Three Essays in Experimental Economics

Austin Edward Bradley

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

Alec Smith, Chair
Sheryl B. Ball
Bradley S. Paye
Ali Habibnia

April 24, 2024
Blacksburg, Virginia

Keywords: Economics, Experimental Economics, Behavioral Economics, Subjective Probability, Gambler's Fallacy, Hot Hand Fallacy, Cheap Talk, Communication, Natural Language Processing, Free-form Communication, Real-Effort, Trust

Three Essays in Experimental Economics

Austin Edward Bradley

(ABSTRACT)

The experiments presented and analysed in this dissertation concern two well-established phenomena in behavioral economics: that human decision makers hold biased beliefs about probability and that free-form communication between economic agents promotes cooperation far in excess of what standard theory predicts. First, Chapter 2 studies subjective probability, focusing on the well-established existence of both the Hot Hand and Gambler's Fallacies — the false expectation of positive and negative autocorrelation, respectively. Both biases are prevalent throughout a wide variety of real-world contexts; what causes a person to favor one over the other? We conduct an experiment in which we observe fully informed subjects switching between the Hot Hand and Gambler's Fallacies when predicting future outcomes of mathematically identical sequences. Subjects exhibit the Gambler's Fallacy when predicting single outcomes but favor the Hot Hand when asked explicitly to estimate probabilities. Connecting our results to existing theory suggests that very subtle changes in framing lead decision makers to employ substantially different approaches to form predictions.

The remainder of this dissertation studies cheap talk communication between human subjects playing incentivised trust games. In Chapter 3, we study free-form communication using a dataset of over 1000 messages sent between participants in a laboratory trust game. We employ Natural Language Processing to systematically generate meaningful partitions of the messages space which we can then examine with established regression approaches. Our investigation reveals features correlated with trust that have not previously been considered. Most notably, highly detailed, specific promises establish trust more effectively than other

messages which signal the same intended action. Additionally, we observe that the most and least trusted messages in our dataset differ starkly in their quality. Highly trusted messages are longer, more detailed, and contain fewer grammatical errors whereas the least trusted messages tend to be brief and prone to errors.

In Chapter 4, we examine whether the difference is message quality affects trust by acting as a signal of effort. We report the results of an experiment designed to test whether promises which require higher levels of effort result in greater trust from their recipients. We find that more costly promises lead recipients to trust more frequently. However, there is no corresponding, significant difference in the trustworthiness of their senders. Further, when asked their beliefs explicitly, recipients do not believe that higher cost promises are more likely to be trustworthy. This presents a potential challenge to our understanding of trust between economic decision makers. If effort increases trust without altering receivers' beliefs, receivers must be concerned with factors other than their own payoff maximization. We conclude by presenting a followup experiment where varying effort cost cannot convey the sender's intentions, however, the results are inconclusive.

Three Essays in Experimental Economics

Austin Edward Bradley

(GENERAL AUDIENCE ABSTRACT)

This dissertation presents three projects in which we examine how human decision makers' choices differ from those predicted by standard economic theory. The experiments we conduct cover two broad topics: the way humans estimate the probability of random events and how communication leads to greater cooperation between agents with potentially conflicting monetary interests. It is well established that humans often hold distorted beliefs about probability. Depending on the direction of their bias, these beliefs are consistent with either the Hot Hand or Gambler's Fallacy. In Chapter 2, we examine the factors which may cause people to change the direction of their bias. Subjects exhibit the Gambler's Fallacy when predicting single outcomes, but favor the Hot Hand when asked explicitly to estimate probabilities. Chapters 3 and 4 study cheap talk communication between decision makers — messages which carry with them no commitment mechanism. It is no surprise to the average person that communication may enhance cooperation and trust between people. Experimental economists have verified this intuition in laboratory experiments and found that free-form communication is particularly effective. However, the precise mechanism through which free-form communication enhances cooperation is unclear. In Chapter 3, we collect a large dataset of free-form messages transmitted between players of an investment game. We then employ Natural Language Processing tools, novel to the Economics laboratory, to parse the unstructured data and identify message features associated with changes in trust and trustworthiness. Chapter 4 continues to examine communication, investigating whether the effort required to a promise affects its perceived or actual trustworthiness. We

find that higher effort promises lead to greater trust, but find no corresponding increase in trustworthiness.

Dedication

"Three Essays in Experimental Economics" is dedicated to my parents, Brian and Helen Bradley; my sister, Lyra; my dear friends, Paul Gemperlein, Sydney Gemperlein, Ben Grove, Abby Miles-Grove, Dan Grove, Nathanael Nulph, Monica Simon, Eric Svitek, Elise Green, and Francesca Ferrante; and finally, my loving girlfriend, Dr. Diana Burden.

Acknowledgments

I would like to take this space to briefly acknowledge those without whom I could not have completed this dissertation. First, I thank my advisor Alec Smith and committee members Sheryl Ball, Brad Paye, and Ali Habibnia for their guidance throughout my years at Virginia Tech. This work also benefited greatly from comments and feedback received while presenting at conferences of the Economic Science Association and Western Economic Association and at seminars held at Virginia Tech and the University of Pisa. Undergraduate research assistants Clay Baldwin, Rachel Casem, Sania Anam Mahmood, and Sydney McCarthy supported this work by manually classifying message data used in Chapter 3. Finally, I would like to thank and acknowledge the time and efforts of the hundreds of Amazon Mechanical Turk workers who volunteered to participate in the experiments that follow.

Contents

List of Figures xi

List of Tables xiii

1 Introduction 1

2 Factors Affecting Perception of Random Events 4

2.1 Introduction . . . 5

2.2 Theory: The Rabin-Vayanos Model . . . 8

2.3 Hypotheses . . . 13

2.4 Methods . . . 14

2.5 Results . . . 15

2.5.1 Descriptive Statistics . . . 15

2.5.2 BMR and FN Comparison . . . 17

2.5.3 Indicator Variable Regressions . . . 19

2.5.4 Grouped Flips . . . 23

2.5.5 Visual Presentation Comparison . . . 25

2.5.6 Alternative Features . . . 27

2.6 Conclusion . . . 30

3 Investigating Trust Using Natural Language Processing 32

3.1 Introduction . . . 33
3.1.1 Background . . . 35
3.2 Methods . . . 38
3.3 Data . . . 40
3.3.1 Message Overview . . . 40
3.3.2 Tokenization and Lemmatization . . . 42
3.4 Results . . . 44
3.4.1 Baseline Approach: Literal Meaning . . . 44
3.4.2 Data Validity: Order Effects . . . 45
3.4.3 Regressions on Tokens . . . 46
3.4.4 Cluster Analysis . . . 49
3.4.5 Results: Affinity Propagation Using Corpus-only Similarity Measure 51
3.4.6 Affinity Propagation Using Knowledge-based Similarity . . . 59
3.4.7 Reducing Cluster Size . . . 65
3.4.8 Reduced Corpus-only Clusters . . . 66
3.4.9 Reduced word2vec Clusters . . . 68
3.4.10 Sentiment . . . 70
3.5 Conclusion . . . 72

4 Effort Cost and Communication 76

4.1 Introduction . . . 77
4.1.1 Background . . . 78

4.1.2 Hypothesis . . . 83
4.2 Experiment 1 . . . 84
4.2.1 Experiment 1: Methods . . . 84
4.2.2 Experiment 1: Results . . . 86
4.2.3 Experiment 1: Discussion . . . 90
4.3 Experiment 2 . . . 92
4.3.1 Experiment 2: Methods . . . 92
4.3.2 Experiment 2: Hypotheses . . . 93
4.3.3 Experiment 2: Results . . . 94
4.3.4 Experiment 2: Discussion . . . 98
4.4 Conclusion . . . 99

Bibliography **102**

Appendices **112**

Appendix A Appendix for Chapter 3 **113**

A.1 Affinity Propagation Algorithm Description . . . 113
A.2 Full Results from Table 3.4: Bag-of-words Model . . . 115

Appendix B Appendix for Chapter 4 **116**

B.1 Pairwise Comparisons for Experiment 1 . . . 116
B.2 Communication Mechanism . . . 117

List of Figures

2.1 RV model simulations 12

2.2 Distribution of Subject level Estimate Variance 17

3.1 Basic Trust Game 35

3.2 MTurk Game Overview 39

3.3 Message Length Distribution 41

3.4 Order Effects 45

3.5 Distribution of Cluster Sizes: Jaccard Distance 52

3.6 Distribution of Cluster Sizes using MCS-w2v Distance 60

4.1 Game tree of a simple trust game. 79

4.2 Simple Trust Game with Message Cost 80

4.3 Trust game with payoffs used in experiment 1 85

4.4 Receiver Choices by Required Effort Level 87

4.5 Sender Actions by Required Effort Level 88

4.6 First Order Receiver Beliefs Following a Promise 89

4.7 Game trees with and without random choice as seen by Receivers 92

(a) Without random Sender choice 92

(b) With random Sender choice 92

4.8 Sender Choices by Required Effort 94

4.9 Receiver Responses when Sender Choices are Random 95

4.10 Receiver Response when Senders Retain Control 96

4.11 Receiver Choices Following a Message 97

4.12 Receiver Choices Following No Message 97

B.1 Senders' interface in both Experiment 1 and 2 117

List of Tables

2.1 Mean Duration by Treatment and Order 16

2.2 Comparison to Frydman-Nave Econ. Decision Task 18

2.3 Comparison to BMR Result 18

2.4 Predictions vs. Indicator Variables by Position 20

2.5 Predictions vs. Indicator Variables by Position with Streak Control 21

2.6 Predictions vs. Grouped Flips 23

2.7 Predictions vs. Grouped Flips with Streak Control 24

2.8 Predictions vs. Grouped Flips with Graph Interactions 25

2.9 Predictions vs. Grouped Flips with Graph Interactions with Streak Control . 26

2.10 Predictions vs. Common Features 28

3.1 Example Messages: Most and Least Effective 40

3.2 Most Common Tokens 42

3.3 Regressions on Baseline Messages 44

3.4 Bag-of-words Model 46

3.5 Example Clusters 52

3.6 Most Trusted Corpus-only Clusters 54

3.7 Least Trusted Corpus-only Clusters 55

3.8 Significant Corpus-only Clusters: Share 56

3.9 Example MCS-word2vec Clusters 62

3.10 Least Trusted MCS-word2vec Clusters 63

3.11 Most Trusted MCS-word2vec Clusters 63

3.12 MCS-word2vec Clusters: Share 65

3.13 Reduced Size Corpus-only Clusters 67

3.14 Reduced MCS-word2vec Clusters 69

3.15 Regressions on Sentiment 71

4.1 Regression Estimates: Choices on Effort 88

A.1 Bag of Words Model: Full 115

B.1 Receiver comparison without message 116

B.2 Receiver comparison with message 116

List of Abbreviations

BMR Benjamin, Moore, & Rabin (2017)

FN Frydman & Nave (2017)

IID Independent and Identically Distributed

MCS Mihalcea, Corley, & Strapparava, (2006)

MTurk Amazon Mechanical Turk

NLP Natural Language Processing

w2v word2vec

Chapter 1

Introduction

Three Essays in Experimental Economics presents the results of several experiments examining common cases where human decision makers differ from the agents who populate longstanding models of economic behavior. In total, the datasets analysed in this dissertation provide a tiny glimpse into the thoughts and actions of over 1500 unique individuals. We study deviations from the assumption of full rationality by examining how the people in our sample form subjective probability estimates in Chapter 2. By studying cheap talk communication in Chapters 3 and 4, we examine how social considerations may cause decision makers to deviate from payoff maximizing self-interest. In furthering our understanding of the decisions of these 1500 people, we refine models of economic behavior to better approximate reality.

In Chapter 2 we investigate two common, opposing probability biases — the Hot Hand and Gambler's fallacies. The Gambler's Fallacy predicts that people incorrectly expect independent and identically distributed (IID) sequences to be negatively correlated while the Hot Hand Fallacy predicts incorrect belief in positive correlation. Both of these patterns are well-established in previous literature, which raises the natural question of what causes one to prevail over the other. We conduct an incentivized experiment in which subjects predict the next outcome of a series of IID coin tosses. When asked to place wagers on individual, binary outcomes, subjects favor the Gambler's Fallacy, on average. When the same subjects, observing identical sequences, report their predictions as frequencies with incentives for accuracy, they exhibit the Hot Hand Fallacy. Our results suggest that decision makers use

a substantially different approach to predict individual outcomes compared to estimating probabilities. Connecting our results to existing theory suggests that the Gambler's Fallacy implicitly affects both forms of forecasting but only explicitly affects predictions of individual outcomes.

Chapter 3 and Chapter 4 examine the effect cheap talk communication has on trust between people. It is well established that communication matters to economic decision makers, promoting cooperation in excess of what is predicted based solely on the players' monetary incentives. Free-form communication is particularly effective compared to messages chosen from a prewritten list. However, the addition of free-form communication also vastly increases the complexity of the resulting data, making analysis difficult. Existing approaches for studying free-form communication rely on manually tagging messages based on ex ante hypotheses about communication mechanisms.

In Chapter 3, we employ Natural Language Processing tools to systematically parse unstructured communication data. In the context of a principal-agent trust game, we collect a large dataset of free-form messages, linked to choices made by their senders and their recipients. We use Natural Language Processing tools to partition the message space with fewer ex ante assumptions about which features may be important. Our analysis enables us to identify numerous, previously unexamined features which correlate to significant differences in trust and trustworthiness. Notable among the features we identify are emphatic, detailed promises, which increase trust more than other messages stating the same intent. Further, implied promises which allude to cooperation without explicitly stating an intended action increase trust.

In addition to the features identified with Natural Language Processing, our game structure enables us to make additional observations about effective communication. Messages in our free-form text data are sent to 10 recipients. This novel feature allows us to identify the

most and least effective messages [1] (those which are trusted most/least frequently by their recipients). Examining the two groups of messages side-by-side reveals stark qualitative differences. Those which are most effective are long, detailed, and grammatically correct whereas the least effective are brief and often contain grammatical or factual errors. Although there are numerous features which may plausibly separate the messages into these groups, we select the effort required to compose each message as the most obvious differentiating factor. This observation motivates Chapter 4 of this dissertation which examines the relationship between real-effort cost, communication, and trust.

Chapter 4 presents the results of two experiments which use the same trust game used in Chapter 3 but with a modified communication mechanism. Free-form text is replaced by a single fixed message promising to split the surplus evenly. Sending the message requires varying amounts of real-effort by positioning sliders to target values. This mechanism allows us to hold the message content constant while varying effort cost. We examine the impact of real-effort cost on both recipients' trust and senders' trustworthiness. Our results show that trust increases with effort but we find no corresponding evidence that effort increases trustworthiness or recipients' beliefs about trustworthiness. Together, these three relationships suggest that 'trust' may be motivated by factors other than the trustor's own payoff maximization.

We conduct a follow-up experiment to test our interpretation directly. Using the same communication mechanism, we introduce the possibility that senders' decisions are overridden and made randomly, removing the possibility that a message's effort cost can convey any payoff-relevant information. The results of this experiment are inconclusive. We do not find a significant effort effect within the treatment group, but also fail to replicate well-established results within the control group.

[1] A random sample of messages in each group are shown in Table 3.1

Chapter 2

Factors Affecting Perception of Random Events

Abstract

The Hot Hand and Gambler's fallacies are both well established biases in probabilistic reasoning, referring to incorrect expectations of positive and negative auto-correlation between outcomes. If both of these opposing patterns of bias arise from the same underlying error in reasoning, what causes one to take precedence over the other? In existing experimental and empirical work, the variety of methods and designs makes it difficult to compare results across studies. In this chapter, we report the results of an experiment in which we ask subjects to predict coin tosses after observing previous outcomes, varying both the question used to elicit their predictions and the visual presentation of the past outcomes. When asked to predict individual outcomes, placing wagers on a single result, subjects act in accordance with the Gambler's Fallacy. When asked to state their beliefs as probabilities, with incentives for accurately estimating the true frequency, they favor the Hot Hand. This presents a challenge to understanding biased reasoning about probability as it implies that people use different subjective models to answer theoretically identical questions.

