Martin's draft of Credible 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] *Experimental Economic Lab, Nanjing Audit University. Email: florali@nau.edu.cn*
[3] *Department of Economics, Virginia Tech. Email: alecsmith@vt.edu*

February 26, 2025

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,

*This paper was previously titled "Threats." We thank the participants at multiple seminars and conferences for helpful comments.

1 Introduction

Communicated threats often make the news. Vladimir Putin (or his Kremlin cronies) repeatedly flagged Russia's potential use of nuclear weapons, Donald Trump claimed he "stopped wars with the threat of tariffs," and in the midst of on-going conflict spokespersons of Israel, Iran, and the U.S. declared that further attacks would be retaliated.[1] The prevalence of threats in politics, bargaining, war, etc., suggest that they matter. Yet the mechanisms through which threats work, and their economic impact, are not well understood.

We provide a behavioral theory under which communicated threats interact with beliefs and emotions to influence behavior. The impact can be dramatic in settings where standard analyses predict no effect. We then test our theory for empirical relevance in an experiment.

Our approach is anchored in the psychology of frustration and anger (F&A). Psychologists argued that F&A influences interactions in profound ways, but few economists explored the topic. We will elaborate shortly; for now, let us just emphasize that in our analysis F&A are *belief-dependent motivations*: Frustrations occur when people encounter bad experiences *unexpectedly*, and this triggers anger and aggression. Our key contribution links the F&A-relevant beliefs to *communication*. Messages may move beliefs and switch motivation and behavior. We use the game form in Figure 1 to illustrate:

[Insert Figure 1 here!]

The players' payoffs reflect amounts of money, not their utilities which are affected by F&A. Specifically, P2's preference between *Share* and *Punish* depends on what P2 believed P1 would do, before P1 made her choice. If P2 expected P1 to choose *Out*, then P2 is surprised and frustrated if P1 chooses *In*. P2 then gets aggressive and prefers *Punish* to *Share*. If P2 instead expected *In*, then he is neither surprised nor frustrated and prefers *Share* to *Punish*. Critically, we augment this analysis to take into account the impact of pre-play communication: Suppose P2 threatens P1 that he would *Punish* were P1 to choose *In*. If P1 believes this, and if P2 believes P1 believes this, then the threat becomes credible as it would enhance P2's frustration were P1 to choose *In*, making P2 prefer *Punish* to *Share*.

[1]The Trump quote is from 9/5/2024 (luncheon, *The Economic Club of New York*). Regarding the Israel-Iran conflict, to pick one example, on 10/1/2024 a Senior White House Official stated that a "direct military attack from Iran against Israel will carry severe consequences for Iran." Psychologists and political scientists have argued that threats are common in bargaining (e.g., Deutsch and Krauss (1960)), politics, and international diplomacy (e.g., Huth and Russett (1984); Guzzini (2013)).

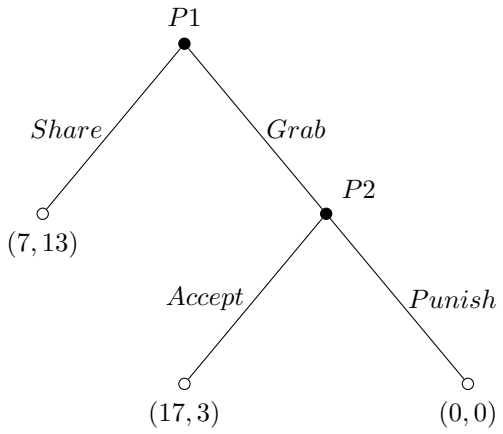


Figure 1. Deterrence Game

These conclusions are unconventional. Suppose, as is standard in traditional game theory, that players' motivations are captured by utilities defined on terminal nodes. Suppose also, as is standard in the literature on cheap talk,[2] that if P2 is allowed to send a pre-play message to P1, then the utilities (across the subgames) are unaffected. Hence, P2's preferences depend neither on message-content nor on his beliefs. Our theory, by contrast, has different implications. The reason can be mathematically identified: Unlike in traditional games, the players' utilities at terminal histories depend on beliefs. Our approach draws on the mathematical framework of so-called *psychological game theory* which allows for that.[3]

When we say "threat" we primarily mean a form of *message*. However, the term can have an alternative definition: a built-in dangerous feature of a situation.[4] In particular, it can be a *choice* that would hurt another player, like *Punish* in Fig. 1 (compare, e.g., Battigalli, Dufwenberg & Smith 2018, Def. 4). We can relate to such definitions as well, and again distinguish our approach as regards implications. Consider the classical game-theoretic literature on equilibrium refinements (e.g., Selten 1975, van Damme 1991). It may be read as addressing threats-as-choices. In particular, it rules out "non-credible threats," choices it

[2] See, e.g., Farrell & Rabin, etc…

[3] Psychological game theory was introduced by Geanakoplos, Pearce & Stacchetti (1989) and further developed by Battigalli & Dufwenberg (2009) who (2022) survey the related subsequent literature.

[4] According to *Collins English Dictionary*: "1. A threat to a person or thing [can be] a danger that something bad might happen to them. A threat is also the cause of this danger. 2. A threat [can also be] a statement by someone that they will hurt you in some way, especially if you do not do what they want."

would not be in a player's interest to carry out. Applied to P2 and choice *Punish* in Fig. 1, the identification is done without reference to P2's beliefs, unlike in our theory.[5]

Schelling (1956, 1958, 1960) discusses communicated threats and their role in deterrence at some length. Klein & O'Flaherty (1992) formalize some of Schelling's ideas. The focus differs from ours: the commitment power of the threatening party (or the credulity of others) is taken for granted, rather than justified with reference to anger & frustration.

Anger is one of the five basic emotions (Ekman, 1992), and all healthy humans experience anger (Averill, 1983, 2012). The "frustration-agression hypothesis" is a classic notion in psychology. It depicts frustration as built up from goal blockage and diminished payoff expectations, which in turn generates anger and aggression (Dollard et al., 1939; Berkowitz, 1989, 2010). Battigalli, Dufwenberg, & Smith (2015; 2019) formalize these ideas, developing a theory which is applicable to a large class of two-stage game forms. They argue that F&A influences pricing, contracting, bargaining, violence, traffic, recessions, contracting, arbitration, terrorism, and politics. Our approach is anchored in Battigalli et al's. However, to address our research questions, their framework needs to be extended. One must move beyond the class of two-stage game forms, and derive new predictions regarding the impact of communicated threats. This is what we do.

We conduct an extensive laboratory study that explores the empirical relevance of our predictions. The design comprises a class of a deterrence game forms – including that of Fig. 1 – that have a strategic structure resembling the chain-store game (Selten, 1978) and the ultimatum minigame (Gale et al., 1995). Across three treatments (motivated in detail in Section 3) we vary whether (and how) free-form messages from P2 to P1 may occur. In traditional analyses, these messages have no impact on behavior; players treat communication as irrelevant. Our theory delivers different predictions. Because of its idiosyncratic features, adequate testing requires us to pay attention to several considerations that otherwise might not be relevant. We flag the issues here: **(i)** When frustration affects utility, players' preferences may become time-inconsistent, so the so-called "strategy method" can not be justified. **(ii)** Our theory delivers predictions about behavior and beliefs; we therefore engage in extensive belief-elicitation. In the communication treatments, beliefs are elicited once before receiving messages and once after receiving messages; therefore, we can observe directly the influence of communication on reported beliefs. **(iii)** The relevant beliefs

[5]Sometimes scholars give *interpretations* involving threat-as-messages, even if these are not explicitly modeled. van Damme (1991, p.4) writes about a game akin to Fig. 1 [translated]: "[P2] threatens [P1] that he will punish [her] by playing [*Punish*] if [she] does not play [*Out*]. ... [T]his threat is not credible since [P2] will not execute it." For another similar take, see Selten (1978, p. 156)

concern not only co-player behavior but also players' *own* future actions (called "plans"). Plans affect own frustration, implying that rational plans may be non-degenerate (i.e., assigning positive probability to more than one choice). **(iv)** Because of (iii), special care must be taken to ensure that our belief-elicitation is incentive-compatible. Using monetary payments to induce prediction accuracy is treacherous. Instead, we develop and defend an honor code + flat-payment protocol. **(v)** Our theory delivers would-be predictions under counterfactual circumstances, which at first glance seem impossible to observe (given (i)). However, a fascinating by-product of (iii) & (iv) is that we can evaluate predictions under the counterfactual circumstances, using the data we elicit regarding players' plans.

