supplementary Figure 13. Greater Anger at Disregarded Threats

When the game reaches the last stage, P2 is equally angry with or without communication (Wilcoxon ranksum: DG p-value = .487, SE p-value = .363). However, depending on the contents of the messages, Player 2 reports different levels of anger with threats and cheap talk. In Supplementary Figure 13, when the game reaches the last stage Player 2 feels slightly more angry when the majority ($> 50\%$) of their messages are threats (Wilcoxon ranksum: DG p-value = .048, SE p-value = .066). This confirms the prediction of the model that threats affect expectations of outcomes, and when expectations are not met, players feel more frustrated with threats compared to cheap talk.

B Social Preference Survey

Along with self-reported anger ratings, we also measure participants's political orientation, risk preferences, and social preferences using selective questions from The Global Preference Survey (Falk et al., 2015). Please refer to questions 4-14 in Supplementary Table 10 for the exact questions.

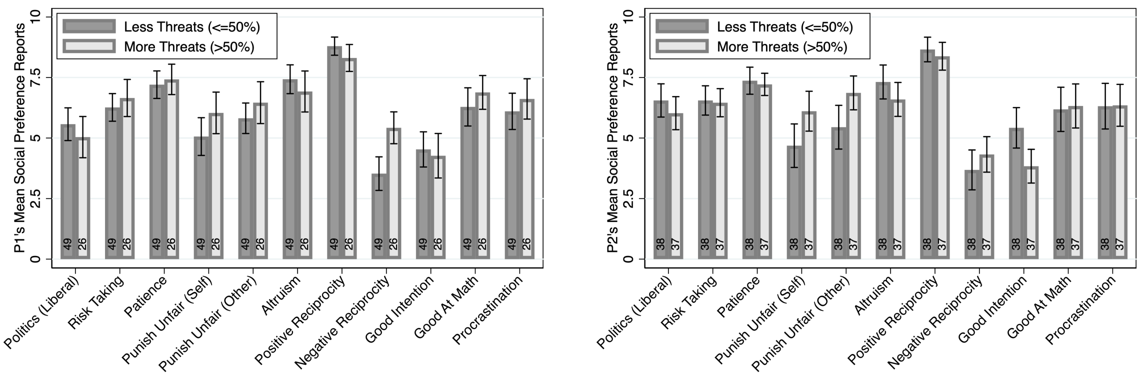
The relationship between self-reported social preferences and the *Punish* rate is depicted



Supplementary Figure 14. Social Preferences and Punish Rate

in Supplementary Figure 14. Political orientation (Wilcoxon ranksum: p-value = .481), risk taking (p-value = .132), patience (p-value = .244), positive reciprocity (p-value = .605), and math skill (p-value = .724) seem to be unrelated with P2's *Punish* rate. Individuals who report higher ratings for altruism (p-value = .043) and good intention (p-value = .028) choose *Punish* less often. Individuals who report higher ratings for punishing unfair offers (both for self (p-value $<$ .001) and others (p-value = .001)), negative reciprocity (p-value = .044), and procrastination (p-value = .035) are more likely to *Punish* P1. However, before we draw the conclusions that individuals with different social preferences behave differently, we need to mention that the above statistical analyses are based on two unbalanced samples. With the specific framing of the survey questions, such as using the terms "willing," "punish," "good cause," etc., participants' self reported social preferences ratings are skewed to one direction.

P1 reports no difference in social preferences between the communication and no communication treatments: political orientation (p-value = .147), risk taking (p-value = .390), patience (p-value = .400), punish unfair offers (both for self (p-value = .442) and others (p-value = .531)), altruism (p-value = .758), positive reciprocity (p-value = .279), negative reciprocity (p-value = .111), good intention (p-value = .513), math skill (p-value = .488), and procrastination (p-value = .807). Whereas, P2 reports more willing to revenge, with communication (p-value = .011). In the communication treatment, P2 is also marginally more liberal (p-value = .071), more willing to punish unfair offer for themselves (p-value = .057), and more willing to punish unfair offer for others (p-value = .087).