2.1 Introduction

Many economic decisions require a person to predict outcomes of random processes. To maximize their expected utility, decision makers must infer the likelihoods of future outcomes from past information. Even under ideal circumstances, this inference process may be highly complex. As a result, either through errors or the use of heuristics, humans decision makers may hold biased beliefs about probability. Numerous experiments have demonstrated that humans differ significantly from the rational Bayesian benchmark, expecting sequences of random data to appear "more random" than the true distribution (see Bar-Hillel and Wagenaar (1991) for a review). For sequences of Independent and Identically Distributed (IID) random variables with binary outcomes, people expect more reversals and thus judge streaks of consecutive outcomes to be excessively unlikely. The expectation of excessive randomness is most commonly attributed to a reliance on the representativeness heuristic to simplify complex inference about random processes — the belief that even short sequences of random data should maintain the features of the long-run distribution (Tversky and Kahneman, 1971). For sequences of IID random variables, we classify errors in judgement as manifestation of either the Gambler's or Hot Hand fallacy. The former is a false expectation of negative correlation between outcomes, and the latter is the unfounded expectation of positive correlation. For flips of a fair coin, representativeness leads to the Gambler's Fallacy because decision makers anticipate reversal of the most recent outcome as it would better approximate the long-run expected frequencies. The Hot Hand also arises from representativeness because the observer over estimates the rarity of streaks of outcomes, leading them to infer the coin must be biased (Gilovich et al., 1985).[1] The existence of opposing patterns of bias is well-established, raising the natural question of what causes one to take precedence

[1]The Hot Hand fallacy draws its name from sports, claiming that players' performances are relatively constant and do not have significant 'hot' or 'cold' streaks. However, this belief may not be a fallacy in the canonical domain. Evidence suggests that players exhibit significant streaks of under/over-performance in professional baseball (Green and Zwiebel, 2018), basketball (Miller and Sanjurjo, 2019), darts (Ötting et al., 2020), golf (Livingston, 2012), and tennis (Jetter and Walker, 2015).

over the other.

Two recent studies in which subjects predict transparently IID binary sequences find that different biases predominate. Benjamin et al. (2017) (BMR) ask their participants to predict the next outcome in a series of coin-tosses after observing a sequence of past outcomes displayed as a string of H's and T's (for Heads and Tails). Subjects bet on the future coin-tosses with asymmetric payoffs for correctly predicting Heads or Tails. In this environment, the Gambler's Fallacy is the dominant bias — subjects believe strongly enough that streaks are more likely to end than persist to overcome the difference in potential payoffs. Frydman and Nave (2017) (FN) use a similar environment — a neutral framing and full transparency about the true IID generating process — and find the opposite: predictions, on average, conform with belief in the Hot Hand Fallacy. Examined together, their results imply a contradiction in subjects' reasoning: they simultaneously believe the next outcome is more likely to be Heads and that the probability of Heads has decreased.

If these opposing biases arise from the same underlying error in reasoning, what causes one to take precedence over the other? In this chapter, we report the results of a laboratory experiment investigating this question. Subjects view sequences of IID coin tosses and predict the next outcome. Motivated by key differences between BMR and FN, we vary both the incentive structure (predicting single outcomes versus estimating probabilities directly) and the visual presentation of previous outcomes (a random-walk graph versus a string of text). Subjects are informed about the true data generating process and that this process is consistent across all treatments. Despite the consistent information and framing, we observe both the Hot Hand and Gambler's fallacies within the same subjects. We connect subjects' responses to the Rabin-Vayanos model of subjective probability (Rabin and Vayanos, 2010), described in detail in Section 2.

It has been established that streak lengths (Asparouhova et al., 2009), human versus purely mechanical data generating processes (Burns and Corpus, 2004; Rao and Hastie, 2023), and

human intentions (Roney and Trick, 2009; Huber et al., 2010) influence the average direction of bias. These factors are meaningful and consistent with how people encounter probability in their day-to-day lives. However, the methods used across these studies vary greatly and present challenges when trying to develop a comprehensive understanding of decision making process. The studies which find subjects switching between the two patterns generally induce this behavior by presenting identical data framed as the result of different generating processes. Although the experimenter knows the data are IID, this is not communicated to participants and changes in framing leave ample opportunity for subjects to infer unintended information. In this study, we build upon these prior results by maintaining a consistent, neutral framing across all treatments and by explicitly and truthfully informing participants that they data they forecast are IID.

Previous studies establish the existence of both the Hot Hand and Gambler's fallacies across a wide variety of domains. Naturally, gamblers themselves often exhibit this bias, avoiding lottery numbers (Clotfelter and Cook, 1993) and roulette outcomes (Sundali and Croson, 2006) which have recently occurred. They continue to avoid recent winners even in parimutuel betting markets where this bias lowers their expected payoffs (Terrell, 1994, 1998). However, gamblers favor the Hot Hand Fallacy with respect to the success or failure of others' predictions. Both in laboratory coin-toss experiments (Huber et al., 2010; Powdthavee and Riyanto, 2015) and in real-world lottery markets (Bou et al., 2016), people are willing to pay for the useless recommendations of recent winners. Similarly, lottery gamblers purchase tickets more frequently from stores that have recently sold winners (Guryan and Kearney, 2008; Lien and Yuan, 2015).

Decisions consistent with the Gambler's Fallacy persist in highly consequential environments. Judges, loan officers, baseball umpires (Chen et al., 2016a), professional soccer goalkeepers (Misirlisoy and Haggard, 2014), and research and development managers (Criscuolo et al., 2021) all exhibit decision patterns consistent with expectations of mean reversion. Belief in the Gambler's Fallacy may explain much of the Disposition Effect, resulting in decision

makers expecting assets to have negatively autocorrelated returns (Jiao, 2017; Weber and Camerer, 1998). However, investors believe in the Hot Hand Fallacy with respect to fund manager's recent returns (Loh and Warachka, 2012) despite evidence against persistent differences in skill (Chen et al., 2000).

In our experiments, we observe both Gambler's Fallacy and Hot Hand behaviors within the same subjects without changes in framing. Altering the incentive structure changes the overall pattern of bias. Subjects conform with the Gambler's Fallacy when predicting single outcomes but follow the Hot Hand Fallacy when reporting their beliefs as frequencies. Changing the visual presentation of previous outcomes has no significant impact on subjects' choices. Reproducing both patterns of bias while minimizing any potential influence of outside information allows us to more closely examine the underlying sources of the contradictory behaviors.

Subjects' choices conform with the Rabin-Vayanos model with full knowledge of the generating process only when predicting single outcomes. When reporting their beliefs as frequencies, their inference is only partially consistent with the Rabin-Vayanos model. Their reported probabilities also imply uncertainty about the true probability of the coin, using the previous outcomes to infer the Bernoulli parameter value. The opposing beliefs across different incentive structures raise important questions about the appropriate methodology for studying subjective beliefs about probability. Extremely small changes in context appear to significantly alter the model humans use to forecast future outcomes.

2.2 Theory: The Rabin-Vayanos Model

We examine our results in the context of the Rabin-Vayanos (RV) model of biased inference. In the full model, subjects predict the outcomes of a first-order autoregressive process with normally distributed errors. Agents infer the model parameters from their observations as

rational Bayesians with one critical exception: they hold an explicit, unchanging belief in the Gambler's Fallacy. They expect recent 'luck' to reverse to maintain local representativeness consistent with the notion of the Law of Small Numbers (Rabin, 2002). The Gambler's Fallacy arises from this belief explicitly and the Hot Hand Fallacy arises implicitly; their belief in the Gambler's Fallacy leads them to believe that long streaks of errors in the same direction are excessively unlikely, resulting in over-inference about the mean of the distribution.

In our context, a series of coin tosses with binary outcomes, we use the Fudenberg et al. (2022) adaptation of the RV model to binary outcomes. The agent observes a sequence of previous coin flips, revealed simultaneously, with each element of the sequence drawn from a Bernoulli distribution with constant parameter θ (the true value being $\theta = .5$, in our case) giving the probability of the coin coming up Heads.

$$s_t \sim Ber(\theta)$$

with a corresponding true error term:

$$\epsilon_t = s_t - \theta$$

In the presence of the Gambler's Fallacy bias, the agent instead believes that each flip follows the distribution:

$$s_t \sim Ber\left(\theta - \alpha \sum_{k=0}^{t} \delta^k (s_{t-k-1} - \theta)\right)$$

with α and δ both being values between 0 and 1 giving the strength of the Gambler's Fallacy bias and the rate of decay over time, respectively. These values are fixed and idiosyncratic to a given agent.

Agents use this biased model of probability to conduct inference and form their predictions about the future. Apart from their persistent belief in the Gambler's Fallacy (α and δ are fixed) they process information as rational Bayesians. With the goal of predicting the next outcome of the sequence, agents apply this reasoning in two steps. First, agents infer the value of the Bernoulli parameter θ using maximum likelihood, selecting the value of $\hat{\theta}$ which best fits the observed sequence.

$$\hat{\theta} = \underset{\theta \in [0,1]}{\arg\max}\ L(\theta|\mathbf{s}) = \prod_{t=1}^{T} \left((2s_t - 1) \cdot \sum_{n=0}^{t-1} \left((s_{t-n} - \theta) \cdot \sum_{k=1}^{n} \alpha^k \delta^n \binom{n}{k} \right) \right)$$

This estimate, $\hat{\theta}$, equates to the agent's long-run beliefs about the sequence, given their distorted, subjective model. Belief in the Gambler's Fallacy leads to *over-inference*, consistent with the Hot Hand by inflating the influence of streaks of consecutive outcomes.

In the second step, agents predict the immediate following outcome, now explicitly incorporating the Gambler's Fallacy distortion, evaluated relative to their long-run belief, $\hat{\theta}$.

$$\hat{\theta_{T+1}} = \hat{\theta} - \alpha \sum_{k=0}^{T} \delta^k (s_{T-k-1} - \hat{\theta})$$

This estimate, $\hat{\theta_{T+1}}$, is the agent's local belief and applies only to the immediate next outcome, accounting for their belief that recent, abnormal 'luck' will soon reverse.

Previous tests of this model and estimates of the α and δ parameters focus solely on the second stage of this prediction process, setting aside the first, long-run inference step under the assumption that subjects are certain of the true generating process apart from their explicit Gambler's Fallacy bias. This is justified because subjects are informed of the true parameter value ($\theta = .5$) and that the outcomes are IID, thus they know the long-run behavior with certainty and have no need to infer. Their long-run beliefs should be constant and equal to the true parameter $\hat{\theta} = \theta$. However, the second step remains and they make

their prediction, accounting for the Gambler's Fallacy distortion measured against the true parameter: $\hat{\theta_{15}} = .5 - \alpha \sum_{k=1}^{14} \delta^k (s_{14-k-1} - .5)$

As the Hot Hand arises from the first step, only the Gambler's Fallacy is possible under this assumption. However, this is inconsistent with the existing literature, finding numerous examples of behavior consistent with the Hot Hand Fallacy in IID environments.

Without the parameter certainty assumption, the model has no closed-form solution for the impact of a particular piece of information on the overall beliefs and behavior of the agent. Because of this, we employ numerical methods to simulate beliefs formed through this model and establish expected results if subjects in our experiments behave consistently with the RV model.

For the case where agents are uncertain about the true Bernoulli parameter, we use numerical methods to estimate beliefs in response to observing each possible sequence of 14 coin tosses for the parameter certainty case, values are calculated directly. We use $\alpha = .2$ and $\delta = .7$ for Gambler's Fallacy parameters based on previous estimates (Benjamin et al., 2017). Using these estimates, we examine the average effect of Heads occurring in each of the 14 possible positions on both the agent's long-run and local beliefs: $\hat{\theta}$ and $\hat{\theta_{15}}$. To do so, we regress the simulated beliefs on dummy variables, each representing Heads in the 14 possible positions.

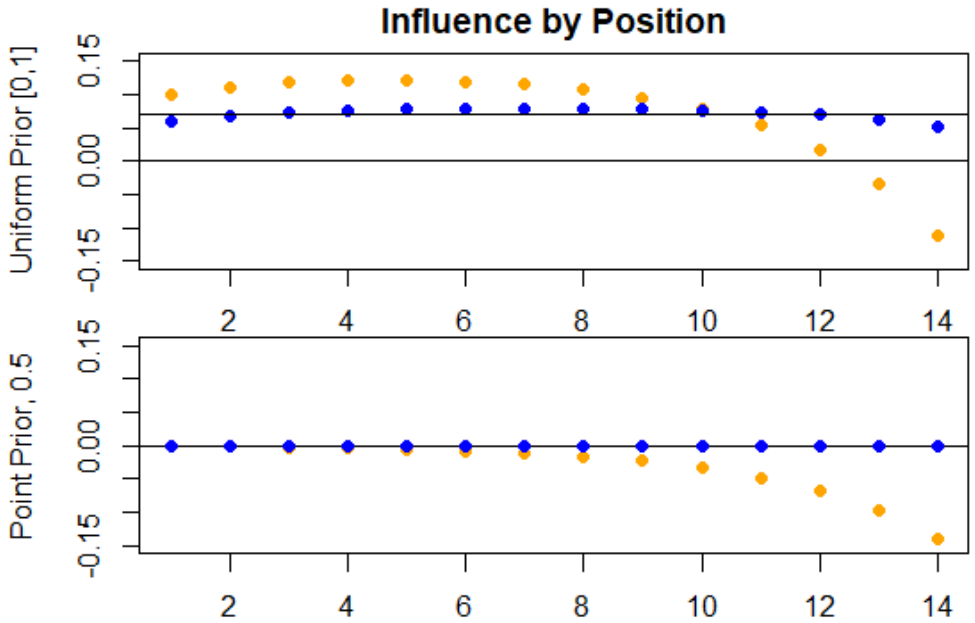
$$\hat{\theta}(\mathbf{s}) = \beta_0 + \sum_{i=1}^{14} s_i \cdot \beta_i$$

The point estimates for the coefficients on each of the 14 positions, both of the overall belief and the local prediction $\hat{\theta}_{15}$, are presented in Figure 2.1.

For a rational Bayesian decision maker, unaffected by the Gambler's Fallacy, the weight assigned to each of the 14 positions would be equal and positive (or zero in the case of parameter certainty). A few key features emerge in the expected responses between the two cases. First, as stated previously, there is no possibility of the Hot Hand Fallacy under

parameter certainty. As such, all coefficient estimates in the local belief model are negative, with the magnitude steadily decaying for less recent outcomes. The long-run belief coefficient is zero for all positions as the agent is assumed to conduct no long-run inference.

Figure 2.1: RV model simulations



Regression model estimates for average effect of a Heads in each position on the decision maker's estimate of the likelihood of Heads. Coefficients for the model estimating long-run beliefs, $\hat{\theta}$ are plotted in blue and coefficients in the model estimating the local belief about the immediate following outcome, $\hat{\theta}_{15}$, are plotted in orange. The black reference line plots the weight placed by a rational Bayesian for comparison.

When uncertain about the Bernoulli parameter, the agent's pattern of inference differs in several key aspects. First, while the influence on the local beliefs remains negative in the positions in the most recent past, they become positive in earlier positions, exceeding the magnitude of the rational Bayesian baseline. The Gambler's Fallacy belief inflates the influence of consecutive outcomes and reduces the influence of reversing outcomes, leading to this over-inference. Further, the influence on long-run beliefs is always positive, in the direction of the Hot Hand. This influence is strongest near the middle of the sequence, exceeding the magnitude of a rational Bayesian, conducting inference without the Gambler's Fallacy distortion. Counter-intuitively, the coefficient estimates for local beliefs are positive and

greater in magnitude than for long-run beliefs. The positive influence of the early outcomes on long-run beliefs also decreases the magnitude of the explicit Gambler's Fallacy effect when agents form their local beliefs.

2.3 Hypotheses

In the context of this model of subjective probability, we conduct an experiment to examine the causes of the opposing biases observed in BMR and FN. To do so, we ask subjects to predict the outcome of a coin-toss after observing a sequence of 14 prior results. We employ a 2 by 2 design, with each factor corresponding to one of the notable design differences between BMR and FN. First, we vary the incentive structure used to elicit subjects' predictions, either asking subjects to either bet on the specific, binary outcome of the next flip as in BMR or asking for an explicit frequency estimate as in FN. Additionally, we vary the visual presentation of the previous outcomes, presenting them as strings of text (ex. HTHT...) as in BMR or as a random walk graph as in FN. Using subjects responses from this game, we test three primary hypotheses.

- Hypothesis 1: The incentive structure used to elicit beliefs matters. Subjects' predictions, on average, imply belief in the Gambler's Fallacy when making binary predictions and imply belief in the Hot Hand Fallacy when reporting probabilities directly.
- Hypothesis 2: Subjects process information differently depending on the visual presentation of previous outcomes. They favor the Hot Hand when shown past flips graphically and the Gambler's Fallacy when shown a string of text. We test this against the null hypothesis that regression coefficients for models fit on data from each presentation type are equal.
- Hypothesis 3: The subjects do not fully believe the stated Bernoulli parameter, leading to the possibility of the Hot Hand Fallacy. We test this against the null hypothesis

that the regression coefficients for all 14 positions are less than or equal to zero; none of the positions are positively correlated with the agent's predictions.

2.4 Methods

Our experiment consists of an online task conducted on Amazon Mechanical Turk (MTurk). This online labor market allows requesters to post small jobs referred to as Human Intelligence Tasks (HITs) which individuals then perform. Typical HITs are extremely brief with many taking only a few seconds to complete. Results from experiments conducted on this platform are similar to those conducted in a laboratory setting (Horton et al., 2011).