No previous experiment had a combined focus on F&A and threats. Persson (2018), Aina et al (2020), and Dufwenberg et al (2024) test aspects of Battigalli et al's F&A model; the former two references do not incorporate communication, while the latter (a companion piece) looks at promises rather than threats (in a different game). There is also a large literature on communication in strategic environments which, however, does not deal with F&A. See Crawford & Sobel (1982, Crawford (1998), Charness & Dufwenberg (2006), and Balliet (2010) for starters. Among the few experiments that study threats, see Rankin (2003), Croson et al. (2003), Masclet et al. (2013), García et al. (2015), and Ellingsen and Johannesson (2004) for work which is very interesting although there is no focus on F&A.

Section **??** develops our theory, Section 4 presents our experimental design, procedures, and hypotheses, Section 5 presents results, and Section 6 concludes.

2 Model

2.1 Definitions

Game form We consider finite two-player extensive game forms with perfect information. H is the set of histories, $Z \subseteq H$ the set of terminal histories, and $Z(h) \subseteq Z$ the set of terminal successors of $h \in H$. $A_i(h)$ is the set of player i's actions at $h \in H$ (taken as singleton if i is not active at h). $\pi_i : Z \to \mathbb{R}$ is i's (monetary) payoff function (not to be confused with i's utility below).

Beliefs At $h \in H$, player i holds beliefs about subsequent play summarized by **belief system** $\alpha_i \in \times_{h \in H} \Delta(Z(h))$.[6] From α_i, derive $\alpha_{i,i} \in \times_{h \in H} \Delta\left(A_i(h)\right)$ and $\alpha_{i,j} \times_{h \in H} \Delta\left(A_j(h)\right)$ which have the mathematical structure of (behavior) strategies. Refer to $\alpha_{i,i}$ as i's "plan" and to $\alpha_{i,j}$ as i's beliefs about j's choices.

Frustration Given $h \in H$, let h' be the history that precedes h, unless h is root in which case $h' = h$. Let $\mathbb{E}[\pi_i | h''; \alpha_i]$ be i's expected payoff evaluated at $h'' \in H$ (calculated using $\alpha_i(\cdot | h'')$). Player i's frustration at $h \in H$ equals

$$\mathrm{F}_i(h; \alpha_i) = \left[\mathbb{E}[\pi_i | h'; \alpha_i] - \max_{a_i \in A_i(h)} \mathbb{E}[\pi_i | (h, a_i); \alpha_i] \right]^+,$$

where $[x]^+ = \max\{x, 0\}$. $\mathrm{F}_i(h; \alpha_i)$ equals the gap, if positive, between i's expected payoff at h' and the highest expected payoff i can obtain at h.

Utility Anger and aggression is triggered by frustration: player i's utility at $h \in H$ is a function $u_{i,h} : Z \times \Delta_i^1 \to \mathbb{R}$ defined by

$$u_{i,h}\left(z, \alpha_i\right) = \pi_i(z) - \theta_i \cdot \mathrm{F}_i(h; \alpha_i) \cdot \pi_j\left(z\right),$$

where $j \neq i$ and $\theta_i \geq 0$ is the parameter that measures how important it is to i to vent her or his frustration by hurting j.[7] The frustration-aggression hypothesis is modeled by assuming that $\theta_i > 0$.

Sequential equilibrium (SE) Our experiment allows subjects to play many times. Players may then (eventually) coordinate on an equilibrium where they hold correct beliefs about each other and maximizes $u_{i,h}$ at every $h \in H$:

Definition: Consider strategy profile $\sigma = (\sigma_i)_{i=1,2}$ and belief systems $(\alpha_i)_{i=1,2}$ such that $\alpha_{i,i} = \sigma_i$ and $\alpha_{i,j} = \sigma_j$ for $j \neq i$. Profile σ is an SE if the following condition holds for

[6] Given a set S, $\Delta(S)$ is the set of probability distributions on S.

[7] The presence of argument α_i in $u_{i,h}$ implies that we have a psychological game; compare Geanakoplos et al. (1989) and Battigalli & Dufwenberg (2009, 2022). If $\theta_i = 0$ one could re-write $u_{i,h}$ without α_i as an argument, without affecting the functional form.

all $i = 1, 2$ and $h \in H$:

$$\sigma_i(a_i|h) > 0 \Rightarrow a_i \in \arg \max_{a_i' \in A_i(h)} \sum_{z \in Z(h)} [\alpha_i(z|(h, a_i')) \cdot u_{i,h}(z, \alpha_i)]$$

An SE equates i's plan and strategy, imposes that i hold correct beliefs about j, and requires that i optimize "locally" at each h taking for granted i's own behavior at $h' \neq h$. Such "local" optimization is appropriate as what i wants to do at h (e.g., when frustrated) may differ from what i might wish would happen at h from the vantage point of $h' \neq h$.[8]

2.2 Deterrence Game

We begin with the deterrence game depicted in Figure 2, where the numbers and variables at the end nodes denote monetary payoffs.[9] There are two players, Player 1 (P1 for short) and Player 2 (P2). The parameters 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*.

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): $(Grab; Accept)$.

[8] Our definition is an adaptation of BD&S's; they limit attention to 2-stage game forms but allow for imperfect information. Battigalli & Dufwenberg (2009) define SE for a larger class of psychological games. They, and we, generalize the SE of Kreps & Wilson (1982). To illustrate in our context: if $\theta_i = 0$ for all i then our SE conforms with backward induction if π_i were each i's utility (follows from the "one-shot-deviation principle;" compare B&D p.17).

[9] Note that the figure depicts a "game form" in the terminology of Battigalli and Dufwenberg (2022).

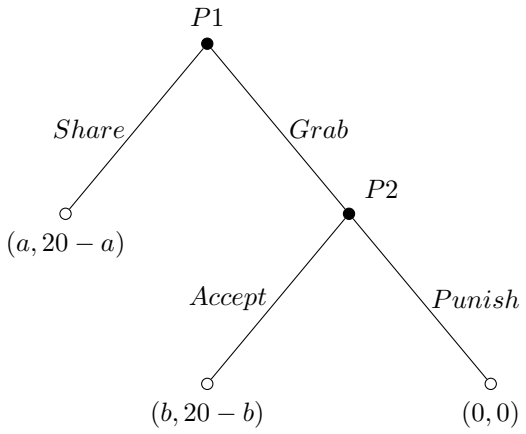


Figure 2. Deterrence Game

2.3 SE of the deterrence game with frustrated anger

The deterrence game has two pure-strategy psychological sequential equilibria (SE) when P2's anger sensitivity parameter θ_2 is sufficiently large. 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.

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.

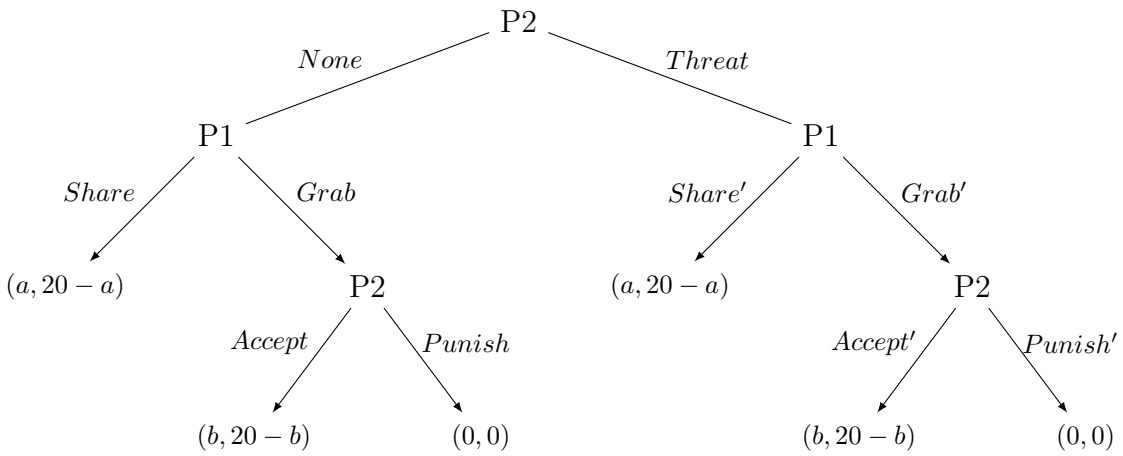


Figure 3. Deterrence game with threats

3 Modeling threat messages with complete information

Consider Figure 3. P2 moves first sending a *Threat* or by choosing *None*. The subgames following both messages are identical.