(a) P1's Reported Social Preferences

(b) P2's Reported Social Preference

Supplementary Figure 15. Social Preferences Reports with Threats vs. Cheap Talk

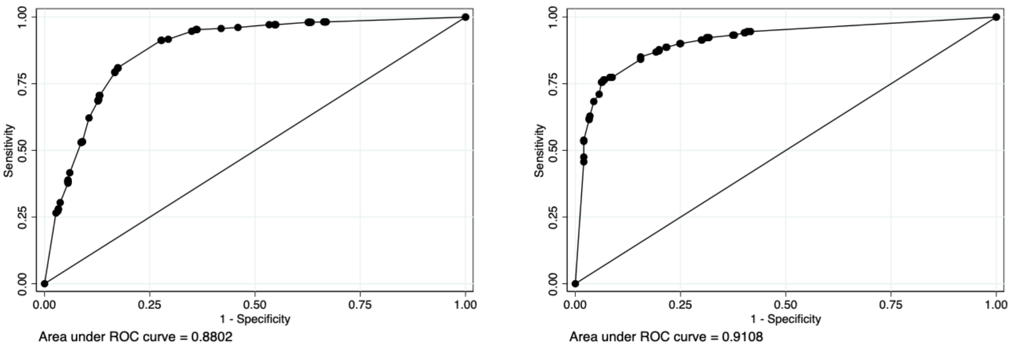
Supplementary Figure 15 illustrates that, in the communication treatment, depending on the message contents, P2 reports different ratings for some social preferences. But P1 again reports the same social preferences with or without threats, except for negative reciprocity (p-value = .001). P2 who reports higher willingness to punish unfair offers (offers for self (p-value = .027) and offers for others (p-value = .022)), to be less altruistic (p-value = .084), and to believe less that people have good intentions (p-value = .005), sends more threats.

Supplementary Table 10. Survey Questions: Anger and Social Preferences

	Questions	Choose 0 if	Choose 10 if
1	How are you feeling if Player 1 chooses Option B (right) in stage 1 in the rounds of short games?	Not angry at all	Very angry
2	How are you feeling if Player 1 chooses Option B (right) in stage 1 in the rounds of long games?	Not angry at all	Very angry
3	How are you feeling if Player 1 chooses Option D (right) in stage 2 after choosing Option B (right) in stage 1 in the rounds of long games?	Not angry at all	Very angry
4	Please describe your political orientation in general	Complete conservative	Complete liberal
5	How willing or unwilling you are to take risks	Completely unwilling to take risks	Very willing to take risks
6	How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future	Complete unwilling to do so	Very willing to do so
7	How willing are you to punish someone who treats you unfairly, even if there may be costs for you?	Complete unwilling to do so	Very willing to do so
8	How willing are you to punish someone who treats others unfairly, even if there may be costs for you?	Complete unwilling to do so	Very willing to do so
9	How willing are you to give to good causes without expecting anything in return?	Complete unwilling to do so	Very willing to do so
10	When someone does me a favor, I am willing to return it.	Does not describe me at all	Describe me perfectly
11	If I am treated very unjustly, I will take revenge at the first occasion, even if there is a cost to do so.	Does not describe me at all	Describe me perfectly
12	I assume that people have only the best intentions.	Does not describe me at all	Describe me perfectly
13	I am good at math.	Does not describe me at all	Describe me perfectly
14	I tend to postpone tasks even if I know it would be better to do them right away.	Does not describe me at all	Describe me perfectly

