The Behavioral and Neural Bases of Social Economic Decision-Making

Zhuncheng (Flora) Li

Dissertation submitted to the Faculty of the
Virginia Polytechnic Institute and State University
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy
in
Economics, Science

Sheryl Ball, Co-Chair
Alec Smith, Co-Chair
Martin Dufwenberg
Eric Bahel
Sudipta Sarangi

March 8, 2019
Blacksburg, Virginia

Keywords: Social Decision-Making, Other-Regarding, Expectation, Communication, Emotion, Psychological Games

The Behavioral and Neural Bases of Social Economic Decision-Making

Zhuncheng (Flora) Li

Abstract

Social economic decision-making considers the well-being and emotions of others. Unlike traditional economics which routinely assumes that individuals care only about their own outcomes, behavioral economics and neuroeconomics offer research strategies which help us explore our social motivations. This dissertation consists of three essays studying the underlying behavioral and neural mechanisms of individuals' social economic decision-making. The analyses focus on investigating experimentally how humans make decisions in three distinct social economic environments.

Chapter 2 examines how individuals react to hold-up when explicit promises are available. Hold-up happens when two parties can form an incomplete contract to cooperate, but the agreement may fall apart due to concerns about the other party gaining bargaining power. We propose that a belief-dependent frustration anger model can explain behavior about investment, cooperation, and costly punishment in a hold-up environment. We show experimentally that communication improves cooperation and increases efficiency. Promises lead to cooperation, and broken promises lead to costly punishment.

Chapter 3 explores threats' deterrence effect and credibility in an ultimatum bargaining environment where two parties can both benefit over trade but have a conflict of interests. We show that a belief-dependent frustration anger model captures the relationship among messages, beliefs, and behavior. Our design permits the observation of communicated threats, credibility, and deterrence. As we hypothesize, messages convey intention to punish the opponents (threats) changes players' expectations, that first movers are largely deterred by the threats and second movers' threats are credible. Threats lead to deterrence and greater propensity for costly punishment.

Chapter 4 investigates the neural basis of individuals' charity donation behavior in a modified dictator game. The right temporoparietal junction (rTPJ) has been associated with social decision-making, but the exact neural mechanism of charitable giving remains unknown. In our experiment, participants allocate money between themselves and a charity in a graphical revealed preference task, that measures both parameterized other-regarding preferences and economic rationality (Monotonicity, WARP, and GARP). We find evidence for a causal role of the rTPJ in determining fairness preferences and economic rationality.

Chapter 2 received support from the Center for Peace Study and Violence Prevention at Virginia Tech.

The Behavioral and Neural Bases of Social Economic Decision-Making

Zhuncheng (Flora) Li

General Audience Abstract

Social economic decision-making considers the well-being and emotions of others. Individuals engage in social economic decision-making on a daily basis, for example, negotiating over an offer, investing or cooperating on a project, bargaining over a purchase, or interacting with friends or strangers. Each of these decisions involves a variety of motivations including money for oneself, the well-being of others, each participants' emotions and future relationships. Because of the complex nature of social economic decisions we need to employ an interdisciplinary research strategy. Behavioral economics applies psychological insights to economic problems and allows us to model the behavior of people who care about more than just money. Neuroeconomics integrates neuroscientific techniques and information about how the brain works to further expand our set of research tools. In this dissertation, we use all of these methods to explore how people make economic decisions in three distinct social scenarios.

All three scenarios are especially intriguing since they represent different ways in which individuals integrate "others" into their own decision-making process. First, hold-up happens when two parties can form an incomplete agreement to cooperate and achieve higher efficiency together, however, the agreement may fall apart due to concerns about the other party gaining more bargaining power. In a historic example, Fisher Body had an exclusive supply agreement with General Motors. When the demand for cars increased sharply, Fisher Body held up General Motors by increasing prices. Second, negotiation is a situation where two parties can both benefit from trade, but they have conflicting interests. Third, individuals who engage in charity donations often sacrifice themselves monetarily to improve well-being of others.

The scientific mission of this dissertation is to advance understanding of how individuals engage in social decision-making. In particular, we examine how communication (promises and threats) influences decision-making involving hold-up and negotiation respectively, and explore the neural mechanism governing altruism and charitable giving. We find evidence that communication enhances cooperation and efficiency in social economic decision-making through by changing expectations about monetary payoffs. In addition, we find evidence that the neural circuits responsible for fair-minded behavior also play a role in regulating economic rationality. This dissertation improves our understanding about how humans engage in social exchanges on both behavioral and neural levels.

Acknowledgments

Though I have many names to whom I want to give thanks, I would like to first recognize my dear husband, Xingjian Liu, who has been supportive and caring throughout my entire doctoral career. Since we first met in my first year at Virginia Tech, he has always been there for me no matter tears or joy.

I would like to thank my parents, Shaoli Zhang and Quanfu Li. Their love surround me since my first cry and their encouragements bring me strength through my study. It has been nine years that we can only reunite once a year. I cannot imagine how hard it is for them. All I can do is to thank them for shaping the person I am today, and I hope that they are proud of my achievement.

I would like to express my special thanks of gratitude to my advisors, Dr. Sheryl Ball and Dr. Alec Smith, who are great advisors and excellent researchers. I feel extremely lucky to have two extraordinary advisors who taught me step by step how to conduct scientific research through their own practice. Sheryl is affirmative and considerate, and Alec is rigorous and inspiring. Together, they make the perfect example leading my academic career. My special thanks also extends to Dr. Martin Dufwenberg and Dr. Benjamin Katz being remarkable coauthors. I have learned a lot from them on how to conduct rigorous research and how to write excellent academic work.

Many people have helped on my way to finish this dissertation. My dissertation committee members, Dr. Sudipta Sarangi and Dr. Eric Bahel have provided many precious insights and comments on improving the dissertation. Weizhe Weng, Dr. Xinde Ji, Xiang Cao, Xiaomeng Zhang, and Dongwoo Lee are great friends who support me through any hardship. I also want to thank the undergraduate and graduate students at the Virginia Tech Economics Laboratory who have helped me to collect all the experimental data. Many other colleagues have also helped me in a way or another. I apologize that I cannot list all of their names.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This dissertation seeks to extend researchers' understanding of humans' social economic decision-making. Social economic decision-making considers the well-being and emotions of others. Individuals engage in social economic decision-making on a daily basis, for example, negotiating over an offer, investing or cooperating on a project, bargaining over a purchase, or interacting with friends or strangers. Each of these decisions involves a variety of motivations including money for oneself, the well-being of others, each participants' emotions and future relationships. Because of the complex nature of social economic decisions, this dissertation consists of three laboratory experiments which utilize methods from economics, neuroscience, and psychology to expand our knowledge of individuals' social decision-making.

One critical aspect of social economic decision-making is the concept of "others." Economists assumed that individuals care about only themselves but not others. With the assumption that people are self-interested, economists did not pay much attention to social economic decision-making, until Rabin (1993) questioned self-interested assumption with a notion of reciprocity, and Fehr and Schmidt (1999) proposed that individuals care also about fairness. Since then, evidence for other-regarding preferences has been well documented (Camerer, 2003; Fehr and Schmidt, 2006; Rotemberg, 2006).

Despite the large literature studying other-regarding preferences, we still do not fully understand how individuals integrate "others" into their utility maximization processes. For example, Rabin (1993) proposes that one's interpretation of others' intentions is a major part of her utility function, whereas Fehr and Schmidt (1999) model individuals who care about the equitable distribution of income. With many studies try to distinguish between distributional preferences and belief-dependent motivations (Nelson Jr, 2004; Sutter, 2007; Xiao and Bicchieri, 2010), the evidence is still not clear. This dissertation explores the role

of belief-dependent motivations in strategic environments involving costly punishment and communication (Chapter 2 and 3). Then we extend to study the mostly unknown neural basis for other-regarding preferences with a neural manipulation technique (Chapter 4).

The three parts of this dissertation study three distinct types of social economic decisions. Chapter 2 explores how explicit promises influence individuals' decision-making in a hold-up situation. Hold-up happens when two parties form an incomplete agreement to cooperate and achieve higher efficiency together, however, the incomplete agreement may fall apart due to concerns about the other party gaining bargaining power. Economic agreements such as trades, bargains, investments, or partnerships can lead to mutual gain. However, in hold-up settings strategic considerations can lead to inefficiency, as concerns about opportunistic behavior by one party can outweigh the possibility of mutual benefit.

We propose that communication in terms of explicit promises can help to mitigate the problem. We derive theoretical predictions by applying the model of frustration and anger of Battigalli et al. (2018) to a three stage hold-up game with costly punishment in the third stage. The basic ideas of the model are 1) decision-makers experience anger when they are frustrated 2) frustration results when material payoffs are less than expected, and 3) anger leads to aggression and the urge to retaliate. This approach requires a formulation of utility where a player's preferences depend both on material payoffs and on expectations about his own and others' behavior. Messages become relevant to the extent that they influence expectations about payoffs, thus linking communication, beliefs, and the willingness to forgo material payoffs to punish others.

We find that communication changes beliefs and raises expectations about payoffs, and promises further raise payoff expectations. Broken promises are rare and are associated with high rates of punishment. Overall, our results are consistent with the idea that the anticipation of costly punishment from angry players leads to increased levels of efficiency and cooperation, and that these effects are stronger when communication is possible.

Chapter 3 focuses on an ultimatum like situation, such as a chain-store game (Selten, 1978) or an ultimatum minigame (Gale et al., 1995), which imitates a simplified negotiation situation. In such games, first mover proposes a split of an endowment, and then the second mover has the opportunity to reject or accept the disadvantageous offer. We argue that communication in term of threats can help achieve cooperation due to its deterrence effect. We build upon the model of frustration and anger of Battigalli et al. (2018), 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.

We design an experiment using a two-person, two-stage deterrence game, which shares the strategic structure of an ultimatum minigame, to examine the relationship between communicated threats and deterrence. As a treatment, we allow free-form messages from the second mover to the first mover. We elicit both players' first order beliefs and their own plans conditional on reaching every stage of the game. In stage one of our experiment, the first mover proposes either a fair split (which is automatically accepted) or a greedy one. If the first mover takes the larger share, then in stage two, the second mover has the option to punish the opponent, so that the initial endowment vanishes.

We find that not only players take the opportunity to send threats to gain bargaining power in the strategic environment, but also threats successfully deter first movers and that second movers tend to follow through on their threats. We also find that beliefs play an essential role in linking communication in the form of threats and behavioral outcomes. All of these findings are consistent with the idea that threats, beliefs, and behavioral outcomes are linked through the mechanism of belief-dependent frustration and anger.

Chapter 4 explores, in a simple charity donation scenario (a modified dictator game), how participants allocate monetary payoffs for self and a local charity. In such situation, individuals sacrifice themselves monetarily to improve the well-being of others. In neuroscience, the structure and function of the right temporoparietal junction (rTPJ) have been associated with both social behavior and also with sensory integration, information processing, and attention allocation. We examine the effect of neuromodulation of rTPJ on other-regarding preferences and rational choice using focal high definition transcranial Direct Current Stimulation (HD-tDCS).

We applied anodal, cathodal, or sham HD-tDCS over the rTPJ to healthy participants during a charitable giving task where participants allocated an endowment of money between themselves and a local food bank. Participants chose allocations either on or under a graphical representation of a budget line. The endowment and the relative price of contributing to the charity were randomly varied across 50 independent trials. Participants and the charity were paid according to one randomly selected decision. We hypothesized that anodal rTPJ stimulation would cause participants to behave more altruistically relative to sham stimulation, and that cathodal rTPJ stimulation would have the opposite effect. We sought to measure the effect of stimulation on both social preferences and also on the consistency and rationality of individuals' choices.

We fit each individual's choice data with a parametric utility function that measures the extent of other regarding behavior. We identify choice consistency and economic rationality

with violations of Monotonicity, the Weak Axiom of Revealed Preference (WARP), and the Generalized Axiom of Revealed Preference (GARP). The parametric utility function estimates indicate that cathodal individuals are on average less fair-minded, while anodal individuals are more fair-minded, relative to sham. We also find that Monotonicity, WARP, and GARP violations are more frequent and severe with cathodal stimulation and less frequent and severe with anodal stimulation, relative to sham.

The three experiments together provide insights about how people make social economic decisions. The evidence suggests that individuals take advantages of using communication to improve cooperation and achieve efficiency. Communication that changes expectations can then influence social economic decision-making. Communication in different forms (promises and threats) has similar but distinct social functions, such that promises lead to cooperation, and broken promises lead to costly punishment, whereas threats are credible and have a strong deterrence effect. We also find evidence that rTPJ causally determines not only other-regarding preferences but also economic rationality.

Chapter 2

Promises and Punishment

Martin Dufwenberg, Flora Li, and Alec Smith

Abstract

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

2.1 Introduction

Communication helps resolve social dilemmas. Even in one-shot complete information games that under conventional assumptions have unique equilibria, laboratory studies typically show that communication results in increased welfare and efficiency.[1] One explanation is that social interactions generate belief-dependent emotions, as modeled in psychological game theory (Geanakoplos et al., 1989; Battigalli and Dufwenberg, 2009). Messages can then credibly move behavior towards greater cooperation by raising payoff expectations and creating commitment power.

The hold-up problem is a social dilemma that arises when relationship-specific investments and incomplete contracts limit the attractiveness of forming welfare-enhancing partnerships. These results typically rely upon the assumption that agents maximize their material self-interest. However, laboratory studies of the hold-up problem often show greater-than-predicted levels of investment and efficiency, pointing to a role for other-regarding preferences.[2]

We identify a new way that communication may foster trust, cooperation, and efficiency in the class of hold-up problems that feature a punishment option. The key idea is that broken promises lead to dashed hopes and frustration, which psychologists associate with anger and aggression (Dollard et al., 1939; Berkowitz, 1989). In informal contracting of a kind germane in many hold-up scenarios, the anticipation of such emotional responses helps to facilitate cooperation and efficiency. We explore, via theory and experiments, the effect of non-binding, pre-play communication on trust, cooperation, deceit, and costly punishment in a class of relevant hold-up games. We derive theoretical predictions by applying the model of frustration and anger of Battigalli et al. (2018) (BDS), and we develop a design appropriate for testing test these predictions in the lab. The basic ideas of the model are 1) decision-makers experience anger when they are frustrated 2) frustration results when material payoffs are less than expected, and 3) anger leads to aggression and the urge to retaliate. This approach requires a formulation of utility where a player's preferences depend both on material payoffs and on expectations about his own and others' behavior. Messages become relevant to the extent that they influence expectations about payoffs, thus linking communication, beliefs, and the willingness to forgo material payoffs to punish others.

In our experimental design, we allow the second-mover to send a pre-play message to the first mover as a treatment. We elicit a variety of beliefs of both the first and second mover

[1] In other settings it is well-known that communication enhances cooperation and welfare. It facilitates collusion in repeated games (McCutcheon, 1997) and allows useful signaling of information (Crawford and Sobel, 1982) or intentions (Farrell, 1987).

[2] See, for example, Hoppe and Schmitz (2011); Dufwenberg et al. (2013) and Haruvy et al. (2018).

that are central to the behavioral theory that we test. This concerns beliefs that participants hold about their own (in the case of first movers) and their co-players' behavior, and these measures allow us to examine how communication influences beliefs, sentiments, and behavior.

A large experimental literature studies cheap talk and deception. Broadly, the evidence suggests that people are somehow averse to lying but make tradeoffs between the costs and benefits of lies (Gneezy, 2005; Dufwenberg et al., 2017). In trust games, promises lead to greater cooperation, and a number of studies support the idea that this relationship results from guilt aversion (Charness and Dufwenberg, 2006, 2011; Battigalli et al., 2013). There is also evidence that deception and broken promises induce greater willingness to engage in costly punishment (Brandts and Charness, 2003; Croson et al., 2003; Sánchez-Pagés and Vorsatz, 2007, 2009; Eisenkopf et al., 2017). Brandts and Charness (2003) find that deception leads to higher punishment rates after unfavorable actions in a simultaneous 2×2 game. Croson et al. (2003) study cheap talk in ultimatum bargaining with two-sided private information, finding that deception affected behavior in repeated ultimatum games. However, none of these literatures explored how communication changes beliefs, how beliefs influence behavior, and how broken promises lead to costly punishment.

A very important precursor to our study is Ellingsen and Johannesson (2004), who also explore communication in a hold-up game. However, since they did not conduct their exercise with BDS' theory in mind, they do not measure all the beliefs that are central to our tests. They suggest that their data is consistent with Fehr and Schmidt's (1999) model of inequality aversion combined with a preference for consistency, and that communication serves to change beliefs about co-player types. This interpretation is quite different from the one we focus on and test for. Later on, we shall return to Ellingsen and Johannesson's findings, and contrast them to ours.

Section 2.2 presents theory. We describe the games that we implement in our experiment. We discuss the application of BDS' model of belief-dependent anger, and the extension needed to incorporate the ideas we have regarding the effect of promises on trust, credibility, and costly punishment. Section 3.3 presents details of the experimental design and implementation, and states hypotheses to be tested. Section 2.4 reports results. Section 2.5 concludes.

2.2 Theory

2.2.1 A hold-up game with punishment

We study a 2-player, 3-stage hold-up game with punishment, as shown in Figure 2.1, where the numbers and variables at end nodes represent monetary payoffs.[3] The variables a and b have the following values: $a \geq 5 \geq b$, and $a \neq b$. In the first stage, Player 1 can go *In* to make an investment of her entire endowment of \$5, or go *Out* to not invest and walk away with her initial endowment. If Player 1 invests, the endowments of both players double, and Player 2 can then propose how to divide the proceeds. To make the problem simple, Player 2 can propose two possible splits. One is to choose *Share*, which is monetarily favorable (or at least as good as the other option) for Player 1. The other is to choose *Take*, which is monetarily favorable for Player 2. If Player 2 *Take*s, Player 1 can then *Reject*, in which case both players receive 0, or *Accept* to settle with a less favorable offer in the third stage. When players care only for monetary payoffs and $b < 5$, the unique subgame perfect Nash equilibrium (SPNE) is $((Out, Accept); Take)$; when $b = 5$ and players care only for monetary payoffs, there are two SPNEs: $((Out, Accept); Take)$ and $((In, Accept); Take)$. This game can also be interpreted as a trust game with punishment, or as an ultimatum bargaining game that the first mover decides to enter. Hold-up situations have been studied with experiments (e.g. Ellingsen and Johannesson, 2004; Dufwenberg et al., 2013; Haruvy et al., 2018), though these authors do not measure beliefs.

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

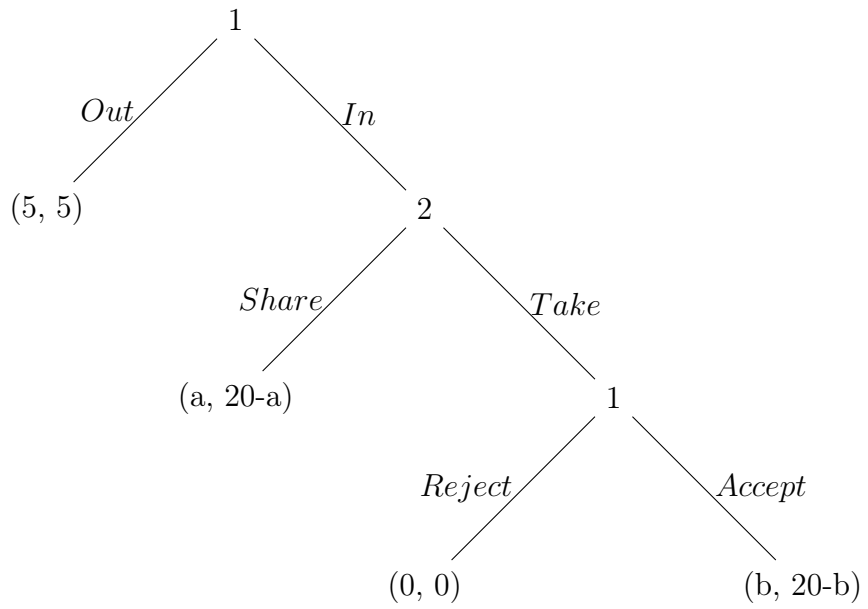


Figure 2.1. A hold-up game with punishment.

2.2.2 Frustration and anger

We focus on the simple anger (SA) proposed by Battigalli et al. (2018).[4] In this model, anger is motivated by frustration, and the tendency to hurt others is proportional to frustration, following the frustration-aggression hypothesis from psychology (Dollard et al., 1939; Berkowitz, 1989). In general, one feels frustrated if one's initial expectation is not met. Frustration is modeled as the gap (if positive) between one's initial expected payoff and the current best possible outcome. At any history h, player 1's frustration is

$$F_1(h;\alpha_1) = \max\left\{\bar{\pi}_1(h_0) - \max_{a_1 \in A_1(h)} \mathbb{E}[\pi_1|h;\alpha_1], 0\right\}, \tag{2.1}$$

where $\bar{\pi}_1(h_0) = \mathbb{E}[\pi_1|h_0;\alpha_1]$ denotes Player 1's expected payoff at the initial history h_0 given his first-order belief α_1 about Player 2's behavior, $a_1 \in A_1(h)$ denotes Player 1's action choice at the history h, so $\max_{a_1 \in A_1(h)} \mathbb{E}[\pi_1|h;\alpha_1]$ gives the maximum possible expected payoff available to Player 1 at the history h.

[4]Only limited version of the original theory is described here. Please refer to Battigalli et al. (2018) for the complete model with variations. Battigalli et al. (2018) propose 3 different versions of the belief-dependent frustration-anger model: 1) Simple anger (SA), 2) Anger from blaming behavior (ABB), and 3) Anger from blaming intentions (ABI). In this paper, we focus on SA; however, one should notice that SA and ABB's predictions coincide in two-player games. Therefore, our design, hypotheses, and results apply to ABB as well.

To capture a simple version of frustrated anger, we assume that Player 1's utility from action a_1 at history h is

$$u_1^{SA}(h, a_1; \alpha_1) = \mathbb{E}[\pi_1|(h, a_1); \alpha_1] - \theta_1 F_1(h; \alpha_1)\mathbb{E}[\pi_2|(h, a_1); \alpha_1], \tag{2.2}$$

where $\theta_i > 0$ is i's sensitivity to anger. A frustrated individual tends to blame and hurt the other player if the cost is low enough. If i is frustrated, i's 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.

In the game defined in Figure 2.1, let the probability that Player 1 assigns to choosing out be $p_1 = \alpha_1(Out|h^0) \in [0, 1]$. Let $q_1 \in [0, 1]$ denote the probability that Player 1 assigns to Player 2 choosing *Share* if stage 2 is realized, i.e. $q_1 = \alpha_1(Share|In)$. and let $r_1 = \alpha_1(Reject|In, Take) \in [0, 1]$ denote the probability that Player 1 assigns to choosing *Reject* conditional on the 3rd stage being reached. We can also define analogously a similar belief system (p_2, q_2, r_2) for Player 2. We further assume that higher order beliefs are correct in the sense that the marginals of the higher order beliefs are equal to the lower order beliefs. In equilibrium, the belief systems of both players coincide, so we may drop the subscripts and generically refer to beliefs p, q, and r.