Our task asks individuals to make predictions about random binary events framed as coin tosses. We use four treatments involving two different incentive structures and two different visual presentations of sequences of flips. Participants are informed that they will be asked questions about random events based about a data-set generated prior to the start of the task. This data-set consists of one million realizations of sequences of 15 simulated tosses of a fair coin. From this pool of sequences, we select one and show the participant the first 14 flips. Participants are then asked to make predictions about the result of the 15th flip. Subjects encounter two incentive structures across Bet and Frequency treatments. Frequency questions ask participants to predict the frequency that they expect Heads to occur and betting questions ask them to bet on the next outcome. This method of sequence generation, the questions used to elicit beliefs, and the instructions are adapted from Benjamin et al. (2017).

When participants are asked to predict a frequency, they are paid based on how close their estimate is to the true frequency across our pool of sequences. Their payoff is calculated as:

$$ECU = 100 - |\text{Estimated frequency} - \text{True frequency}|$$

with ECU being converted to final payment at a rate of 3000 ECU = $1.00.

When participants bet on the next outcome, the payoffs for correctly predicting heads or correctly predicting tails are asymmetric to reveal strictly biased beliefs. The payoff for correctly predicting heads is always 50+x ECU and 50-x ECU for tails, with x being randomly drawn from -5 to 5. This incentive structure is also adapted from BMR.

We use two methods to present participants the 14 flip sequences. In 'Text' treatments, participants are shown a string of H's and T's. In 'Graph' treatments, the previous flips are converted into a graphical representation where heads is shown as an upward move of one unit on a time series graph and tails as a downward movement. This presentation method is designed to closely follow the environment found in the economic decision task of FN.

In designing the graphical presentation, we take care to avoid features that may impact subject's beliefs about the generating mechanism. Specifically, we avoid labeling the axis and we ensure that there is ample space left between the maximum and minimum values observed. This is done to avoid giving the impression that the sequence is approaching an upper or lower bound of possible values which may push participants towards the Gambler's Fallacy. This also avoids the risk that participants may assign special significance to particular numerical values e.g. believing that a sequence is unlikely to cross from positive to negative values.

2.5 Results

2.5.1 Descriptive Statistics

Our data include observations from 201 individuals. Of those 201 subjects, 18 of them gave the unique expected value maximizing prediction for every question. Participants in our study earned $1.00 for participation as well as a bonus based on the accuracy of their predictions during the experiment. A typical participant earned $1.25 in bonus payment.

The task included 60 questions and the average total participation time was 11.3 minutes.

Each subject predicts the 15th outcome through four blocks of 15 questions. Their participation time during the prediction task is summarized in Table 2.1. The mean overall time spent during each 15 question block is 88 seconds for a total average participation time of approximately 6 minutes, not including consent forms and initial instructions.

Table 2.1: Mean Duration by Treatment and Order

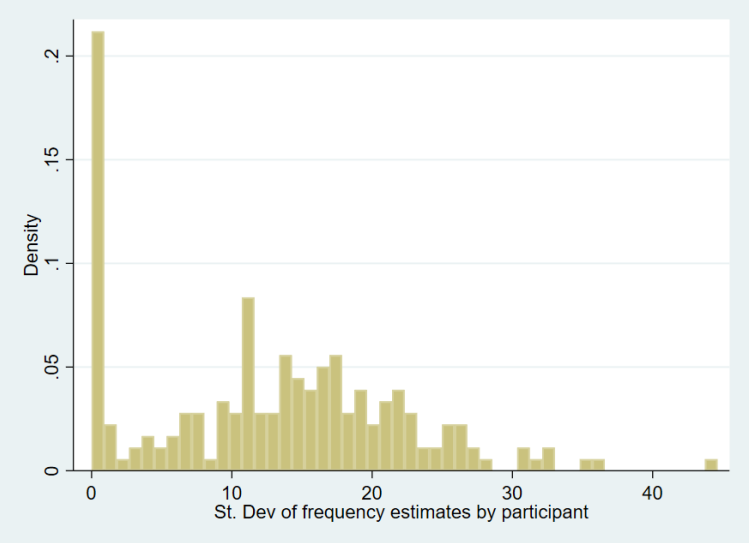
	Order of treatment				
Treatment	1	2	3	4	Total
Graph Bet	150.3	50.1	44.1	55.3	72.8
Graph Freq	151.7	131.0	127.8	147.3	138.4
Text Bet	80.8	56.7	39.9	21.2	50.7
Text Freq	117.6	88.0	89.5	64.9	91.3
Total	121.9	85.4	73.6	72.2	88.3

Average participation time in seconds for each treatment and order encountered

Subjects progress through each block in less time as they progress through the study, having gained familiarity with the task. On average, subjects spend approximately 50% longer on Graph treatment blocks relative to Bet treatments and nearly twice as long answering each frequency question as they do in each betting question. Much of this is likely due to the differences in interfaces with frequency questions requiring subject to type their responses and graph questions requiring scrolling past large images.

Because there is a unique, payoff maximizing choice for all questions, we also briefly verify that our subjects are sufficiently susceptible to bias to warrant analysis. To do so, we examine responses to all frequency questions, grouped at the subject level, and calculate the standard deviation of each participant's reported estimates. Figure 2.2 plots a histogram of these standard deviations.

Figure 2.2: Distribution of Subject level Estimate Variance



Distribution of the standard deviation of frequency estimates across all prompts for each participant.

The modal subject reports the expected-payoff maximizing response of 50% for every frequency question. However, this account for only a fifth of the total subject pool. The remaining subjects all show some degree of bias, altering their frequency estimates in response to the different sequences of past coin flips.

2.5.2 BMR and FN Comparison

Our experiment is primarily motivated by the contrasting results from BMR and FN, given their apparent similarity. Although both ask subjects to predict outcomes of a transparently random process with incentive structures penalizing biased beliefs, BMR finds that most subjects follow the Gambler's Fallacy whereas the Hot Hand Fallacy predominates in FN. With the goal of investigate the source of this discrepancy, we first confirm that the results from our equivalent treatments produce the same patterns of bias. In our design, Bet Text and Graph Frequency approximate the methods in BMR and FN respectively.

Table 2.2: Comparison to Frydman-Nave Econ. Decision Task

	Frequency Estimate
Ending Streak	0.365**
	(2.65)
Constant	48.12***
	(119.02)
Observations	2998

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

t statistics in parentheses. OLS regression estimating frequency of Heads predictions on the length and direction of the final streak.

Table 2.3: Comparison to BMR Result

	Predicts Heads
Total Heads Shown	-0.0134**
	(-3.23)
Bet Payoff for Heads	0.0474***
	(11.52)
Constant	-1.744***
	(-8.37)
Observations	3011

* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$

t statistics in parentheses. Linear probability model estimating the likelihood a subject predicts Heads based on the number of observed heads, controlling for the payoff for correctly predicting Heads.

Fitting the same regression specifications used by the authors of each of the previous studies, our data agree with both of their respective conclusions. Subjects in our Bet Text treatments, most similar in design to BMR, follow the Gambler's Fallacy with respect to streaks of outcomes at the end of the sequence. There is a negative correlation between the length and direction of the final streak of outcomes and subjects' predictions. This relationship is significant at the 5% level, after controlling for the differences in incentives. We also note, at this point, the strength of the incentive effects. Despite the small size — biased responses lower expected payoffs by a fraction of a cent — subjects respond significantly.

Our Graph-Frequency treatment is most similar to the FN environment, where the authors found that the Hot Hand Fallacy predominates. We fit the same regression specification used by FN and present the results in Table 2.2. Our results agree with the FN findings, consistent with the Hot Hand Fallacy. When evaluating the average direction of bias with respect to the concluding streak of the sequence, we find that streaks are positively and significantly correlated with the length and direction of the final streak of outcomes. Subjects' frequency estimates imply that streaks are more likely to continue than they are to reverse.

These initial tests confirm that our game designs are similar enough to the previous experiments to reproduce the opposing patterns of bias. Further, variation between our treatments results in subjects switching the direction of their bias. This occurs with identical participants observing identical sequences across the four treatments. The remainder of the results examine these changes in detail.

2.5.3 Indicator Variable Regressions

We next examine the overall pattern of inference by position with our subjects' choices as we did in our simulations. We use a mixed effects model with standard errors clustered at the subject level. For each of the types of questions, we fit a separate regression model regressing either the subjects' frequency estimates or outcome prediction on indicator variables. For

Frequency treatments, we fit:

$$\hat{\theta}(\mathbf{s}) = \beta_0 + \sum_{i=1}^{14} s_i \cdot \beta_i$$

For betting questions, we also include the weight of the payoffs in favor of Heads:

$$\hat{\theta}(\mathbf{s}) = \beta_0 + \sum_{i=1}^{14} s_i \cdot \beta_i + \beta_{bw} \cdot \text{Bet Weight}$$

The results of these regressions are presented in Table 2.4

Table 2.4: Predictions vs. Indicator Variables by Position

	Freq. Graph t-stat	Freq. Text t-stat	Bet Graph t-stat	Bet Text t-stat
Flip 1	1.81*	0.079	0.43	1.59
Flip 2	0.94	2.22**	1.46	1.66*
Flip 3	2.12**	1.84*	1.61	0.050
Flip 4	-0.94	0.51	-2.61***	-0.81
Flip 5	1.06	0.56	-2.09**	0.62
Flip 6	2.80***	2.56**	0.12	0.28
Flip 7	2.86***	1.98**	-0.22	0.10
Flip 8	3.81***	1.79*	-1.63	0.40
Flip 9	3.89***	2.91***	-0.35	-0.23
Flip 10	4.81***	2.37**	0.83	0.61
Flip 11	1.21	2.28**	-1.01	-1.18
Flip 12	1.65*	1.68*	-2.00**	-2.02**
Flip 13	0.47	-1.02	-1.47	-3.25***
Flip 14	1.43	-0.56	-0.79	-3.01***
Bet weight			3.84***	4.59***
Observations	2624	2622	2627	2637

Mixed effect estimator; clustered by participant

$* : p < 0.10, ** : p < 0.05, *** : p < 0.01$

Examining the coefficient estimates for each model, we first note a stark difference between Frequency and Bet treatments. With Frequency treatments, all significant coefficient estimates ($p < 0.05$) are positive, with the strongest estimates located near the middle of the sequence. This is inconsistent with the Rabin-Vayanos model's prediction of agent's fore-

casts for the immediately following flip. Regardless of whether subjects believe the coin is fair with certainty, decision makers are expected to have their decisions correlate negatively with flips in the recent past.

The regression specification above assumes that the information in each position has a linear effect on the overall decision. However, this is inconsistent with many previous interpretations of the Hot Hand and Gambler's fallacies which place special significance on streaks of outcomes. We estimate the same regression models as above, with the addition of a variable for the length and direction of the streak of consecutive outcomes at the tail end of the sequence. Sequences ending in Heads are encoded as positive values and streaks of Tails are encoded as negative. Table 2.5 presents the results.

Table 2.5: Predictions vs. Indicator Variables by Position with Streak Control

	Freq. Graph t-stat	Freq. Text t-stat	Bet Graph t-stat	Bet Text t-stat
Flip 1	0.29	-0.62	0.61	1.63
Flip 2	0.91	2.16**	1.42	1.62
Flip 3	2.31**	1.94*	1.61	0.056
Flip 4	-1.33	0.24	-2.60***	-0.97
Flip 5	0.58	0.30	-2.08**	0.61
Flip 6	3.30***	2.77***	0.041	0.18
Flip 7	2.94***	2.01**	-0.019	0.35
Flip 8	2.49**	1.08	-1.68*	0.28
Flip 9	3.63***	2.76***	-0.38	-0.29
Flip 10	3.70***	1.73*	0.93	0.80
Flip 11	0.41	1.66*	-1.10	-1.32
Flip 12	1.11	1.35	-1.99**	-1.99**
Flip 13	0.43	-1.06	-1.53	-3.28***
Flip 14	0.54	-1.00	-0.83	-3.06***
Streak	1.76*	1.08	0.44	0.60
Bet weight			3.03***	3.53***
Observations	2624	2622	2627	2637

Mixed effect estimator for a linear probability model in betting treatments and reported frequencies in frequency treatments; clustered by participant. Only t-statistics are reported. $* : p < 0.10, ** : p < 0.05, *** : p < 0.01$

Including the streak control does not alter the overall results in the regressions model. In the

frequency treatments, regardless of the visual presentation method, all significant coefficient estimates are positive, consistent with the Hot Hand Fallacy. As in the previous regression specifications, the significant effects are concentrated around the middle of the sequence. Flips near the beginning and end of the sequence have no statistically significant impact on subjects' frequency estimates.

In betting treatments, the added streak variable does not alter the tendency towards the Gambler's Fallacy observed in the previous specification. All significant coefficient estimates are negative across both visual presentations. In text treatments, these coefficients are concentrated near the end of the sequence. There is no clear pattern in the location of the significant coefficients in the graphical treatments.

With and without the controls for streaks, the results have the same implications for our hypotheses. First, the evidence supports Hypothesis 1: the incentive structure used to elicit beliefs matters. When subjects bet on individual outcomes, all significant coefficient estimates are negative, consistent with the Gambler's Fallacy. In treatments where subjects predict frequencies, all significant coefficient estimates are positive, consistent with the Hot Hand. Second, we find no evidence supporting Hypothesis 2. There is no evidence of differences in responses depending on whether Text or Graph presentations are used. We test this further in the following subsection. Finally, the differing responses to the changing incentive structures provide evidence regarding Hypothesis 3: the subjects' choices are inconsistent with parameter certainty under the RV model. In Frequency questions, all significant coefficient estimates are positive. This is only possible in the RV model if decision makers attempt to infer the model parameters given the observed information. In Betting questions, we find no evidence to reject the assumption of parameter certainty. All significant coefficient estimates are negative, consistent with the RV model's agents' belief in reversion to the mean without uncertainty about the model parameters.

2.5.4 Grouped Flips

The results from the indicator variable regressions reveal diverging responses to the same sequences depending on incentive structure. In frequency questions, predictions are positively correlated with flips near the middle of the sequence. The patterns are less clear in betting questions. In both Bet-Graph and Bet-Text treatments, all significant coefficient estimates are negative with negative estimates primarily near the end of the sequence. However, we do not have the statistical power to confidently establish this pattern in the Bet-Graph treatment.

To address this, we consider an alternative specification, combining positions in the sequence into groups of four, reducing the number of independent variables in our regression model. In this new regression specification, we use the number of Heads in the first four positions, the middle four positions (6 through 9) and final four positions. By aggregating the flips together, we increase the power we have to detect the patterns observed in the indicator variable regressions: negative significant coefficients near the end of the sequence in betting questions and positive, significant coefficients in the middle in frequency treatments. The results are presented in Table 2.6.

Table 2.6: Predictions vs. Grouped Flips

	Freq. Graph t-stat	Freq. Text t-stat	Bet Graph t-stat	Bet Text t-stat
First 4	1.82*	3.04***	-0.71	1.04
Middle 4	6.45***	5.20***	-0.55	1.32
Last 4	2.43**	1.66*	-3.73***	-6.84***
Bet weight			6.48***	7.01***
Observations	2624	2622	2627	2637

Mixed effect estimator for a linear probability model in betting treatments and reported frequencies in frequency treatments; clustered by participant. Only t-statistics are reported. $*p < 0.10, **p < 0.05, ***p < 0.01$

The alternative specification agrees with the results of the indicator variable regressions. In

frequency question treatments, regardless of visual presentation, the significant coefficient ($p < 0.05$) estimates are all positive, consistent with belief in the Hot Hand. For betting questions, the significant coefficient estimates are all negative, excluding the significant positive estimate for the bet weight control. One notable difference in this model is that we get a much clearer picture of the main source of the Gambler's Fallacy tendencies in the Bet-Graph treatments compared to the individual position regressions. Bet-Graph treatments follow the same patterns as Bet-Text treatments, with significant coefficient estimates consistent with the Gambler's Fallacy in the final four flips but no significant response to earlier flips in the sequence.

As in the indicator variable regression model, we estimate our grouped model again with a control variable for the length and direction of the ending streak. The results for each treatment are presented in Table 2.7.

Table 2.7: Predictions vs. Grouped Flips with Streak Control

	Freq. Graph t-stat	Freq. Text t-stat	Bet Graph t-stat	Bet Text t-stat
First 4	1.38	3.02***	-0.72	1.03
Middle 4	6.84***	5.17***	-0.57	1.29
Last 4	1.34	1.62	-3.71***	-6.79***
Streak	3.11***	-0.16	-0.23	-0.23
Bet weight			6.34***	6.85***
Observations	2624	2622	2627	2637

Mixed effect estimator for a linear probability model in betting treatments and reported frequencies in frequency treatments; clustered by participant. Only t-statistics are reported. $*p < 0.10, **p < 0.05, ***p < 0.01$

Including a control variable for streak length does not alter the primary results from the previous regression specifications. In frequency questions, all significant coefficient estimates are positive with the strongest effect near the middle of the sequence. In betting questions, all significant coefficient estimates are negative with only flips in the final positions affecting subjects' decisions.