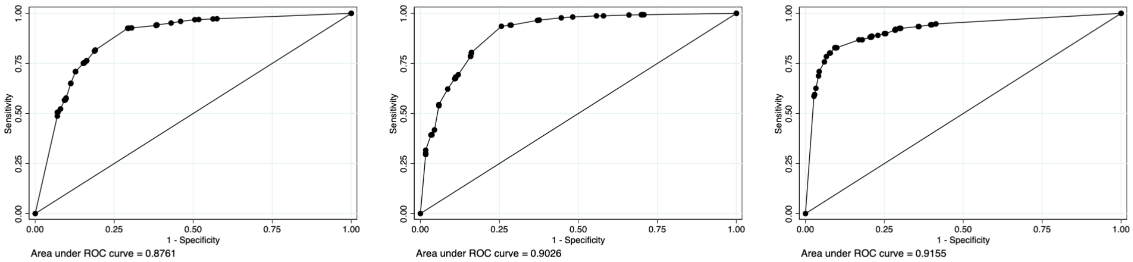
C Gender Differences

D Belief Elicitation



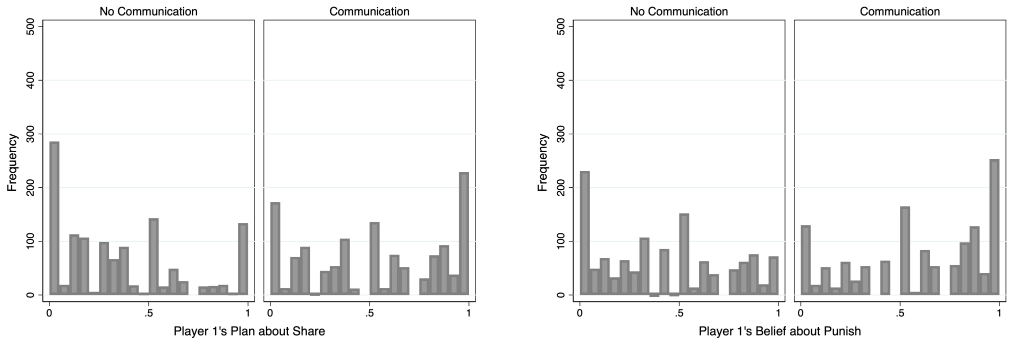
(a) P1's Plan about *Grab* **(b)** P2's Plan about *Punish*

Supplementary Figure 16. Reported Plan Predicts Own Behaviors - Deterrence Games



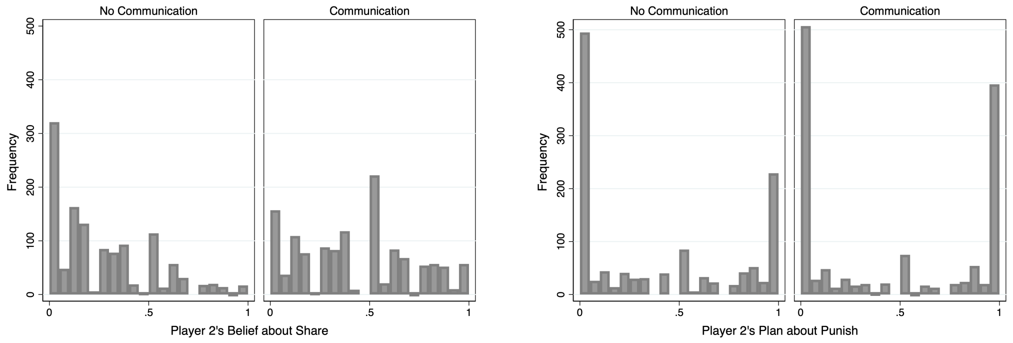
(a) P1's Plan on *Grab* (St1) **(b)** P1's Plan on *Grab* (St2) **(c)** P2's Plan on *Punish*

Supplementary Figure 17. Reported Plan Predicts Own Behaviors - Staggered Entry Games



(a) P1's Plan about *Share* **(b)** P1's Belief about *Punish*

Supplementary Figure 18. P1's Reported Beliefs Histograms



(a) P2's Belief about *Share* **(b)** P2's Plan about *Punish*

Supplementary Figure 19. P2's Reported Beliefs Histograms

E Instructions

Below are the instructions for the communication treatment. The no communication treatment instructions are identical except for the two paragraphs mentioning messages.

Experiment Instruction

Welcome to the experiment. The purpose of this experiment is to study how people make decisions in a particular situation. Please feel free to ask a question at any time by raising your hand. Please do not speak to other participants during the experiment. Cell phones are not allowed during the entire experiment.

Your will receive $10 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.

The experiment consists 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.**

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

There are two different games in the experiment, the short game and the long game.

The **Short Game** consists of two stages. The picture below may help and will be shown in each round. Player 1's payoffs are listed above Player 2's payoffs. The payoffs will change in each round. The game proceeds as follows:

- Player 1 goes first and must decide between **A** and **B**.
  - If **A** is chosen, the game ends with the payoffs specified for that round.
  - 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, the game ends and both players receive $0.

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

Prior to the start of each short game, 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 (at the discretion of the experimenter, who will monitor the messages). 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 **Long Game** consists of three stages. The picture below may help and will be shown in each round. The payoffs will change in each round. 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 with the payoffs specified for that round.
  - If **B** is chosen, the game proceeds to stage 2.
- If Player 1 chooses **B**, Player 1 must decide between **C** and **D**.
  - If **C** is chosen, the game ends with payoffs specified for that round.
  - If **D** is chosen, the game proceeds to stage 3.
- If Player 1 chooses **D**, Player 2 must decide between **E** and **F**.
  - If **E** is chosen, the game ends with payoffs specified for that round.
  - If **F** is chosen, the game ends and both players receive $0.