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

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

For intermediate values of θ_1 there are two psychological sequential equilibria in pure strategies, the efficient one and the inefficient one. Figure 2.2 shows the equilibria as a function of the anger sensitivity of Player 1. In the experiment, participants play multiple rounds of the

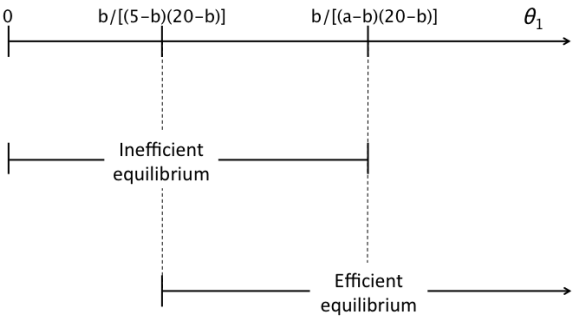


Figure 2.2. Sequential equilibria as a function of the anger sensitivity θ_1 of Player 1.

game with randomly matched opponents with feedback, which allows participants to learn towards the equilibria.

2.2.3 Communication

In this particular hold-up game with punishment, with standard preferences, communication affects neither expectations nor behavior. Whereas, belief-dependent social emotion models, such as the frustration-anger model, predict that particular messages can influence both expectation and behavior. When accompanied by pre-play freeform communication from Player 2 to Player 1, this 3-stage hold-up game with punishment allows for Player 1 to show trust, for Player 2's promises to be made kept or broken, and for the possibility of costly punishment in response to opportunistic behavior by Player 2. Promises are the messages that convey intention to cooperate, such as "If you go *In*, I will *Share*." Belief-dependent frustration-anger model predicts that, with promises, Player 1 is more likely to invest, Player 2 is more likely to keep her promises with anticipation of anger, and Player 1 is more likely to punish broken promises due to a high level of frustration.

2.3 Experiment

To study the effect of communication on trust and punishment we implemented the game in Figure 2.1 in a laboratory experiment. We used a within-subjects design where subjects played

variations of the game over multiple rounds with anonymous random matching partners, and with communication as a treatment variable.

2.3.1 Procedures

The experiment was programed using z-Tree (Fischbacher, 2007) and conducted at the Virginia Tech Economics Laboratory. A sample of the experiment instructions is reproduced in the Appendix. We conducted a total of 11 sessions, with 200 total participants.[5] Each session included between 14-20 participants with an average of 18.4 per session. Sessions took about 1.75 hours to complete.

At the beginning of each experiment session, participants were randomly assigned to the role of either Player 1 or Player 2, which remained fixed throughout the experiment. Before each round, participants were randomly and anonymously matched with a partner of the opposite role (i.e., we used stranger matching). After Session 4, we increased the show-up fee from \$5 to \$10 to improve turnout. After the experiment, participants were paid according to the outcome of one randomly selected round of play. Excluding the showup fee, participants earned an average of \$12.24.

Each session involved 20 rounds separated into two blocks: 10 rounds of communication, and 10 rounds where no communication was allowed. At the end of each round, both players received feedback on the outcome of that round. We counterbalanced the order of the communication block across sessions, so that in 5 of the 11 sessions the first 10 rounds involved messages from player 2 to player 1, and the no-message block followed; the other 6 sessions experienced the no-message block first.

In the communication blocks, the only restrictions on message sending were that the message had to be less than 140 characters long, and to retain confidentiality, individuals were not allowed to reveal their identity in the message. At the end of the experiment participants were paid according to the outcome of one randomly selected round out of the 20 played.

To vary the decision problem from round to round, in each block, participants played 10 different variations on the game in Figure 2.1 in a random order. The game variations are shown in Table 2.1, where all the numbers are in dollars. A change of the parameter b indicates changing the price for punishment, and we vary the price for punishment from 1 to 5. The difference $a - b$ indicates the "take amount": either $a - b = 4$ to indicate a low take amount, or $a - b = 10$ to indicate a high take amount. The payoff splits in Stage 2 and Stage

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

3 are asymmetric, such that $a \neq 10$, to reduce the saliency of an equal split.

Table 2.1. Experiment design – game variations.

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

2.3.2 Belief Elicitation

In a survey of the literature on experiments that compare the strategy method (where players make conditional decisions for each possible history) and the direct response method (where play unfolds sequentially) Brandts and Charness (2011) report that the strategy method leads to substantially lower levels of punishment. Because we are interested specifically in costly punishment, we therefore employ the direct response method in our experiment. The direct response method, however, is problematic for measuring beliefs: incentivizing truthful beliefs by e.g. a scoring method creates a spillover effect where players are rewarded to move to the next stage of the game. In response to this issue, we do not offer monetary payoffs for incentivized beliefs. Rather, participants are rewarded with $5 to answer mandatory belief elicitation questions, and they have the opportunity to pledge that their reported beliefs will be truthful at the beginning of the experiment.

In the experiment we measured the first-order beliefs that participants held about their own (in the case of first movers) and their co-players' behavior. We elicited Player 1's beliefs regarding the likelihood of choosing *Out* (p), Player 1's conditional first order beliefs of Player 2's probability of choosing *Share* (q), and Player 1's own plan of choosing *Reject* (r) conditional on entering the 3rd stage. To examine how messages influence beliefs, in the communication treatment we measure Player 1's beliefs both before and after messages are received. If Player 1 chose *In*, we elicited Player 2's second order belief about q and first order belief about the conditional probability that Player 1 will choose *Reject* after Player 2 made a decision on the 2nd stage.

When employing the direct response method in a sequential game with belief elicitation, there is a potential conflict between the incentives for behavior and for reporting truthful beliefs (e.g. Rutström and Wilcox, 2009; Blanco et al., 2010).[6] We are interested in studying behavior and beliefs in response to events that might trigger so-called "hot" emotions such as anger, so we avoid the use of the strategy method in favor of observing direct responses during sequential play (Brandts and Charness, 2011). This approach has the advantage that decisions to *Share* (for Player 2) or *Reject* (for Player 1) are implemented for certain. However, when employed in conjunction with belief elicitation that uses e.g. a proper scoring rule, the direct response method potentially generates incentives to, for example, continue the game in order to receive payment for a reported belief in a future stage, or to choose behavior consistent with a previously reported belief. The problem is exacerbated when eliciting beliefs about a player's own future behavior.[7] Trautmann and Kuilen (2015) find that flat fee incentives perform almost as well as more complicated methods for eliciting beliefs such as proper scoring rules. We therefore do not employ a scoring rule, but offer a flat fee incentive for players to report their true probabilistic beliefs.

2.3.3 Hypotheses

Based on the experimental design, we derived the following testable hypotheses:

Hypothesis 2.1. *Communication increases the frequency of cooperative outcomes and improves efficiency.*

We expect communication to increase the frequency of cooperative outcomes and improves efficiency, consistent with a number of studies of communication and cooperation (Charness and Dufwenberg, 2006; Balliet, 2010; Battigalli et al., 2013), and studies of communication and efficiency (Blume and Ortmann, 2007; Avoyan and Ramos, 2017; Fehr and Sutter, 2016).

Hypothesis 2.2. *Communication influences Player 1's reported beliefs in the direction of increased likelihood of investment, cooperation, and costly punishment.*

As we hypothesize that communication serves a channel for changing expectations, we predict that Player 1's reported beliefs change after receiving a message.

Hypothesis 2.3. *Player 1's higher 1st order belief about Player 2's probability of cooperation leads to higher rates of Reject choices.*

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

[7] See also the discussion of incentivizing own beliefs in Toussaert (2018), who addresses this issue by eliciting beliefs about a "similar other." Because we are interested in the prediction that one's *own* beliefs may be the relevant variable for anger and costly punishment, we do not employ methods that involve proxies such as similar others or the average belief in the room (e.g. Charness and Dufwenberg, 2006).

This hypothesis is motivated by the frustration-anger model, which assumes that diminished payoff expectations make aggression and costly punishment more attractive. Hypotheses 2.2 and Hypothesis 2.3 connect communication and costly punishment via the effect of communication on beliefs.

Motivated by the results of Charness and Dufwenberg (2006) and the subsequent literature, we also hypothesized that the content of the free-form messages would play an important role in connecting communication with behavior via beliefs. In particular, we predicted that messages including promises (statements of intent to choose *Share*) would change beliefs and plans in the direction of increased investment, cooperation, and punishment.

Hypothesis 2.4. *Communication influences beliefs via promises; cheap talk (non-promise messages) has no impact on beliefs.*

We predicted that Player 1 would report a lower probability of subsequently choosing *Out*, would expect *Share* with higher probability, and would be report a greater probability of choosing *Reject* after *Take* when messages contained promises.

Hypothesis 2.5. *Broken promises lead to higher rejection rate, and promises lead to higher cooperation rate compare to non-promise messages.*

We predicted that the effect of promises on beliefs would carry through to behavior. In particular, an implication of the frustration-anger model is that if promises are believed and then broken, the higher initial expectation of cooperation generates greater frustration and leads to a higher likelihood of rejection in the 3rd stage.

2.4 Results

We begin our examination of the results by measuring the effect of communication on game outcomes and behavior. Next, motivated by the model of belief-dependent anger, we look at the relationships between communication, beliefs, and behavior. We then investigate message content, focusing on how promises affect beliefs and behavior.

2.4.1 The Effect of Communication on Outcomes

Communication has a strong effect on efficiency and cooperation. Figure 2.3 compares outcomes from the no-message and message blocks, pooling the data from all sessions. The

Share rate is higher in communication treatment (60.10% vs. 46.70%), see Table 2.2. A 1-sided Fisher's exact test confirms that the cooperation rate is higher in the communication treatment (p-value$<$0.001). This result is consistent with the belief dependent models of frustrated anger and guilt aversion and with Hypothesis 2.1, that communication increases cooperative outcomes. A chi-squared test shows that the communication treatment has a significant effect on the distribution of outcomes (terminal histories) (p$<$0.001). The conditional *Reject* rate is also higher in the communication treatment (40.20% vs. 35.93%), but this difference is not significant (p-value=0.197, 1-sided Fisher's exact test).

Table 2.2. The effect of communication.

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

There is a significant difference between sessions with communication first and sessions with communication second. Figure 2.4 shows that there is a persistent effect of communication on outcomes. In the first 10 periods, there is a significant difference in cooperation rate between communication-first and communication-second sessions. This difference disappears in period 11-20. This suggests that the communication effect is so strong that after being exposed to the communication environment, participants behave as if they are still sending and receiving messages, even in the no-communication treatment.

Because of the persistent effect of communication, we examine the distribution of outcomes after restricting the sample to include only the first 10 periods. Figure 2.5 shows the effects of communication on the distribution of outcomes in the first 10 rounds only, when the no-communication group has no experience with messages. The figure demonstrates a much stronger effect of communication. The mean fraction of *Share* outcomes in the communication treatment in the first 10 rounds is 58.8%, close to the overall mean, but in the first 10 rounds without communication, the *Share* rate is only 35.8%. The difference demonstrates that communication has a strong and persistent effect.

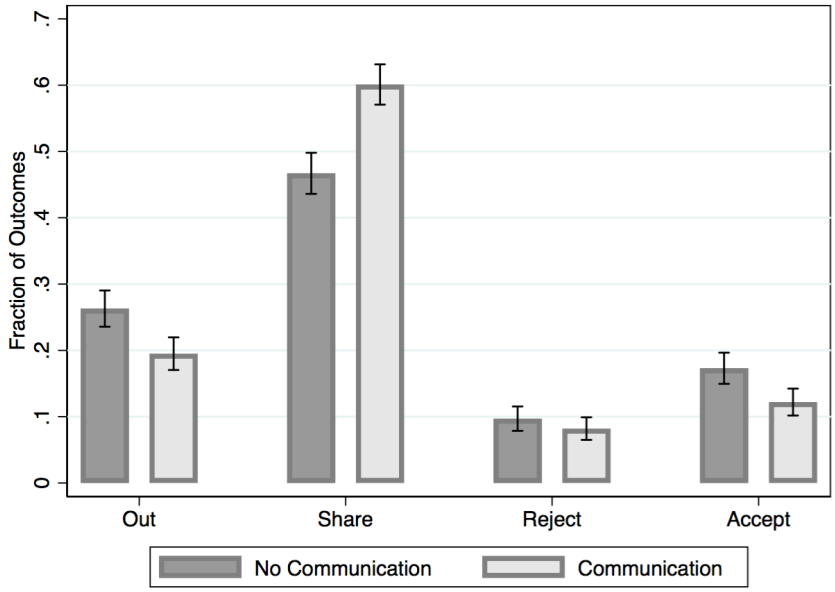


Figure 2.3. Outcomes and Communication Summary.

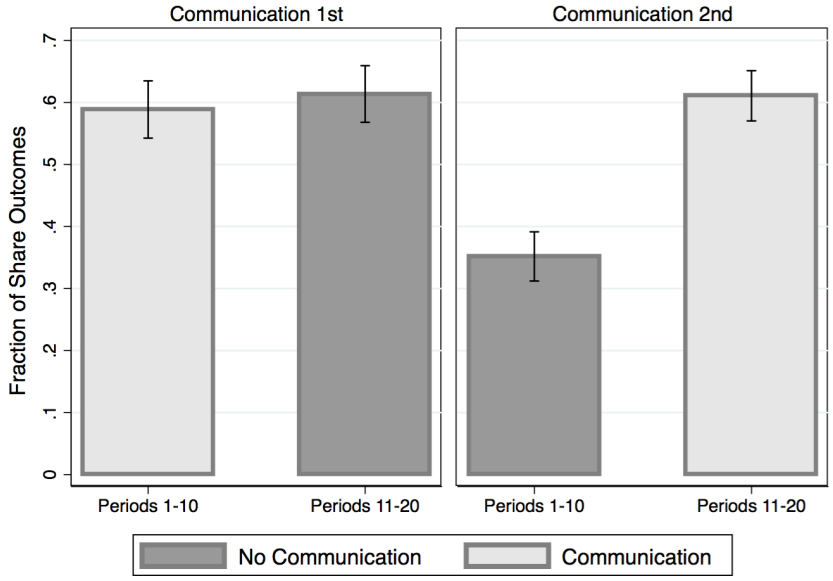


Figure 2.4. Persistent Communication Effect.

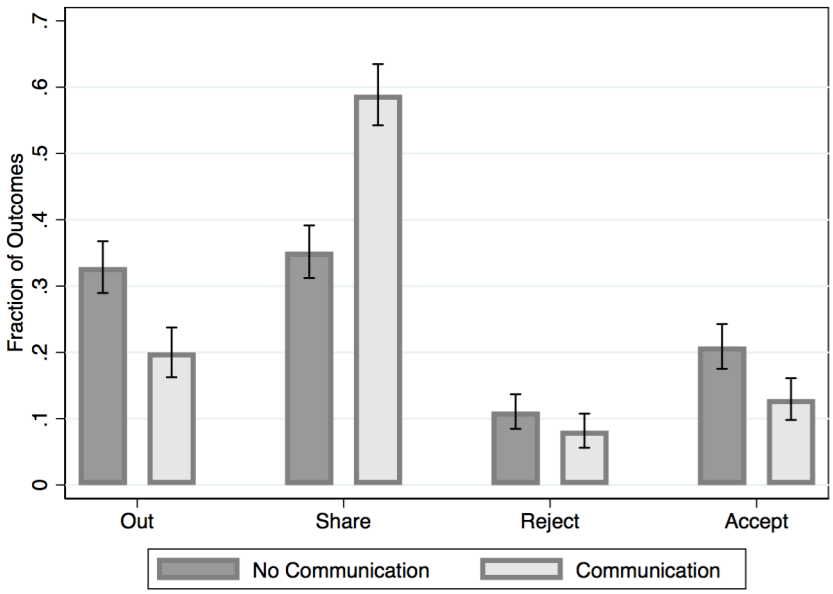


Figure 2.5. First 10 Period Outcome Summary.

We also compare average payoffs in each treatment. Figure 2.6 shows that Player 2's average payoffs are insignificantly higher in the communication treatment; however, Player 1's average payoffs significantly increase \$0.71 from no-communication to communication treatment. This suggests that social welfare or efficiency increases if communication is allowed. This result is consistent with our Hypothesis 2.1, that communication improves efficiency.

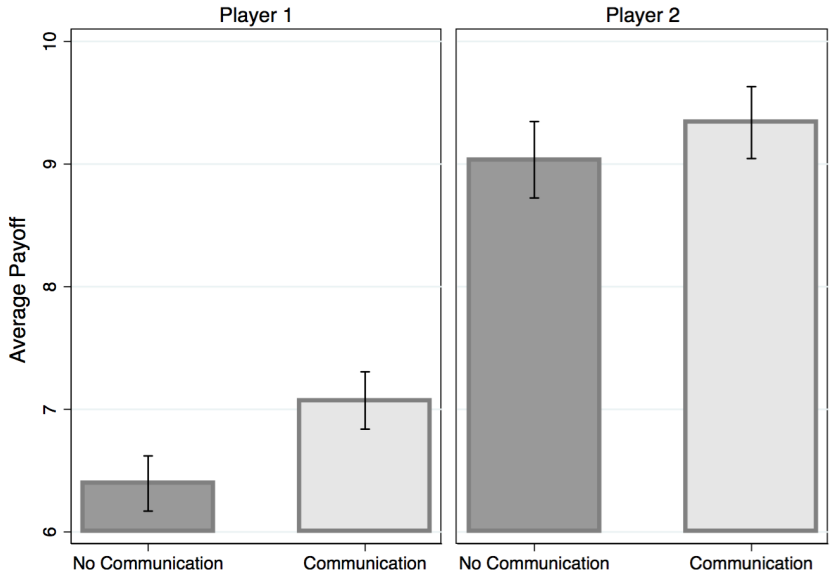


Figure 2.6. Average Payoff by Player Type and Communication Treatment.

Figure 2.7 shows that communication treatment affects Player 1's reported beliefs, and this result is consistent with our Hypothesis 2.2. Player 1 believes that Player 2 will cooperate with higher probability (1st order belief about *Share*) when communication is allowed. Communication treatment affects Player 1's own plans as well. With communication, Player 1 believes that she is less likely to play *Out* but more likely to *Reject* if 3rd stage is reached. Two sample Mann-Whitney-Wilcoxon rank-sum tests confirms that Player 1's beliefs are significantly different in the communication treatment and the no-communication treatment ($p < 0.001$ for own plan about *Out*, $p < 0.001$ for 1st order belief about *Share*, and $p = 0.003$ for own plan about *Reject*).

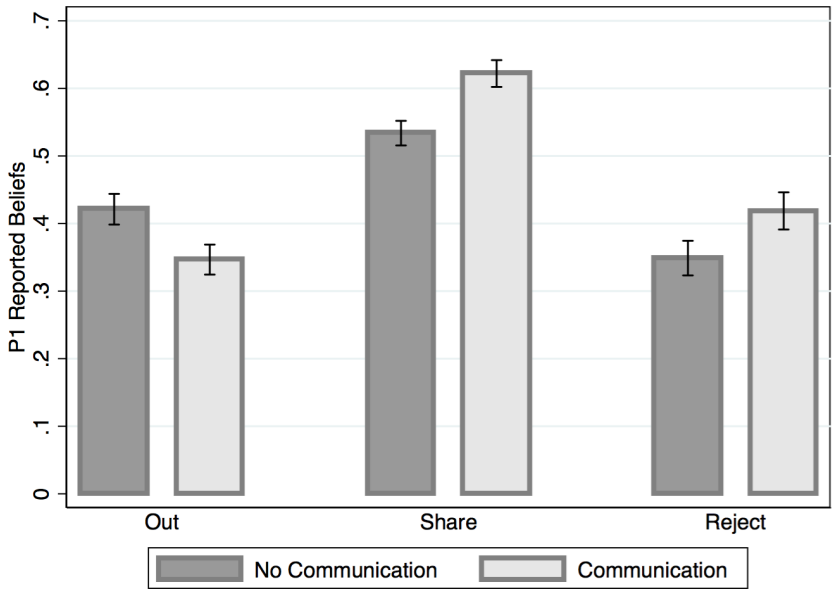


Figure 2.7. Communication Influences P1's Reported Beliefs.

2.4.2 Beliefs and Plans

To study the role of communication and beliefs in driving costly punishment, we run fixed effect logistic regressions with the dependent variable reject = 1 if Player 1 rejects the offer in stage 3, and reject = 0 if Player 1 accepts the offer in stage 3 in Table 2.3. The key to evaluating the belief-dependent model is to actually look at how beliefs about *Share* influence behavior (Hypothesis 2.3), since models of self interest and of distributional preferences imply that these beliefs should have no impact on behavior in the 3rd stage of the game, after controlling for the cost of punishment (in the form of the payoff from *Accept*).

Column A examines the factors driving *Reject* choices when we control for the game permutation with the variables "Payoff from *Accept*" (*i.e.*, the cost for punishment) and "High Take," which equals 1 if it is a High Take game ($a - b = 10$), and equals 0 otherwise ($a - b = 4$). The coefficients on "Communication" and "High Take" are not significant. However, the coefficients on "Payoff from *Accept*" and "Period" are significantly different from 0. When we further control for Player 1's 1st order beliefs about *Share* (column B), again, "Communication" and "High Take" are not significant. However we find a significant relationship between Player 1's belief about *Share* and decision to reject the offer after *Take*. A 10% increase in "Belief about *Share*" increases Player 1's chance of rejecting by 2.488%, which is consistent with Hypothesis 2.3. In column C, when using session controls, none of the variables included in regression B is significant except for "Payoff from *Accept*." Regression D uses "Promise"

Table 2.3. Logistic Regressions – Determinants of P1's *Reject* Choice.

	A mfx / se	B mfx / se	C mfx / se	D mfx / se	E mfx / se
Payoff from *Accept*	-0.1985***	-0.1814***	-0.0806**	-0.1680***	-0.1604*
	(0.0219)	(0.0245)	(0.0411)	(0.0283)	(0.0915)
High Take	0.0442	0.0769	0.0170	0.1407**	0.0728
	(0.0584)	(0.0606)	(0.0349)	(0.0717)	(0.0978)
Communication	0.0174	0.0111	0.0053		
	(0.0494)	(0.0483)	(0.0234)		
Period	0.0137***	0.0129***	0.0058	0.0116	0.0155
	(0.0044)	(0.0045)	(0.0040)	(0.0071)	(0.0265)
Belief about *Share*		0.2488**	0.0684	0.4949***	0.3771
		(0.1173)	(0.0934)	(0.1397)	(0.3470)
Promise				0.0692	0.0564
				(0.2282)	(0.5434)
Observations	474	474	474	204	204
AIC	539.6	533.9	462.9	230.2	186.1
BIC	560.4	558.9	483.7	250.1	202.7
Session controls	No	No	Yes	No	Yes
Subject controls	No	No	No	No	No

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

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

instead of "Communication". "High Take" becomes significant along with "Payoff from *Accept*" and "Belief about *Share*." In column E, when using session control, none of the variables included in regression D is significant except for "Payoff from *Accept*."

There is a natural selection bias inherent in observing *Reject* decisions after *Take*: Player 2 is more likely to select *Take* when the probability of *Reject* is low. Therefore, we use Player 1's *Reject* plan (Player 1's reported belief about *Reject* at the start of the game) as a proxy for Player 1's actual behavior in 3rd stage. Since not all games reach the 3rd stage, but all Player 1s reported their plans if games reach to 3rd stage, this proxy allows us to study Player 1s choices without selection bias.

In Table 2.4, we employ fixed effects linear regressions to study the determinants of Player 1's reported *Reject* plan (divided by 100, to scale between 0 and 1). When controlling only for the games, the effect of "Communication" is marginally statistically significant in affecting Player 1's plan to reject (column A). When further controlling for "Period" in column B,

Table 2.4. Linear Regressions – Determinants of P1's *Reject* Plan.