In our model, players cannot be frustrated at the initial node. However, preplay messages might change beliefs and cause frustration. Below, we model this explicitly.

3.1 Equilibrium Analysis (with slow play)

With Slow play (fast reference updating) we need to first identify the SE of the subgame. First, the selfish SE exists with Slow Play: At the root ($h = \emptyset$) P1 expects $20 - b$. After *Grab*, P2 still expects $20 - b$, is not frustrated, and chooses to *Accept*. Second, the deterrence SE still exists with Slow Play. At the root P1 expects $20 - a$. If P2 deviates to choose *Grab*, the relevant reference belief for P1 is the one she held at the root. Thus P1 is frustrated after *Grab* and chooses *Punish*. More generally, Slow play and Fast play coincide in 2-stage games.

3.1.1 Deterrence game

Theorem 1. *An equilibrium of the deterrence game with communication (Figure 3) exists whereby Player 2 sends a threat and*

i) Play follows the deterrence equilibrium in the subgame after the threat

ii) Play follows the selfish equilibrium after no threat (off path)

In short, threat messages are effective deterrents.

Proof. First, consider the deterrence game, and allow for preplay messages as in Figure 3. Consider the candidate SE where P1 sends a *Threat*, and play follows the deterrence SE after *Threat* and follows the selfish SE after *None*. At the root P1 expects the highest possible payoff of $20 - b$, from *Share*. That is, consider the non-degenerate action profile $((Grab, Share'), (Threat, Accept, Punish'))$. We assume that beliefs are consistent with this profile, and check for sequential rationality to verify that this profile is an SE.

We have already demonstrated that the deterrence SE and the selfish SE exist in each subgame. Here we assume that the deterrence SE follows *Threat* and the selfish SE follows *None*.

After *Threat*, P2 expects *Share*. If P1 deviates to select *Grab'*, then P2's frustration will be $(20 - a) - (20 - b) = b - a > 0$ (as in the deterrence SE). Therefore, with sufficiently large θ_2, P2 will select *Punish'*, so P1 will prefer to *Share*.

After *None*, P2 expects *Grab*, and so after $(None, Grab)$, P2 expects payoff $20 - b$, and hence is not frustrated, and chooses *Accept*. Note that this is the case event though P2 had expected payoff $(20 - a)$ at h^0; after $(None, Grab)$, the relevant reference belief is P2's expected payoff after $None$, where P2 expects $(20-b)$. Finally, P1 has no incentive to deviate to *Share*, since $b > a$. Thus, each players' behavior in each of the subgames is sequentially rational.

At the root (h^0), P2 compares the payoff from $None$ to that from $Threat$ and chooses $Threat$, since again $(20-a)-(20-b) = b-a > 0$. Thus, $((Grab, Share'), (Threat, Accept, Punish'))$ is an SE. When the reference belief updates after each action, the arrival of a threat changes beliefs and influence play. □

3.1.2 Other equilibria with slow play

Neither player can be frustrated in equilibrium

For the game in Figure 3, it is natural to ask if receiving a threat might frustrate a Player 1. Thus we are first interested in asking whether there are equilibria where P1 becomes frustrated after receiving a threat, and hence chooses *Grab*, negating the effect of the threat. More formally, can $((None, Accept, Punish'), (Grab, Grab))$ be an SE? In this conjectured equilibrium, there is mutual frustration off the equilibrium path. However, as we will see, this cannot be an SE.

Theorem 2. *There is no pure-strategy SE of the deterrence game with threats involving mutual frustration.*

Proof. First, along the equilibrium path: P1 cannot be frustrated after receiving no threat (*None*). P2 expects *Grab*, and so is not frustrated when it is selected by P1. Because P2 is not frustrated, she prefers the material payoff $b > 0$ from *Accept*.

Next suppose that a deviation (or a tremble) by P2 leads to the history $h = Threat$, and P1 gets the move. Given beliefs, the best payoff P1 can now get is now a; the higher b payoff is unattainable since P2 will choose to *Punish* in the conjectured equilibrium. So $F_1(Threat; \alpha_1) = b - a > 0$, and P1 experiences positive frustration. However:

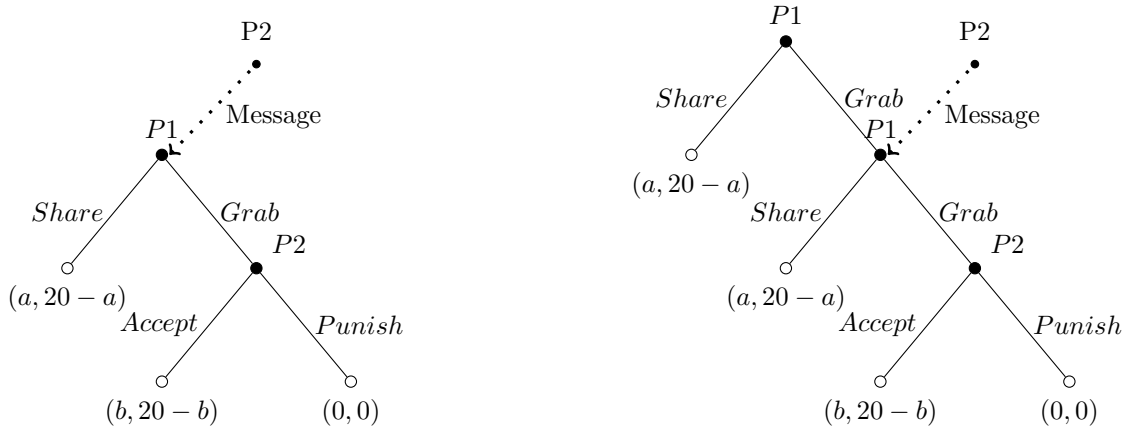
$$
\begin{aligned}
&u_1((Threat, Grab); \alpha_1) = 0 \\
&u_1((Threat, Share); \alpha_1) = a - \theta_1(b-a)(20-a) \\
&\text{so P1 prefers } Grab \text{ if} \\
&0 > a - \theta_1(b-a)(20-a) \\
&\theta_1 > \frac{a}{(b-a)(20-a)}
\end{aligned}
$$

Suppose this condition holds and $\theta_1 > \frac{a}{(b-a)(20-a)}$ and $h = (Threat, Take)$. In the conjectured equilibrium, P2 initially expected $20 - b$ [**note to martin: this next part changes from fast play**], but after *Threat*, P2 expects 0 (note that deviations and/or trembles do not change P1's beliefs). After $(Threat, Grab')$ P2 can still achieve $20 - b$, so $F_2((Threat, Take); \alpha_2) = \max 20 - b, 0 = 0$. Since $F_2(\cdot) = 0$, P2 will deviate to *Accept'*. Thus $((None, Accept, Punish'), (Grab, Grab'))$ is not an SE. □

4 Experiment

4.1 Design

We use a between-subject design where the treatment variable is pre-play communication.[10] 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 4(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 4. 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 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.

[10] Dufwenberg et al. (2021) showed that communication effect is persistent throughout the whole session. Therefore, we employ a between-subject design for communication treatment in this paper.

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 **??**. 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.[11]

Table 1. Game Variations

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

4.2 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 were incentivized via a flat fee.[12] 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

[11] 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.

[12] Other works employing this method include Toussaert (2018); Ameriks et al. (2007) and Dufwenberg et al. (2021).

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 5.