2.5.5 Visual Presentation Comparison

Using the grouped flip specification, we also test explicitly for differences across graphical treatments. We include a dummy variable for the visual treatment, interacted with the other variables to test whether certain features in the model have unequal impacts depending on the visual presentation. Table 2.8 presents the results.

Table 2.8: Predictions vs. Grouped Flips with Graph Interactions

	(1) Freq. Questions	(2) Betting Questions
Graph	-1.702	0.00852
	(-0.73)	(0.24)
First 4	1.023***	0.00173
	(2.64)	(0.24)
Middle 4	1.602***	0.0101
	(4.17)	(1.19)
Last 4	0.497	-0.0534***
	(1.06)	(-4.61)
Graph × First 4	-0.304	-0.0129
	(-0.59)	(-1.37)
Graph × Middle 4	0.239	-0.0174*
	(0.51)	(-1.68)
Graph × Last 4	0.477	0.0202*
	(0.95)	(1.82)
Bet weight		0.0505***
		(7.73)
Observations	5995	6013

Mixed effect estimator for a linear probability model in betting treatments and reported frequencies in frequency treatments; clustered by participant. Only t-statistics are reported. t statistics in parentheses $*p < 0.10, **p < 0.05, ***p < 0.01$

Including the interaction between the graphical treatment and the groups of four positions, we find no evidence that visual presentation affects choices. The previous results hold with regard to the treatment types — significant, negative coefficients for the last four flips in betting questions and positive, significant coefficients throughout in frequency questions.

There is no significant impact from the interactions of a dummy variable indicating graphical presentations and the variables for total heads in the first, middle, and final four positions.

Table 2.9: Predictions vs. Grouped Flips with Graph Interactions with Streak Control

	(1) Freq. Questions	(2) Betting Questions
Graph	-1.204	0.00922
	(-0.52)	(0.26)
First 4	1.041***	0.000846
	(2.61)	(0.12)
Middle 4	1.656***	0.00960
	(4.28)	(1.14)
Last 4	0.469	-0.0545***
	(0.99)	(-4.68)
Streak	0.0442	-0.00150
	(0.26)	(-0.59)
Graph × First 4	-0.497	-0.0129
	(-0.94)	(-1.37)
Graph × Middle 4	0.236	-0.0176*
	(0.50)	(-1.70)
Graph × Last 4	0.113	0.0212*
	(0.20)	(1.85)
Graph × Streak	0.360	-0.00128
	(1.31)	(-0.46)
Bet weight		0.0514***
		(7.52)
Observations	5995	6013

Mixed effect estimator for a linear probability model in betting treatments and reported frequencies in frequency treatments; clustered by participant. t statistics in parentheses $*p < 0.10, **p < 0.05, ***p < 0.01$

As in the indicator variable regression specifications, we estimate the grouped flip model with a dummy variable indicating graphical treatments, including a control variable for streaks. The results are presented in Table 2.9.

Including streaks in the model does not reveal any significant differences between the visual presentation methods. There are no significant coefficient estimates at the 5% level for any of the interaction terms which include the Graph variable.

The results from the regressions involving grouped flips agree with the results from the indicator variable regressions. First, the evidence from the grouped regressions supports Hypothesis 1, the inference patterns are substantially different in Frequency treatments compared with Bet treatments. We cannot directly compare the coefficients in the two models as the dependent variables differ, but the significant effects differ in direction, not just in magnitude. With betting questions, all significant coefficient estimates are negative and consistent with the Gambler's Fallacy while all significant coefficients are positive, consistent with the Hot Hand with frequency questions. Second, we continue to find no evidence of differences between the visual treatments. We fail to find any statistically significant difference in coefficients fit using Text treatments relative to Graphical treatments. Finally, regarding Hypothesis 3, we find the same mixed results found in the indicator variable regressions. In betting questions, we find no evidence inconsistent with the RV model under parameter certainty. All significant coefficients are negative the GB and TB treatments, consistent with decision makers reacting only to the expectation that "luck" reverses without conducting any parameter inference. However, the evidence in frequency treatments is inconsistent with the RV model under parameter certainty. All significant coefficient estimates are positive, which requires that decision makers are inferring the model parameter, implying that they are not certain about the true values.

2.5.6 Alternative Features

Our regression specifications thus far have deviated from the previous literature by considering each flip instead of summary statistics describing potentially relevant features of the sequence. The indicator variable regression models reveal a clear discrepancy between beliefs

Table 2.10: Predictions vs. Common Features

	(1) Freq. Graph	(2) Freq. Text	(3) Bet Graph	(4) Bet Text
Reversals	0.339	-0.940	-0.00214	-0.0433**
	(0.38)	(-1.09)	(-0.10)	(-2.06)
Heads	1.100	0.0975	-0.0189	-0.0529***
	(1.48)	(0.14)	(-1.10)	(-3.06)
Streak	0.522	-0.0195	0.0201	0.0433***
	(1.25)	(-0.05)	(1.62)	(3.45)
Reversals × Heads	0.0176	0.184	0.000317	0.00820**
	(0.14)	(1.49)	(0.09)	(2.32)
Reversals × Streak	-0.0334	-0.0237	-0.00571**	-0.0109***
	(-0.47)	(-0.34)	(-2.27)	(-4.30)
Bet weight	0	0	0.0518***	0.0466***
	(.)	(.)	(10.83)	(9.68)
Observations	2998	2997	3002	3011

Mixed effect estimator for a linear probability model in betting treatments and reported frequencies in frequency treatments; clustered by participant. t statistics in parentheses $*p < 0.10, **p < 0.05, ***p < 0.01$

depending on the question used to elicit subject beliefs. However, this assumes an independent, linear effect by each flip in the sequence. This is inconsistent with much of the previous literature which directs attention towards streaks of outcomes and number of reversals in a sequence in addition to the total number or Heads or Tails. To address this, we estimate an alternative regression model, replacing the indicator variables with these common features and their interactions. As before, we also include the weight advantaging heads in betting questions as a control variable.

Overall, where we find significant coefficient estimates in these new regression specifications, they agree with the findings from the indicator variable model — Frequency treatments tend towards the Hot Hand Fallacy while Bet treatments favor the Gambler's Fallacy. For

Frequency treatments, the total number of Heads is significantly, positively correlated with the reported frequency estimates for Heads in the Frequency-Graph treatment at the 5% level. There are no significant coefficient estimates in the Frequency-Text treatment.

Bet treatments are more complex. The Gambler's Fallacy is still the predominating bias, but is the result of several factors interacting. In the Bet-Graph treatment, the only feature, other than the control variable for the weight of incentives, significant at the 5% level is the interaction between the number of reversals and the length of the final streak. The estimated effect is negative, consistent with previous findings about the Gambler's Fallacy. The longer the streak, and the more varied the sequence before the beginning of the streak, the more likely the decision maker is to expect the streak to end.

In Bet-Text treatments, we find more features with significant coefficient estimates, several of which appear in contrast to the Gambler's Fallacy pattern previously observed. The length of the concluding streak is significant at the 5% level and positive. However, this is counteracted by the interaction between streak length and number of reversals, which is significant at the 1% level and negative. In our sample, the mean number of reversals in a sequence is 6.6 with a minimum of 4. Even at the minimum, the magnitude of the interaction coefficient exceeds the streak alone coefficient. The Gambler's Fallacy remains the dominant direction of bias. The number of heads has a significant, negative effect which is mitigated by the significant positive impact of the interaction between number of heads and total reversals. Given their respective magnitudes, the hot hand only dominates with respect to the number of heads, when the number of reversals is high.

This alternative regression specification allows us to examine the non-linear features previously found to influence biased probability estimates. Overall, the estimates, when significant conform with the previous, indicator variable regression model but do little to reveal a meaningful alternative understanding. Significant estimates in frequency questions favor the Hot Hand while the model conforms with the Gambler's Fallacy for betting questions

for the typical sequence. However, beyond this most general agreement, these estimates do not suggest a coherent alternative model. What we can say with relative confidence is that these regressions provide further support for the key finding in our indicator variable model — subjects approach frequency and betting questions using different methods for each environment.

2.6 Conclusion

In this chapter, we investigate the formation of subjective beliefs about probability. Specifically, we examine the factors which influence the direction of this bias towards either the Gambler's Fallacy or the Hot Hand Fallacy. To do so, we conduct an incentivized experiment where subjects predict future outcomes of an IID random process. Our subjects are transparently and unambiguously informed of the true generating process which allows us to examine subjects' beliefs about the random process free from any noise introduced by changing framings. Based on the opposing results of BMR and FN, we vary the incentives used to elicit beliefs as well as the visual presentations of past outcomes. In doing so, we successfully produce behavior consistent with both the Gambler's and Hot Hand fallacies within the same individuals, responding to the same sequences of data.

We find consistent evidence that the incentive structure used to elicit subjects' beliefs changes the overall tendency from the Gambler's Fallacy to the Hot Hand Fallacy. When predicting individual outcomes (betting whether the next outcome will be Heads or Tails) decision makers act in accordance with the Gambler's Fallacy. When giving their beliefs directly as probabilities, they favor the Hot Hand. Subjects switch between the Hot Hand and Gambler's Fallacies even though subjects are reacting to identical information across both all treatments, implying directly contradictory beliefs. They appear to simultaneously hold the beliefs that Heads is more likely than Tails and that Tails is more likely than Heads. We find no substantial evidence that the visual presentation method affects subjects' choices.

Overall, the differences between the two elicitation questions used suggest a serious challenge to the current understanding of probability biases. In response to identical information, subjects in our experiments react in discordant ways depending on the way they are asked to report their beliefs. The difference between these questions is, purposefully, extremely subtle. In Bet treatments, we ask subjects to predict the immediate next outcome. In frequency treatments, we ask subjects to predict the frequency of the immediate next flip being heads out of all of the cases where the first 14 flips are the same. We repeat this to emphasize that the two questions are as close to identical as possible. We ask subjects to make their predictions for the same future event given identical information and histories. We do not ask subjects to estimate the long-run frequency of heads if a single sequence continues indefinitely, which would make our results consistent with the existing understanding of the Gambler's Fallacy where the influence of the 'luck' distortion decreases over time.

Despite the questions' similarity, the results suggest that two discrepancies arise in subjects' implied beliefs. First, betting questions lead respondents to report beliefs consistent with the Rabin-Vayanos model, reporting the belief for the single, immediately following flip. The Gambler's Fallacy predominates and there is no evidence that they doubt the fairness of the coin — they show no tendency to positively correlate their beliefs with previous flips. In contrast, subjects respond to frequency questions in a way consistent with the overall, long-run beliefs for agents in the Rabin-Vayanos model. Their beliefs are correlated positively with previous flips and these effects are strongest near the middle of the sequence — the Hot Hand is the prevailing bias. Further, this behavior implies that subjects doubt the fairness of the coin.

These results reveal a significant challenge towards understanding biased beliefs about probability. The features which are significant to decision makers, and the ways which they respond to them, are difficult to predict ex-ante. Fully understanding these common distortions requires further, detailed investigation.

Chapter 3

Investigating Trust Using Natural Language Processing

Abstract

Communication matters in economic decision making, even in games with simple, unique equilibria. However, the exact mechanism through which this effect occurs remains uncertain. Previous studies have only been able to handle ad hoc hypotheses due to their relatively small collections of messages. In this chapter, we employ Natural Language Processing (NLP) on data from a large-scale trust game run on Amazon's MTurk system with one-way communication from trustees to trustors. NLP methods, combined with our data set of over 1000 messages, allow us to approach the data with minimal substantive assumptions to investigate the mechanisms underpinning trust. The structure of our data-set clearly links trustors' choices to their received messages, allowing us to identify new features that increase the perceived trustworthiness of a message. Additionally, knowing the ultimate decisions of the agents, we identify features of messages which predict that they are likely not trustworthy.

3.1 Introduction

Communication matters in economic decision making. This remains true even in games with unique, easily discerned equilibria. In these settings, standard theory predicts that, since all parties are fully informed about the payoffs for each outcome, any costless signal sent between players will contain no useful information. In the absence of any sort of commitment mechanism, any rational, payoff-maximizing individual understands that a message they receive does not alter the underlying incentives and thus would not change their behavior. However, experimental evidence consistently shows that this is not the case (see Crawford (1998) for a review).

The increase in cooperation is particularly pronounced when decision makers communicate using free-form messages. In trust games, free-form promises increase trust much more than prewritten messages with the same literal meaning (Charness and Dufwenberg, 2010). Furthermore, a modest communication effect persists even when communication is restricted to only small talk, irrelevant to the experimental task (Fiedler and Haruvy, 2009). Although the impact of free-form communication is well established, the reason for the increase in cooperation remains unknown. The flexibility afforded by free-form communication also increases complexity, making it difficult to study. Free-form messages provide a rich source of information on decision makers' choice processes. However, this complexity also means that there is immediately apparent way to sort through the immense message space.

In this chapter, we investigate whether Natural Language Processing techniques provide an effective means of uncovering new factors which drive the outsized impact of free-form communication. We study one-way messaging data in the context of a principal-agent trust game adapted from Charness and Dufwenberg (2006) where Senders (agents) compose pre-play messages to Receivers (principles). Receivers may earn more by trusting the Sender but run the risk of the Sender keeping all potential gains for themselves, leaving the Receiver

with nothing. This anonymous text-based messaging is similar to what people encounter everyday in an ever-increasing share of interactions occurring online. With the large dataset of experimental communication data and the associated choices, we employ a variety of Natural Language Processing techniques, allowing us to identify new features associated with changes in both trust and trustworthiness.

Previous literature establishes the impact of communication in a wide variety of settings including bargaining (Ellingsen and Johannesson, 2004), principal–agent trust games (Charness and Dufwenberg, 2006; Ben-Ner et al., 2007; Chen and Houser, 2017), and Oligopoly games (Fonseca and Normann, 2012; Cooper and Kühn, 2016). As an extreme example, communication remains influential even in dictator games, free of potential strategic considerations, (Andreoni and Rao, 2011). The consistent theme across these settings is that communication promotes cooperation and efficiency.

The fact that cheap talk messages impact choices at all presents a significant obstacle in developing complete theories of behavior; determining what makes a message effective is crucial to furthering this understanding. Previous studies have proposed a wide variety of mechanisms through which cheap talk messages can influence the behavior of others. Potential factors include: an aversion to lying (Lundquist et al., 2009), a reduction in social distance between subjects (Buchan et al., 2006; Charness and Gneezy, 2008), and the content of the messages independent of personal relationships (Mohlin and Johannesson, 2008). In each of these cases, the findings arise from comparing different permissible levels of communication, not an analysis of message contents. The full magnitude of the impact of communication cannot be explained solely as the result of game-relevant signals (Charness and Dufwenberg, 2010).

Machine learning offers a new set of tools and techniques we may use to parse unstructured communication data. Previous laboratory experiments employ machine learning to discover novel findings in complex environments to inform theory (Camerer et al., 2019; Fudenberg

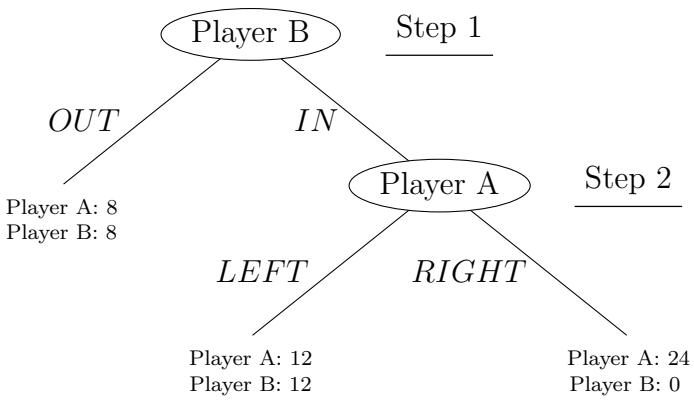
and Liang, 2019). If properly processed, similar techniques can be used on text data. Natural Language Processing (NLP) has already provided insights into economic decision making, analyzing the impact of investor sentiment on the movements of stock prices (Tetlock, 2007). Much like our setting, this sentiment has no formal informational value yet still contains predictive power for future price movements. In the economics laboratory, NLP is still in its relative infancy. However, the existing literature has shown its potential both to streamline data classification (Penczynski, 2019; Tebbe and Wegener, 2022) and to produce new findings (Hanaki and Ozkes, 2022; Bursztyn et al., 2023; Andres et al., 2023).

The remainder of the chapter proceeds as follows. The remainder of this section presents relevant theory followed by the experimental procedures we use to generate a large data-set of free-form text messages. We then provide a detailed description of this data set along with the methods applied to make this unstructured data useful in Section 2. Section 3 presents our analysis of the impact of features we identify in our message corpus. Finally, we summarize our findings and conclude in Section 4.