	A coef / se	B coef / se	C coef / se	D coef / se	E coef / se
Payoff from *Accept*	-0.0723***	-0.0736***	-0.0677***	-0.0741***	-0.0690***
	(0.0082)	(0.0082)	(0.0068)	(0.0078)	(0.0086)
High Take	0.0068	0.0087	0.0272***	0.0208	0.0377***
	(0.0109)	(0.0080)	(0.0080)	(0.0158)	(0.0105)
Communication	0.0698*	0.0552***	0.0449**		
	(0.0362)	(0.0171)	(0.0198)		
Period		0.0122***	0.0118***	0.0196***	0.0196***
		(0.0015)	(0.0017)	(0.0039)	(0.0038)
Belief about *Share*			0.1230***	0.1122**	0.2232***
			(0.0361)	(0.0530)	(0.0527)
Promise				0.0064	-0.0089
				(0.0219)	(0.0325)
Constant	0.5623***	0.4445***	0.3561***	0.3408***	0.2527***
	(0.0557)	(0.0505)	(0.0433)	(0.0834)	(0.0830)
Observations	2000	2000	2000	1000	1000
AIC	682.198	561.399	538.041	121.620	1062.800
BIC	704.601	589.403	571.646	151.067	1092.247
Session controls	No	No	No	No	Yes
Subject controls	Yes	Yes	Yes	Yes	No

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

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

"Communication" becomes highly significant. After including "Belief about *Share*" in column C, all variables are significant including "High Take." When replacing "Communication with "Promise" in regression D and E with either session or subject control, all other variables are significant but not "Promise." Communication significantly influences Player 1's reported plans to *Reject*, but promises have no effect on Player 1's reported plan. This result is consistent with the notion that beliefs are the channel by which communication influences behavior.

2.4.3 Promises

To examine the relationship between message content, beliefs, and behavior, we manually coded messages as promises if they follow the pattern of "If you choose *In*, I will *Share*."

Using this approach, we identify 32% of messages as promises, and the median number of promises per session was 32.2%.

Figure 2.8 shows that promises do have a strong effect on Player 1's reported beliefs. Promises increase Player 1's belief about Player 2's cooperative behavior (1st order belief about *Share*). Promises also influence Player 1's beliefs about their own actions (beliefs about *Out* and *Reject*). Player 1s report that they will be less likely to choose *Out*, but will be more likely to punish Player 2 after receiving a promise. If a message is coded as a promise, Wilcoxon signed-rank tests show a significant difference in Player 1's reported beliefs before and after a message (p-value < 0.0001) for all three measured beliefs. However, for messages coded as cheap talk, Wilcoxon signed-rank tests return insignificant results indicating that beliefs before and after a non-promise message are not significantly different (0.1952 for p, 0.2866 for q, and 0.7294 for r). The result indicates that promises have a significant effect upon beliefs, but that non-promise messages (cheap talk) have an insignificant effect, consistent with Hypothesis 2.4.

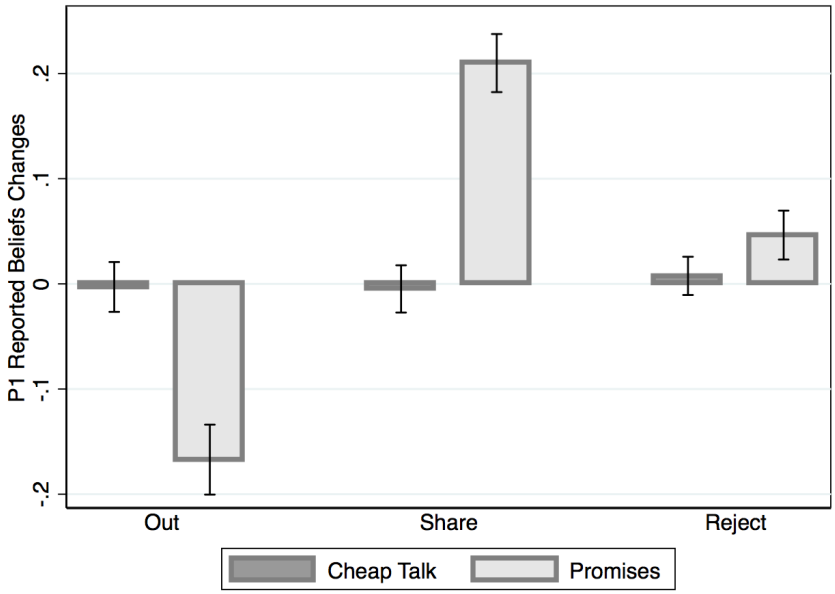
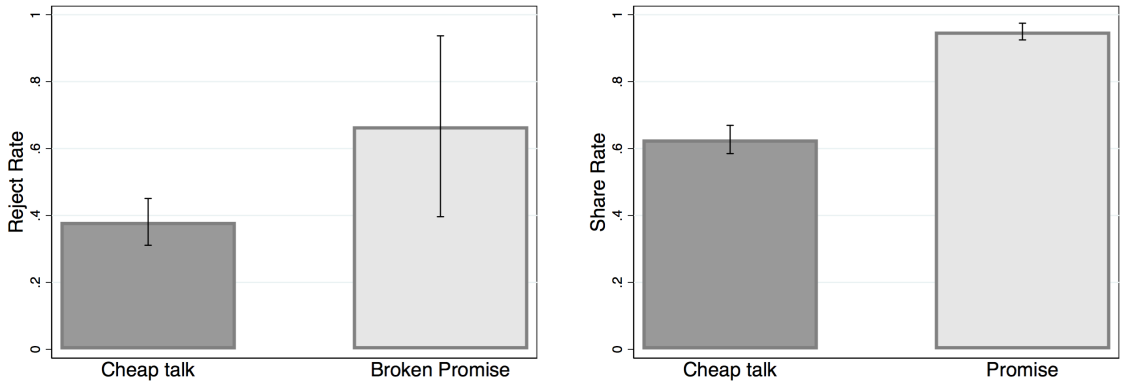


Figure 2.8. Belief Change After Receiving A Message.

To further test Hypothesis 2.5, we look at how promises influence behavior. We observe much higher conditional *Share* and *Reject* rates when a promise is made, consistent with Hypothesis 2.5 that promises foster cooperation, but broken promises leads to higher level of punishments. The effect of promises is greater than the effect of communication, and messages other than promises have no effect on behavior.



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

Figure 2.9. Broken Promises and Kept Promises.

The result shown in Figure 2.9(b) are consistent with the frustration-anger model in that if Player 2 anticipates Player 1's belief change following a promise, Player 2 will be motivated to choose *Share* to avoid punishment. When comparing Player 2's behavior after cheap talk vs. promises, the rate of *Share* choices is significantly higher following a promise (1-sided Fisher's exact test: p-value=0.000). The result holds for individual games. Supplementary Figure 2.11(b) shows that the *Share* rate for promises is higher across all 10 games compared to games with messages categorized as cheap talk.

Figure 2.9(a) also shows that the rate of *Reject* choices is higher when a promise is broken compared to the rate of *Reject* choices when messages are categorized as cheap talk, consistent with Hypothesis 2.5. A one-sided Fisher's exact test confirms that the conditional rate of *Reject* choices after a broken promise is significantly higher than in games with messages categorized as cheap talk (p-value=0.023).

2.5 Discussion

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

Experimental and behavioral economists convincingly argued that models of social preferences are needed to explain human behavior, but little such work factors in anger and frustration. One may wonder if doing so is necessary. For example, can models of inequity aversion (e.g. Fehr and Schmidt, 1999; Bolton and Ockenfels, 2000) explain our results? One implication of inequity aversion is that if player 1 ever rejects a high offer in the 3rd stage, then she/he would never accept a lower offer, regardless of communication or beliefs. Using this idea we can classify subjects into four categories, shown in Table 2.5. "IA Violation" represents subjects whose behavior is inconsistent with inequity aversion: they either reject a higher and accept a lower offer, or they both reject and accept the same offer (e.g. rejecting an offer of 3 in one period and accepting 3 in another). "Inequity Averse" subjects' behavior is always consistent with inequity aversion, "Self-interest" refers to players who always accept any offer, and "Unclassified" are subjects that faced fewer than two different offers.

Table 2.5. Classification of Player 1 behavior.

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

Table 2.5 indicates that 36% of subjects are inconsistent with either self-interest or inequity aversion, 28% of subjects demonstrate behavior consistent with inequity aversion, while 33% of subjects behave as if they care only for material self-interest. Moreover, the number of subjects whose behavior is inconsistent with inequity aversion or self-interest increases when subjects have more decisions in the 3rd stage. This suggest that inequity aversion cannot explain the behavior of at least one-third of our participants, and models that allow for non-consequential behavior such as that in Battigalli et al. (2018) may be needed to fully capture the range of behavior demonstrated.

Another strand of models addresses subjects' tendency to honor promises, e.g. work on guilt aversion (see (Charness and Dufwenberg, 2006); compare (Battigalli and Dufwenberg, 2007)) or a direct preference to honor a promise (e.g. Vanberg, 2008). Related models help explain why communication increases the frequency of Share choices, but our results indicate that frustration and anger in our game has additional effects. First, models of a tendency to honor promises first cannot explain the behavioral results we observe in the 3rd stage of the games, regarding increased rates of punishment when promises are breached. Second, the overall rates of Share choices that we observe are much higher than in comparable studies that do not allow for a punishment stage (e.g. Charness and Dufwenberg, 2006).

Finally we emphasize that we have explored but one side of the messages that may be

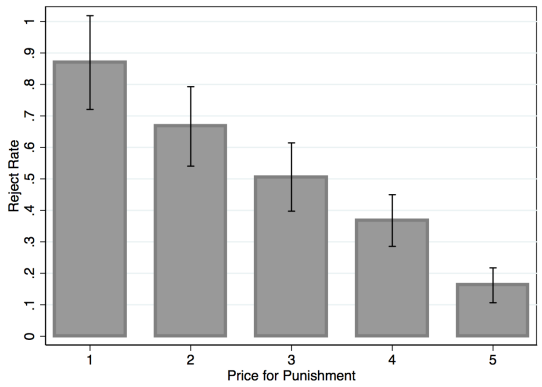
conjectured to matter in three-stage investment (or hold-up) games. Namely, we focused on communication from the second-mover to the first-mover, and relevance for our anger-and-frustration hypotheses concerned promises of trustworthy behavior. Our analysis left out communication from the first-mover to the second-mover, a case in which the relevance for our anger-and-frustration hypotheses would concern threats that punishments will be meted out if trustworthy behavior is not observed. We propose that analyzing threats is a very interesting topic, and in fact one which we address in a companion paper (Dufwenberg, Li, and Smith, 2018b).

2.6 Appendices

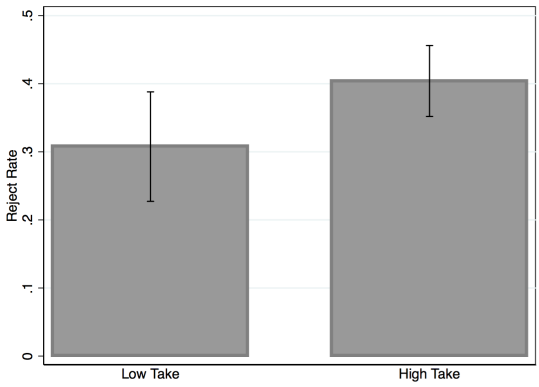
2.6.1 Supplementary Graphs and Tables

Supplementary Table 2.6. The effect of Promises on Outcomes

	Out	Cooperation	Rejection	Acceptance	Total
	23	281	10	5	319
Promises	7.21%	88.09%	3.13%	1.57%	100.00%
			66.67%	33.33%	100.00%
	176	344	74	127	721
Cheap Talk	24.41%	47.71%	10.26%	17.61%	100.00%
			36.82%	63.18%	100.00%
	199	625	84	132	1040
Total	19.13%	60.10%	8.08%	12.69%	100.00%
			38.89%	61.11%	100.00%

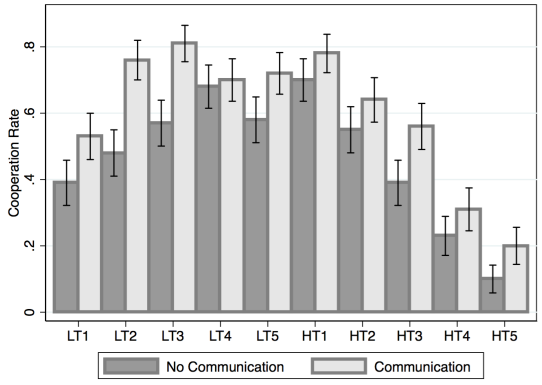


(a) Rejection rate decreases with price for punishment

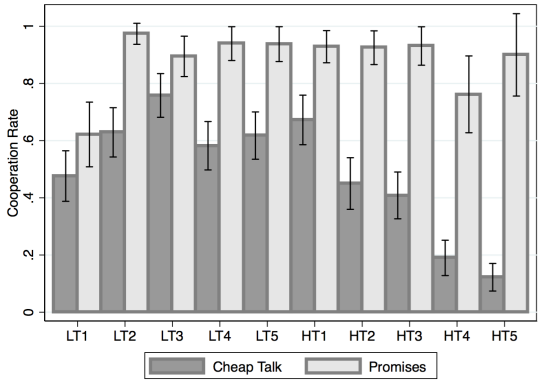


(b) Rejection rate increases with level of Take Rate

Supplementary Figure 2.10. Rejection rate by game structure

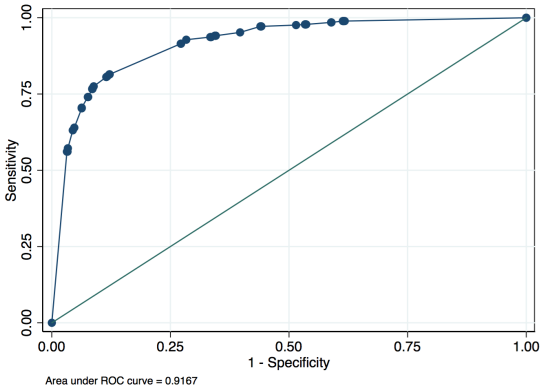


(a) High cooperation with communication

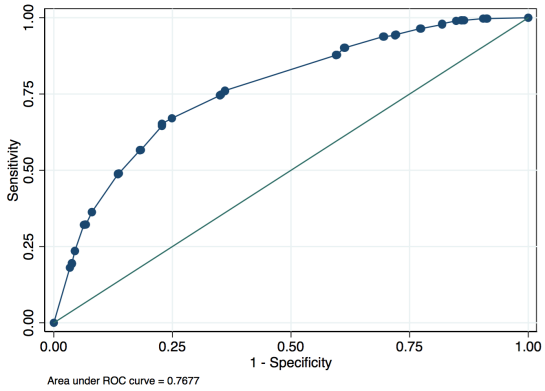


(b) High cooperation with promises

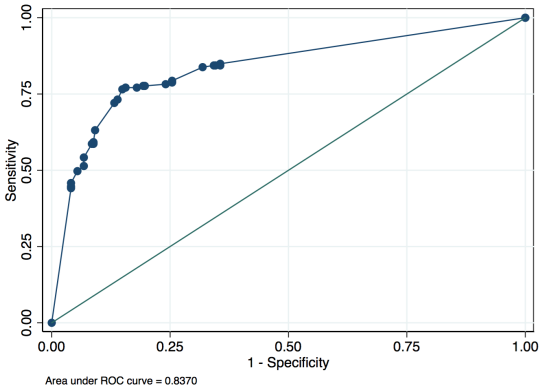
Supplementary Figure 2.11. Cooperation Rate by Communication and Promises



(a) P1's Plan about *Out*



(b) P2's Plan about *Share*



(c) P1's Plan about *Reject*

Supplementary Figure 2.12. Reported Plan Predicts Own Behaviors

2.6.2 Instructions

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

Part I Instructions

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

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

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

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

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

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

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

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

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

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

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

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

Part II Instructions

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

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

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

As before, the game consists of three stages. The picture below may help and will be shown in each round. Payoffs will change in each round, so please familiarize yourself with the picture. Player 1's payoffs are listed above Player2's payoff. The game proceeds as follows:

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

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

Chapter 3

Threats

Martin Dufwenberg, Flora Li, and Alec Smith

Abstract

We study communication, deterrence, and costly punishment in a laboratory experiment. Our design permits us to examine the relationship between communicated threats, credibility, and beliefs. We show that a theoretical model of belief-dependent anger captures the relationship between messages, beliefs, and behavior. The model predicts that threats can serve as credible commitment devices that lead to increased propensity for costly punishment. Our experimental results support the theory, demonstrating that threats change beliefs and payoff expectations and lead to greater levels of punishment. Communicated threats deter co-players from exploiting the strategic environment to their advantage.

3.1 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), and much of the work in early game theory 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. Threats 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.

We design an experiment using a two-person, two-stage deterrence game to examine the relationship between communicated threats and deterrence. This game has the same strategic structure as the chain-store stage game (Selten, 1978) and the ultimatum minigame (Gale

et al., 1995).[1] In stage one of our experiment, the first mover (P1) proposes either a fair split (which is automatically accepted) or a greedy one. If P1 takes 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 where P1 must choose the greedy offer twice, and in the communication treatment of this "staggered entry" game P2 sends messages after the first stage. 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.

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 reject offers 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., 2018a), we study the relationship between promises and costly punishment. The results in that paper are consistent with the 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. 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, 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

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

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 3.2. We present the experiment design details, experiment procedure, and derived hypotheses in Section 3.3. Section 4.3 presents results, and Section 4.4 concludes.

3.2 Deterrence, Anger, and Threats

3.2.1 Deterrence Game

We focus on the deterrence game depicted in Figure 3.1, 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 *Take* to keep a larger share to herself and let P2 make the next decision. If the game continues to stage 2, P2 can either *Accept* the proposed offer, or *Reject* the proposed offer 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 *Take*.

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

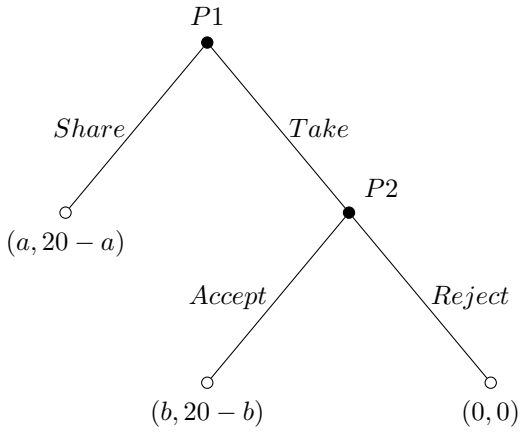


Figure 3.1. Deterrence Game

3.2.2 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 3.1, 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

$$F_2(h;\alpha_2) = \left[\bar{\pi}_2(h_0) - \max_{a_2 \in A_2(h)} \mathbb{E}[\pi_2|h;\alpha_2]\right]^+, \tag{3.1}$$

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

$$u_2^{SA}(h, a_2; \alpha_2) = \mathbb{E}[\pi_2|(h, a_2); \alpha_2] - \theta_2 F_2(h;\alpha_2)\mathbb{E}[\pi_1|(h, a_2); \alpha_2], \tag{3.2}$$

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 3.1, let the probability that P1 assigns to choosing *Take* be $p_1 = \alpha_1(Take|h^0) \in [0, 1]$. Let $q_1 \in [0, 1]$ denotes the probability that P1 assigns to P2 choosing *Reject* if stage 2 is realized, i.e. $q_1 = \alpha_1(Reject|Take)$. 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 $(Share; Reject)$ 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 *Reject* the offer if $\theta_2 > \frac{20-b}{(b-a)b}$. The unique SPE $(Take; Accept)$ consists another SE. When P2 expects $(Take; Accept)$, her initial monetary payoff is $20 - b$. If P1 chooses *Take*, P2 experiences 0 frustration. P2 chooses *Accept* with all possible θ_2.

3.2.3 Threats

To study threats, we allow communication as a treatment. In the experiment, P2 can send a free-form message to P1 in the 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 *Reject*. An explicit threat looks like "if you choose *Take*, I will *Reject*." If P2 fails to deter, and P1 chooses to *Take*, 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 *Reject* when deterrence fails.

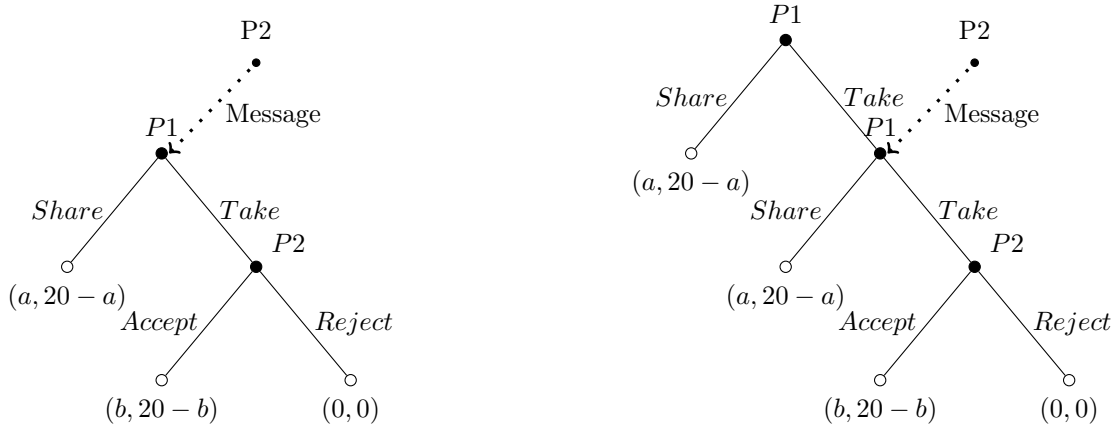
3.3 Experiment

3.3.1 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 3.2(b). The only difference between the two games is that in the staggered entry game P1 has to choose *Take* 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 *Take* in the first stage of the staggered entry game. In the staggered entry games, P1's *Take* 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

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

and 2.



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

Figure 3.2. 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 $m_1 = \alpha_1(Take|h^0)$ is the probability P1 assigns to choosing *Take* herself in stage 1, $p_1 = \alpha_1(Take|Take)$ is the probability P1 *Take*s again in stage 2, and $q_1 = \alpha_1(Reject|Take, Take)$ is P1's 1st order belief on P2's *Reject* 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 3.2. The payoff structures are described in Table 3.1, 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 3.1. 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]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.

3.3.2 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 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 *Take* (p_1), P1's 1st order belief of P2 choosing *Reject* (q_1) conditional on reaching 2nd stage, P2's 1st order belief about P1 choosing *Take* (p_2), and P2's conditional plan of *Reject* (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 *Take* (m_1) in stage 1, her own plan about choosing *Take* (p_1) in stage 2 conditional on reaching the stage, and 1st order belief about P2's conditional probability of choosing *Reject* (q_1). P2 reported 1st order beliefs on 1st and 2nd stage conditionally (m_2, p_2), and her own plan of choosing *Reject* (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 3.3.

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. (2018a).