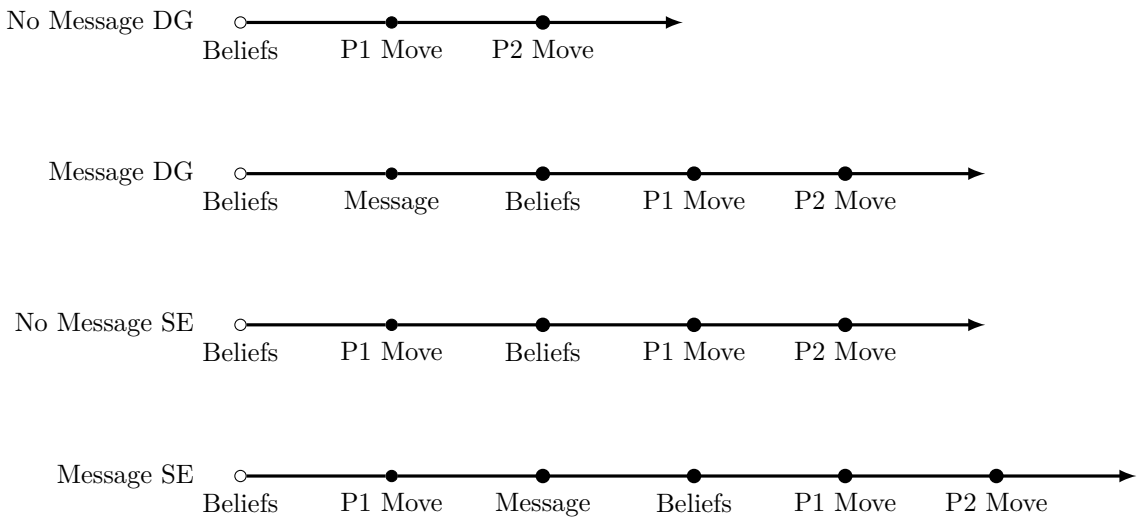


Figure 5. Experiment Timeline

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

4.3 Hypotheses

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

Based on previous studies on communication (Ellingsen and Johannesson, 2004; Charness and Dufwenberg, 2006; Vanberg, 2008; Dufwenberg et al., 2021), we expect that

Hypothesis 1. *Communication has a strong deterrence effect, and promotes higher social welfare.*

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. 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 deterrence and 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.

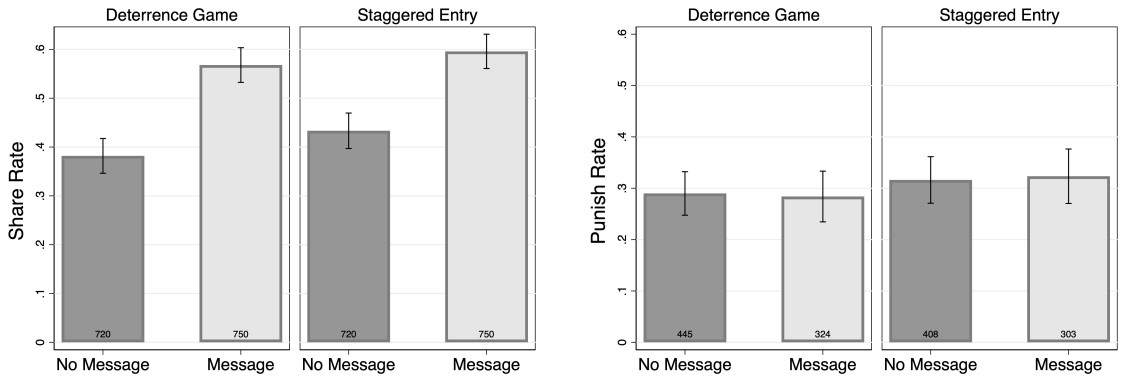
Hypothesis 5. *Players eventually reach to one of the two Sequential Equilibrium* ($\{Share, Punish\}$ *and* $\{Grab; Accept\}$)*. Threats select* $\{Share; Punish\}$ *to be reached more often.*

5 Results

This section is organized as follows: Section 5.1 summarizes the overall behavioral treatment effect on deterrence and costly punishment, and test Hypothesis 1. Section 5.2 presents behavioral outcome on threats vs. cheap talk to test Hypothesis 2. Section 5.3 constructs the belief-dependent motivation, and discusses relationship between threats and expectations (Hypothesis 3)and relationship between beliefs and behavior (Hypothesis 4). Section 5.4 presents the behavioral tendency on equlibrium selection (Hypothesis 5).

5.1 The Effect of Communication

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 6(a), with the vertical bar representing the 95% confidence interval.



(a) P1's *Share* Rate **(b)** P2's *Punish* Rate

Figure 6. Outcome Summary with Communication Treatment Effect

At first glance the communication treatment does not seem to have an effect on P2's

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.

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 6(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. (2021) 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

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

	P1's *Share* Choice		P1's *Share* Plan	
	A coef / se	B coef / se	C coef / se	D 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.

(between subjects) we report the results from linear regressions that pool the data for a given 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 non-parametric results, when regressing P1's *Share* choice (Table 3), communication increases *Share* rate significantly, and when regressing P2's *Punish* choice (Table **??**) , 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 18 & 19 (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 7. In total, communication helps to increase welfare (P1's and P2's payoffs combined) by \$1.05 (1-sided Fisher's exact, $p < .001$).

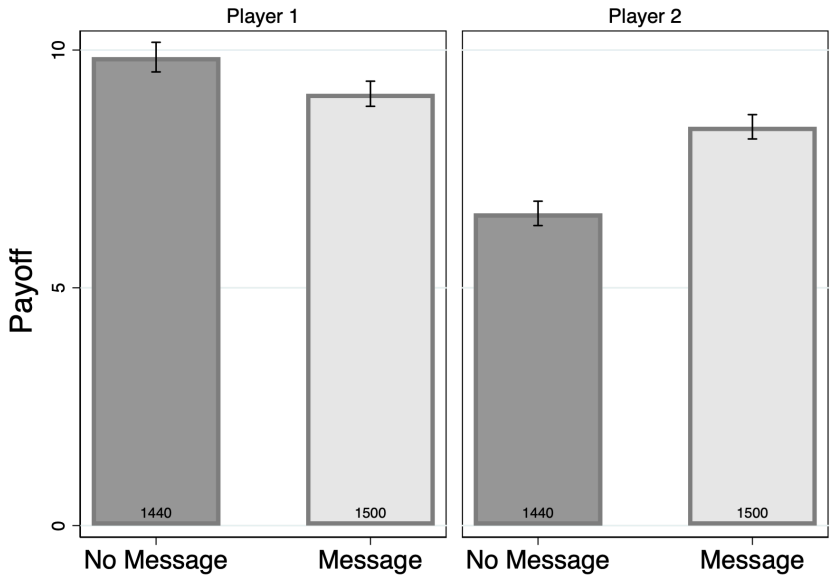


Figure 7. Payoff Distribution

5.2 The Credibility of Threats

To examine the effect of message contents on behavior (Hypothesis 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 8 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 4, 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 4. 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 4 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 2, that threats result in a higher *Share* rate in both games. These results are graphically presented in Figure 9(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 4 demonstrates that for the deterrence games, the conditional *Punish* rate is significantly

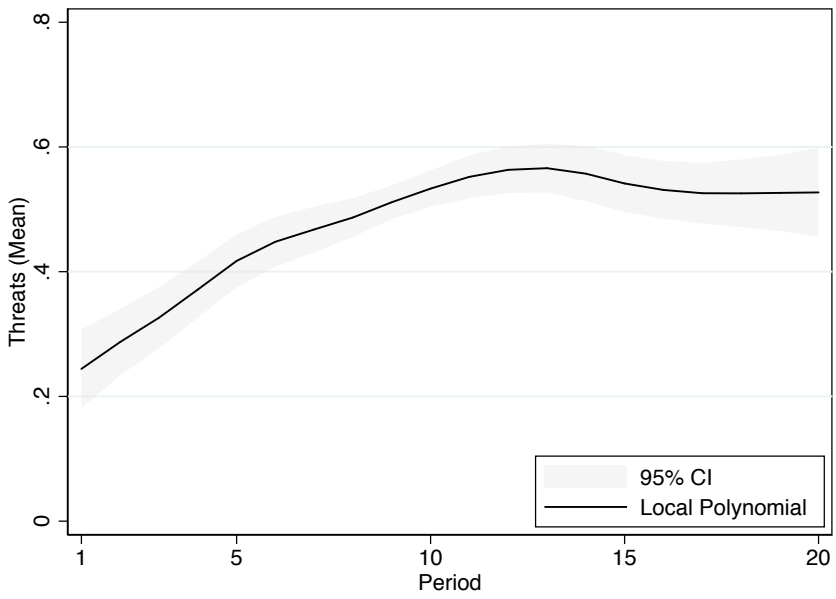


Figure 8. Number of Threats in Each Period

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 9(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.