3.1.1 Background

In this study, we examine a simple trust game with the following form:

Figure 3.1: Basic Trust Game



The standard subgame perfect equilibrium from backward induction is trivial. Player A's maximize their monetary payoff by choosing Right if given the opportunity, leaving Player B's with nothing. Knowing this, Player B's maximize their own expected payoffs by distrusting Player A's and playing Out in the first stage. Thus, for the game to allow an alternative outcome, players must have considerations outside of their own monetary gain.

Traditionally, when communication is added to a game, authors omit the messaging step from their formal representation of the game. We pause here to note, however, that when the game allows free-form text communication, the game remains technically finite. Given a maximum message length of L characters with 256 potential ASCII choices for each, the action space for the Sender's pre-play message contains 256^L possible choices. This is intractably large and this observation alone does little to further our understanding of communication and the vast majority of this space is filled with random, unintelligible messages. However, it allows us to more carefully formulate the task at hand when we study the contents of subjects' messages. The current state of the art is to manually partition this message space into categories corresponding to available actions in the game (promises to choose Left, requests to choose In, etc.). Our goal is to expand upon this by utilizing more granular partitionings using NLP to capture a greater portion of the message space's complexity.

To represent the existing approach for studying free-form communication, we define a baseline partition of the message space we call literal meaning. The literal meaning of a message is a mapping from the message space to possible actions in the game. We chose this as our benchmark because it can be generalized to other environments besides our trust game to reasonably separate the basic, game-relevant signals from the "purely free-form" elements of a message. Given the actions in our trust game, there are nine possible messages: In, Out, or none for the Receiver's actions and Left, Right, or none for the Sender's actions. We omit Out and Right because they are rarely used, leaving us with four possible messages. Senders may signal their intent to choose Left, suggest the Receiver choose In, both, or neither. Using Natural Language Processing, we examine whether this literal meaning, based

solely of available actions, provides a sufficiently full account of the information contained in free-form messages.

Literal meaning forms the benchmark against which we examine the value of NLP. In the analyses that follow, we employ a variety of NLP tools and techniques to generate different reductions of the message space. Across the different formulations of the message space, we also include the literal meanings of a message as control variables. In doing so, when we identify message features which significantly correlate with subjects' choices, we find evidence of predictive information not yet considered.

In the trust game we examine, there are two choices which may correlate with message choice: the Receiver's choice between In and Out and the Sender's choice between Left and Right. We refer to each of these decisions as Trust and Share, respectively.

Within this framework, we define our research hypotheses.

- Hypothesis 1: There are significant predictors of trust after controlling for literal meaning.

- Hypothesis 2: There are significant predictors of trustworthiness after controlling for literal meaning.

In addition, we examine whether the features we identify accurately and effectively signal the Sender's intentions to the Receiver.

- Hypothesis 3: Any significant predictors of trustworthiness are also significant predictors of trust.

3.2 Methods

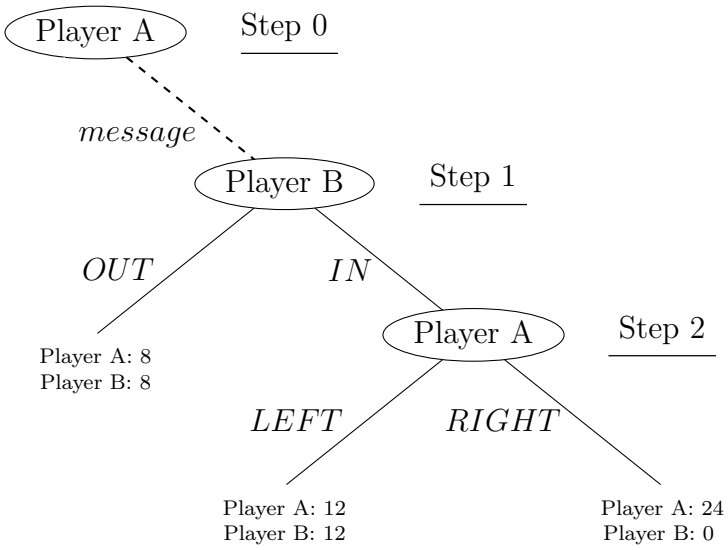
We recruited our subjects on Amazon Mechanical Turk (MTurk), using task software programmed in the oTree framework for Python (Chen et al., 2016b). MTurk allows requesters to post Human Intelligence Tasks (HITs) that can be performed for a small fee by a large pool of workers. Numerous previous studies conducted on MTurk have yielded high-quality results consistent with those from traditional laboratory settings (Snowberg and Yariv, 2021). To further ensure the quality of our subject pool, we restrict eligibility to include only accounts that have at least a 99% approval rate and more than 1,000 submitted HITs, which corresponds to approximately the 75th percentile of MTurk workers (Robinson et al., 2019). Participants earn a $1.00 participation fee as well as a bonus payment based on their choices. The typical bonus payment is $1.10. The median completion time is approximately 8 minutes.

After accepting the HIT, participants provide informed consent and receive the instructions for the study followed by a brief comprehension quiz. Participants are required to correct any incorrect answers before proceeding to the study. Once all questions are answered correctly, participants proceed and play 10 rounds of a trust game first introduced by Berg et al. (1995) with instructions adapted closely from Charness and Dufwenberg (2006).

The game consists of two players — Player A, the Sender and Player B, the Receiver – — and two steps. In the first step, Player B chooses between two actions: In or Out. If they select Out, both players receive 8 tokens of experimental currency and the round ends; otherwise, the game continues to step 2. In step 2, Player A chooses between Left or Right. Choosing Left results in both players receiving 12 tokens while choosing Right results in a payoff of 24 tokens for Player A and 0 for Player B. Figure 3.2 provides an overview of this game.

Prior to step 1, Player A may send a message to influence Player B's decision. We pur-

Figure 3.2: MTurk Game Overview



posefully provide no guidance to subjects on how they should use their message to avoid potential experimenter demand effects. This message is delivered to Player B prior to their decision between In and Out.

Each subject composes only one message as Player A which is sent to all 10 Player Bs they are paired with. Multiple responses allow us to capture a more detailed measure of trust. We assign pairings sequentially, with each participant acting as Player B for the 10 players immediately preceding them and as Player A for the 10 players following them.

Our method of implementing this trust game has two distinct benefits over previous studies. First, conducting our experiment on MTurk allows us to recruit a large pool of subjects rapidly, at low cost. As a result of this, our current dataset is quite large relative to previous experimental communication studies. Second, collecting multiple reactions to each message gives us a more precise measurement of its effectiveness.

3.3 Data

3.3.1 Message Overview

Our data consist of 1004 unique messages and their corresponding actions collected during the summer of 2020. For each message, we measure both trust and trustworthiness. Trust is measured as the frequency the recipients of a particular message choose to play In while trustworthiness is the binary Left/Right choice of the sender; choosing Left and splitting the pool of tokens is trustworthy. To summarize the general content and variation within our messages, Table 3.1 shows a selection of the most effective and least effective messages — always and never trusted, respectively. We briefly note that there is an immediately apparent qualitative difference between these groups. Effective messages are longer and display a higher level of effort and understanding on average than ineffective messages. For the sake of space, only a partial selection of effective messages is shown.

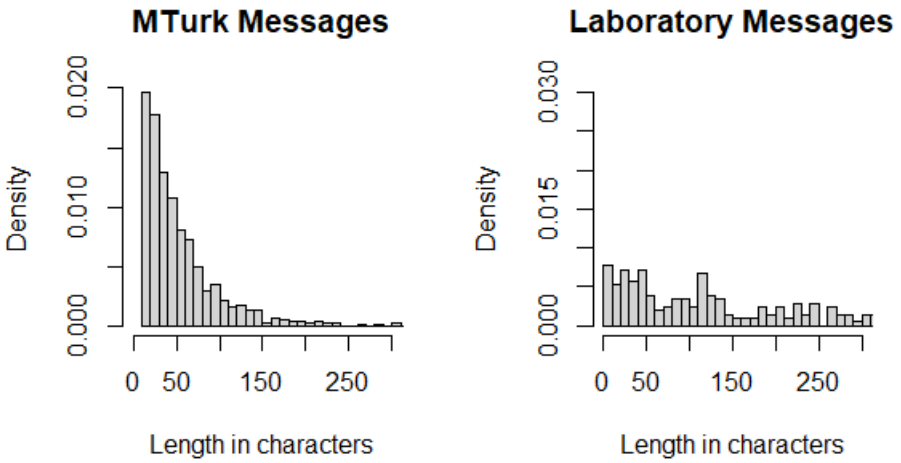
Table 3.1: Example Messages: Most and Least Effective

Type	Messages
Most Trusted	I'm going LEFT, so please select IN. This will maximize the earnings for both of us equally. choose IN so that we will both get 12 tokens Choose "In" and I'll make sure we both get 12 tokens! Hi there. I would understand why you would choose OUT, but I promise to make the choice ... I promise you that I will always choose "left" so that we both get 12 points each. Helloo!!! Let's make this work for both of us in the best possible way! Have a great day! Hey, I'm not a greedy person and in these situations, I always choose the most beneficial option for ... You can be sure that i will choose left, as it is the best option for both us. Please go ahead and select 'In' I'm good for the 12. I believe that there are real people here, and I'm completely down with splitting the 24 and ... I'm picking left so if you want a few extra tokens, go for in. Totally understand if you don't but that's my choice. I'm always going to choose "left" because I think it's dickish to screw another player over. I will maximize both our outcomes ie LEFT if you stay in.
Least Trusted	I'm going to choose Right i get 24 you get 0 U GO ON STRITE Hello Let's play the game I want choose Player A hi, how are you doing? Hey. let's work together, let's both choose out. I choose out choose out

A sample of messages trusted by all/none of their recipients.

In our data, the median message length is 40 characters with a mean of 54.5, so our messages are considerably shorter on average than those seen when similar procedures are conducted in a laboratory setting — median length of 110 and mean of 137.6, (Charness and Dufwenberg, 2006). Figure 3.3 presents the distribution of length in our sample compared to this laboratory setting

Figure 3.3: Message Length Distribution



A comparison of the distribution of length in our messages collected through MTurk and messages from a similar, in-person experiment reported in Charness & Dufwenberg (2006).

To ensure the validity of our data, we check that our subjects' behavior conforms with those in previous laboratory experiments. We manually label promises with the help of four research assistants. We label a message a a promise if three out of the four labeled a message as such. For a message to be classified as a promise we require not only that the message contain a statement of a future action, but also that is be coherent within the context of the game. Out of the hundreds of unique messages, a limited number of participants promised to do actions which are either impossible, i.e. "I will choose to be player B," or promised to play Right and take the entire pool for themselves. Although these are clearly promises in a grammatical sense, we do not include them in our classification. Out of 1004 total messages, 389 are promises. We also manually identify requests for In.

3.3.2 Tokenization and Lemmatization

Table 3.2: Most Common Tokens

Tokens	Frequency	Tokens	Frequency	Tokens	Frequency
i	751	so	131	'll	72
choose	586	u	131	best	72
in	498	player	120	please	69
left	393	's	116	more	67
you	365	that	114	maximize	66
to	359	can	106	point	65
will	341	make	102	our	59
be	338	out	97	each	56
the	291	hi	97	most	55
both	264	hello	96	game	54
we	253	12	95	all	54
and	244	play	94	b	52
a	235	do	88	always	51
go	176	token	86	money	51
if	175	'm	82	pick	51
get	174	fair	81	with	50
of	153	this	81	have	49
for	152	it	77	right	48
let	137	good	74	earn	46

To work with our text data, subjects' messages must first be processed into collections of tokens, each representing an individual or short series of words. With natural language, this preprocessing step presents significant challenges as we simultaneously want to make the problem tractable by reducing the number of unique words and retain as much of the richness of the original messages as possible. This preprocessing is conducted using the Natural Language Toolkit package for Python, (Loper and Bird, 2002). To do this, we utilize the WordNet lemmatizer, (Fellbaum, 2010) to reduce the various conjugations and declensions into their base forms. This lemmatizer performs best when the part of speech

of a words are included, so we also employ Treebank part of speech tagger (Marcus et al., 1993). The end result of these two algorithms is that each message is split into an ordered set of tokens, each token being a word and part of speech pair.

Usually, messages are further reduced by removing stop-words which provide little additional information about the contents of a document. This is a useful step when trying to categorize documents based on their subject matter, where the typical stop-words would be widely shared between the documents in a corpus. However, we omit this step as standard stop-word dictionaries would remove a great deal of information relevant to our specific environment. The messages in our corpus are largely all about the same topic – the game at hand. As such, our variation largely stems from different rhetorical approaches around the same topic. Words like "I", "we", and "you" would all be filtered out by standard stop-word dictionaries, but are clearly important in our environment. Table 3.2 presents the number of occurrences of the most frequent tokens in our data.

In addition to the collection of tokens representing each message, we construct a measure of the sentiment of each message. In stock market data, investor sentiment has predictive value of future returns (Tetlock, 2007). Using the OpinionFinder Subjectivity Lexicon (Wilson et al., 2005), we count the number of positive, negative, and total words in each message. This positive and negative word classification requires some minor adjustments to better fit our environment. For example, "right" is identified as a positive word in the sentiment dictionary. As this is an action label in our game, this is removed from the subjectivity dictionary.

3.4 Results

3.4.1 Baseline Approach: Literal Meaning

Our goal in this chapter is to examine whether using Natural Language Processing to analyse the content of participant messages yields any insights that would otherwise not have been obvious. To do so, we create a baseline message model which maps the free-form text messages to their literal meanings, based solely on the game's action space. Before proceeding with the NLP techniques, we first establish the validity of our baseline model by regressing subjects' choices onto dummy variables representing the possible literal meanings.

Table 3.3 below summarizes the results of both the Trust and Share (trustworthiness) models.

Table 3.3: Regressions on Baseline Messages

	Trust	Share
LEFT	0.219^{***}	0.225^{***}
	(0.015)	(0.039)
IN	0.149^{***}	0.041
	(0.018)	(0.046)
BOTH	0.253^{***}	0.186^{***}
	(0.016)	(0.041)
CONSTANT	0.462^{***}	0.532^{***}
	(0.009)	(0.022)
N	976	969

Regression coefficient estimates for the literal meaning baseline. $* : p < .05, ** : p < .01, *** : p < .001$

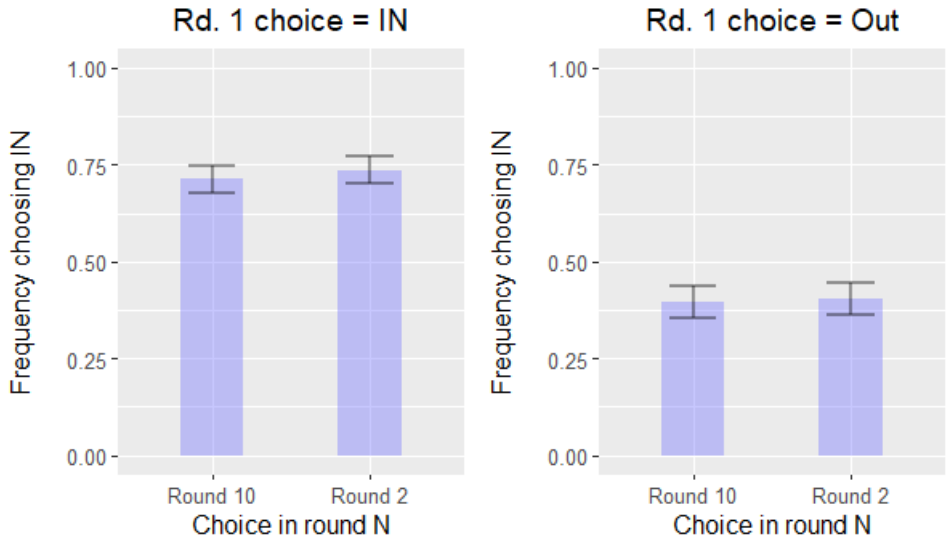
The results agree with what has been seen in previous studies involving communication. In the trust model, where the dependent variable is the frequency with which recipients of a given message choose the In action, statements of intention to choose Left by the Sender, requests that the recipient play In, and combinations of the two have significant positive effects relative to messages which contain none of these features. In the Share model, using

the likelihood that the Sender chooses Left after sending a given message as the dependant variable, we find that players who send messages promising to choose Left, sharing the tokens, are significantly more likely to do so. Requesting that the recipient of a message choose In without promising Left has no significant relationship with the Sender's behavior.

3.4.2 Data Validity: Order Effects

Additionally, we check to see if there are significant order effects based on prior messages received. To do this, we divide our subject pool into two groups based on their behavior as Receivers. One group includes all subjects who chose In in the first round, with the other including all who chose Out. We then compare the frequency with which each group chose In in round 2 and in round 10. The assumption is that if there are substantial order effects, the In/Out choice in the second round will be significantly correlated with the first round choice, however, the choice in round 10 will not be.