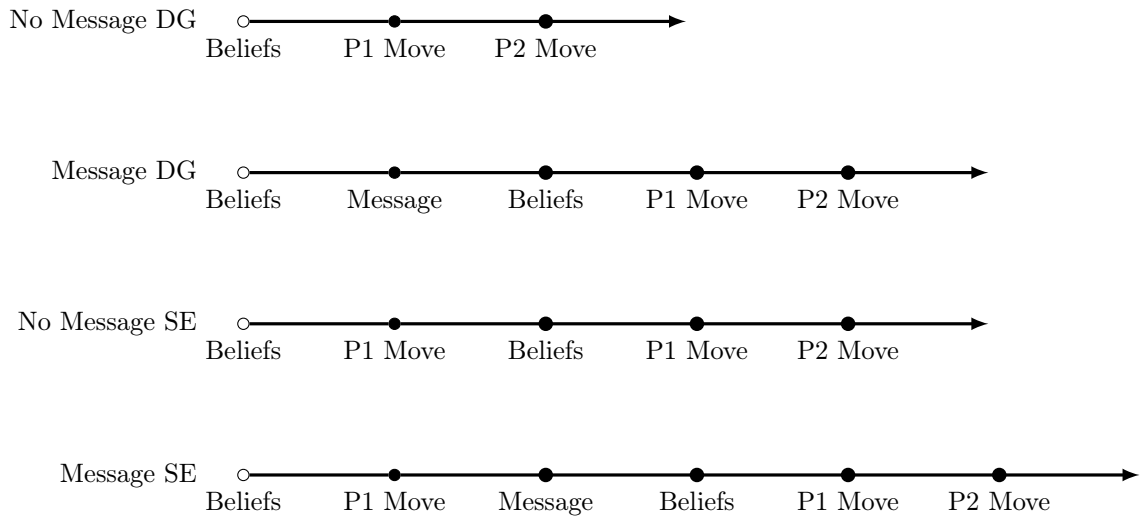


Figure 3.3. Experiment Timeline

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

3.3.3 Hypotheses

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

Hypothesis 3.1. *Threats lead to a higher rate of deterrence.*

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

Hypothesis 3.2. *Threats lead to a higher rate of costly punishment.*

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 *Reject*. We expect to observe more *Reject* outcomes with threats when reaching to stage 2, relative to messages involving no threats (cheap talk).

Hypothesis 3.3. *Communication in the form of threats drives the effect of messages on beliefs.*

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

Hypothesis 3.4. *The effect of threats on behavior is belief-dependent.*

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.

3.4 Results

This section is organized as follows: Section 3.4.1 summarizes the overall behavioral results on treatment effect. We focus on analysis of the communication treatment effect on cooperation and costly punishment. Section 3.4.2 presents results on threats vs. cheap talk. In Section 3.4.2, we conduct non-parametric and regression analyses to test Hypothesis 3.1 and 3.2. Section 3.4.3 tests Hypothesis 3.3 and 3.4 regarding participants belief-dependent motivations.

3.4.1 The Effect of Communication on Cooperation & Costly Punishment

Overall, we find that communication has a strong deterrence effect. Table 3.2 summarizes the outcomes of each game using session-level averages. First, when communication is not allowed, P2 chooses *Reject* 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 the 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 3.2. This result is also illustrated in Figure 3.4(a), with the vertical bar representing the 95% confidence interval.

At first glance the communication treatment does not seem to have an effect on P2's *Accept* vs. *Reject* choices, as shown in Table 3.2. When focusing only on P2's behavior in the last stage, we notice a slightly higher but non-significant *Reject* rate in the communication

Table 3.2. Communication Treatment Effect on Behavior

	P1's *Share* Rate			P2's *Reject* 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.

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 3.4(b). We see roughly the same *Reject* rate in both treatments in the deterrence and the staggered entry games. Although we do not see a clear difference in P2's *Reject* behavior comparing the different treatments, we cannot simply conclude that communication impacts only P1 and not P2. Dufwenberg et al. (2018a) 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 *Reject*, 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 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 the communication treatment effect. Consistent with previous non-parametric

Table 3.3. 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*** (0.007)	-0.169*** (0.005)	-0.089*** (0.007)	-0.089*** (0.005)
Staggered Entry	0.041* (0.022)	0.041** (0.017)	-0.050** (0.019)	-0.050*** (0.013)
Communication		0.171*** (0.017)		0.180*** (0.013)
Constant	0.984*** (0.027)	0.899*** (0.026)	0.739*** (0.025)	0.649*** (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.

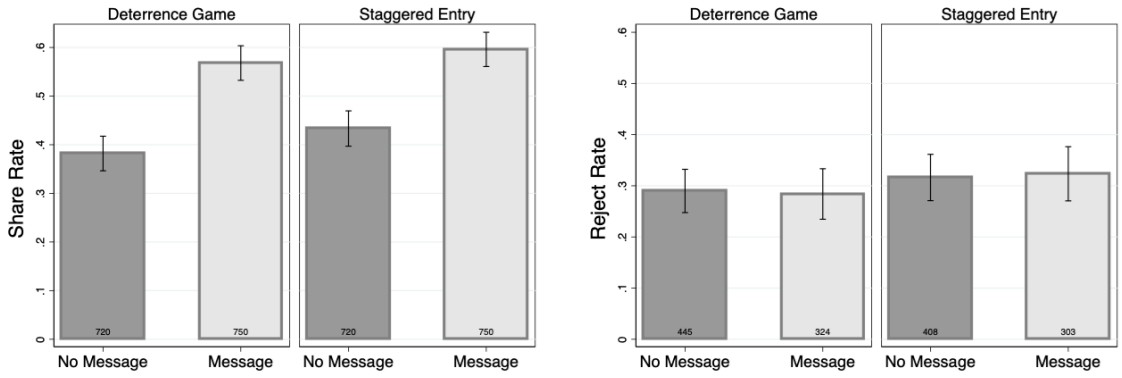
Note: We ran linear probability regressions for P1's *Share* Choice and linear regressions for P1's *Share* Plan. Data for each game are aggregated at the session level.

Table 3.4. Regression results – The Effect of Communication on P2's *Reject* Choice and Plan

	P2's *Reject* Choice		P2's *Reject* Plan	
	A coef / se	B coef / se	C coef / se	D 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 *Reject* Choice and linear regressions for P2's *Reject* Plan. Data for each game are aggregated at the session level.



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

Figure 3.4. Outcome Summary with Communication Treatment Effect

results, when regressing P1's *Share* choice (Table 3.3), communication increases *Share* rate significantly, and when regressing P2's *Reject* choice (Table 3.4) , communication does not seem to affect *Reject* 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 Supplementary Figures 3.14 & 3.15 (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 *Reject* decisions. In addition, the coefficient on "Staggered Entry" becomes marginally significant. P2 reports that she is more likely to choose *Reject* 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 3.5. In total, communication helps to increase welfare (P1's and P2's payoffs combined) by \$1.05 (1-sided Fisher's exact, $p < .001$).

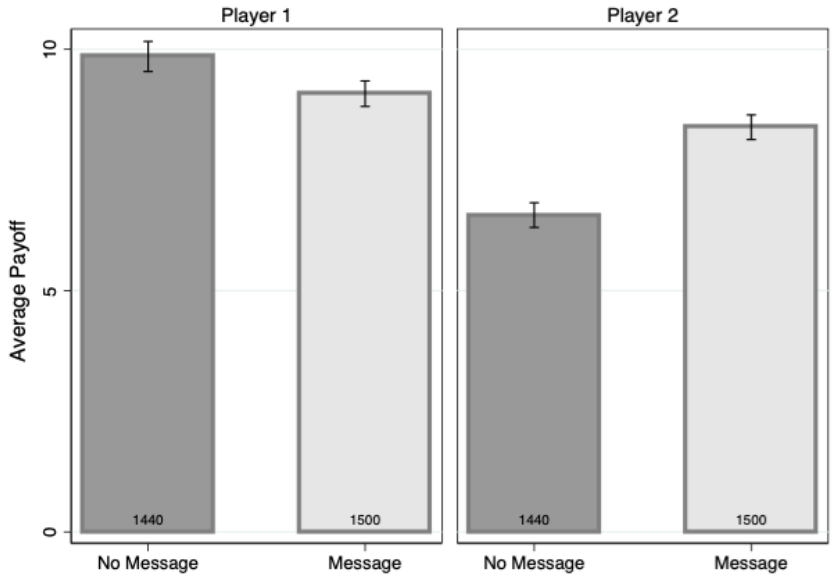


Figure 3.5. Payoff Distribution

3.4.2 The Credibility of Threats

To examine the effect of message contents on behavior (Hypotheses 3.1 and 3.2), 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 *Take*, I will *Reject*." 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 3.6 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 3.5, 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 3.5. The Effect of Threats on Behavior

Deterrence Game		Share	Accept	Reject	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	Reject	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 3.5 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 3.1, that threats result in a higher *Share* rate in both games. These results are graphically presented in Figure 3.7(a), with the vertical bars representing the 95% confidence intervals.

To test Hypothesis 3.2, we examine P2's behavior with both threats and cheap talk. Table 3.5 demonstrates that for the deterrence games, the conditional *Reject* rate is significantly higher with threats (43.07% vs. 17.65%, 1-sided Fisher's exact, $p < .001$). The same result

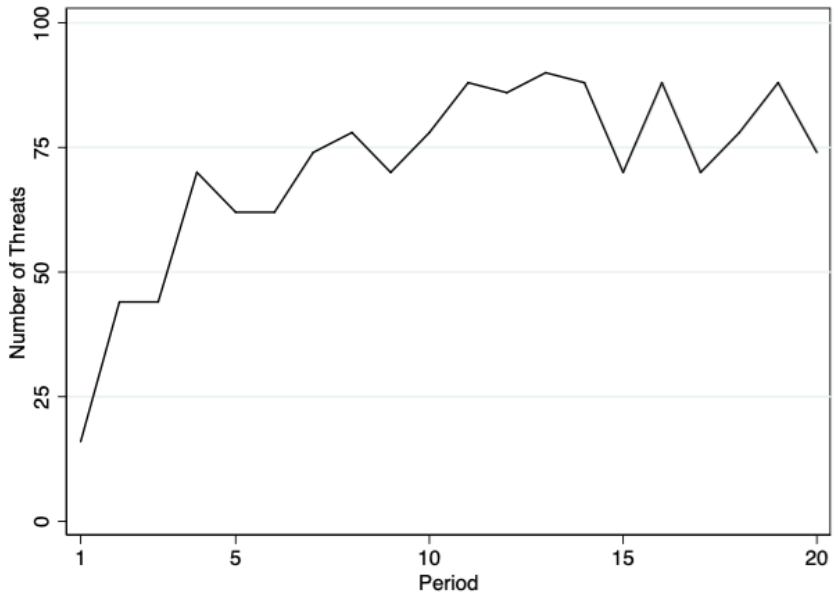
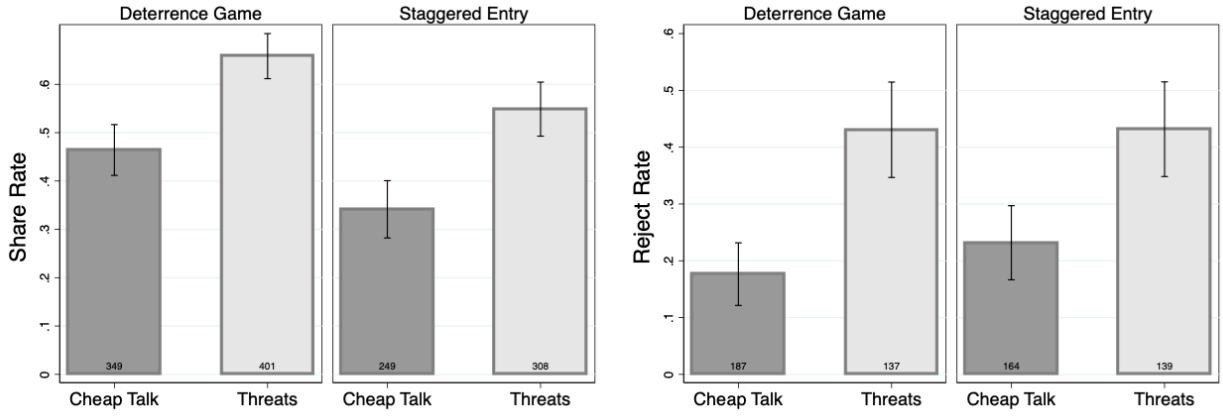


Figure 3.6. Number of Threats in Each Period

holds for the staggered entry games (43.17% vs. 23.17%, 1-sided Fisher's exact, $p < .001$). Figure 3.7(b) demonstrates that P2 *Reject*s more often when sending a threat instead of sending cheap talk. This is consistent with our Hypothesis 3.2 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 3.6. "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 *Reject* choices. In addition, we observe in these regression analyses that our staggered entry procedure produces higher rates of both *Share* and *Reject* choices.



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

Figure 3.7. Outcome Summary Comparing Threats vs. Cheap Talk

Table 3.6. Logistic Regressions – Effect of Threats on Players' Behavior

	P1's *Share* Choice		P2's *Reject* Choice	
	A coef / se	B coef / se	C coef / se	D coef / se
Payoff from *Accept*	-0.859*** (0.052)	-0.872*** (0.054)	-0.475*** (0.054)	-0.523*** (0.067)
Staggered Entry	0.134* (0.077)	0.179** (0.082)	0.229** (0.112)	0.214* (0.124)
Period	0.050*** (0.007)	0.044*** (0.006)	0.036 (0.024)	0.013 (0.023)
Threats		0.418*** (0.161)		1.230*** (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.

3.4.3 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 3.3) and that threats would work through the mechanism of belief-dependent frustration and anger to generate a self-fulfilling effect on behavior (Hypothesis 3.4). To test Hypothesis 3.3, 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 3.4.

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 3.7 presents summary statistics for self-reported beliefs (both players' beliefs about *Share* and *Reject*) recorded after messages are received, and Supplementary Figures 3.16 and 3.17 (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 3.4.1, self-reported beliefs and plans are good predictors of participant behavior (see Supplementary Figures 3.14 & 3.15 in the Appendix for ROC analyses).

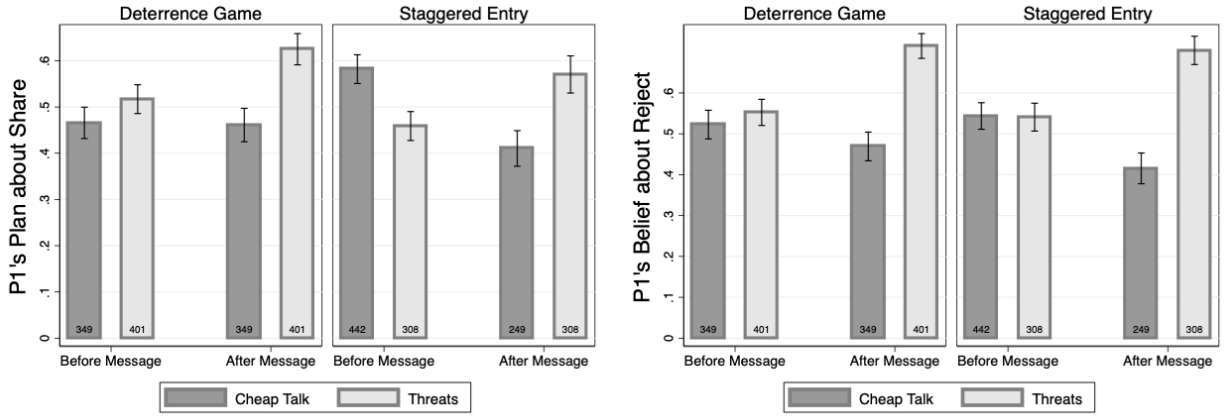
For both players and for both types of games, the effect of communication on reported beliefs is driven by the messages containing threats (Figures 3.8 & 3.9), consistent with Hypothesis 3.3.

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 3.8(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 3.7. 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: *Reject*	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: *Reject*	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 the 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 *Reject*

Figure 3.8. P1's Reported Beliefs

We note a similar pattern in P1's reported 1st order beliefs about P2's *Reject* choices. Figure 3.8(b) shows that P1s' reported 1st order belief about *Reject* 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 *Reject* 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.

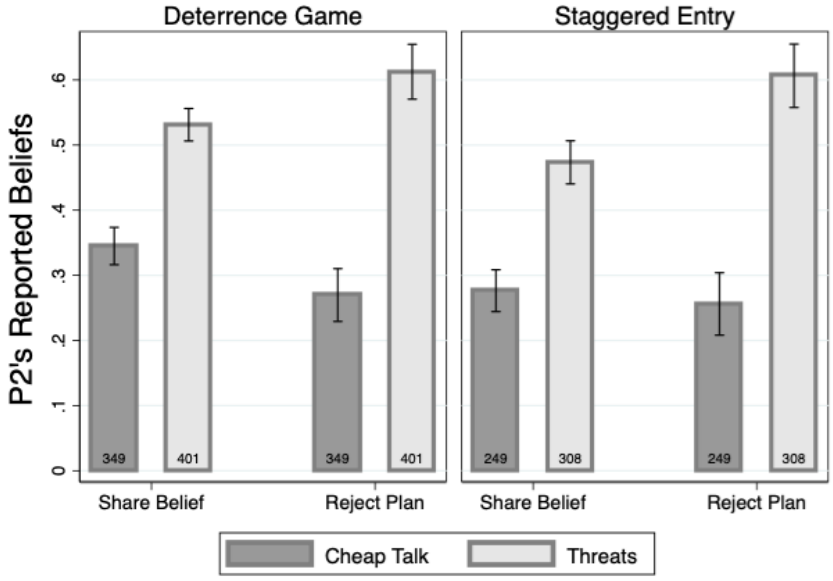


Figure 3.9. P2's Reported Beliefs

Figure 3.9 demonstrates that on average, P2 reports a higher 1st order belief about *Share*, and a higher probability to choose *Reject* 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 3.3.

We also run logistic regressions to test Hypothesis 3.4, focusing on whether participants' 1st order beliefs are associated with P1's choice between *Share* and *Take* and P2's choice between *Reject* and *Accept*. In Table 3.8, 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 *Reject* and plan to *Share* is positively associated with P1's *Share* choice. For the communication treatment sample, comparing Table 3.6 regression model B to Table 3.8 regression model H, the effect of threats diminishes after adding P1's 1st order belief about *Reject*. 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 3.8 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 3.4.

Table 3.9 presents logistic regressions with subject level controls in order to illustrate the relationship between P2's reported beliefs and P2's choice of *Reject*. We study this relationship again on three samples: the full sample, the no communication treatment sample, and the communication treatment sample. As in Table 3.8, 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 *Reject*. Even after controlling for "Threats" (model H) in the communication treatment sample, P2's 1st order belief about *Share* shows a strong association with *Reject* 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 *Reject* (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 *Reject* predicts P2's actual *Reject* choice well, and that it is reasonable to treat P2's plan as a close proxy for P2's choice.

Table 3.8. 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: *Reject*		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 3.9. Logistic Regressions – Effect of Beliefs on P2's *Reject* 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: *Reject*			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.

3.5 Conclusion

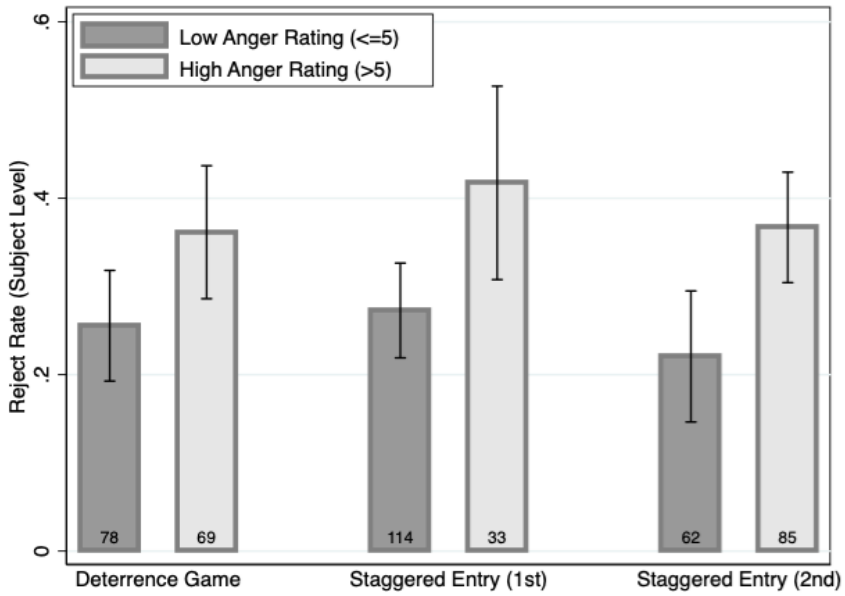
In this paper, we study the relationship between threats, credibility, and costly punishment, deriving theoretical predictions from the model of belief-dependent anger of Battigalli et al. (2018). When combined with the notion that communicated messages influence beliefs, our model implies that threats 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.

3.6 Appendices

3.6.1 Self-Reported Anger

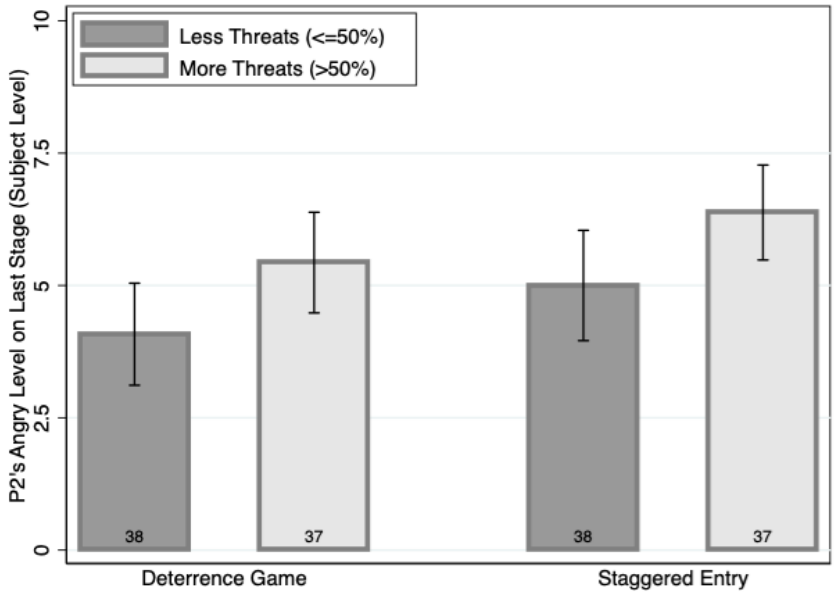
After the experiment concludes, we elicited self-reported measures of anger from participants assigned the role of Player 2. We are able to examine whether an individuals' level of anger is correlated with their behavior. Various studies have shown that the ultimatum game induces negative emotions especially 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 *Take* in the deterrence games, 2) If P1 chose *Take* in the 1st stage of the staggered entry games, and 3) If P1 chose *Take* in the 2nd stage of the staggered entry games. Questions 1-3 in Supplementary Table 3.10 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 3.10. Greater Anger with Higher Reject Rate

In Supplementary Figure 3.10, 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 *Reject* 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 *Take* on the

2nd stage, individuals report higher anger ratings, compared to when opponents choose *Take* on the 1st stage in the staggered entry games (1 sided t-test p-value $< .001$). P2's anger builds up with opponent's *Take* actions, and this might be the reason why P2 is more likely to *Reject* in the staggered entry games than in the deterrence games.

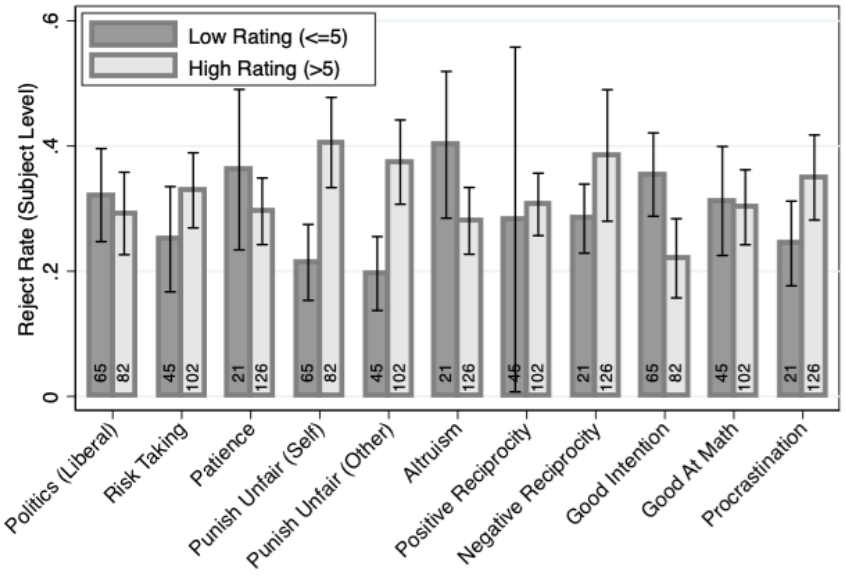


Supplementary Figure 3.11. Greater Anger at Disregarded Threats

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 3.11, 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.