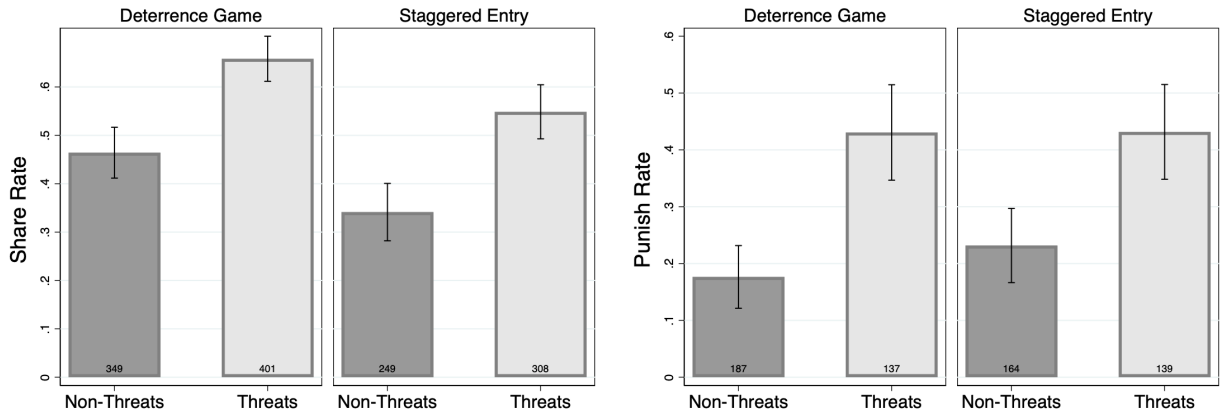
Using only the communication data, we examine the effect of threats on behavior with subject level fixed effect logistic regressions in Table 5. "Payoff from *Accept*" and the indicator variable "Staggered Entry" are used to control for individual games, and "Period" is used to control for extent of 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.

Table 5. 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.



(a) P1's *Share* Rate

(b) P2's *Punish* Rate

Figure 9. Outcome Summary Comparing Threats vs. Cheap Talk

5.3 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 6 presents summary statistics for self-reported beliefs (both players' beliefs about *Share* and *Punish*) recorded after messages are received, and Figures 20 and 21 (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 5.1, self-reported beliefs and plans are good predictors of participant behavior (see Figures 18 & 19 in the Appendix for ROC analyses).

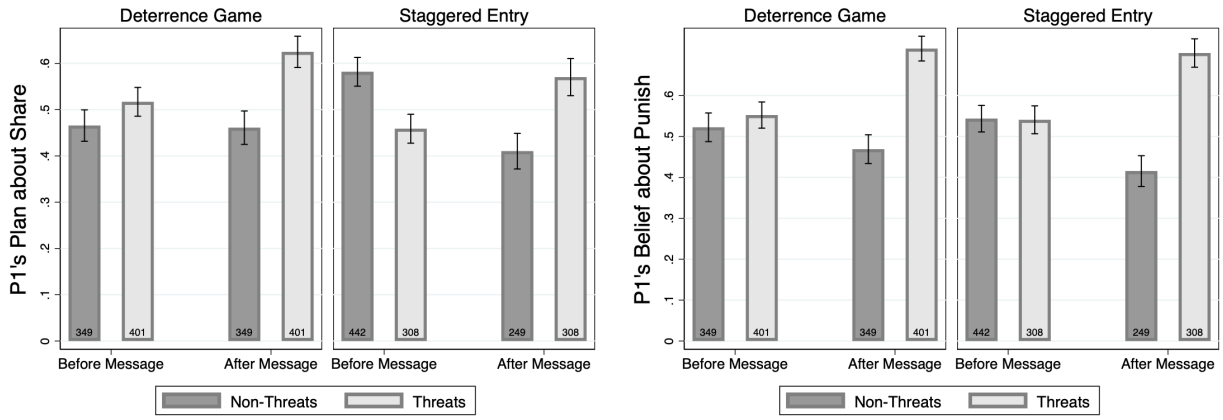
Table 6. Summary Statistics – Reported Beliefs

	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.

For both players and for both types of games, the effect of communication on reported beliefs is driven by the messages containing threats (Figures 10 & 11), 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 10(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.



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

Figure 10. P1's Reported Beliefs

We note a similar pattern in P1's reported 1st order beliefs about P2's *Punish* choices. Figure 10(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.

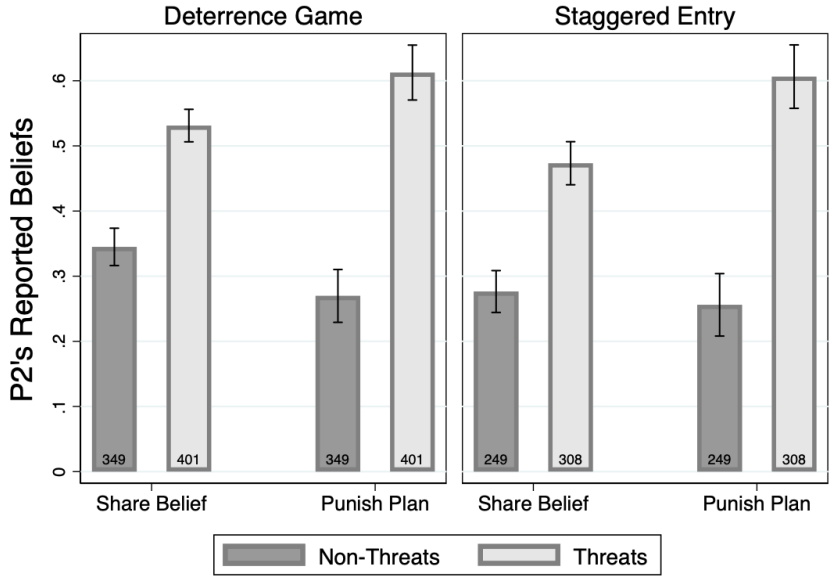


Figure 11. P2's Reported Beliefs

Figure 11 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 7, 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 5 regression model B to Table 7 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 7 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 8 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 7, 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 7. 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 8. 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.

5.4 Sequential Equilibrium Selection

In Figure 12, we see a pattern for convergence in either of the Sequential Equilibrium ($\{Share; Punish\}$ and $\{Grab; Accept\}$). In addition, we see a equilibrium selection over $\{Share; Punish\}$ as well. The non-equilibrium outcome *Punish* happens least frequently; 15.24% of the games end with *Punish*, with no notable change in rate throughout the experiment. Figure 12 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 $p < .001$).

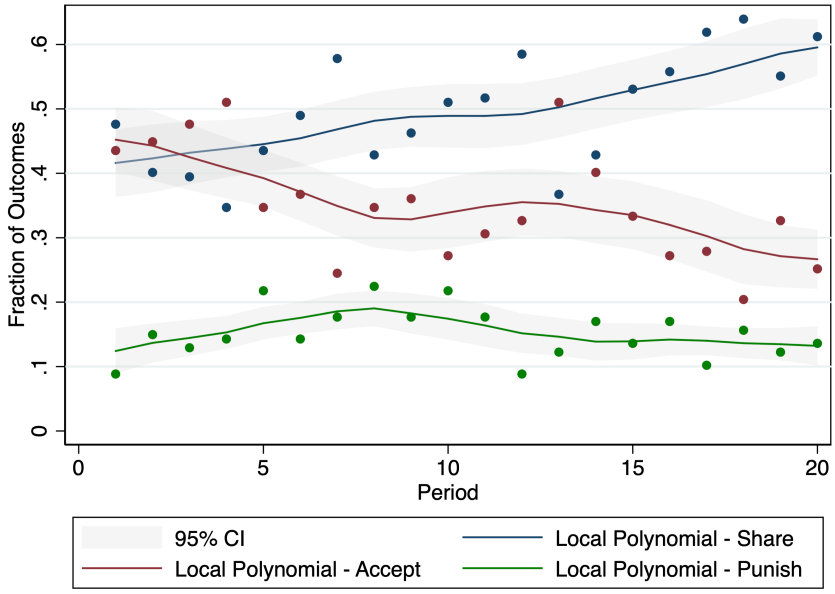


Figure 12. Equilibrium Convergence

[**zzNOTE**]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.