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

In each of the Long Games, if Player 1 chooses **B**, and before the game proceeds to stage 2, 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 (at the discretion of the experimenter, who will monitor the messages). 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.

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 the 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.

References

Aina, C., Battigalli, P., and Gamba, A. (2018). Frustration and anger in the ultimatum game: An experiment.

Ameriks, J., Caplin, A., Leahy, J., and Tyler, T. (2007). Measuring self-control problems. *The American Economic Review*, 97(3):966–972.

Averill, J. R. (1983). Studies on anger and aggression: Implications for theories of emotion. *American Psychologist*, 38(11):1145–1160.

Averill, J. R. (2012). *Anger and aggression: An essay on emotion.* Springer Science & Business Media.

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

Battigalli, P., Dufwenberg, M., and Smith, A. (2017). Frustration and anger in games.

Berkowitz, L. (1989). Frustration-aggression hypothesis: Examination and reformulation. *Psychological Bulletin*, 106(1):59–73.

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

Bradbury, J. W. and Vehrencamp, S. L. (1998). *Principles of Animal Communication.* Sinauer.

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.

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

Crawford, V. (1998). A survey of experiments on communication via cheap talk. *Journal of Economic Theory*, 78:286–298.

Crawford, V. P. and Sobel, J. (1982). Strategic information transmission. *Econometrica: Journal of the Econometric Society*, 50(6):1431–1451.

Croson, R., Boles, T., and Murnighan, J. K. (2003). Cheap talk in bargaining experiments: lying and threats in ultimatum games. *Journal of Economic Behavior & Organization*, 51(2):143–159.

Deutsch, M. and Krauss, R. M. (1960). The effect of threat upon interpersonal bargaining. *The Journal of Abnormal and Social Psychology*, 61(2):181–189.

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

Dufwenberg, M., Li, F., and Smith, A. (2018). Promises and punishment.

Ekman, P. (1992). An argument for basis emotions. *Cognition & Emotion*, 6(3-4):169–200.

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

Falk, A., Becker, A., Dohmen, T., Enke, B., Huffman, D., and Sunde, U. (2015). The nature and predictive power of preferences: Global evidence. *IZA Discussion Paper No. 9504*.

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.

Gale, J., Binmore, K. G., and Samuelson, L. (1995). Learning to be imperfect: The ultimatum game. *Games and Economic Behavior*, 8(1):56–90.

García, L. A. P., Aguilar, C. A. C., and Muñoz-Herrera, M. (2015). The bargaining power of commitment: An experiment of the effects of threats in the sequential hawk–dove game. *Rationality and Society*, 27(3):283–308.

Grecucci, A., Giorgetta, C., van't Wout, M., Bonini, N., and Sanfey, A. G. (2013). Reappraising the ultimatum: an fmri study of emotion regulation and decision making. *Cerebral Cortex*, 23(2):399–410.

Güth, W. and Kocher, M. G. (2014). More than thirty years of ultimatum bargaining experiments: Motives, variations, and a survey of the recent literature. *Journal of Economic Behavior & Organization*, 108:396–409.

Guzzini, S. (2013). *Realism in International Relations and International Political Economy: the continuing story of a death foretold*. Routledge.

Huth, P. and Russett, B. (1984). What makes deterrence work? cases from 1900 to 1980. *World Politics*, 36(4):496–526.

Manning, A. and Dawkins, M. S. (1998). *An introduction to animal behaviour*. Cambridge Univeristy Press.

Masclet, D., Noussair, C. N., and Villeval, M.-C. (2013). Threat and punishment in public good experiments. *Economic Inquiry*, 51(2):1421–1441.

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

Rankin, F. W. (2003). Communication in ultimatum games. *Economics Letters*, 81:267–271.

Schelling, T. C. (1956). An essay on bargaining. *The American Economic Review*, 46(3):281–306.

Schelling, T. C. (1958). The strategy of conflict. prospectus for a reorientation of game theory. *Journal of Conflict Resolution*, 2(3):203–264.

Selten, R. (1978). The chain-store paradox. *Theory and Decision*, 9(2):127–159.

Smith, J. M. and Price, G. R. (1973). The logic of animal conflict. *Nature*, 246:15–18.

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

Xiao, E. and Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences*, 102(20):7398–7401.