3.6.2 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 3.10 for the exact questions.

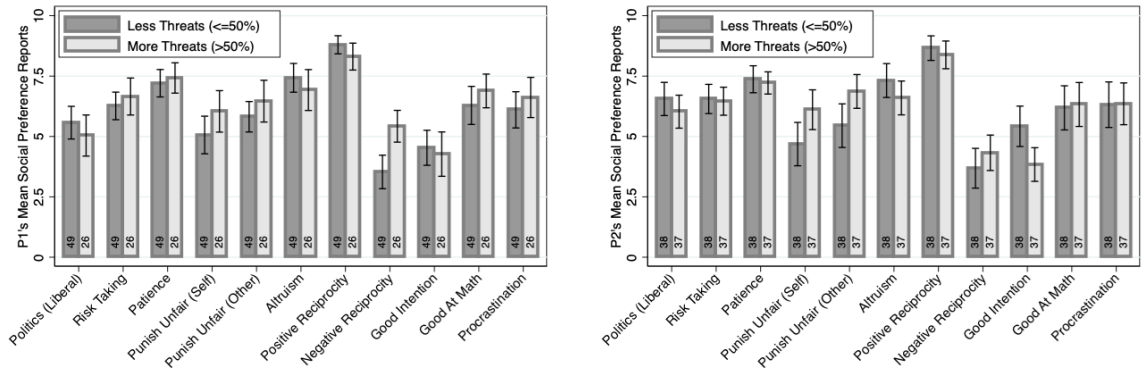


Supplementary Figure 3.12. Social Preferences and Reject Rate

The relationship between self-reported social preferences and the *Reject* rate is depicted in Supplementary Figure 3.12. 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 *Reject* rate. Individuals who report higher ratings for altruism (p-value = .043) and good intention (p-value = .028) choose *Reject* 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 *Reject* offers. 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

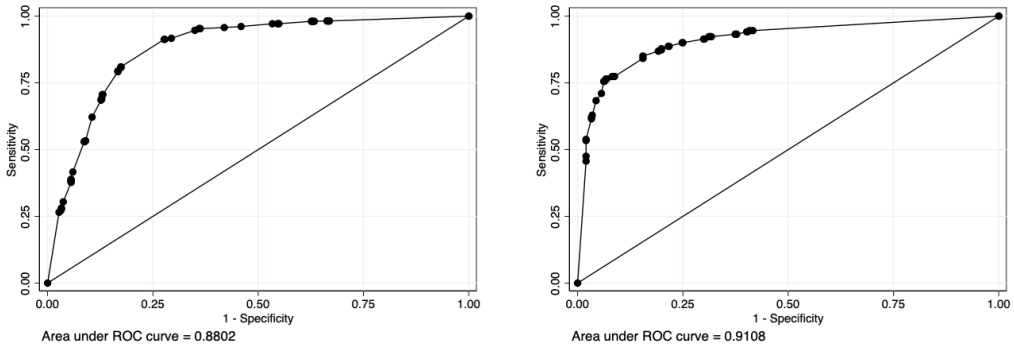
Supplementary Figure 3.13. Social Preferences Reports with Threats vs. Cheap Talk

Supplementary Figure 3.13 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 3.10. Survey Questions: Anger and Social Preferences

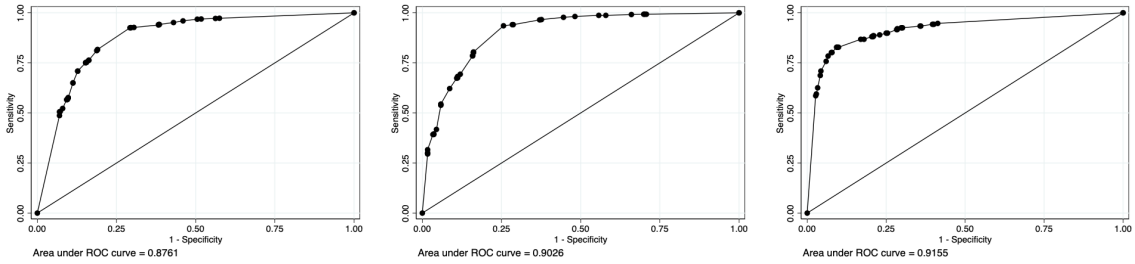
	Questions	Choose 0 if	Choose 10 if
1	How are you feeling if Player 1 chooses Option B (right) in stage 1 in the rounds of short games?	Not angry at all	Very angry
2	How are you feeling if Player 1 chooses Option B (right) in stage 1 in the rounds of long games?	Not angry at all	Very angry
3	How are you 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

3.6.3 Belief Elicitation



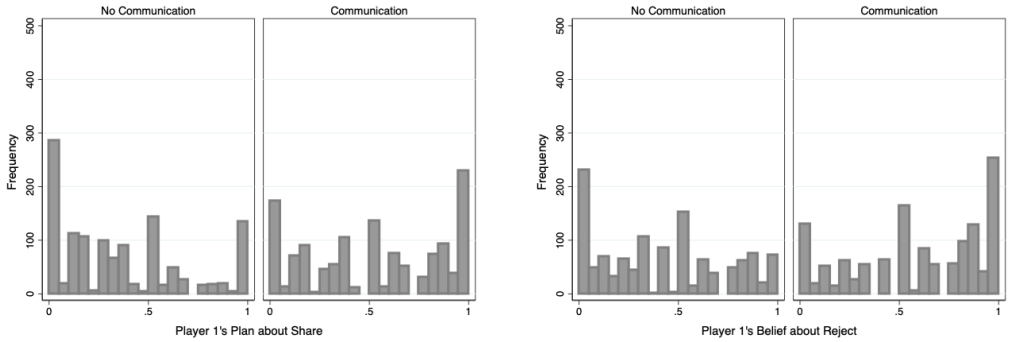
(a) P1's Plan about *Take* **(b)** P2's Plan about *Reject*

Supplementary Figure 3.14. Reported Plan Predicts Own Behaviors - Deterrence Games



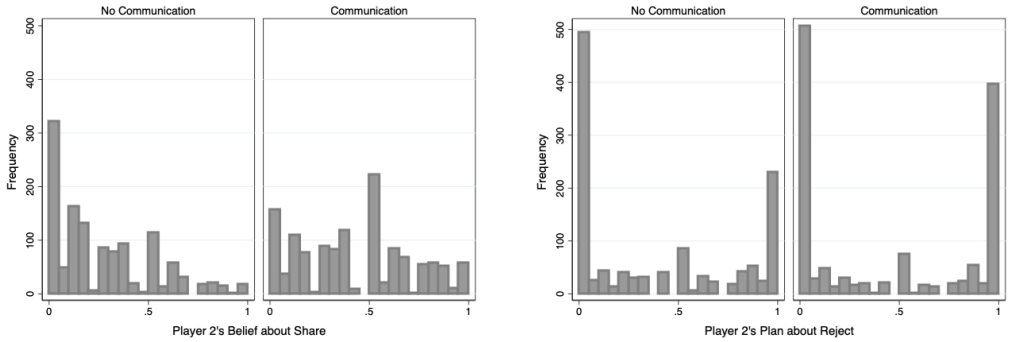
(a) P1's Plan on *Take* (St1) **(b)** P1's Plan on *Take* (St2) **(c)** P2's Plan on *Reject*

Supplementary Figure 3.15. Reported Plan Predicts Own Behaviors - Staggered Entry Games



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

Supplementary Figure 3.16. P1's Reported Beliefs Histograms



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

Supplementary Figure 3.17. P2's Reported Beliefs Histograms

3.6.4 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.

Chapter 4

Neuromodulation of Other-Regarding Preferences via HD-tDCS over the Right Temporoparietal Junction

Flora Li, Sheryl Ball, Xiaomeng Zhang, and Alec Smith

Abstract

We examine the effect of focal high definition transcranial Direct Current Stimulation (HD-tDCS) over the brain's right temporoparietal junction (rTPJ) on other-regarding preferences and rational choices. We hypothesized that anodal rTPJ stimulation would cause participants to behave more altruistically relative to sham stimulation, and that cathodal rTPJ stimulation would have the opposite effect. We measure the effect of stimulation on both social preferences and also on the consistency and rationality of individuals' choices. Consistent with prior studies we find that rTPJ plays an important role in social behavior: HD-tDCS stimulation over rTPJ modulates other-regarding preferences. We also find that behavior was less rational in the cathodal condition, and more rational in the anodal condition, relative to sham. Thus we conclude that rTPJ plays a role not only in other-regarding behavior but also that rTPJ activity more broadly affects the consistency and rationality of choices. The results suggest that comprehensive theories of rTPJ function in social behavior should account for the multifaceted role that rTPJ plays in processing sensory information.

4.1 Introduction

Humans often sacrifice to benefit unrelated or even unknown others, even when there is no possibility of future interaction. Such pro-social behavior is evolutionarily stable (Hamilton, 1964; Bowles, 2006; Alger and Weibull, 2013), is important for the functioning of large social groups (Fehr and Fischbacher, 2003), and is economically rational (Eckel and Grossman, 1996; Andreoni and Miller, 2002). A large and growing literature identifies prosocial behavior with overlapping patterns of activity in both the brain's reward and valuation systems and in a social cognition network that includes the medial prefrontal cortex, the anterior insula, the temporal sulcus and the temporoparietal junction (Singer et al., 2006; Harbaugh et al., 2007; Chang and Sanfey, 2011; Morishima et al., 2012; Hutcherson et al., 2015a,b; Tusche et al., 2016).

In humans, noninvasive brain stimulation (NBS) permits the identification of the causal role of specific brain regions in cognition and behavior. A number of studies have used this approach to study social behavior (Cattaneo et al., 2011; Sellaro et al., 2015; Marini et al., 2018). The role of the TPJ in social cognition has been studied extensively via NBS (Young et al., 2010; Santiesteban et al., 2012; Donaldson et al., 2015; van Elk et al., 2017). However, the TPJ also performs many other functions (Mitchell 2007), such as integrating and processing sensory information and regulating attention (Corbetta and Shulman, 2002; Binder et al., 2009; Krall et al., 2015; Gohil et al., 2016). Furthermore, because decisions result from the complex interrelationship of neuronal activity in a diffuse network, it is important to establish that the modulation of neural activity in a particular region results in decisions that are consistent with rational choice.

In this paper we report the results of an NBS study that measures the effect of focal electric stimulation over the temporoparietal junction on both altruistic preferences and choice quality. In our experiment human subjects participated in a graphical revealed preference task while undergoing high-definition transcranial direct current stimulation (HD-tDCS) over the right temporoparietal junction (rTPJ). In the task participants allocated money between themselves and a charity while facing varying prices and budgets. Participants were randomly assigned to anodal, sham, or cathodal stimulation. This random assignment and use of a paradigm expressly designed to measure the consistency of choices permits us to not only identify the directional effect of stimulation of the rTPJ on preferences, but also to measure the consistency of the choices participants made, a measure of economic rationality. This makes it possible to check whether brain stimulation affects behavior in a consistent manner.

Through NBS, we show that rTPJ plays two distinct roles in social decision-making. First, stimulation over the rTPJ affects participant's choices about how to divide money between

themselves and a charity. Fewer participants in the cathodal treatment choose equal splits, consistent with previous results (Güroğlu et al., 2011; Morishima et al., 2012). Furthermore, we find that rTPJ is involved in regulating individuals' choice consistency: the choices of the participants in the anodal treatment are more consistent with economic rationality, calculated via Afriat's Critical Cost Efficiency Index (CCEI). Thus, we find that focal anodal stimulation over the rTPJ simultaneously results in more fair-minded and more rational behavior.

4.2 Materials and Methods

4.2.1 Participants

We recruited 104 healthy participants from the university community. We analyzed data from 102 participants (age: max 65, min 18, mean 22.64, SD 7.19; gender: 55 males, 47 females).[1] Out of 102 participants 7 were university staff or faculty, 15 were graduate students, and the rest were undergraduate students. We obtained informed consent from all participants, and all participants completed a pre-experiment safety screening questionnaire to ensure they were tDCS compatible. The experimental protocol was approved by the Institutional Review Board of Virginia Tech.

4.2.2 Transcranial Direct Current Stimulation Treatments

During the behavioral task, participants received either anodal, cathodal, or sham HD-tDCS. This is a non-invasive neural manipulation technique with no more than minimum risk (Villamar et al., 2013). Stimulation was delivered over the right temporoparietal junction (CP6) with an intensity of 2mA using a 4x1 ring electrode montage. Figure 4.1.A shows the simulation of the anodal stimulation, and the current modeling software (Soterix and SimNibs) indicates the precise stimulation location. We used a neuroConn DC-Stimulator MC (München, Germany) to administer the stimulation.

The tDCS stimulation was turned on after participants read the instructions and completed three practice trials. then participants were required to wait for 2 minutes to begin the behavioral task to ensure that they were comfortable and allow the effects of the stimulation

[1]One participant was dropped from the study due to an adverse event (Li et al., 2018). The other participant was excluded due to excessive conductive gel application, therefore we suspected that there was no electric current running through the target area.

to equilibrate. Participants completed the task at their own pace, and stimulation was terminated when participants finished the task. The average stimulation duration is 11.79 minutes, which is well within the safety limit of tDCS best practices (Poreisz et al., 2007; Bikson et al., 2009). Of the 102 participants, 34 received anodal tDCS stimulation (age: max 47, min 18, mean 22.65, SD 5.76; gender: 17 males, 17 females), 34 received cathodal tDCS stimulation (age: max 53, min 18, mean 22.38, SD 7.00; gender: 21 males, 13 females), and 34 received sham tDCS stimulation (age: max 65, min 18, mean 22.88, SD 8.84; gender: 17 males, 17 females). A post-experiment questionnaire indicated that participants could not distinguish among different tDCS treatments (see the supplementary materials for details).

4.2.3 Experiment Design and Procedures

The behavioral task is a graphical revealed preference paradigm where participants allocate money between themselves and a charity. The task was programmed with Matlab and Psychtoolbox (Brainard, 1997; Pelli, 1997; Kleiner et al., 2007). This behavioral task is simple, flexible, and designed to encourage rational decisions. Similar task designs have been used in behavioral experiments to study decision quality and altruism, risk preferences, charitable giving, political behavior, and stress (Andreoni and Miller, 2002; Fisman et al., 2007; Choi et al., 2014; Fisman et al., 2017; Cettolin et al., 2018). In the task participants allocated an endowment of tokens between themselves and a local food bank, Feeding America Southwest Virginia (FASWVA), during 50 independent trials. Keeping tokens for themselves was referred to as Keep, and allocating tokens to FASWVA was referred to as Pass. In each trial, participants could choose allocations either on or under a graphical representation of a budget line (Figure 4.1.B). The endowment and relative price of contributing to the charity were randomly varied across 50 trials, such that the budget line intersected with at least one axis at 50 or more tokens, with a maximum intercept value of 100 tokens, using the same design as in Fisman et al. (2017). After participants made each allocation, feedback was shown on the next screen. Participants received $20 compensation. In addition, both the participant and the charity received additional earnings based on the allocation from one randomly selected trial, where each allocated token was worth $0.50.

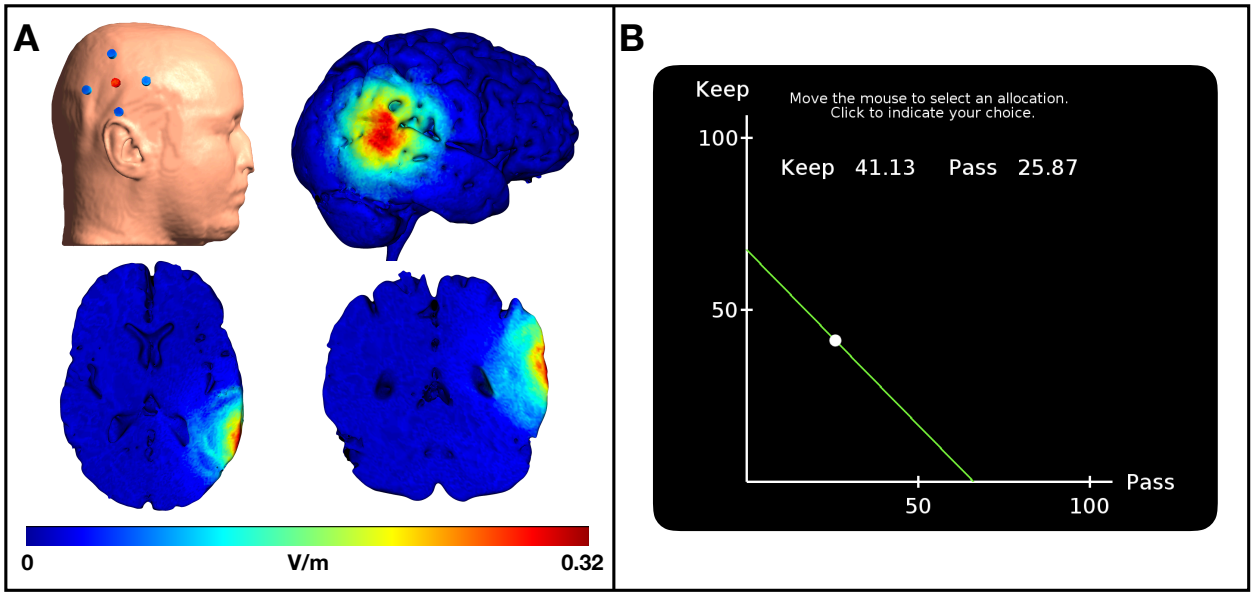


Figure 4.1. Current Modeling and Experiment Task. **A.** Current Modeling. Simulation of the norm of the electric field (V/m) induced by anodal stimulation over CP6. The current modeling shows the stimulation pattern on the surface of the brain with a sagittal view (top right), and how far the stimulation reaches on an axial slice (bottom left) and a coronal slice (bottom right). **B.** Experiment Task. Participants choose an allocation on or under the budget line (green line). Points on and under the line represent all feasible allocation of tokens. Subjects do not have enough tokens to choose above the line. Allocations under the line are inefficient. Decisions are made by moving a mouse and clicking to choose a desired point. While participants move the mouse, they can see a white dot is moving and can view allocation information on the top of the screen.

4.2.4 Model and Analysis

Other-Regarding Preference

We fit each individual's choice data with a parametric constant elasticity of substitution (CES) utility function that measures the extent of other regarding behavior.

$$u(\pi_k, \pi_p) = \left(\alpha\pi_k^{\rho} + (1-\alpha)\pi_p^{\rho}\right)^{\frac{1}{\rho}}$$

In the CES utility function, π_k and π_p represent tokens allocated to self and the charity respectively. The parameter $\alpha \in [0, 1]$ measures individual's other-regarding behavior, where $\alpha = 0$ means complete selflessness, $\alpha = 1$ means complete selfishness, and $\alpha = 0.5$ means equal division (see Figure 4.2.A). The parameter $\rho \in [-\infty, 1]$ describes the curvature of an individual's indifference curve, and captures the tradeoff between maximizing the number of tokens distributed and fair division. When ρ approaches 1, the indifference curve approaches a straight line, and participants heavily favor allocations that give every token to either the participant or the charity. When ρ approaches $-\infty$, the indifference curve approaches an

L-shaped or Leontief indifference curve, and selected allocations occur closer to the center of the budget line.

The budget constraint is $\frac{\pi_k}{K} + \frac{\pi_p}{P} \leq 1$, where K refers to the maximum possible tokens to self, and p refers to possible maximum tokens to the charity. We can then solve the expenditure function: [2]

$$\frac{\pi_k}{K} = \frac{\left(\frac{\alpha}{1-\alpha}\right)^{\frac{1}{1-\rho}}}{\left(\frac{\alpha}{1-\alpha}\right)^{\frac{1}{1-\rho}} + \left(\frac{K}{P}\right)^{\frac{\rho}{\rho-1}}}$$

For each individual i we use maximum likelihood with $n = 1, \ldots, 50$ decisions to obtain parameters α_i and ρ_i.

$$\frac{\pi_{k,i}^n}{K_i^n} = \frac{\left(\frac{\alpha_i}{1-\alpha_i}\right)^{\frac{1}{1-\rho_i}}}{\left(\frac{\alpha_i}{1-\alpha_i}\right)^{\frac{1}{1-\rho_i}} + \left(\frac{K_i^n}{P_i^n}\right)^{\frac{\rho_i}{\rho_i-1}}} + \varepsilon_i^n$$

Rationality

Utility functions are derived from individuals' preferences. When an individual's choices are consistent with the requirements of economic rationality, we can conclude that she is maximizing utility.

There are various measures of rationality, and we can identify three major rationality measures based on our experimental design. We measure choice consistency and economic rationality by measuring 1) violations of the Monotonicity Axiom (Monotonicity), 2) violations of the Weak Axiom of Revealed Preference (WARP), and 3) violations of the Generalized Axiom of Revealed Preference (GARP). Figure 4.2 Panel B, C, and D show graphical representations of the three rationality measures which are described in more detail below. For decision bundles $x_1(\pi_k, \pi_p), x_2(\pi_k, \pi_p), x_3(\pi_k, \pi_p)$ we define:

Directly Revealed Preference ($\succsim^D$): $x_1 \succsim^D x_2$ if $x_1, x_2 \in B$, and $x_1 \in C(B)$.

Indirectly Revealed Preference ($\succsim^I$): $x_1 \succsim^I x_2$ if $x_1 \succsim^D x_3$ and $x_3 \succsim^D x_2$.

[2] The step-by-step derivation can be found in the supplementary materials.

Monotonicity requires that, for $x_1(\pi_k, \pi_p), x_2(\pi_k, \pi_p) \in B$, if $x_1 \gg x_2$, then $x_1 \succ x_2$. Utility maximization and Monotonicity require individuals to choose consumption bundles on the budget constraint. Any under the line consumption bundle x would violate Monotonicity.[3] The severity of Monotonicity violations is tested by calculating the absolute distance to the budget constraint: the greater the distance is, more severe the violation is. Monotonicity implies that indifference curves are thin, downward sloping, and do not cross each other.

WARP states that, if $x_1(\pi_k, \pi_p) \succsim^D x_2(\pi_k, \pi_p)$ then we do not have $x_2 \succsim^D x_1$. If both consumption bundles x_1, x_2 are available for two different budget constraints B, B', and an individual chooses x_1 under B, and chooses x_2 under B', then she violated WARP. WARP is necessary for the existence of a strictly convex utility function.

GARP states that, if $x_1(\pi_k, \pi_p) \succsim^I x_2(\pi_k, \pi_p)$, then we do not have $x_2 \succsim^D x_1$. Violation of Transitivity implies violation of GARP, therefore GARP rules out preference cycles. The severity of GARP violations is measured using Afriat's (1972) Critical Cost Efficiency Index (CCEI), where CCEI $\in [0, 1]$. CCEI measures how far budget constraints must be shifted to avoid a GARP violation. CCEI = 1 represents no GARP violation, and CCEI = 0 represents the most severe GARP violation. GARP is both a necessary and sufficient for the existence of a well-behaved utility function.