Figure 3.4: Order Effects



Participant choices in the second and tenth rounds, conditional on their decision in the first round. 95% confidence intervals are plotted in black.

We find no evidence of substantial order effects (presented in Figure 3.4). For both groups, the frequency of subjects choosing In is not significantly different between rounds two and ten. However, we do observe evidence of subject-level heterogeneity. Those who chose In (or Out) in the first round tend to do so throughout the duration of the game.

3.4.3 Regressions on Tokens

Table 3.4: Bag-of-words Model

Tokens	Trust	T-stat	Trustworthiness	T-stat
choose	-0.023	(-1.512)	-0.099*	(-2.259)
in	0.068***	(3.545)	0.036	(0.662)
left	0.060**	(3.278)	0.083	(1.608)
to	0.029*	(2.069)	0.088*	(2.187)
we	0.047*	(2.202)	0.138*	(2.306)
and	0.031*	(1.998)	0.046	(1.051)
go	-0.013	(-0.732)	-0.114*	(-2.326)
for	0.008	(0.398)	0.106*	(1.974)
u	0.048*	(2.007)	0.003	(0.043)
's	0.001	(0.038)	0.159*	(2.363)
that	0.042*	(2.043)	0.043	(0.742)
can	0.046*	(2.084)	-0.143*	(-2.275)
make	0.066**	(3.015)	0.010	(0.165)
out	-0.101***	(-4.933)	0.087	(1.509)
fair	0.052*	(2.316)	0.113	(1.793)
this	0.063**	(2.730)	0.099	(1.510)
maximize	0.067*	(2.247)	-0.148	(-1.692)
always	0.053*	(2.068)	0.136	(1.869)
right	-0.130***	(-4.614)	-0.224**	(-2.767)

Each coefficient is associated with a dummy variable indicating a message includes the relevant word. Only tokens with significant coefficient estimates are included. Control variables for messages' literal meaning are omitted from the table. $* : p < .05, ** : p < .01, *** : p < .001$

The literal meaning model of communication provides us with a reasonable baseline against which we can compare results using Natural Language Processing. In our first NLP model, we

augment our literal message model to include the top 50 most commonly used tokens in our message corpus. By using this model, with control variables included for the literal meaning categories in the baseline model, we examine which, if any, of the most commonly used words are associated with higher (or lower) than expected levels of trust and trustworthiness — the likelihood that recipients of a message choose In and senders choose Left, respectively.

Table 3.4 presents the results of the Trust and Share frequency regressed on the token dummy variables. As this model assumes that each message is an unordered collection of independent tokens, we refer to this as the bag-of-words model. For brevity, we omit tokens that are not significant in either of the two models from the table. The full results are included in the appendix.

Examining the Trust model first, the tokens with the most significant effects are those which correspond to action labels on the game tree, even after including the literal meaning controls. Each of these tokens effect on the likelihood the recipient of a message chooses In is in the expected direction with “In” and “Left” having positive effects significant at the 1% level and “Out” and “Right” having negative effects significant at the 1% level. In addition, there are several other tokens with significant, positive effects in the Trust model. “Make” and “This” are both significant at the 1% level and positive. The coefficient estimates for the tokens “To”, “We”, “And”, “u”, “That”, “Can”, “Fair”, “Maximize”, and “Always” are all positive and significant at the 5% level.

With less statistical power from fewer observations, there are fewer significant tokens in the Share model. In the Share model, the only Action label with a significant coefficient estimate is “Right” which has a significant, negative effect at the 1% level. In addition, “choose” “go” and “can” are negatively associated with the likelihood senders choose Left and significant at the 5% level. The coefficient estimates for “To,” “We,” “For,” and “ ’s,” are positive and significant at the 5% level.

This approach shows that there is indeed additional information relevant to predicting play-

ers' choices remaining even after considering the bare meaning of the messages. For many of these tokens, the interpretation of their coefficient estimates is clear. The significance of the Action labels is, perhaps, somewhat unsurprising. However, their significance even after controlling for promises to choose Left and requests to choose In suggests that these sort of statements are most effective when they directly and unambiguously reference the intended or requested action.

For the remaining significant tokens, properly interpreting the results requires us to return to our message corpus to examine the context in which they appear.

Having only one decision associated with each message (instead of ten in the Trust model) we find that there are fewer tokens significantly associated with changes in the likelihood their authors chose Left. Of the action labels present in the game, Right, is the only one with a significant coefficient estimate in the Share model, predicting a much lower likelihood of the sender choosing Left. Although the negative coefficient is coherent with the messages which contain the Right token, it is rather unexpected. Stating an intention to choose Right does not make sense in the usual understanding of the trust game wherein the message senders' attempt to persuade their partners that they are likely to choose Left. Given the incongruity between stating intent to choose Right and the usual strategy of persuasion, it is tempting to write these messages off as mistakes, written by participants who did not understand the game's instructions. However, this would not explain why these Senders follow through on their promises to choose Right. This, instead, suggests that there is a subset of participants who are strongly averse to lying but not necessarily averse to unfair outcomes.

The four tokens with positive coefficient estimates in the Share model fall into two categories. The first of these contains the plural first-person pronouns "We" and " 's" which is used as for contractions containing "us" almost exclusively. The remaining two positive share coefficient tokens are "To" and "For" which are primarily used in messages that detail a reason why the recipient should trust the sender.

Apart from "Right," the remaining significant tokens in the Share choice model do not have readily apparent interpretations in the trust game. Interpreting these tokens, "Go," "Can," and "Choose," requires that we examine them in the context in which they appear. Messages using the "can" token are typically used by Senders to provide a reason why the Receiver should choose In by describing the outcome. A typical message of this type ends with "... so we can get the most tokens." "Go" messages are used to convey relatively weak promises, using more casual language than other promises. "Choose" messages, although often used in promises, are more often used as commands for the Receiver to choose In.

Comparing the results in each model, we do not find significant discrepancies that indicate effective deception. Tokens indicative of effective deception would be those that have significant, negative coefficients in the Share model but positive, significant coefficients in the Trust model. Only one token, "Can," fits this definition. Given the large number of regressors in the model, this is indistinguishable from random chance. Therefore, we draw no conclusions from this regression model about potential signs of effective deception.

3.4.4 Cluster Analysis

What the above analysis shows, in short, is that there are features in our corpus of messages that have strong correlations with both trust and trustworthiness choices even after controlling for literal meaning. With this approach — running OLS regressions on dummy variables for common tokens — we establish an initial result that literal meaning does not capture all the useful information contained in this free-form text data. These significant tokens suggest new classes of messages worth further investigation. However, this method of treating each message as linear combinations of independent words leaves out much of the complexity present in natural language.

To address this shortcoming, we employ cluster analysis to identify groups of similar messages which we may then aggregate and examine further. Cluster analysis produces groups of

messages such that the within-group similarity is greater than the inter-group similarity. The specifics of this task vary greatly based on the chosen measure of similarity and objective function comparing intra- and inter-group similarity. This permits us a great deal of freedom to choose a similarity measure which captures more of the complexity of natural language than in our previous approach.

Previous authors have employed cluster analysis on laboratory data including subject choices (Ellingsen et al., 2018; Grimm et al., 2021; Woods et al., 2022) and eye tracking data (Polonio and Coricelli, 2019; Devetag et al., 2016).

Our approach involves two steps. First, we employ Affinity Propagation, (Frey and Dueck, 2007) to perform our cluster analysis and group our messages. Affinity Propagation allows for both the number of clusters and the number of elements assigned to each to vary, which we require in the absence of ex ante knowledge of the shape of the message space. Flexible cluster size also allows non sequitur messages to separate themselves into smaller clusters instead of forcing themselves into unsuitable clusters to achieve an arbitrary number of groups. Each cluster also forms around an exemplar message, giving us a natural starting point when we manually interpret and label each cluster. The Affinity Propagation algorithm is described in Appendix A.

Following this aggregation, we then employ standard OLS, fitting a linear probability model to test which areas of our message space have significantly higher rates of trust (greater likelihood Receivers choose In) and trustworthiness (greater likelihood Senders choose Left). As with our bag-of-words token regressions, we add a manually coded control variable for literal meanings.

Conducting cluster analysis with affinity propagation requires us to define a similarity matrix $\mathbf{S}$ for the preference of each element for potential exemplars. Given this degree of flexibility, we conduct cluster analysis with two different similarity measures, one using solely the information in our corpus of text (corpus-only approach) and the other informed by outside

natural language knowledge (knowledge-based approach). By using multiple similarity measure, we test the robustness of our conclusions against changes in our clustering procedures. If cluster analysis is a useful tool for us to process free-form laboratory text data, we expect the results to be relatively consistent across these choices.

In the Corpus-only approach, we define our similarity matrix $\mathbf{S}_J$ by taking each message, again, as an unordered collection of tokens and calculating the Jaccard index of overlap between sets. For messages m_i and m_k, both being sets of tokens:

$$s(i,k)_{\,i\neq k} = j(m_i, m_k) = \frac{|m_i \cap m_k|}{|m_i \cup m_k|} \; ; \; s(k,k) = \text{Median}(\mathbf{J})$$

Applying this measure to each pair in our corpus yields a complete graph in matrix form; each element $j(i,j)$ representing the similarity of message i to message j and vice versa; a higher value implying greater similarity.

3.4.5 Results: Affinity Propagation Using Corpus-only Similarity Measure

Affinity Propagation identifies 112 clusters of similar messages in our corpus. The median cluster size is 8 messages with a mean of 9.1. Figure 3.5 presents the distribution of the number of messages assigned to each cluster.

We conduct cluster analysis with the goal of identifying key features in our corpus, without a priori knowledge of the number of expected clusters or their contents. Instead, we must investigate each cluster and assess whether or not the clusters share a common meaning to a sufficient degree. We must also manually assign a label to each cluster, describing the common theme shared among the messages assigned to it, using the exemplar element for each cluster as our starting point. In Table 3.5 below, we present two example clusters

and their exemplars, demonstrating the similarity between the messages grouped by affinity propagation.

Figure 3.5: Distribution of Cluster Sizes: Jaccard Distance

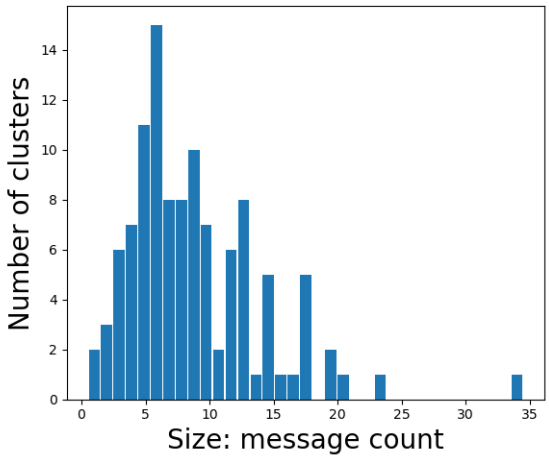


Table 3.5: Example Clusters

Cluster	Message (**exemplar**)
55	**I believe in fair play**
55	I believe in equality
55	I am playing in
55	I believe in playing fair and having an equal outcome.
55	I believe in fair, even-steven financial deals. I like a win-win situation. ...
55	I believe in being fair with people.
58	**Hey, I want to make sure we both get the highest amount of bonus!...**
58	Hopefully we both want to get the highest points possible
58	I will make sure we get the same amount of tokens
58	Hello. I hope we both make a good amount of bonus money today! Good luck!
58	Hi there! I always split these 50-50 because I would want someone to do the ...
58	Hello. Lets make this to where we both can get the largest bonuses out of th...
58	To make sure you guys get a higher bonus, please choose IN in each step so w...

In both example clusters, the elements are sufficiently similar to lead to a clear substantive interpretation. Cluster 55 messages state that the sender believes in fairness, which is cap-

tured well by the exemplar message, "I believe in fair play." Messages in cluster 58 state the author's hopes for the game outcome, speaking in terms of the amount of tokens. Again, the exemplar, "Hey, I want to make sure we both get the highest amount of bonus..." represents the typical message in the cluster well.

Regressing Trust Choices on Clusters

Our primary research question is whether free-form message text contains additional useful information after accounting for a message's literal meaning. To do so, we fit an OLS regression model regressing players' choices (Receivers' frequency choosing In and Senders' likelihood of choosing Left following a given message) on dummy variables for each of the four possible literal meanings. Following this, we regress the residuals from the literal meaning regressions on dummy variables for cluster assignments. Using this two step approach, the coefficients and t-statistics reported are for tests for heteroskedacity in the literal meaning model. Messages in clusters with significant, positive coefficient estimates are trusted (trustworthy) more often than predicted by literal meaning alone.

We first examine the most effective messages in gaining the recipient's trust. For each cluster, we include both the exemplar message and a brief description of the contents of the common themes.

Again, we are cautious of over-inference, and avoid delving too deeply into any individual cluster. Instead, we focus on the common features across these groups of highly effective messages. The most effective messages largely fall into two broad categories. First, two of the top ten clusters (numbers 78 and 45) follow a common three part structure — explicitly requesting the receiver plays In, promising to choose Left, and stating the positive result should both these choices be made. These messages conform with our definition of a promise, but go a step further in describing the outcome and requesting specific action from the Receiver. An additional two clusters (19, 58, and 96) do not follow this three-part formula

Table 3.6: Most Trusted Corpus-only Clusters

Cluster	**Description** Exemplar Message	Size	Trust (t-stat)	Share (t-stat)
84	**Promises 'what's best for everyone'**	13	0.19	0.01
	Hi, I will make the choice that is beneficial for all of us :)		(4.2)	(0.06)
78	**Requests In w 'please', promises Left, specifies '12 token' outcome**	13	0.16	-0.07
	Please choose "IN" so that we can both get 12 points. I will choose "LEFT."		(3.48)	(-0.51)
96	**Requests In so both can 'maximize'**	9	0.19	-0.18
	If you choose In we can both maximize our tokens.		(3.19)	(-1.09)
58	**Suggests that both 'get the highest bonus'**	7	0.19	-0.03
	Hey, I want to make sure we both get the highest amount of bonus! ...		(2.96)	(-0.19)
40	**"Let's try and be fair"**	6	0.20	0.06
	lets try and be fair		(2.95)	(0.29)
103	**Promises with justification. Also states explicit outcome**	15	0.11	0.16
	I want to be as fair as possible for both of us. I am not greedy ...		(2.6)	(1.31)
81	**Promises 'most for both of us'**	12	0.12	0.32*
	Trying to make the most for both of us		(2.49)	(2.38)
45	**Request In, promise Left, & specifies outcome for 'both of us'**	18	0.10	0.11
	Hello, I will chose left if you decide to go in. I want this to be a win/win ...		(2.45)	(0.95)
75	**A greeting plus a promise of the 'most' or 'best' outcome**	5	0.18	0.06
	Hey! Let's try and make as much as we can for everyone!		(2.43)	(0.29)
19	**Promises left with an explicitly stated 'reason to trust'**	5	0.17	0.09
	I know you have to trust me, but I promise to pick left for both our benefits.		(2.33)	(0.45)

The top ten most significant, positive coefficients when regressing Trust on cluster dummies, controlling for promise labels and length, along with the corresponding coefficient estimates in our Share model. Clusters are described with a manual description in bold as well as each cluster's exemplar message chosen by the Affinity Propagation algorithm. $* : p < .05, ** : p < .01, * * * : p < .001$

but are also detailed promises.

Second, four of the ten clusters are 'soft promises' (clusters 103, 40, 81, 84). These messages fail to meet our definition of a promise, requiring the author to makes a clear statement of a future intended action. Instead, they make softer claims: "Let's try to be fair," for example. Although these are not strictly promises in the grammatical sense, the implication is clear in the context of the game. Recipients of these messages interpret them as statements of intent to choose Left even without an explicit action indicated.

Table 3.7: Least Trusted Corpus-only Clusters

Cluster	Exemplar Message	Size	Trust (t-stat)	Share (t-stat)
97	choose out	6	-0.29	0.05
			(-4.29)	(0.24)
32	choose right!	10	-0.21	-0.34*
			(-4.05)	(-2.27)
41	I chose out.	12	-0.18	-0.08
			(-3.50)	(-0.59)
34	hello..let's play	10	-0.15	0.13
			(-2.62)	(0.80)
1	Please Choose In	15	-0.10	-0.02
			(-2.20)	(-0.13)
86	i will be fair	10	-0.11	-0.01
			(-2.03)	(-0.07)
29	I think you should choose In	8	-0.12	-0.34*
			(-2.02)	(-2.03)
21	I will choose Left	24	-0.07	0.00
			(-2.01)	(0.04)

Significant, negative coefficients when regressing Trust on cluster dummies. Controls for literal meaning and length are omitted from the table. Clusters are described with each cluster's exemplar message chosen by the Affinity Propagation algorithm. For each cluster, the exemplar message adequately represents the messages included. $* : p < .05, ** : p < .01, * * * : p < .001$

None of the increased rates of trust in these clusters have a corresponding significant increase in the author's trustworthiness. However, we cannot reliably interpret this as evidence of

effective deception. Our design focuses on the In/Out decision and thus has lower statistical power to estimate Left/Right coefficients; each message has ten readers but only one author.