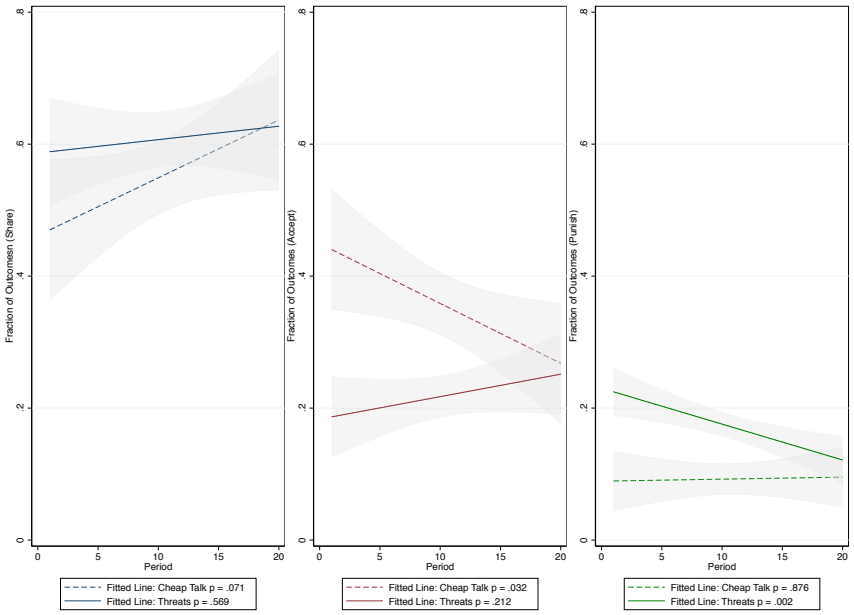


Figure 13. Equilibrium Convergence

6 Conclusion

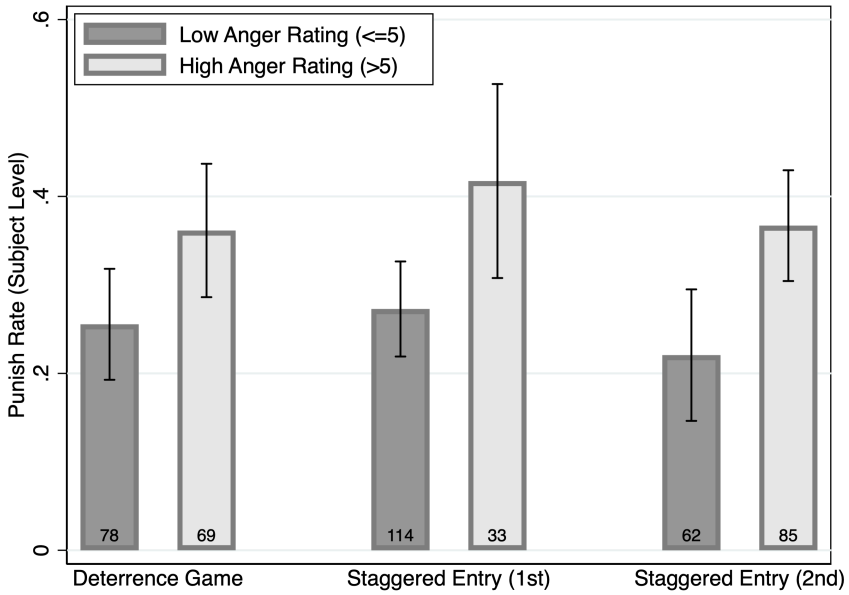
In this paper, we study the relationship between threats, credibility, and costly punishment, deriving theoretical predictions from the model of belief-dependent anger of Battigalli et al. (2015, 2019). When combined with the notion that communicated messages influence beliefs, our model implies that 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

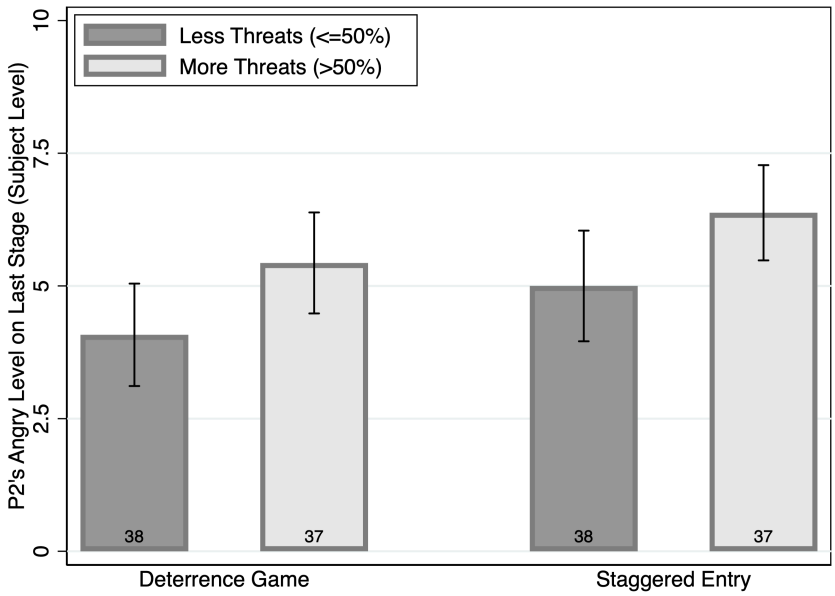
After the experiment concludes, we elicited self-reported measures of anger from participants assigned the role of Player 2. We are able to examine whether an individuals' level of anger is correlated with their behavior. Various studies have shown that the ultimatum game induces negative emotions especially 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 9 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 14. Greater Anger with Higher Punish Rate

In Supplementary Figure 14, 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.



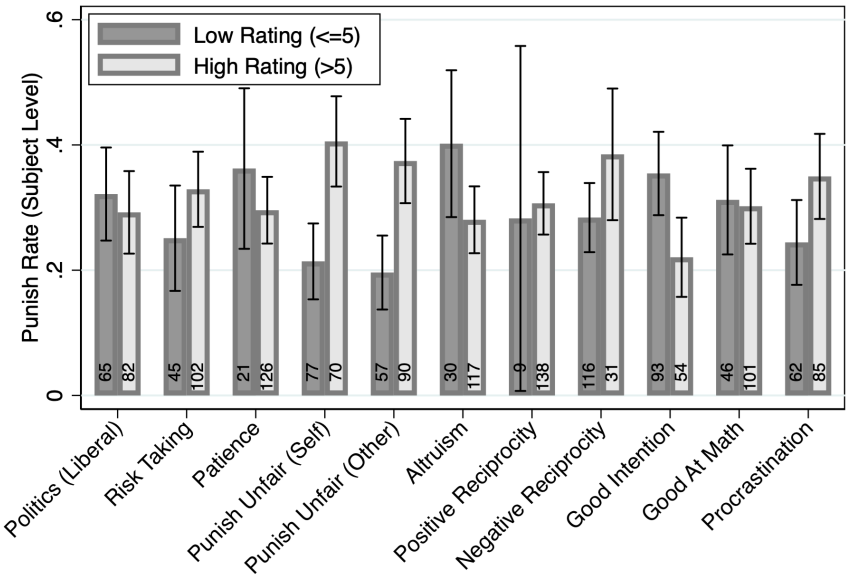
Supplementary Figure 15. Greater Anger at Disregarded Threats

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 15, 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 9 for the exact questions.