[3]It might be challenging for some participants to aim on the budget line; therefore, only if the distance from the allocation to the budget line is greater than \$0.05, that allocation is counted as one violation of Monotonicity.

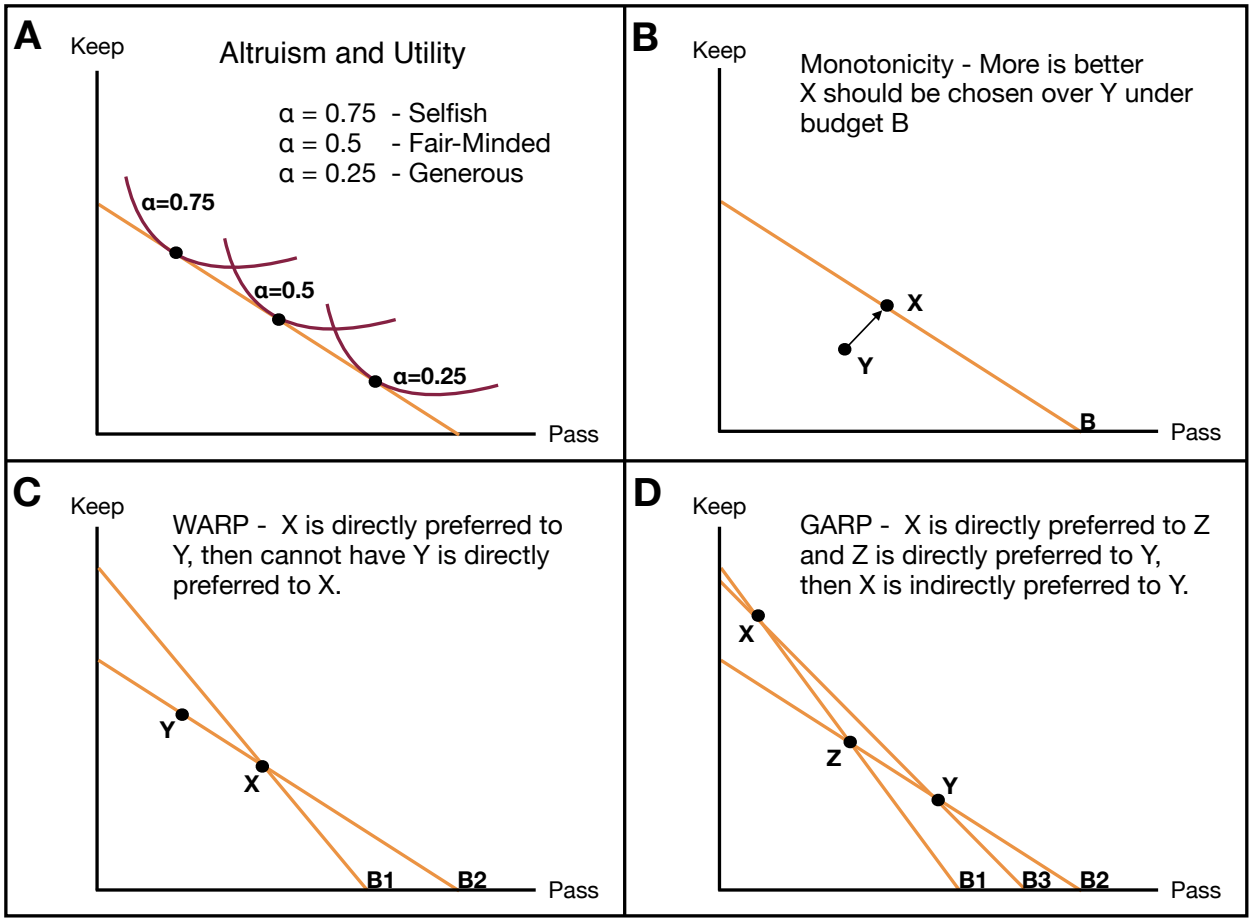


Figure 4.2. Theoretical Modeling. **A.** Example of CES Utility Function (our structural model for other-regarding preferences). CES Utility Function: $u(\pi_k, \pi_p) = \left(\alpha\pi_k^{\rho} + (1-\alpha)\pi_p^{\rho}\right)^{\frac{1}{\rho}}$, where α stands for weight on self, $1-\alpha$ stands for weight for charity, and ρ controls the curvature of the utility function. Here ρ is fixed and the decisions of three individuals with different preferences α =0.75 (selfish), 0.5 (fair-minded), and 0.25 (generous) are illustrated. **B.C.D.** Rationality and Choice Consistence Measurements. **B.** Monotonicity: For X and Y available under budget B, if X is strictly greater than Y, then X is preferred to Y. For monotonicity to hold, an individual can not choose any allocation under the budget line. The absolute distance from the chosen bundle to the budget line measures the severity of a Monotonicity violation. **C.** WARP: If X is directly revealed preferred to Y, then we do not have Y directly revealed preferred to X. WARP implies that if an individual chooses X over Y when the budget line is B1, she can not choose Y over X when the budget line is B2. **D.** GARP: If X is indirectly revealed preferred to Y, then we do not have Y directly revealed preferred to X. GARP implies that if an individual chooses X over Z when the budget line is B1, and Z over Y when the budget line is B2, she can not choose Y over X when the budget line is B3. Violation of Transitivity implies a violation of GARP.

4.3 Results

4.3.1 Donation Behavior

Since the number of tokens to be divided varies across trials, we evaluate individual's charity donation behavior by comparing Percent Kept $=\pi_k/K$, which equals 1 when an individual is completely selfish. Our sham result is comparative to previous work with a similar task: our sham participants on average kept 65.75% of the total tokens, where participants in Fisman et al. (2017) kept 65% of their tokens. Figure 4.3.A. shows that individuals' average Percent Kept in anodal treatment is marginally smaller than that in sham treatment (t test: p = .0735), while the cathode treatment is not different from either anode (t test: p = .4291) or sham (t test: p = .2804).

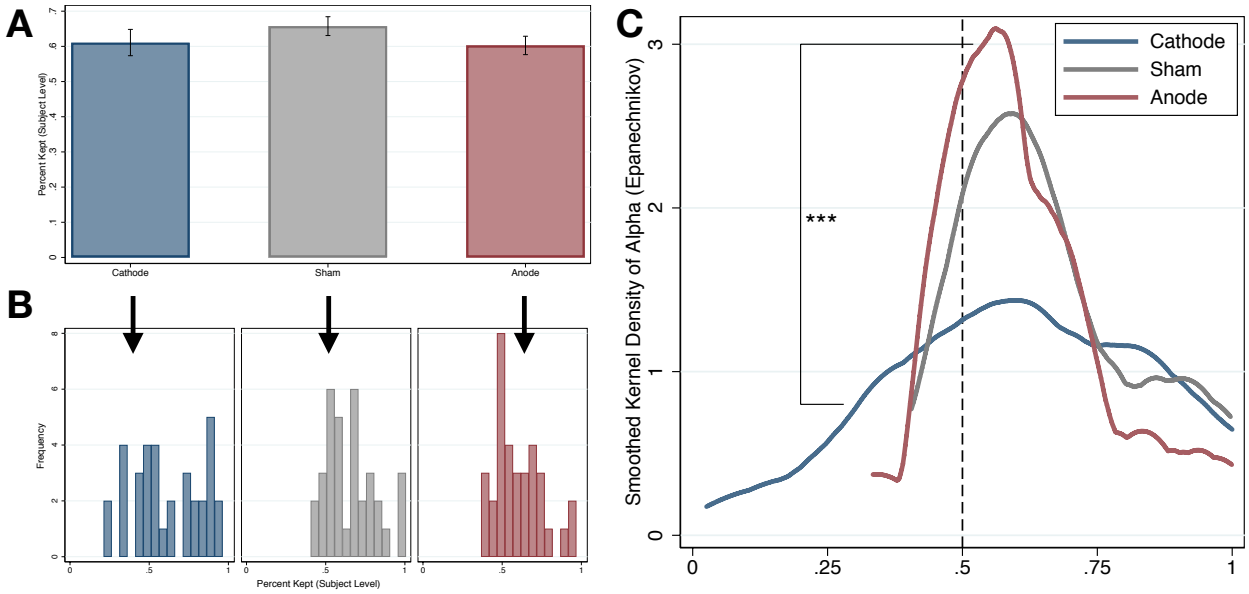


Figure 4.3. Other-Regarding Preferences Results. * if p-value < 0.1, ** if p-value < 0.05, and *** if p-value < 0.01. Error bars denote standard error of the mean (SEM). **A.B.** Relationship between tDCS treatments and charity giving behavior. Percent Kept $= \pi_k/K$ measures how much an individual kept relative to the initial endowment. **A.** Relationship between tDCS treatment and mean of Percent Kept. Wilcoxon ranksum tests fail to reject the null hypothesis that the medians are the same across all treatments (anode vs. cathode p = .9511, anode vs. sham p = .1136, and sham vs. cathode p = .6359). **B.** Relationship between tDCS treatments and distribution of Percent Kept. Histograms show that the distribution of Percent Kept are different across tDCS treatments. A test for equality of standard deviations shows that anodal distribution is the most concentrated and the cathodal distribution is the least concentrated (p = .0220). **C.** Relationship between tDCS treatments and CES selfishness parameter α, a measure of the value each participant places on their own payoff versus the charity ($\alpha = 1$, pure self-interest, $\alpha = 1$ pure altruism). Anodal stimulation results in a concentration of α values over 0.5; a test for equality of standard deviations (anode vs. cathode) rejects the null hypothesis of no difference (p = .0013).

The sample mean does not tell the complete story, however. When we look at a histogram of percent of tokens kept (Figure 4.3.B), we find that while the data have similar means

the distributions are significantly different (Epps-Singleton (ES) test: p = .0241). Behavior in the anodal treatment is fairly concentrated around the distribution mean, while results from the cathodal treatment have no obvious peak. Not surprisingly, therefore, the standard deviation from the cathodal treatment is larger than that from the anodal treatment (Standard Deviation (SD) test: p = .0220).

We next turn to the parametric structural CES model to estimate the parameter α, which measures selfishness, across the three treatments. We find that the structural model delivers very similar results as the analysis of individuals' percent kept. In Figure 4.3.C, the average of parameter α does not differ across three treatments (t test: anode $<$ cathode p = .2257, anode $<$ sham p = .9273, and sham $<$ cathode p = .8317).

Similar to the result for amount kept noted above, we find that both the distribution shape and standard deviation of α are different. More anodal individuals have α closer to 0.5 (fair-minded), and less anodal individuals have α closer to 0 (complete selfless) and 1 (complete selfish). Anodal distribution of α is slim, bell-shape, and very concentrated around 0.5, whereas cathodal distribution looks more like a uniform distribution (ES test: p = .0214). Distribution of α in sham is also different from that in cathode (ES test: p = .0767), but is not different from that in anode (ES test: p = .5803). The standard deviation test shows that anodal distribution is more concentrated, and cathodal distribution less concentrated, relative to sham (SD test: anode $<$ cathode p = .0013, anode $<$ sham p = .0819, and sham $<$ cathode p = .0481). We show that stimulation over rTPJ induces individuals' preferences for fairness.

4.3.2 Rationality Violations

We report results on both the frequency and severity of rationality violations. Figure 4.4 Panel A, B, and C shows how frequently participants violate Monotonicity, WARP, and GARP, respectively. We observe a consistent pattern for all three rationality measures: the fewest violations in occur in the anodal condition, and the most violations occur in the cathodal condition (t test: Monotonicity p = .0981, WARP p = .0279, and GARP p = .0294). We also observe that, for all three rationality measures, the cathode condition is not different from sham (t test: Monotonicity p = .2129, WARP p = .3802, and GARP p = .3072), but there are fewer WARP and GARP violations in the anode condition than that in sham (t test: Monotonicity p = .2588, WARP p = .0492, and GARP p = .0687).

Figure 4.4.D shows that GARP violations in cathodal treatment are not only the most frequent but also the most severe. CCEI in the anodal treatment is significantly higher than

in the cathodal (t test: p = .0102) and sham (t test: p = .0057) treatments, but CCEI in sham and in cathode are not different from each other (t test: p = .4942). Violation of Monotonicity shows a similar result, which can be found in the supplementary materials. The consistent pattern of both frequency and severity of rationality violations suggest that anodal tDCS over the rTPJ causes individuals to make both more rational and more consistent decisions.

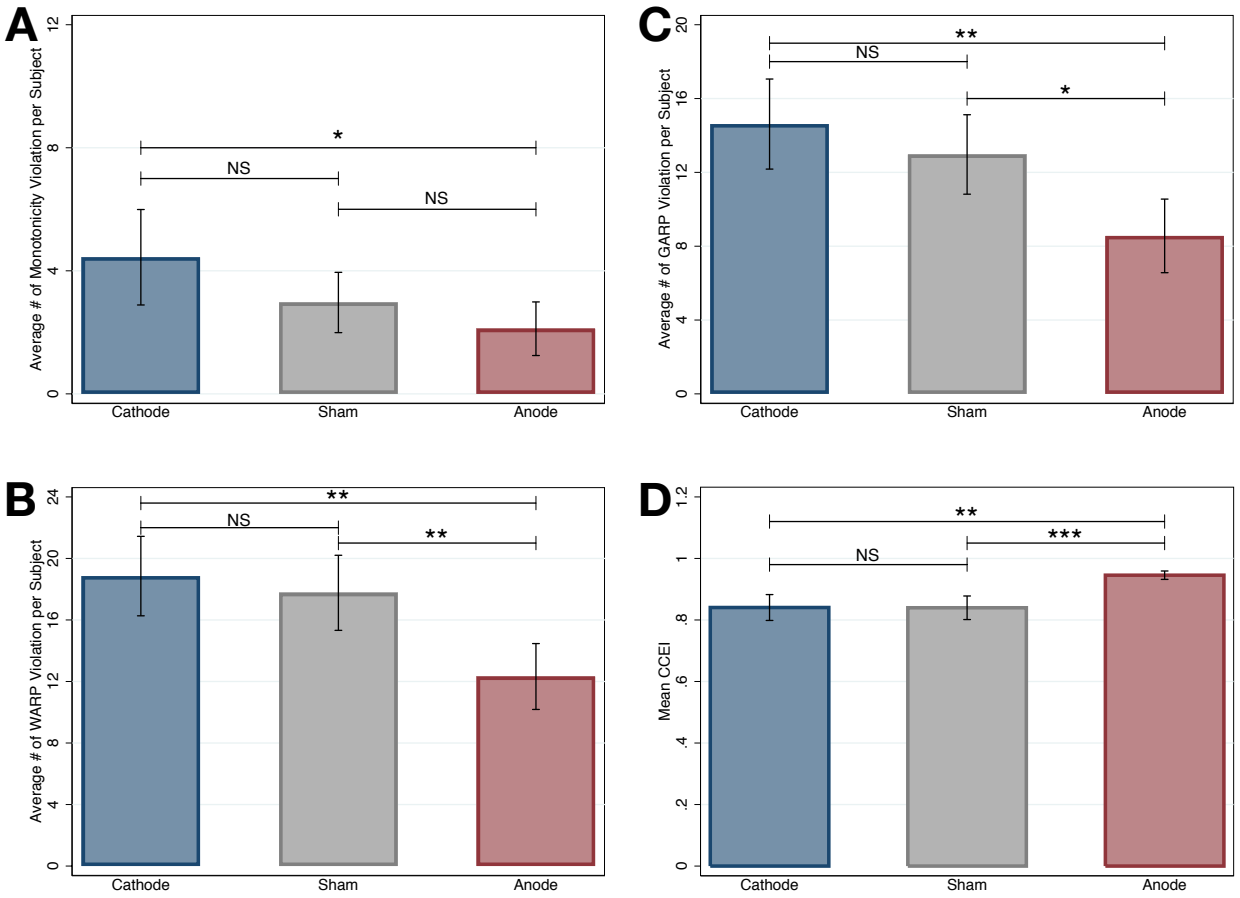


Figure 4.4. Rationality and Choice Consistency Results. * if p-value < 0.1, ** if p-value < 0.05, and *** if p-value < 0.01. Error bars denote standard error of the mean (SEM). **A.** The number of Monotonicity violations per subject, by treatment. Anodal participants have fewer Monotonicity violations; a t-test of the difference in means (anode vs. cathode) rejects the null hypothesis of no difference (p = .0981). **B.** The number of the Weak Axiom of Revealed Preference (WARP) violations per subject, by treatment. There is fewest WARP violations in the anodal treatment; a t-test of the difference in means (anode vs. cathode) rejects the null hypothesis of no difference (p = .0279). **C.** The number of the Generalized Axiom of Revealed Preference (GARP) violations per subject, by treatment. Anodal participants violate GARP least frequently. A t-test of the difference in means (anodal vs. cathodal) rejects the null hypothesis of no difference (p = .0294). **D.** Mean Critical Cost Efficiency Index (CCEI) values, by treatment. CCEI is a measure of the severity of violations of GARP. Violations of GARP in the anodal treatment is the least severe; a t-test of the difference in means (anodal vs. cathodal) rejects the null hypothesis of no difference (p = .0102).

4.4 Discussion

To summarize, using NBS we find that rTPJ plays two simultaneous roles in social decision-making: 1) modulating preferences for fairness and 2) regulating choice quality. We argue that based on current evidence, preferences for fairness and choice quality are two distinct functions of rTPJ. This raises the possibility that when previous NBS studies modulate fairness concerns in social decision-making, they also may affect economic rationality.

4.4.1 Other-Regarding Preferences

We show that rTPJ is causally involved in modulating preferences for giving. When we look at the donation behavior results, we notice that the significant result is driven by that cathodal participants exhibit a different distribution of donation behavior compared to both sham and anodal participants, using an Epps-Singleton test. This result supports the previous evidence showing that rTPJ is involved in making fair decisions in social context (Güroğlu et al., 2011; Morishima et al., 2012).

Our experiment measures the effect of HD-tDCS over rTPJ on behavior. Participants' charitable giving decisions might be driven by either their intrinsic altruistic preferences or their inference of experimenters' and charity's intention. In our experimental design, we cannot distinguish between pure altruism (Hutcherson et al., 2015a) and theory of mind (Young et al., 2010), and both concepts have been associated with rTPJ. Another limitation results from our between-subjects design. limiting our conclusions to the population level. We plan to explore these concerns in future research.

4.4.2 Rational Choices

We show that rTPJ plays a role in regulating individuals' rational preferences. We observe a systematic pattern indicating that anodal participants are more rational compared to both sham and cathodal participants. This is the first evidence that neuromodulation over rTPJ can lead to more rational decisions.

The rTPJ is also involved in processing sensory information. This brings into question the possibility that TPJ influences social behavior or rationality through neural computations that govern attention, perspective-taking, or other mechanisms. Our evidence at least suggests

that rationality results are not driven by hemifield neglect (Ptak and Schnider, 2011).[4] Our experimental design, however, restricts us from investigating further.

Economics literature suggests that revealed preferences can be affected by individual's attention (Masatlioglu et al., 2012; Caplin and Dean, 2015). It is also well established that rTPJ is closely associated with visual processing (Corbetta and Shulman, 2002; Corbetta et al., 2008). Therefore, it is reasonable to suspect that our rationality results are driven by the attentional differences caused by tDCS over rTPJ. It is possible that even though the whole area under the budget line is available for selections, but participants only pay attention intentionally (strategy) or unintentionally (induced by tDCS) to a subset of the choice set. To further investigate this matter, we propose to study the same paradigm using tDCS over rTPJ and eyetracker concurrently.

We cannot completely rule out the possibility that 1) the primary other-regarding result drives a secondary rationality result, or vice versa, or 2) rationality result does not exist without the charitable giving background, that is saying rTPJ involves in rationality only with social contexts. In future work, we plan to further study the role of the rTPJ in governing rational choices in both social and non-social contexts.

[4]Details in supplementary materials. Hemifield neglect refers individual's failure to be aware of objects to one side of their field of vision, even when their vision itself is normal. Left hemifield neglect can happen with impairment of rTPJ. Supplementary Figure 4.8.B and 4.8.C show that there is no difference in the horizontal and vertical distance of participant's allocation decisions to the budget line, as might be expected if hemifield neglect had influenced their choices.

4.5 Supplementary Document

4.5.1 CES Utility Specification

The charity giving task can be specified as a utility maximization problem using a CES utility function. In this utility maximization problem, π_k and π_p represent tokens allocated to self and the charity respectively. K refers to possible maximum amount of tokens to self, and P refers to possible maximum amount of tokens to the charity.

$$\max_{\pi_k,\pi_p} \left(\alpha\pi_k^\rho + (1-\alpha)\pi_p^\rho\right)^{\frac{1}{\rho}} \quad \text{s.t} \quad \frac{\pi_k}{K} + \frac{\pi_p}{P} \leq 1 \tag{4.1}$$

To solve the problem, we first form the associated Lagrangian

$$\mathcal{L}(\pi_k, \pi_p, \lambda) \equiv \left(\alpha\pi_k^\rho + (1-\alpha)\pi_p^\rho\right)^{\frac{1}{\rho}} + \lambda\left(1 - \frac{\pi_k}{K} + \frac{\pi_p}{P}\right) \tag{4.2}$$

Assuming monotonicity in preferences, the budget constraint will hold with equality at the solution. Assuming an interior solution, the Kuhn-Tucker conditions coincide with the ordinary first order Lagrangian conditions and the following equations must hold at the solution values π_k, π_p, λ:

$$\frac{\partial\mathcal{L}}{\partial\pi_k} = \alpha\pi_k^{\rho-1}\left(\alpha\pi_k^\rho + (1-\alpha)\pi_p^\rho\right)^{\frac{1}{\rho}-1} - \frac{\lambda}{K} = 0 \tag{4.3}$$

$$\frac{\partial\mathcal{L}}{\partial\pi_p} = (1-\alpha)\pi_p^{\rho-1}\left(\alpha\pi_k^\rho + (1-\alpha)\pi_p^\rho\right)^{\frac{1}{\rho}-1} - \frac{\lambda}{P} = 0 \tag{4.4}$$

$$\frac{\partial\mathcal{L}}{\partial\lambda} = 1 - \frac{\pi_k}{K} - \frac{\pi_p}{P} = 0 \tag{4.5}$$

Combining equations (4.3) and (4.4), we get

$$\lambda = K\alpha\pi_k^{\rho-1}\left(\alpha\pi_k^\rho + (1-\alpha)\pi_p^\rho\right)^{\frac{1}{\rho}-1} = P(1-\alpha)\pi_p^{\rho-1}\left(\alpha\pi_k^\rho + (1-\alpha)\pi_p^\rho\right)^{\frac{1}{\rho}-1} \tag{4.6}$$

Solving equation (4.5), we get

$$\pi_p = P - \frac{P}{K}\pi_k \tag{4.7}$$

We can get the expenditure function using equations (4.6) and (4.7).