Examining the least effective clusters (those with the most significant negative coefficients in the Trust regression) we can again identify two broad themes. As with the most effective clusters, we also present each cluster's coefficient in the Share regression, modeling the Senders' choices given the message they sent.

Table 3.8: Significant Corpus-only Clusters: Share

Cluster	**Description** Exemplar Message	Size	Share (t-stat)	Trust (t-stat)
79	**"I like this game"**	8	0.46	0.06
	I like this game.		(2.77)	(1.07)
35	**Suggestion to 'work together'**	11	0.37	0.10*
	Hiya. I hope we can work together.		(2.61)	(1.98)
81	**Promises 'most money for both of us'**	12	0.32	0.12*
	Trying to make the most for both of us		(2.38)	(2.49)
55	**"I believe in fair play"**	6	0.46	0.16*
	I believe in fair play		(2.38)	(2.3)
9	**"Good player"**	8	-0.29	-0.03
	GOOD PLAYER		(-1.73)	(-0.47)
29	**"I think you should choose In"**	8	-0.34	-0.12*
	I think you should choose In		(-2.03)	(-2.02)
61	**"Going left"**	6	-0.43	-0.06
	Going left		(-2.26)	(-0.94)
32	**"choose right!"**	10	-0.34	-0.21***
	choose right!		(-2.27)	(-4.05)
52	**Appeal to 'maximization'**	9	-0.45	0.02
	let's maximize OUR bonuses		(-2.84)	(0.43)

The most significant, positive coefficients when regressing Share on cluster dummies. Controls for literal meaning and length are omitted from the table. Clusters are described with a manual description in bold as well as each cluster's exemplar message chosen by the Affinity Propagation algorithm. $* : p < .05, ** : p < .01, *** : p < .001$

As suggested by the token regressions, Receivers distrust messages that seem to go against the Senders' best interests (clusters 32, 97, and 41). There is no immediately apparent

explanation for sending a message stating your intent to play Right or asking the reader to choose Out, although these messages do not violate the rules of the game.

The second theme among the most distrusted messages is newly revealed by our Cluster Analysis; brief messages containing greetings but no other content (cluster 34) have highly significant, negative coefficients after controlling for literal meaning. The remaining message clusters (1, 86, 29, and 21) are all the briefest possible messages one can send that have a meaning corresponding to game actions. Compared to messages with the same literal meaning, their recipients are significantly less likely to choose In.

Regressing Share choices on clusters

Moving to the Senders' Left/Right choices, we fit an OLS model, using dummy variables for each message's cluster assignment and human-coded literal meaning as independent variables to estimate the likelihood that its author chooses to share the tokens (choose Left in step 2). As a consequence of our experiment's design, this model has significantly less statistical power than our Trust model; there are ten trust choices for each message but only one associated Share choice. We focus primarily on relating the Share coefficient estimates to their Trust model counterparts, asking whether Receivers' choices are rational responses to accurate predictions about Senders' behavior. Table 3.8 presents clusters with the most significant positive and most significant negative Share model coefficient estimates.

The clusters with the significant coefficient estimates in the Share model repeat themes seen in the Trust model. The majority of clusters with significant, positive coefficient estimates are soft promises (clusters 35, 55, and 81). The Sender implies they will be trustworthy, but does not explicitly state an intended action. Additionally, messages in which the Sender states that they "like this game" are significantly correlated with choosing Left. Every Sender of messages in cluster 79 chose Left in step 2.

Also similar to the Trust model, the messages in these least trustworthy clusters are almost

all noticeably lower quality than the typical message in our corpus. This includes both nonsensical messages like those in cluster 32 and the briefest possible promises or requests in cluster 61 and 29. The cluster with the strongest negative coefficient estimate is the exception to this pattern of poor quality — cluster 52, which contains explicit appeals to payoff maximization. Unlike the low-quality message clusters, there is no corresponding significant negative coefficient in our Trust model. Recipients of these appeals to maximization continue to choose In, entrusting the tokens to their partner at the same rate as other messages with the same literal meaning.

Considering both the Trust and Share models formed from our corpus-only clusters, we find five broad classes of messages which warrant further study. Our primary goal is to study the determinants of trust and we find two themes associated with increased trust and two associated with decreased likelihood of trust. Both detailed promises — explicitly promising an action, giving their partner a call to action, and detailing the promised outcome — and soft promises (e.g. "I think fairness is important,") are associated with increased trust. These messages tend to have corresponding positive coefficient estimates in out Share model, however, we cannot make strong claims about this agreement due to the Share model's relative lack of statistical power. In contrast, Receivers trust poor-quality messages and brief greetings significantly less often. Our coefficient estimates in the Share model generally support their skepticism, with estimates agreeing in direction if not also in significance.

The final notable cluster, appeals to maximization, provides our only evidence of possible effective deception. Despite our Share model estimates indicating that these messages are the least likely to be trustworthy, we find no significant effect in our Trust model, despite it having much greater statistical power.

3.4.6 Affinity Propagation Using Knowledge-based Similarity

In the preceding section, we show the potential of cluster analysis as a tool for making sense of unstructured text data. From over 1000 messages, Affinity Propagation successfully aggregates individual messages into clusters which have substantial within-cluster similarly. These clusters allow us to conduct careful, limited inference and identify previously unknown message features which impact their recipients' choices.

However, these results are not without caveats. The clusters themselves are sensitive to our choice of similarity measure and other tuning parameters in the affinity propagation model. The Jaccard index as a similarity measure is both easily interpretable and based solely on the text contained in our message corpus. This simplicity makes it an attractive baseline similarity measure but, in choosing this measure, we ignore the large body of existing Natural Language Processing work. To illustrate this weakness, consider the brief greeting clusters, 34 and 4 in the above model (exemplar messages "Hello, let's play" and "Hi, how are you?" respectively.) Although the within-cluster similarity is satisfactory, the division between the two is based solely on the use of "Hi" versus "Hello" to convey nearly identical information. We make no claim about this specific partition being the definitive description of the message space, but this over-clustering raises concerns about whether our results are robust to our specification of similarity measure.

We test the robustness of our corpus-only Affinity Propagation results by conducting the same clustering and regression methods but using an alternative similarity. This Knowledge-based approach is trained on existing information outside of our corpus of messages. In this second approach, we measure the similarity between two messages using an approach first introduced by (Mihalcea et al., 2006) to compare short text documents. For two sets of

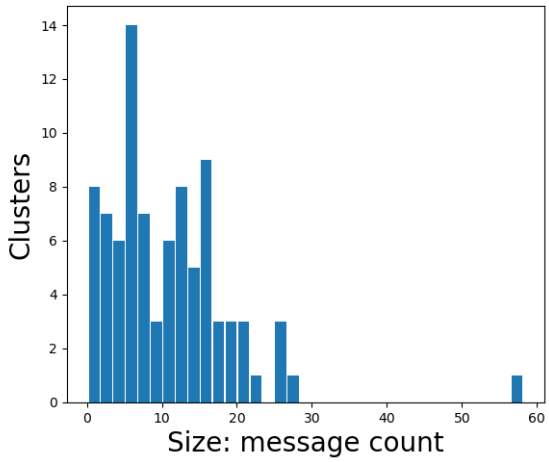
words T_1 and T_2, the MCS similarity is:

$$S_{mcs} = \frac{\sum_{w \in T_1} \text{maxsim}(w, T_2) \cdot \text{idf}(w)}{2 \sum_{w \in T_1} \text{idf}(w)} + \frac{\sum_{w \in T_2} \text{maxsim}(w, T_1) \cdot \text{idf}(w)}{2 \sum_{w \in T_2} \text{idf}(w)}$$

maxsim(w,T) is the maximum semantic similarity between a word w and any word in the sentence T, the precise details of which vary depending on the choice of semantic similarity measure. We measure word-to-word semantic similarity with the Google word2vec neural network, (Mikolov et al., 2013). After training, word2vec takes individual words as inputs and outputs a vector representation reduction of latent meaning. The network we implement here is pre-trained on a corpus of articles available through Google News, using the Gensim python package (Rehurek and Sojka, 2011). The resultant neural network encodes individual words into a 300-dimension vector; the similarity between any two words is the cosine distance between their word2vec vector forms.

idf(w) is the inverse document frequency for word w, $idf(w) = ln(\frac{N}{n(w)})$ where N is the total number of messages in our corpus and $n(w)$ is the number of messages containing w (Sparck Jones, 1972). A higher idf value for a word implies greater specificity and thus receives greater emphasis when calculating similarity.

Figure 3.6: Distribution of Cluster Sizes using MCS-w2v Distance



After calculating the similarity matrix using the MCS and w2v, we again apply affinity propagation, resulting in 89 unique clusters. The mean size is 11.3 messages with a median size of 10. Figure 3.6 shows the distribution of cluster sizes.

Table 3.9 presents two example clusters formed by Affinity Propagation using the MCS-w2v measure of message similarity. As in the corpus-only approach, there is a high degree of within-cluster similarity around a common theme. For the two selected examples, cluster 3 and 59, we label them as greetings and emphatic promises to choose Left respectively. As before, the exemplar elements serve as reliable summaries of each cluster's contents. We also observe that the greater flexibility afforded by the Knowledge-based similarity measure achieves the desired result of softening the boundaries between clusters while maintaining within-group similarity. Cluster 3 now includes short greetings regardless of whether they use "Hi" or "Hello."

With the new, reduced-size clusters, we again fit Ordinary Least-Squares models, regressing Trust and Share choices on dummy variables for each message's literal meaning. Following this, we regress the residuals of the Trust and Share models on dummy variable for each cluster. Our goal is to assess the robustness of our previous approach and examine whether the same communication strategies are correlated with changes in behavior across the two partitions of the message space. Table 3.11 and Table 3.10 present the most significant positive and most significant negative coefficient estimates in the Trust model.

Compared to the Corpus-only model, we find fewer clusters with significant, positive coefficients the Knowledge-based Trust model. However, the same message features stand out as the most effective. The clusters significant at the 5% level follow the same two patterns we observed in the previous model. We see detailed promises represented in cluster 39 and soft promises in cluster 31 and 15. Additionally, we see clusters featuring suggestions to cooperate without explicitly promising an action or requesting one from its recipient (clusters 64, and 66). Finally, cluster 32, which includes appeals to 'maximization' has a significant, pos-

Table 3.9: Example MCS-word2vec Clusters

Cluster	Message (**exemplar**)
3	**Hi, how are you?**
3	a bit confussed but attemp either way. hi
3	HELLO! HOW ARE YOU?
3	Are you ready?
3	Hello, what are you going to choose?
3	hi how are you
3	HI HOW ARE YOU
3	hi hoe are you
3	hi, how are you?
3	hello how are you doing
3	Hello, how are you doing?
59	**I will always choose LEFT.**
59	I will always choose Left. I want to split the bonus in a fair way.
59	Hello. Let's try to maximize our earnings. I will always choose LEFT as I be...
59	I will always choose LEFT. That means we are both assured of getting a fair ...
59	I always split evenly with my fellow turkers.
59	I will always play Left if you play In.
59	I will choose Left. I always try to split earnings with other workers when I...
59	I'll always choose Left.
59	I promise to always play left so we all benefit.

Table 3.10: Least Trusted MCS-word2vec Clusters

Cluster	Exemplar Message	Size	Trust (t-stat)	Share (t-stat)
80	I choose out	16	-0.20	0.02
			(-4.49)	(0.14)
27	choose right!	15	-0.19	-0.40**
			(-4.17)	(-3.13)
61	it really good and nice	6	-0.21	-0.20
			(-2.9)	(-1.06)
42	out player a-8 player -b 8	7	-0.16	0.32’
			(-2.41)	(1.79)
74	Please choose IN.	27	-0.08	0.01
			(-2.37)	(0.13)
85	I AM PLAYER A	8	-0.15	-0.16
			(-2.37)	(-0.98)
3	Hi, how are you?	21	-0.08	0.05
			(-2.06)	(0.47)

Significant negative coefficient estimates when regressing In/Out choices on cluster dummies using a knowledge-based measure of similarity. Controls for literal meaning and length are omitted from the table. Clusters are described with each cluster’s exemplar message chosen by the Affinity Propagation algorithm. $* : p < .05, ** : p < .01, *** : p < .001$

Table 3.11: Most Trusted MCS-word2vec Clusters

Cluster	**Description** Exemplar Message	Size	Trust (t-stat)	Share (t-stat)
31	**State intent to do ‘best for both of us’**	26	0.11	0.09
	I’ll do the best for both of us!		(3.12)	(0.93)
66	**“Let’s work together”**	15	0.11	0.13
	Lets work together		(2.53)	(1.06)
64	**Suggest ‘both earning as much as possible’**	5	0.18	0.06
	Hey! Let’s try and make as much as we can for everyone!		(2.33)	(0.29)
32	**Appeal to ’maximization’**	6	0.10	-0.35**
	lets maximize		(2.08)	(-2.79)
39	**Propose ‘maximizing both’ player’s earnings**	5	0.09	-0.06
	If you choose IN, I will choose LEFT to maximize both of our earnings.		(1.96)	(-0.53)
15	**“I believe in equality”**	5	0.17	0.35
	I believe in equality		(1.94)	(1.49)

Significant positive coefficient estimates when regressing In/Out choices on cluster dummies using a knowledge-based measure of similarity. Controls for literal meaning and length are omitted from the table. Clusters are described with each cluster’s exemplar message chosen by the Affinity Propagation algorithm. $* : p < .05, ** : p < .01, *** : p < .001$

itive association with trust. Notably, this cluster also has a significant, negative correlation with the likelihood their authors choose Left. Cluster 32 is the only cluster in our model with significant coefficients in both models in opposing directions.

The least trusted messages in the knowledge-based model are effectively identical to those found in the corpus-based model. The bulk of these clusters contain messages that suggest the author does not understand the game. They state intent to either choose actions that are impossible within the bound of the game, e.g. "I choose player A," or state their intention to choose Right. Additionally, we again see that recipients of short, simple greetings are significantly more likely to distrust their sender, even relative to other non-promise messages.

In the Share model (Table 3.12) we see once again that the broad themes from our model using a knowledge-based measure of similarity match with the previous, corpus-only approach. The clusters associated with higher likelihoods of their senders choosing Left are detailed promises (cluster 73) specifying an outcome should the recipient choose In.

The clusters with significant, negative coefficients also support our conclusions from the Corpus-only approach. Senders who state their intention to play Right tend, strongly, to do so and recipients respond appropriately to these messages by choosing Out. Our only significant evidence of effective deception, explicit appeals to payoff maximization, is also robust to the alternative measure of similarity. The authors of these messages choose Right 34 percentage points more often, however, there is no significant change in their recipients' In/Out choices.

Overall, the conclusions from our knowledge-based approach agree with the previous, Corpus-only method. Although the resulting partitions are, indeed, sensitive to the choice of similarity measure, the same message features stand out in both the Trust and Share models. Receivers of emphatic, detailed promises are significantly more likely to choose In, relative to the typical promise. Similarly, soft promises, which imply an intention to choose Left without an explicit statement of intent, yield greater trust than other non-promise messages.

Table 3.12: MCS-word2vec Clusters: Share

Cluster	**Description** Exemplar Message	Size	Share (t-stat)	Trust (t-stat)
73	**Promises Left conditional on In & Specifies the outcome** If you choose in, I will choose left.	59	0.12* (1.97)	0.00 (0.02)
32	**Appeal to Maximization** lets maximize	14	-0.35 (-2.79)	0.10* (2.08)
27	**Promise to choose Right** choose right!	15	-0.40 (-3.13)	-0.19*** (-4.17)

The most significant coefficient estimates when regressing Left/Right choices on cluster dummies using a Knowledge-based measure of similarity. Controls for literal meaning and length are omitted from the table. Clusters are described with a manual description in bold as well as each cluster's exemplar message chosen by the Affinity Propagation algorithm. $* : p < .05, ** : p < .01, *** : p < .001$

With less statistical power, the Share model provides less insight into Senders' behavior, with two coefficient estimates significant at the 5% level. This includes both promises to choose Right and appeals to payoff maximization, both identified in the Corpus-only model. Despite the strong, negative coefficient estimate in the Share model, Receivers are more likely to trust an appeal to payoff maximization than other messages with the same literal meaning. There are no clusters with significant, positive coefficient estimates.

3.4.7 Reducing Cluster Size

Given the size of our corpus, containing a little over 1000 unique messages, having nearly one cluster for every ten messages raises concerns of over-clustering. We again reiterate that our purpose requires only that the within cluster similarity is sufficiently high. Our purpose is to identify new, relevant message features, not to determine a definitive partitioning of the message space which captures all of the potential complexity and nuance of written communication. Still, with these cluster sizes being implications of largely arbitrary rule-of-thumb parameter choices, we want to examine whether our results are robust to different tuning parameter decisions.