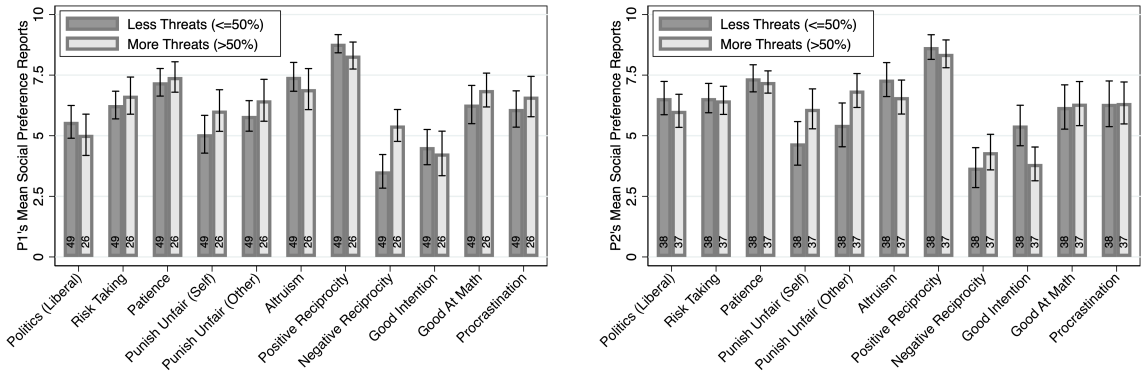
The relationship between self-reported social preferences and the *Punish* rate is depicted



Supplementary Figure 16. Social Preferences and Punish Rate

in Supplementary Figure 16. 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 17. Social Preferences Reports with Threats vs. Cheap Talk

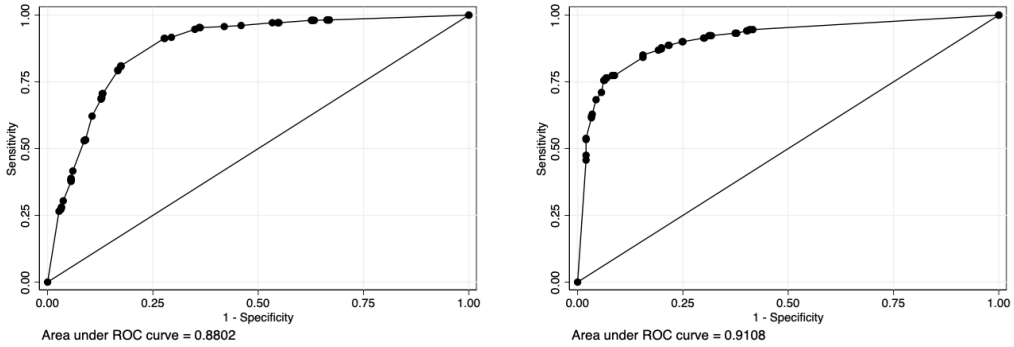
Supplementary Figure 17 illustrates that, in the communication treatment, depending on the message contents, P2 reports different ratings for some social preferences. But P1 again reports the same social preferences with or without threats, except for negative reciprocity (p-value = .001). P2 who reports 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 9. 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

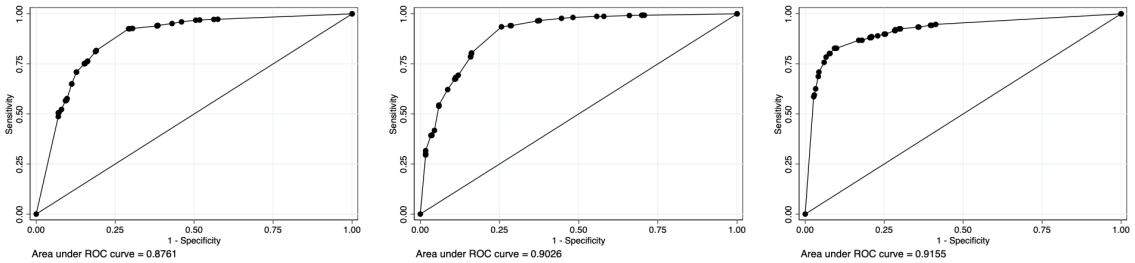
C Gender Differences

D Belief Elicitation



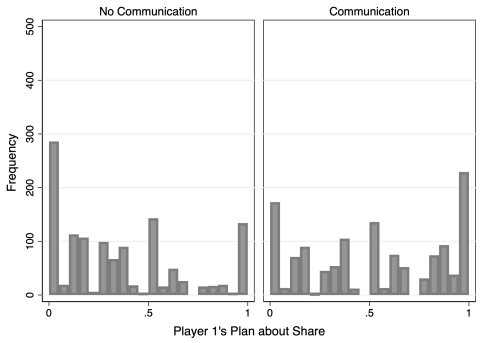
(a) P1's Plan about *Grab* **(b)** P2's Plan about *Punish*

Supplementary Figure 18. Reported Plan Predicts Own Behaviors - Deterrence Games

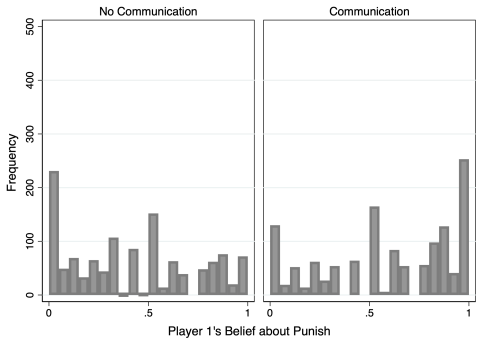


(a) P1's Plan on *Grab* (St1) **(b)** P1's Plan on *Grab* (St2) **(c)** P2's Plan on *Punish*

Supplementary Figure 19. Reported Plan Predicts Own Behaviors - Staggered Entry Games

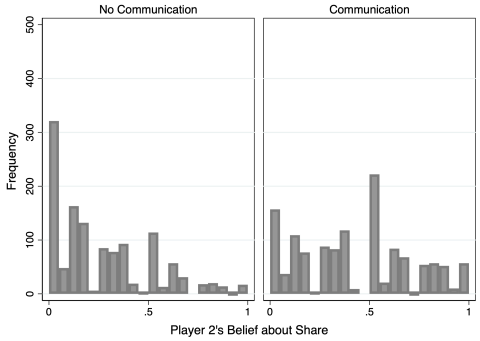


(a) P1's Plan about *Share*

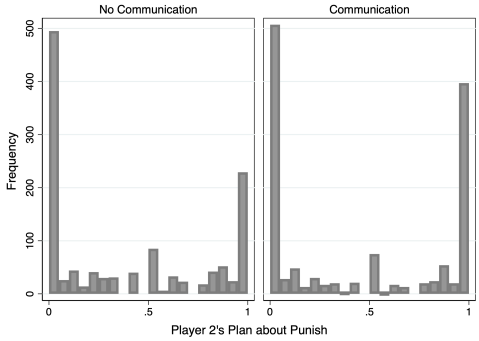


(b) P1's Belief about *Punish*

Supplementary Figure 20. P1's Reported Beliefs Histograms



(a) P2's Belief about *Share*



(b) P2's Plan about *Punish*

Supplementary Figure 21. 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.