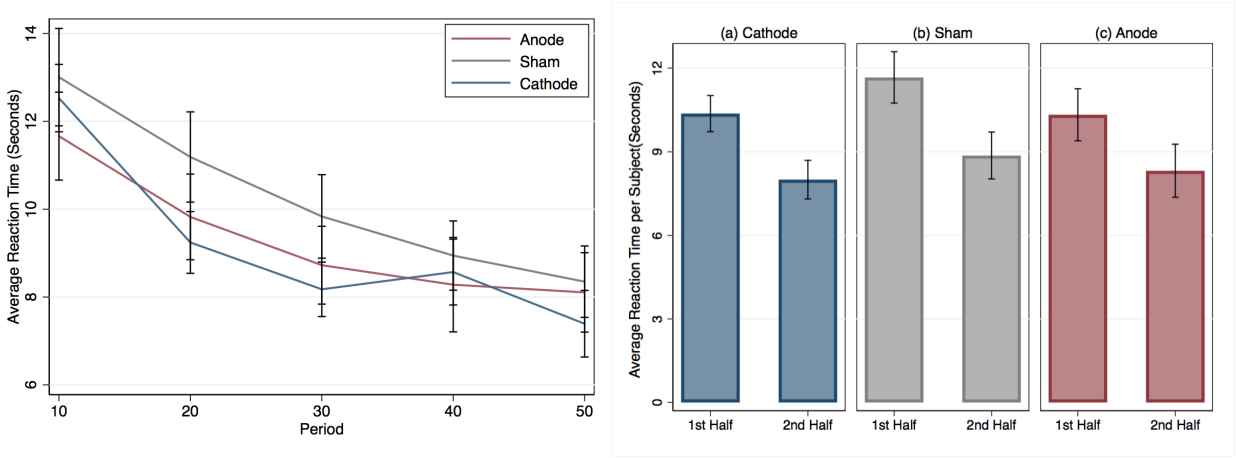
$$
\begin{aligned}
K\alpha\pi_k^{\rho-1} &= P(1-\alpha)\pi_p^{\rho-1} && \text{from (4.6)}\\
\left(\frac{\pi_p}{\pi_k}\right)^{\rho-1} &= \frac{K\alpha}{P(1-\alpha)}\\
\frac{\pi_p}{\pi_k} &= \left(\frac{K}{P}\cdot\frac{\alpha}{1-\alpha}\right)^{\frac{1}{\rho-1}}\\
\frac{P-\frac{P}{K}\pi_k}{\pi_k} &= \left(\frac{K}{P}\cdot\frac{\alpha}{1-\alpha}\right)^{\frac{1}{\rho-1}} && \text{from (4.7)}\\
\frac{P}{\pi_k}-\frac{P}{K} &= \left(\frac{K}{P}\right)^{\frac{1}{\rho-1}}\cdot\left(\frac{\alpha}{1-\alpha}\right)^{\frac{1}{\rho-1}}\\
\frac{P}{\pi_k} &= \frac{P}{K}+\left(\frac{K}{P}\right)^{\frac{1}{\rho-1}}\cdot\left(\frac{\alpha}{1-\alpha}\right)^{\frac{1}{\rho-1}}\\
\pi_k &= \frac{P}{\frac{P}{K}+\left(\frac{K}{P}\right)^{\frac{1}{\rho-1}}\cdot\left(\frac{\alpha}{1-\alpha}\right)^{\frac{1}{\rho-1}}}\\
\frac{\pi_k}{K} &= \frac{\frac{P}{K}}{\frac{P}{K}+\left(\frac{K}{P}\right)^{\frac{1}{\rho-1}}\cdot\left(\frac{\alpha}{1-\alpha}\right)^{\frac{1}{\rho-1}}}\\
&= \frac{\frac{P}{K}\cdot\left(\frac{\alpha}{1-\alpha}\right)^{\frac{-1}{\rho-1}}}{\frac{P}{K}\cdot\left(\frac{\alpha}{1-\alpha}\right)^{\frac{-1}{\rho-1}}+\left(\frac{K}{P}\right)^{\frac{1}{\rho-1}}}\\
&= \frac{\left(\frac{\alpha}{1-\alpha}\right)^{\frac{-1}{\rho-1}}}{\left(\frac{\alpha}{1-\alpha}\right)^{\frac{-1}{\rho-1}}+\left(\frac{K}{P}\right)^{\frac{1}{\rho-1}+1}}\\
&= \frac{\left(\frac{\alpha}{1-\alpha}\right)^{\frac{1}{1-\rho}}}{\left(\frac{\alpha}{1-\alpha}\right)^{\frac{1}{1-\rho}}+\left(\frac{K}{P}\right)^{\frac{\rho}{\rho-1}}} && (4.8)
\end{aligned}
$$

Equation (4.8) is the expenditure function and can be translated into our final structural model.

4.5.2 Reaction Time

Researchers typically use choice (e.g. charity donation behavior) data to reveal individuals' preferences; however, choice data carries no information on the processes of reaching to such decision. Reaction time data on the other hand is especially important to help understand individuals' decision processes.

Here, we present some results on the reaction time data.



A. Reaction Time over 10 Rounds **B.** Reaction Time over 25 Rounds

Supplementary Figure 4.5. Reaction Time Decreases Over Time. p-values in parentheses. Error bars denote standard error of the mean (SEM). For all three treatments, reaction time decreases over time. **A.B.** Reaction Time over 10 Rounds shows reaction time decreases over time. **ranksum test**: 1st half reaction time $>$ 2nd half. Anode (.0430), Sham (.0137), Cathode (.0119). Sham participants on average take longer time to make decisions. **ranksum test**: Anode $<$ Sham (.0001), Cathode $<$ Sham (.0003), Anode $\neq$ Cathode (.6675)

Supplementary Table 4.1 shows the determinants of individuals' reaction time, and the regressions helps to explain why participants in sham condition have longest reaction time. Monotonicity violations increase cathodal participants' reaction time but have no impact on sham or anodal participants. Violations of GARP do not affect reaction time in all three treatments. WARP violations decrease anodal participants' reaction time, but have no influence on cathodal and sham participants. Percept Kept influences reaction time through a quadratic form instead of a linear form, which means that in all three treatments, participants use more time to make a fair offer compared to generous or selfish offers. In general, sham participants take longer to make a fair offer compared to cathodal and anodal participants.

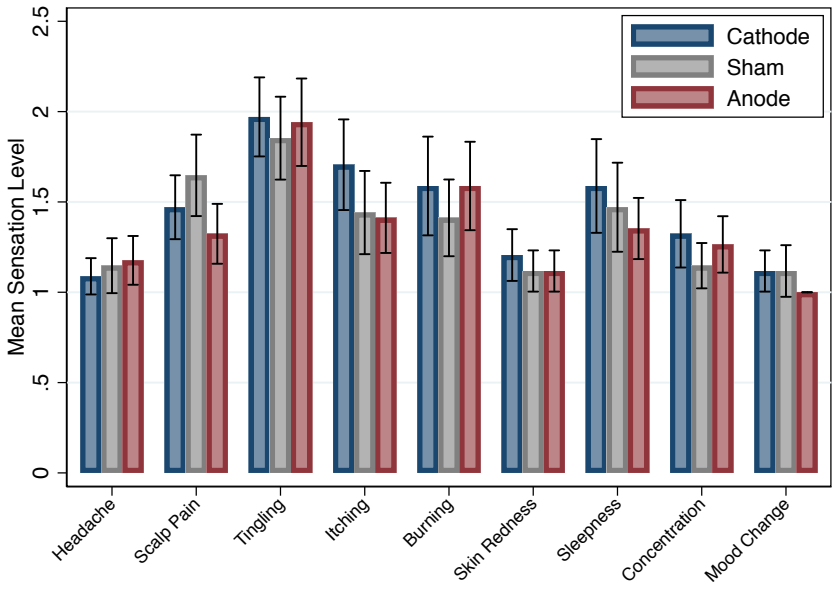
Table 4.1. Regressions – Determinants of Reaction Time

	Cathode coef / se	Sham coef / se	Anode coef / se
Round	-0.107***	-0.110***	-0.091***
	(0.015)	(0.020)	(0.015)
Budget Slope (degree)	-0.041***	-0.040***	-0.048***
	(0.010)	(0.010)	(0.010)
Percent Kept	8.779**	14.928***	9.149***
	(3.645)	(3.582)	(2.543)
Percent Kept Squared	-9.322***	-13.944***	-8.972***
	(3.419)	(3.185)	(2.341)
violation of Monotonicity	1.753**	0.122	0.185
	(0.828)	(0.815)	(0.899)
violation of WARP	0.416	-0.436	-0.568*
	(0.365)	(0.409)	(0.306)
violation of GARP	-0.173	0.336	0.033
	(0.474)	(0.503)	(0.383)
Constant	12.527***	12.124***	12.369***
	(1.054)	(1.020)	(1.213)
Observations	1700	1700	1700
BIC	10702.610	11108.933	10497.779
Subject controls	Yes	Yes	Yes

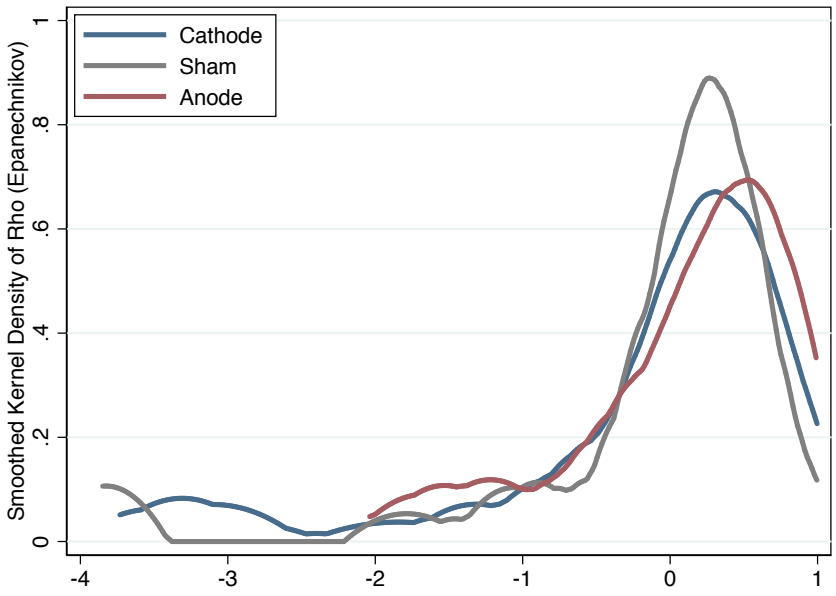
* $p < 0.1$, ** $p < 0.05$, *** $p < 0.01$. Standard errors in parentheses.

Note: We ran fixed effect regressions for participants's reaction time for in each treatment.

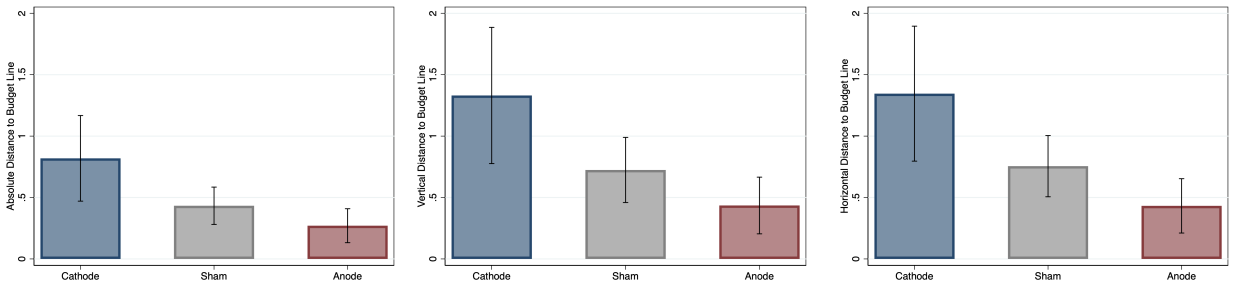
4.5.3 Supplementary Figures



Supplementary Figure 4.6. Participants' Self Reported tDCS Sensation. p-values $< .05$ are repaorted with Wilcoxon ranksum test in parentheses. Error bars denote standard error of the mean (SEM). Participants self-report tDCS sensation levels in the post-experiment survey. 1 = Absent, 2 = Mild, 3 = Moderate, and 4 = Severe. Sham participants report more Scalp Pain than anodal participants. A t-test of the difference in means (anodal vs. sham) rejects the null hypothesis of no difference (p = .0322). Anodal participants report less Mood Change than cathodal participants. A t-test of the difference in means (anodal vs. cathodal) rejects the null hypothesis of no difference (p = .0407). These two results are surprising due to the direction of the effect. For all the other ratings, participants report very similar sensation levels across three treatments.



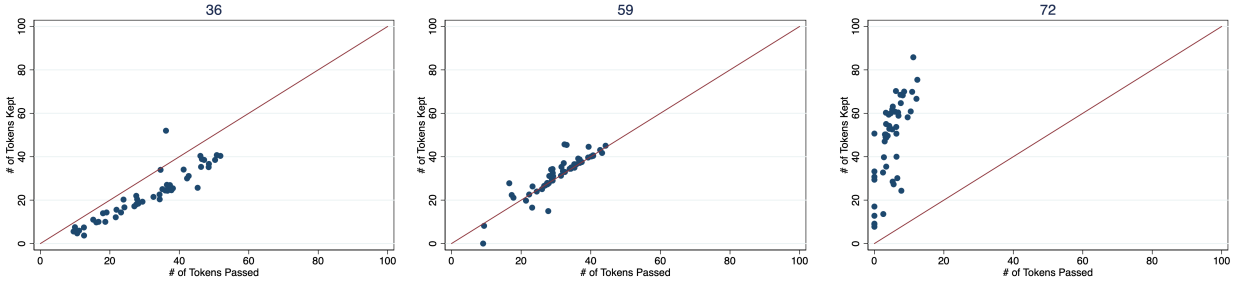
Supplementary Figure 4.7. Distribution of ρ. Individuals with $\rho < -9$ are dropped due to range limitation in the graph (4 participants, 2 cathode, 1 sham, and 1 anode), but all analyses include all participants. The distribution of ρ across treatments have different standard deviations. Anodal stimulation results in a concentration of *rho*; a test for equality of standard deviations (anode vs. cathode) rejects the null hypothesis of no difference ($p < .0001$).



A. Absolute Distance **B.** Vertical Distance **C.** Horizontal Distance

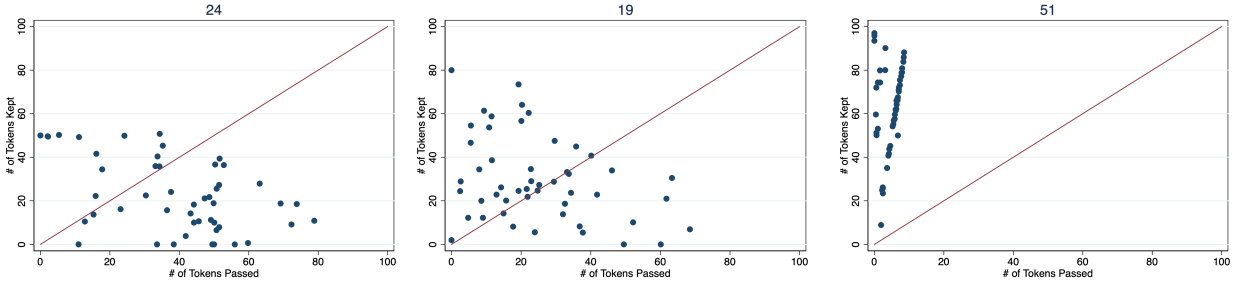
Supplementary Figure 4.8. Severity of Monotonicity Violations measured as absolute, vertical, and horizontal distance to the budget line. p-values in parentheses. Error bars denote standard error of the mean (SEM). **A.** Absolute distance to budget line is used to measure Monotonicity violation severity. Violations of Monotonicity in the anodal treatment is the least severe; a t-test of the difference in means (anodal vs. cathodal) rejects the null hypothesis of no difference ($p = .0742$). **B.** Vertical distance to budget line is the shortest in the anodal treatment; a t-test of the difference in means (anodal vs. cathodal) rejects the null hypothesis of no difference ($p = .0704$). **C.** Horizontal distance to budget line is the shortest in the anodal treatment; a t-test of the difference in means (anodal vs. cathodal) rejects the null hypothesis of no difference ($p = .0639$) **B.C.** We do not observe hemifield neglect caused by tDCS over rTPJ; a t-test of the difference in means (vertical distance vs. horizontal distance) cannot rejects the null hypothesis of no difference (anodal $p = .8714$, sham $p = .7722$, cathodal $p = .8421$).

4.5.4 Representative Sample Decisions



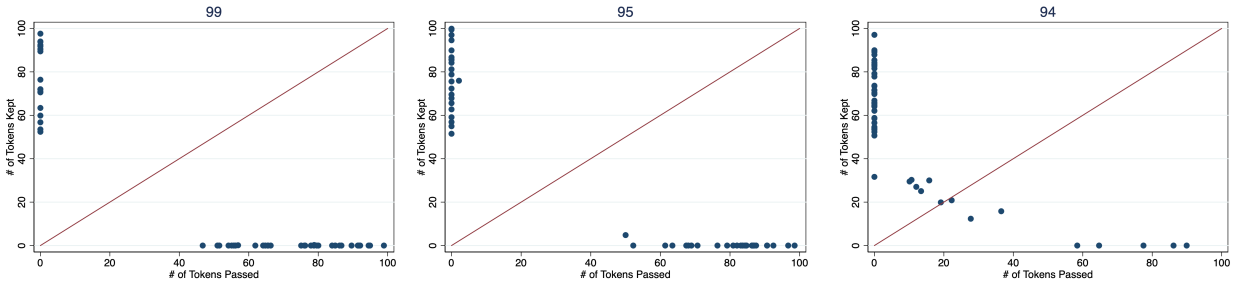
A. # 36, $\alpha = .026, \rho = -9.550$ **B.** # 59, $\alpha = .619, \rho = -11.666$ **C.** # 72, $\alpha = 1, \rho = -3.235$

Supplementary Figure 4.9. Rawlsians with $\rho \to -\infty$



A. # 24, $\alpha = .3537, \rho = .092$ **B.** # 19, $\alpha = .546, \rho = .043$ **C.** # 51, $\alpha = .931, \rho = -.047$

Supplementary Figure 4.10. Cobb-Douglas with $\rho \to 0$



A. # 99, $\alpha = .450, \rho = .9967$ **B.** # 95, $\alpha = .509, \rho = .991$ **C.** # 94, $\alpha = .726, \rho = .827$

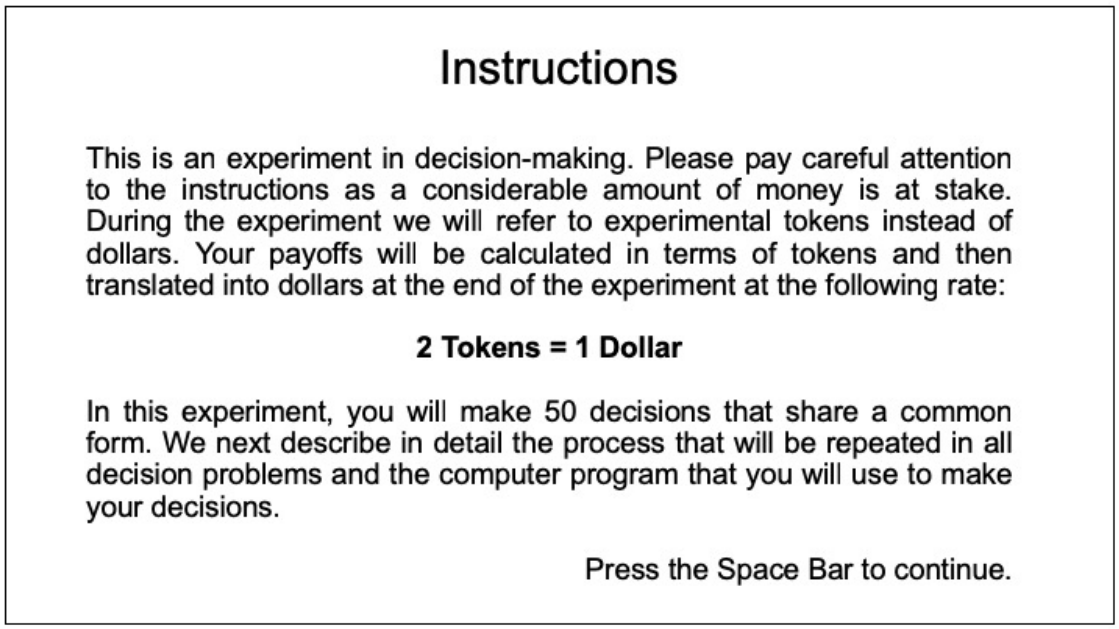
Supplementary Figure 4.11. Utilitarians with $\rho \to 1$

Supplementary Figures 4.9, 4.10, and 4.11 show 50 decisions from 9 participants. The red line represents an equal split of the tokens. Supplementary Figure 4.9 shows individuals with Rawlsian utility functions where $\rho \to -\infty$. These individuals have L-shaped or Leontief

utility. Supplementary Figure 4.10 shows individuals who have Cobb-Doulas utility functions, which are bowed in and smooth. Supplementary Figure 4.11 shows individuals with Utilitarian utility functions, which is represented with a straight line, therefore, those individuals prefer allocations where the budget line touches either of the axes. Panels labeled A show allocations from generous individuals, panels labeled B show allocations from individuals with strong preferences for fairness, and panels labeled C show allocations from selfish individuals.

4.5.5 Instructions

The original instructions have black backgrounds. For ease of reading, colors are inverted. The original budget lines were displayed in green, here they appear pink.



In each decision, you will be asked to allocate tokens between yourself and a local charity: Feeding America Southwest Virginia (FASWVA). The following 3 slides present some information about FASWVA from the charity's website.

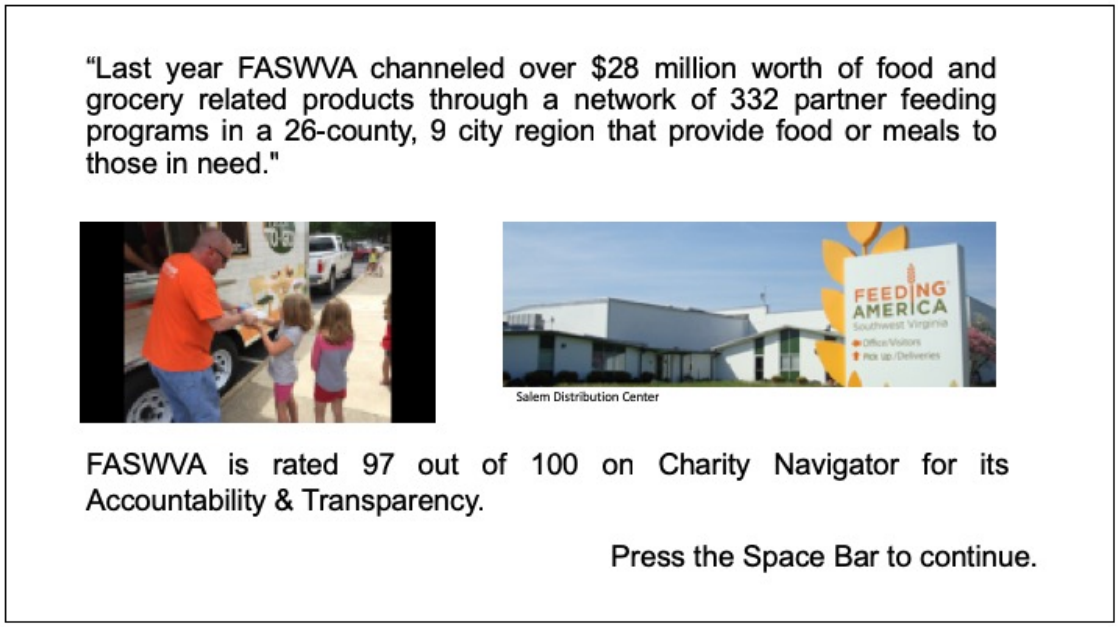
Press the Space Bar to continue.

FEEDING AMERICA
Southwest Virginia

"The primary function of FASWVA is to manage a Food Bank to secure large quantities of food for the hungry of Southwest Virginia. The Food Bank is an affiliate member of Feeding America and for the last three decades the Food Bank's ultimate mission has remained the same: eliminate hunger in the region."

Press the Space Bar to continue.


FEEDING AMERICA
Southwest Virginia
Serving 26 counties &10 cities in Southwest Virginia
SALEM DISTRIBUTION CENTER
Counties: Bedford • Botetourt • Carroll • Craig • Floyd • Franklin • Giles • Henry • Montgomery • Patrick • Pittsylvania • Pulaski • Roanoke • Wythe
Cities: Bedford • Danville • Martinsville • Radford • Roanoke • Salem
ABINGDON DISTRIBUTION CENTER
Counties: Bland • Buchanan • Dickenson • Grayson • Lee • Russell • Scott • Smyth • Tazewell • Washington • Wise
Cities: Galax • Norton • Bristol
ALLEGHANY HIGHLANDS DIRECT DISTRIBUTION PROGRAM
County: Alleghany
City: Covington
Alleghany
Craig
Botetourt
Giles
Roanoke
Bedford
Buchanan
Montgomery
Dickenson
Tazewell
Bland
Pulaski
Franklin
Wise
Russell
Wythe
Floyd
Pittsylvania
Smyth
Carroll
Lee
Scott
Washington
Grayson
Patrick
Henry
Press the Space Bar to continue.


"Last year FASWVA channeled over $28 million worth of food and grocery related products through a network of 332 partner feeding programs in a 26-county, 9 city region that provide food or meals to those in need."
FEEDING AMERICA
Southwest Virginia
Office/Visitors
Pick up /Deliveries
Salem Distribution Center
FASWVA is rated 97 out of 100 on Charity Navigator for its Accountability & Transparency.
Press the Space Bar to continue.

We will refer to the tokens that you allocate to yourself as tokens that you **Keep**, and tokens that you allocate to the charity as tokens that you **Pass** to the charity.

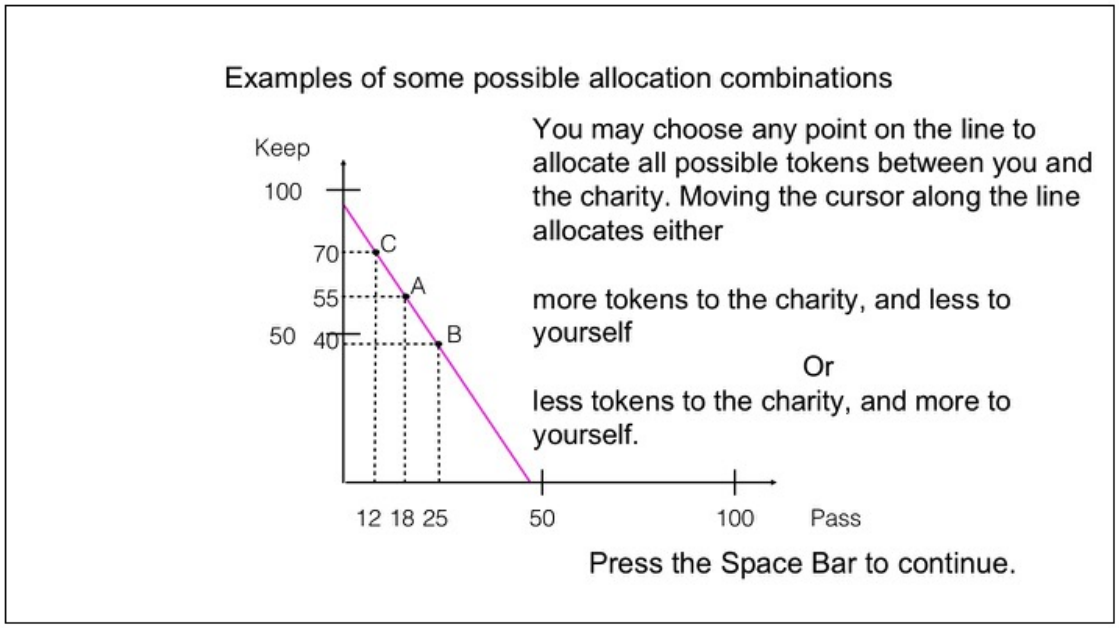
Each decision will involve choosing a point on or under a line representing possible token allocations to you (**Keep**) and the charity you chose (**Pass**). In each decision, you may choose any combination of tokens to Keep and Pass – in other words, any combination of tokens to yourself and tokens to the charity – that is on or under the line.

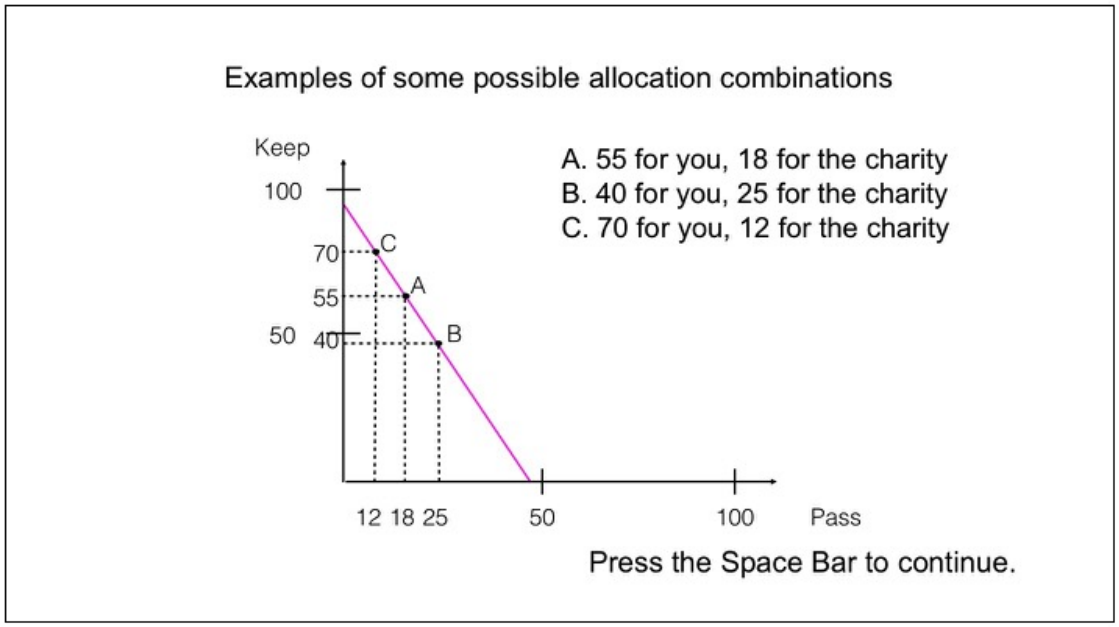
Press the Space Bar to continue.

In each graph, **Keep** corresponds to the vertical axis and **Pass** corresponds to the horizontal axis; the points on or under the diagonal lines in the graphs represent possible token allocations to **Keep** (tokens you to you) and **Pass** (tokens you to a charity) that you might choose.

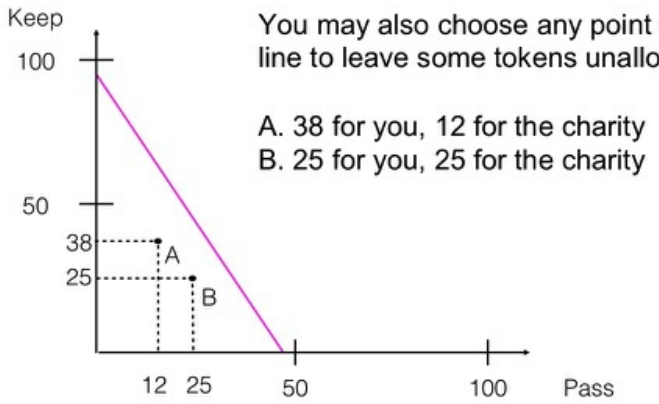
Choosing a point on the diagonal lines means that you allocate all possible tokens between you and the charity. However, by choosing a point under the diagonal lines, you do not only allocate tokens between you and a charity, but also leave some tokens unallocated. You will see an example on the next page.

Press the Space Bar to continue.


Examples of some possible allocation combinations
Keep
100
70
55
50
40
C
A
B
12 18 25
50
100
Pass
You may choose any point on the line to allocate all possible tokens between you and the charity. Moving the cursor along the line allocates either
more tokens to the charity, and less to yourself
Or
less tokens to the charity, and more to yourself.
Press the Space Bar to continue.


Examples of some possible allocation combinations
Keep
100
70
55
50
40
C
A
B
12 18 25
50
100
Pass
A. 55 for you, 18 for the charity
B. 40 for you, 25 for the charity
C. 70 for you, 12 for the charity
Press the Space Bar to continue.

Examples of some possible allocation combinations

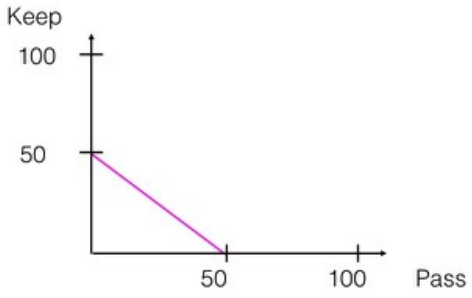


You may also choose any point under the line to leave some tokens unallocated.

A. 38 for you, 12 for the charity
B. 25 for you, 25 for the charity

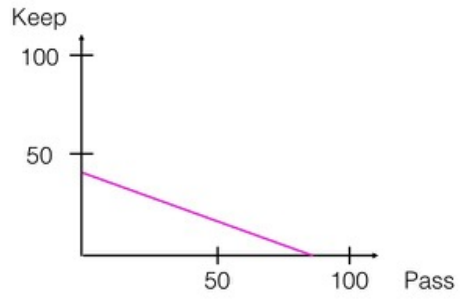
Press the Space Bar to continue.

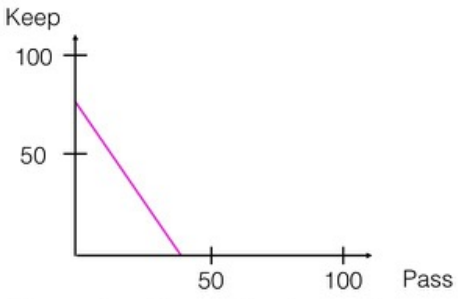
If, for example, the diagonal line runs from 50 tokens on the **Keep** axis to 50 tokens on the **Pass** axis, you could choose to keep all 50 tokens for yourself, or pass all 50 tokens to the other person, or anything in between.



Press the Space Bar to continue.

Most of the decision problems will involve flatter or steeper lines: if the line is flatter, one less token for yourself means more than one additional token is passed to the other person; if the line is steeper, one less token kept means less than one additional token passed to the other person.





Press the Space Bar to continue.

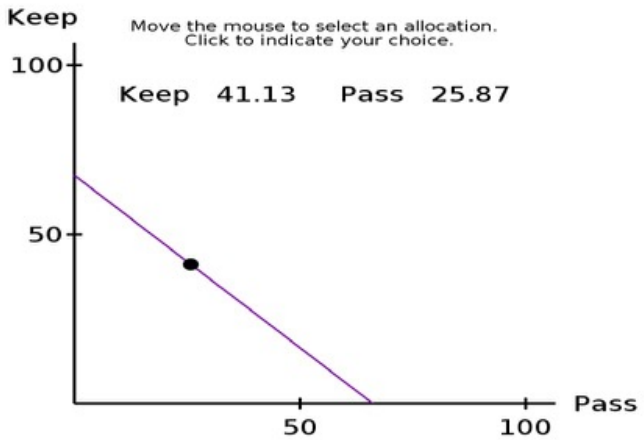
After each decision is made, a feedback page will show you exactly how much you chose to Keep and how you choose to Pass. The feedback page will be shown for up to 10 seconds, or until you press the Space Bar.

Press the Space Bar to continue.

Each of the 50 decision problems will start by having the computer select a diagonal line at random. All of the lines that the computer will select will intersect with at least one of the axes at 50 or more tokens, but will not intersect either axis at more than 100 tokens. The lines selected for you in different decision problems are independent of each other and depend solely upon chance.

Press the Space Bar to continue.

This is exactly what a decision trial looks like.



Press the Space Bar to continue.

At the end of the experiment, you will be privately paid with cash for one randomly selected trial from all 50 trials. For the paid trial, the experimenter will transfer the donation you made to the charity under your supervision.

Press the Space Bar to continue.

This is the end of instructions. Next, you are going to have 3 practice trials. Practice trials look exactly the same as the real experiment. Practice trials will help you getting familiar with the experiment. Practice trials do not count for your payment.

If you are ready for the practice trials,
Press the Space Bar to continue.

Bibliography

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

Alger, I. and Weibull, J. W. (2013). Homo moralis—preference evolution under incomplete information and assortative matching. *Econometrica*, 81(6):2269–2302.

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

Andreoni, J. and Miller, J. (2002). Giving according to garp: An experimental test of the consistency of preferences for altruism. *Econometrica*, 70(2):737–753.

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.

Avoyan, A. and Ramos, J. (2017). A road to efficiency through communication and commitment.

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

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

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

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

Battigalli, P., Dufwenberg, M., and Smith, A. (2018). Frustration and anger in games.

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

Bikson, M., Datta, A., and Elwassif, M. (2009). Establishing safety limits for transcranial direct current stimulation. *Clinical neurophysiology: official journal of the International Federation of Clinical Neurophysiology*, 120(6):1033.

Binder, J. R., Desai, R. H., Graves, W. W., and Conant, L. L. (2009). Where is the semantic system? a critical review and meta-analysis of 120 functional neuroimaging studies. *Cerebral Cortex*, 19(12):2767–2796.

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

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

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

Bowles, S. (2006). Group competition, reproductive leveling, and the evolution of human altruism. *Science*, 314(5805):1569–1572.

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

Brainard, D. H. (1997). The psychophysics toolbox. *Spatial vision*, 10:433–436.

Brandts, J. and Charness, G. (2003). Truth or consequences: An experiment. *Management Science*, 49(1):116–130.

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.

Camerer, C. F. (2003). *Behavioral game theory, experiments in strategic interaction*. Princeton University Press.

Caplin, A. and Dean, M. (2015). Revealed preference, rational inattention, and costly information acquisition. *American Economic Review*, 105(7):2183–2203.

Cattaneo, Z., Mattavelli, G., Platania, E., and Papagno, C. (2011). The role of the prefrontal cortex in controlling gender-stereotypical associations: a tms investigation. *NeuroImage*, 56(3):1839–1846.

Cettolin, E., Dalton, P. S., Kop, W., and Zhang, W. (2018). Cortisol meets garp: The effect of stress on economic rationality.

Chang, L. J. and Sanfey, A. G. (2011). Great expectations: neural computations underlying the use of social norms in decision-making. *Social Cognitive & Affective Neuroscience*, 8(3):277–284.

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

Charness, G. and Dufwenberg, M. (2011). Participation. *The American Economic Review*, 101(4):1211–1237.

Choi, S., Kariv, S., Müller, W., and Silverman, D. (2014). Who is (more) rational? *American Economic Review*, 104(6):1518–1550.

Corbetta, M., Patel, G., and Shulman, G. L. (2008). The reorienting system of the human brain: from environment to theory of mind. *Neuron*, 58(3):306–324.

Corbetta, M. and Shulman, G. L. (2002). Control of goal-directed and stimulus-driven attention in the brain. *Nature Reviews Neuroscience*, 3(3):201.

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.

Donaldson, P. H., Rinehart, N. J., and Enticott, P. G. (2015). Noninvasive stimulation of the temporoparietal junction: A systematic review. *Neuroscience & Biobehavioral Reviews*, 55:547–572.

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

Dufwenberg, M., Li, F., and Smith, A. (2018b). Threats.

Dufwenberg, M., Servátka, M., and Vadovič, R. (2017). Honesty and informal agreements. *Games and Economic Behavior*, 102:269–285.

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

Eckel, C. C. and Grossman, P. J. (1996). Altruism in anonymous dictator games. *Games and Economic Behavior*, 16(2):181–191.

Eisenkopf, G., Gurtoviy, R., and Utikal, V. (2017). Punishment motives for small and big lies. *Journal of Economics & Management Strategy*.

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*.

Farrell, J. (1987). Cheap talk, coordination, and entry. *RAND Journal of Economics*, 18(1):34–39.

Fehr, D. and Sutter, M. (2016). Gossip and the efficiency of interactions. *IZA Discussion Paper No. 9704*.

Fehr, E. and Fischbacher, U. (2003). The nature of human altruism. *Nature*, 425(6960):785.

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

Fehr, E. and Schmidt, K. M. (2006). Fehr, ernst, and klaus m. schmidt. "the economics of fairness, reciprocity and altruism–experimental evidence and new theories. *Handbook of the economics of giving, altruism and reciprocity*, 1:615–691.

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

Fisman, R., Jakiela, P., and Kariv, S. (2017). Distributional preferences and political behavior. *Journal of Public Economics*, 155:1–10.

Fisman, R., Kariv, S., and Markovits, D. (2007). Individual preferences for giving. *American Economic Review*, 97(5):1858–1876.

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.

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

Gneezy, U. (2005). Deception: The role of consequences. *The American Economic Review*, 95(1):384–394.

Gohil, K., Hahne, A., and Beste, C. (2016). Improvements of sensorimotor processes during action cascading associated with changes in sensory processing architecture–insights from sensory deprivation. *Scientific Reports*, 6:28256.

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.

Güroğlu, B., van den Bos, W., van Dijk, E., Rombouts, S. A., and Crone, E. A. (2011). Dissociable brain networks involved in development of fairness considerations: understanding intentionality behind unfairness. *NeuroImage*, 57(2):634–641.

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.

Hamilton, W. D. (1964). The genetical evolution of social behaviour. ii. *Journal of theoretical biology*, 7(1):17–52.

Harbaugh, W. T., Mayr, U., and Burghart, D. R. (2007). Neural responses to taxation and voluntary giving reveal motives for charitable donations. *Science*, 316(5831):1622–1625.

Haruvy, E., Katok, E., Ma, Z., and Sethi, S. (2018). Haruvy, ernan, elena katok, zhongwen ma, and suresh sethi. "relationship-specific investment and hold-up problems in supply chains: theory and experiments. *Business Research*, pages 1–30.

Hoppe, E. I. and Schmitz, P. W. (2011). Can contracts solve the hold-up problem? experimental evidence. *Games and Economic Behavior*, 73(1):186–199.

Hutcherson, C. A., Bushong, B., and Rangel, A. (2015a). A neurocomputational model of altruistic choice and its implications. *Neuron*, 87(2):451–462.

Hutcherson, C. A., Seppala, E. M., and Gross, J. J. (2015b). The neural correlates of social connection. *Cognitive, Affective, & Behavioral Neuroscience*, 15(1):1–14.

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

Kleiner, M., Brainard, D., Pelli, D., Ingling, A., Murray, R., and Broussard, C. (2007). What's new in psychtoolbox-3. *Perception*, 36 ECVP Abstract Supplement 14.

Krall, S. C., Rottschy, C., Oberwelland, E., Bzdok, D., Fox, P. T., Eickhoff, S. B., Fink, G. R., and Konrad, K. (2015). The role of the right temporoparietal junction in attention and social interaction as revealed by ale meta-analysis. *Brain Structure and Function*, 220(2):587–604.

Li, F., Ball, S., Katz, B., and Smith, A. (2018). Case report of syncope during a transcranial direct current stimulation experiment in a healthy adult participant. *Brain Stimulation: Basic, Translational, and Clinical Research in Neuromodulation*, 11(5):1201–1202.

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

Marini, M., Banaji, M. R., and Pascual-Leone, A. (2018). Studying implicit social cognition with noninvasive brain stimulation. *Trends in cognitive sciences*, 22(11):1050–1066.

Masatlioglu, Y., Nakajima, D., and Ozbay, E. Y. (2012). Revealed attention. *American Economic Review*, 102(5):2183–2205.

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

McCutcheon, B. (1997). Do meetings in smoke-filled rooms facilitate collusion? *Journal of Political Economy*, 105(2):330–350.

Morishima, Y., Schunk, D., Bruhin, A., Ruff, C. C., and Fehr, E. (2012). Linking brain structure and activation in temporoparietal junction to explain the neurobiology of human altruism. *Neuron*, 75(1):73–79.

Nelson Jr, W. R. (2004). Equity or intention: it is the thought that counts. *Journal of Economic Behavior & Organization*, 48(4):423–430.

Pelli, D. G. (1997). The videotoolbox software for visual psychophysics: Transforming numbers into movies. *Spatial vision*, 10:437–442.

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

Poreisz, C., Boros, K., Antal, A., and Paulus, W. (2007). Safety aspects of transcranial direct current stimulation concerning healthy subjects and patients. *Brain research bulletin*, 72(4-6):208–214.

Ptak, R. and Schnider, A. (2011). The attention network of the human brain: relating structural damage associated with spatial neglect to functional imaging correlates of spatial attention. *Neuropsychologia*, 49(11):3063–3070.

Rabin, M. (1993). Incorporating fairness into game theory and economics. *American Economic Review*, pages 1281–1302.

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

Rotemberg, J. J. (2006). Altruism, reciprocity and cooperation in the workplace. *Handbook of the economics of giving, altruism and reciprocity*, 2:1371–1407.

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

Sánchez-Pagés, S. and Vorsatz, M. (2007). An experimental study of truth-telling in a sender-receiver game. *Games and Economic Behavior*, 61(1):86–112.

Sánchez-Pagés, S. and Vorsatz, M. (2009). Enjoy the silence: an experiment on truth-telling. *Experimental Economics*, 12(2):220–241.

Santiesteban, I., Banissy, M. J., Catmur, C., and Bird, G. (2012). Enhancing social ability by stimulating right temporoparietal junction. *Current Biology*, 22(23):2274–2277.

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.

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

Sellaro, R., Derks, B., Nitsche, M. A., Hommel, B., van den Wildenberg, W. P., van Dam, K., and Colzato, L. S. (2015). Reducing prejudice through brain stimulation. *Brain Stimulation*, 8(5):891–897.

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

Singer, T., ben Seymour, O'doherty, J. P., Stephan, K. E., Dolan, R. J., and Frith, C. D. (2006). Empathic neural responses are modulated by the perceived fairness of others. *Nature*, 439(7075):466.

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

Sutter, M. (2007). Outcomes versus intentions: On the nature of fair behavior and its development with age. *Journal of Economic Psychology*, 28(1):69–78.

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

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

Tusche, A., Böckler, A., Kanske, P., Trautwein, F.-M., and Singer, T. (2016). Decoding the charitable brain: empathy, perspective taking, and attention shifts differentially predict altruistic giving. *Journal of Neuroscience*, 36(17):4719–4732.

van Elk, M., Duizer, M., Sligte, I., and van Schie, H. (2017). Transcranial direct current stimulation of the right temporoparietal junction impairs third-person perspective taking. *Cognitive, Affective, & Behavioral Neuroscience*, 17(1):9–23.

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

Villamar, M. F., Volz, M. S., Bikson, M., Datta, A., DaSilva, A. F., and Fregni, F. (2013). Technique and considerations in the use of 4x1 ring high-definition transcranial direct current stimulation (hd-tdcs). *Journal of visualized experiments: JoVE*, 77.

Xiao, E. and Bicchieri, C. (2010). When equality trumps reciprocity. *Journal of Economic Psychology*, 31(3):456–470.

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

Young, L., Camprodon, J. A., Hauser, M., Pascual-Leone, A., and Saxe, R. (2010). Disruption of the right temporoparietal junction with transcranial magnetic stimulation reduces the role of beliefs in moral judgments. *Proceedings of the National Academy of Sciences*, 107(15):6753–6758.