F Questionnaire

Player 2 Self-Reported Anger Rating (Communication Treatment)

1. How are you feeling if Player 1 chooses Option B (right) in stage 1 in the rounds of short games after receiving your message? Please indicate your answer on a scale from 0 to 10. A 0 means "not angry at all," and a 10 means "very angry"
2. How are you feeling if Player 1 chooses Option B (right) in stage 1 in the rounds of long games? Please indicate your answer on a scale from 0 to 10. A 0 means "not angry at all," and a 10 means "very angry"
3. How are you feeling if Player 1 chooses Option D (right) in stage 2 after receiving your message (such that Player 1 chose Option B (right) in stage 1) in the rounds of long games? Please indicate your answer on a scale from 0 to 10. A 0 means "not angry at all," and a 10 means "very angry"

Player 2 Self-Reported Anger Rating (Non-Communication Treatment)

1. How are you feeling if Player 1 chooses Option B (right) in stage 1 in the rounds of short games? Please indicate your answer on a scale from 0 to 10. A 0 means "not angry at all," and a 10 means "very angry"
2. How are you feeling if Player 1 chooses Option B (right) in stage 1 in the rounds of long games? Please indicate your answer on a scale from 0 to 10. A 0 means "not angry at all," and a 10 means "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? Please indicate your answer on a scale from 0 to 10. A 0 means "not angry at all," and a 10 means "very angry"

Socioeconomic Survey

1. Gender
   (a) Male
   (b) Female

(c) Other

2. Age

3. Are you Hispanic or Latino?

   (a) Yes

   (b) No

4. How would you describe yourself?

   (a) American Indian or Alaska Native

   (b) Asian

   (c) Black or African American

   (d) Native Hawaiian or Other Pacific Islander

   (e) White

   (f) Caucasian

5. How many years have you been at Virginia Tech?

   (a) 1

   (b) 2

   (c) 3

   (d) 4

   (e) 5

   (f) 6

   (g) More than 6

6. Do you regularly attend religious services?

   (a) Yes

   (b) No

7. What is your household income relative to other students at Virginia Tech?

   (a) Significantly higher

   (b) Somewhat higher

(c) About the same

(d) Somewhat lower

(e) Significant lower

8. In addition to school, do you

   (a) Work at a full-time job?

   (b) Work at a part-time job?

   (c) Do not have a job.

9. Are you a honor student?

   (a) Yes

   (b) No

10. What is your Major/College? (If more than one applies, pick the one that you consider to be your primary Major or College)

    (a) Economics (either in COB or COS)

    (b) Agriculture and Life Sciences

    (c) Architecture and Urban Studies

    (d) Business other than Economics

    (e) Engineering

    (f) Liberal Arts and Human Sciences

    (g) Natural Resources and Environment

    (h) Science other than Economics

11. How many Economics classes have you taken at the university level?

    (a) None

    (b) 1

    (c) 2

    (d) 3

    (e) 4 or more

12. What is your MOTHER's current occupation? If she is retired or deceased, please list her most recent occupation.

13. Please indicate the highest level of education your MOTHER completed:

    (a) Some high school

    (b) High school diploma or equivalent

    (c) Some college or associate degree

    (d) B.A, B.S., or other bachelor degrees

    (e) M.A., M.S., M.B.A., or other master degrees

    (f) M.D., J.D., PhD, or other doctoral degrees

    (g) Other

14. What is your FATHER's current occupation? If he is retired or deceased, please list his most recent occupation.

15. Please indicate the highest level of education your FATHER completed:

    (a) Some high school

    (b) High school diploma or equivalent

    (c) Some college or associate degree

    (d) B.A, B.S., or other bachelor degrees

    (e) M.A., M.S., M.B.A., or other master degrees

    (f) M.D., J.D., PhD, or other doctoral degrees

    (g) Other

16. Please describe your political orientation in general, using a scale from 0 to 10, where 0 means you are "complete conservative" and 10 means you are "complete liberal."

Risk and Social Preferences

1. How willing or unwilling you are to take risks, using a scale from 0 to 10, where 0 means you are "completely unwilling to take risks" and 10 means you are "very willing to take risks."

2. How willing are you to give up something that is beneficial for you today in order to benefit more from that in the future? Please again indicate your answer on a scale from 0 to 10. A 0 means "completely unwilling to do so," and a 10 means "very willing to do so."

3. How willing are you to punish someone who treats you unfairly, even if there may be costs for you? Please again indicate your answer on a scale from 0 to 10. A 0 means "completely unwilling to do so," and a 10 means "very willing to do so."

4. How willing are you to punish someone who treats others unfairly, even if there may be costs for you? Please again indicate your answer on a scale from 0 to 10. A 0 means "completely unwilling to do so," and a 10 means "very willing to do so."

5. How willing are you to give to good causes without expecting anything in return? Please again indicate your answer on a scale from 0 to 10. A 0 means "completely unwilling to do so," and a 10 means "very willing to do so."

6. How well does the following statement describe you as a person? "When someone does me a favor, I am willing to return it." Please indicate your answer on a scale from 0 to 10. A 0 means "does not describe me at all," and a 10 means "describes me perfectly."

7. How well does the following statement describe you as a person? "If I am treated very unjustly, I will take revenge at the first occasion, even if there is a cost to do so." Please indicate your answer on a scale from 0 to 10. A 0 means "does not describe me at all," and a 10 means "describes me perfectly."

8. How well does the following statement describe you as a person? "I assume that people have only the best intentions." Please indicate your answer on a scale from 0 to 10. A 0 means "does not describe me at all," and a 10 means "describes me perfectly."

9. How well does the following statement describe you as a person? "I am good at math." Please indicate your answer on a scale from 0 to 10. A 0 means "does not describe me at all," and a 10 means "describes me perfectly."

10. How well does the following statement describe you as a person? "I tend to postpone tasks even if I know it would be better to do them right away." Please indicate your answer on a scale from 0 to 10. A 0 means "does not describe me at all," and a 10 means "describes me perfectly."

11. Is there anything else you would like to tell the experimenters about this experiment?

References

Ameriks, J., Caplin, A., Leahy, J., and Tyler, T. (2007). Measuring self-control problems. *The American Economic Review*, 97(3):966–972. 12

Averill, J. R. (1983). Studies on anger and aggression: Implications for theories of emotion. *American Psychologist*, 38(11):1145–1160. 3

Averill, J. R. (2012). *Anger and aggression: An essay on emotion*. Springer Science & Business Media. 3

Battigalli, P. and Dufwenberg, M. (2022). Belief-dependent motivations and psychological game theory. *Journal of Economic Literature*, 60(3):833–882. 6

Battigalli, P., Dufwenberg, M., and Smith, A. (2015). Frustration and anger in games. 3, 30

Battigalli, P., Dufwenberg, M., and Smith, A. (2019). Frustration, aggression, and anger in leader-follower games. *Games and Economic Behavior*, 117:15–39. 3, 30

Berkowitz, L. (1989). Frustration-aggression hypothesis: Examination and reformulation. *Psychological Bulletin*, 106(1):59–73. 3

Berkowitz, L. (2010). Appraisals and anger: how complete are the usual appraisal accounts of anger? In *International handbook of anger*, pages 267–286. Springer. 3

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. 12

Charness, G. and Dufwenberg, M. (2006). Promises and partnership. *Econometrica*, 74(6):1579–1601. 14

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. 4

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. 1

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. 3

Dufwenberg, M., Li, F., and Smith, A. (2021). Promises and punishment. 11, 12, 14, 16

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

Ellingsen, T. and Johannesson, M. (2004). Promises, threats and fairness. *The Economic Journal*, 114(495):397–420. 4, 14

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*. 13, 32

Fischbacher, U. (2007). z-tree: Zurich toolbox for ready-made economic experiments. *Experimental Economics*, 10(2):171–178. 12

Gale, J., Binmore, K. G., and Samuelson, L. (1995). Learning to be imperfect: The ultimatum game. *Games and Economic Behavior*, 8(1):56–90. 3

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. 4

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. 31

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. 31

Guzzini, S. (2013). *Realism in International Relations and International Political Economy: the continuing story of a death foretold*. Routledge. 1

Huth, P. and Russett, B. (1984). What makes deterrence work? cases from 1900 to 1980. *World Politics*, 36(4):496–526. 1

Masclet, D., Noussair, C. N., and Villeval, M.-C. (2013). Threat and punishment in public good experiments. *Economic Inquiry*, 51(2):1421–1441. 4

Persson, E. (2018). Testing the impact of frustration and anger when responsibility is low. *Journal of Economic Behavior & Organization*, 145:435–448. 4

Rankin, F. W. (2003). Communication in ultimatum games. *Economics Letters*, 81:267–271. 4

Selten, R. (1978). The chain-store paradox. *Theory and Decision*, 9(2):127–159. 3

Toussaert, S. (2018). Eliciting temptation and self-control through menu choices: a lab experiment. *Econometrica*, 86(3):859–889. 12

Vanberg, C. (2008). Why do people keep their promises? an experimental test of two explanations. *Econometrica*, 76(6):1467–1480. 14

Xiao, E. and Houser, D. (2005). Emotion expression in human punishment behavior. *Proceedings of the National Academy of Sciences*, 102(20):7398–7401. 31