To address this concern, we conduct the same cluster analysis followed by OLS regression procedure as above except that we adjust our tuning parameter, a message's affinity towards itself, to reduce the number of resulting clusters. Specifically, we select this parameter with the objective of maximizing the BIC of the resulting linear model independently for both the Coprus-only and word2vec similarity measures.

This optimization reduces the number of clusters in the Corpus-only and word2vec model to 27 and 16 down from 112 and 89 respectively. As before, using cluster assignments as dummy variables, we fit a linear probability model, estimating the frequency recipients choose Left, trusting the sender, and the frequency with which senders choose Left, sharing the tokens if given the opportunity. Figures 3.13 and 3.14 present the coefficient estimates for the Corpus-only and Knowledge-based models, respectively.

3.4.8 Reduced Corpus-only Clusters

Out of the 27 clusters grouped using only information found in our corpus, 6 clusters (6, 10, 11, 18, 20, & 25) have positive, significant coefficient estimates in the trust model. We again see the two common themes from the previous cluster regressions. Clusters 10, 11, 18, and 25 include detailed promises, requesting In and specifying an outcome in addition to the stated intention to choose Left. The remaining two — clusters 6 and 20 — contain suggestions to work together, implying both a request for In and an intent to play Left without explicitly stating this.

In contrast, five clusters are associated with lower trust frequencies among their recipients (clusters 0, 1, 3, 9, and 12). These messages fall into two distinct themes. The first group consists of errors, messages that are either entirely nonsensical or that are not in line with a basic understanding of the game structure (clusters 3 and 9). Additionally, clusters 0 and 12 contain the briefest possible requests that the recipient choose In. The remaining cluster (cluster 1) contains brief greetings with no other content. Brief greetings and nonsense

Table 3.13: Reduced Size Corpus-only Clusters

Cluster	Size	**Description** Exemplar Message	Trust (t-stat)	Share (t-stat)
0	44	**Brief request for In**	-0.06*	0.03
		Please Choose In	(-2.07)	(0.45)
1	21	**"Hi, how are you?"**	-0.08*	-0.06
		Hi, how are you?	(-2.07)	(-0.61)
3	18	**"I choose right"**	-0.18***	-0.30**
		I choose right.	(-4.37)	(-2.70)
5	22	**Good luck / good morning**	-0.04	-0.25*
		good luck!	(-1.15)	(-2.52)
6	27	**Suggestions to 'work together'**	0.12***	0.20*
		I hope we can work together to win.	(3.57)	(2.28)
9	25	**"I choose out"**	-0.20***	0.04
		I chose out.	(-5.46)	(0.41)
10	44	**Request In, promise Left, & specify the 'maximum/best' outcome**	0.06*	0.03
		If you choose IN, I will choose LEFT to maximize both of our earnings.	(2.40)	(0.45)
11	103	**Request In, promise Left, & specify '12 token' outcome**	0.05**	0.05
		If you choose In I will choose Left and we will both get 12 tokens	(2.84)	(1.01)
12	37	**Brief promise of Left and request for In**	-0.07*	-0.11
		Choose IN. I'll choose LEFT.	(-2.33)	(-1.47)
17	83	**Promise Left conditional on In**	-0.01	0.14**
		I will be choosing left if you choose in	(-0.74)	(2.75)
18	64	**Request In, promise Left, specify '12 token' outcome & uses 'please'**	0.05*	-0.11'
		Please choose "IN" so that we can both get 12 points. I will choose "LEFT."	(2.22)	(-1.89)
19	18	**"I like this game"**	0.02	0.34**
		I like this game.	(0.41)	(2.92)
20	42	**States intent to make the 'most for both of us'**	0.09***	0.25***
		Trying to make the most for both of us	(3.39)	(3.47)
22	24	**Suggests 'best payoff' without a request or promise**	0.01	-0.24*
		ALL THE BEST	(0.16)	(-2.36)
24	34	**"I am Player A"**	-0.03	-0.18*
		I AM PLAYER A	(-1.04)	(-2.25)
25	48	**Request In, promise Left, and specifies 'most money'**	0.06*	0.06
		You should choose IN so I can choose LEFT and we both make the most money.	(2.30)	(0.93)

Coefficient estimates when regressing Left/Right and In/Out choices on cluster dummies using the corpus-only measure of similarity. Controls for literal meaning and length are omitted from the table. Clusters are described with a manual description in bold as well as each cluster's exemplar message chosen by the Affinity Propagation algorithm. Only clusters with significant coefficient estimates in either model are shown for brevity. $* : p < .05, ** : p < .01, *** : p < .001$

messages are also found in the previous cluster specifications.

In the reduced corpus-only Share model, we largely find agreement between the two models, with the signs of significant coefficient estimates in the Share model agreeing, at least in direction with those in the Trust models. There are no clusters which have coefficient estimates significantly different from zero in the Share model with significant, opposing coefficient estimates in the Trust model.

In the Share model, four clusters have significant positive coefficient estimates. Clusters 6 and 20 both consist of suggestions to "work together" or statements of intent to do so. Cluster 17 includes promises to choose Left, conditional on the Receiver's choice to choose In. This conforms to the previous results showing that detailed promises outperform simple ones. Finally, messages in which the Sender states that they "like this game" (cluster 19) are significantly more likely to come from trustworthy senders than other messages with no mention of game actions.

Negative, significant clusters in the reduced, corpus-only Share model also agree with the results of the previous approaches. Nonsensical messages, promising to choose Right or to choose Player A are significantly, negatively associated with their senders choosing Left (clusters 3 and 24). Senders of messages in cluster 22 state their intent to "maximize" and are significantly more likely to choose Right.

3.4.9 Reduced word2vec Clusters

Clustering using the word2vec similarity measure yields fewer clusters compared to the Corpus-only approach. As intended, this approach mitigates some of the potential over-clustering seen in the first approach and we have fewer groups of closely related messages divided into multiple categories. After this reduction, the same general themes remain common among the messages in clusters with significant coefficient estimates, both positive and

Table 3.14: Reduced MCS-word2vec Clusters

Cluster	Size	**Description** Exemplar Message	Trust (t-stat)	Share (t-stat)
0	43	**Request for In including 'please'**	-0.05'	0.05
		Please Choose In	(-1.89)	(0.66)
1	46	**"I choose Player A"**	-0.09***	0.09
		I choose player A	(-3.48)	(1.31)
2	140	**Requests In, promises Left, & specifies the 'best' outcome**	0.02	0.04
		I will choose left if you choose IN	(1.56)	(1.01)
3	26	**Brief Promises for Right**	-0.14***	-0.36***
		choose right	(-3.96)	(-3.81)
4	115	**Brief promises for Left only**	-0.02	-0.03
		i will choose left	(-0.97)	(-0.57)
5	113	**States intent for what's 'best for everyone'**	0.05**	0.01
		Hello, I want to do what's best for the both of us so I'm ...	(3.24)	(0.24)
6	34	**Greeting including 'good luck'**	-0.02	-0.03
		good luck!	(-0.56)	(-0.38)
7	69	**Greeting including 'let's play'**	-0.03	-0.02
		Hello Let's play the game	(-1.4)	(-0.42)
8	70	**"Let's make the most money"**	0.03	-0.08
		lets go in	(1.59)	(-1.47)
9	69	**Requests In, promises Left, & specifies '12 token' outcome**	0.05*	0.06
		If you choose In I will choose Left and we will both get 1...	(2.34)	(1.1)
10	80	**Request for In only**	0.02	-0.01
		Choose The In	(0.89)	(-0.13)
11	29	**Proposal to 'work together'**	0.11***	0.08
		Lets work together	(3.51)	(0.95)
12	64	**Promises Left only**	-0.0	-0.11'
		im going left	(-0.19)	(-1.77)
13	38	**Brief greetings**	-0.05'	0.02
		HELLO! HOW ARE YOU?	(-1.7)	(0.29)
14	37	**Promises Out**	-0.15***	0.05
		I choose out	-(5.07)	(0.67)
15	37	**States intent to be 'fair'**	0.05'	0.13'
		I will be fair	(1.77)	(1.71)

Coefficient estimates when regressing Left/Right and In/Out choices on cluster dummies using the word2vec measure of similarity. Controls for literal meaning and length are omitted from the table. Clusters are described with a manual description in bold as well as each cluster's exemplar message chosen by the Affinity Propagation algorithm. ' : $p < .1, * : p < .05, ** : p < .01, * * * : p < .001$

negative, in the Trust model. Of the 16 clusters, 4 have significant positive coefficient estimates, controlling for literal meaning (clusters 5, 9, and 11). These clusters capture the same patterns observed in the models using the previous similarity measures. Clusters 5 includes 'soft promises' in which the Sender implies that they will choose Left without explicitly stating intended actions. Similarly, messages in cluster 11 suggest that the players "work together." This is also a soft promise, but also implies a request that the Receiver choose In as well. Finally, cluster 9 includes detailed promises, with requests for In and a specified outcome of "12 tokens."

The clusters with significant, negative coefficient estimates in the reduced word2vec Trust model are again familiar. All significant, negative clusters (clusters 1, 3, and 14) are either incoherent (cluster 1 messages promise to "choose Player A") or make little strategic sense in the context of the game. Clusters 3 and 14 promise to choose Right or request the Receiver chooses Out, respectively.

In the Share model, there is one cluster with a significant coefficient estimate, positive or negative. Messages in cluster 3 are promises to choose Right. Consistent with our previous results, Senders of these messages keep their word and choose Right at higher frequencies.

3.4.10 Sentiment

Using the bag-of-words and affinity propagation approaches allow us to gain valuable insights to the message topics and their association with trust behaviors. Although this has already allowed us to discover new potentially impactful features of persuasive messages, there remains a great deal of nuance left out by this reduction. To capture a portion of this additional nuance, we conduct further analysis into our messages by examining their sentiment.

To do this, we re-code our set of tokens into positive, negative, and neutral words based on

Table 3.15: Regressions on Sentiment

	Trust	Share	Trust w Control	Share w Control
Positive Words	0.004	0.014	0.006	0.012
	(0.006)	(0.015)	(0.006)	(0.015)
Negative Words	-0.036**	-0.047	-0.006	-0.033
	(0.013)	(0.033)	(0.012)	(0.033)
Total Words	0.010***	0.006*	0.005***	0.003
	(0.001)	(0.002)	(0.001)	(0.003)
Left_Only			0.195***	0.210***
			(0.015)	(0.040)
In_Only			0.135***	0.029
			(0.017)	(0.047)
Both			0.198***	0.144**
			(0.017)	(0.045)
CONSTANT	0.473***	0.545***	0.413***	0.498***
	(0.010)	(0.025)	(0.010)	(0.027)
N	976	969	976	969

Regression coefficient estimates for the number of tokens with positive or negative sentiment as defined by the OpinionFinder Subjectivity Lexicon. $* : p < .05, ** : p < .01, * * * : p < .001$

the classifications provided by the OpinionFinder Subjectivity Lexicon (Wilson et al., 2005). As with previous steps, we must alter this procedure slightly to better fit our environment, removing 'right' from the list of words with positive sentiment as it has a neutral meaning as a potential action in our game; all usages of this word in our data refer to the game action. We construct three variables for each message: a count of positive words, negative words, and a control for total word count.

We regress trust and trustworthiness on positive, negative, and total word count both with and without controlling for a message's literal meaning. Table 3.15 presents the results of these regressions.

Without controlling for literal meaning, we find that messages with negative sentiment have a significant negative relationship with trust, with each additional positive word being associated with a 3.6 percentage point decrease in trust frequency. There is a additional, positive impact of increased message length on trust. However, after controlling for literal meaning,

the significance of this sentiment effect disappears. Neither positive nor negative words have a significant impact on trust or trustworthiness. In the data from our trust game, message sentiment is not a useful predictor of behavior.

3.5 Conclusion

Free-form communication increases cooperation between economic decision makers. However, it is unclear why this effect is so pronounced relative to similar information conveyed through limited communication mechanisms. Investigating this phenomenon is difficult as the same free-form structure which allows for increased cooperation also makes any data collected difficult to parse. Free-form communication increases the action space available to Senders to an intractably large magnitude. Analysing the contents of free-form messages requires that researchers divide the space of all possible actions into a more manageable partition. Typically, the message space is reduced based on human coders assigning each message a meaning corresponding to actions available in the game. In this chapter, we employ Natural Language Processing to generate alternative partitions of the messages space. We then examine whether these alternative message structures reveal any new features of effective communication which would not otherwise have been captured by a message model based solely on the relevant game's actions.

Our first approach, we consider each message as a series of independent tokens, reducing the message space into a manageable collection of binary variables that we can then approach with standard OLS regression methods. Using this approach, we establish a baseline result that, after controlling for the literal meaning of a message, our message data contain information useful in predicting trust and trustworthiness. Explicit reference to promised/requested actions ("Left" or "In") results in greater trust compared to less-explicit promises or requests. Further, emphatic promises containing the word "Always" yield greater trust than

other promises and declarations of "Fairness" yield greater trust than other non-promises. The bag-of-words approach yields useful insights into free-form communication but comes with obvious drawbacks. The assumption of linear independence is unrealistic and interpreting results requires one to examine tokens in their original context to give meaning to the regression results. To address the drawbacks of the bag-of-words approach, we conduct cluster analysis, better preserving the complex relations between tokens.

In our second approach, we conduct cluster analysis using Affinity Propagation to divide the message space into groups with similar content. For Affinity Propagation to group similar messages, we must select a distance measure which quantifies the similarity between pairs of messages. We perform Affinity Propagation multiple times with different tuning parameters and distance measures. For distance measures, we use the Jaccard index of set overlap and a measure based on the Google word2vec similarity measure between pairs of words. The Jaccard index is calculated solely based on information within our collection of messages whereas word2vec incorporates existing NLP knowledge from external corpora.

The specifics vary depending on the distance measure and tuning parameters used, but across our four Affinity Propagation groupings, several common themes emerge in the most/least effective clusters after controlling for literal meaning. Promises and requests which specify an outcome (e.g., "12 tokens" or "the most bonus for both of us") are trusted more often than other comparable messages. We also find that clusters containing 'soft promises' — messages such as "I like being fair" which imply cooperation without an explicitly stated intention — elicit trust more often than other non-promise messages. Together, these results suggest that promises are not binary, but instead very in strength. Detailed, emphatic promises perform better than simple promises while implicit promises increase trust, but to a lesser extent than explicit ones.

In contrast, messages consisting only of greetings result in lower rates of trust as do low quality messages (message which contain grammatical errors, are brief, or display apparent

confusion about the game). This is notable because brief, low quality messages are not significantly less trustworthy than others with the same literal meaning. Receivers are either inferring incorrect information about the senders of poor quality messages or considering factors beyond their own monetary payoffs when choosing between In and Out.

With much less statistical power (one sender versus ten receivers of each message) our insights into Senders' choices are limited. However, when we do find clusters significantly correlated with higher/lower rates of trustworthiness, they tend to also be significant predictors of trust in the same direction. One notable exception to this rule are appeals directly to optimization, e.g. "Let's maximize our payoff." Appeals to optimization are significantly less trustworthy, but Receivers do not choose In at a significantly lower rate. This is the only evidence we find of potential, effective deception in our analysis.

Finally, we use sentiment analysis to examine the effect of a message's tone on trust. We find no significant correlation between a message's tone and trust or trustworthiness.

The Natural Language Processing techniques we use require delicate interpretation and we must limit the conclusions we draw from our analyses. This work does not claim to establish a definitive model of free-form communication. Our cluster analysis successfully groups sufficiently similar messages together to allow us to interpret each cluster, but the specifics rely heavily on largely arbitrary choices of similarity measures. Our approach is intended to demonstrate a process for systematically parsing unstructured message data to identify features worth further investigation.

In our trust game data, Natural Language Processing using cluster analysis suggests that future research into free-form communication should consider that promises may come in varying strengths. Detailed, emphatic promises outperform simple statements of intention while implied, 'soft' promises increase trust to a lesser extent. The techniques we employ are not well suited to definitively establish that the strength of a promise matters in building trust, but our results may motivate additional research. Receivers' aversion to low-quality

messages also warrants further investigation.

Overall, what the analyses we report establish is that free-form communication data contain useful information beyond their basic, literal meanings. Using Natural Language Processing, we can effectively parse unstructured message data to identify new features associated with higher rates of cooperation. Although the approaches we use do not allow us to draw definitive conclusions about the impact of specific message features, the results suggest directions for future research. Larger studies employing the same tools used here can use Natural Language Processing to generate hypotheses about relevant communication features which can then be tested with greater precision in follow-up experiments.

Chapter 4

Effort Cost and Communication

Abstract

Cheap talk may come in many different forms that are not equally cheap. Different signals may require different costs to the sender but still carry identical information to the receiver. We report the results of an experiment designed to test whether promises which require higher levels of effort result in greater trust from their recipients. We find that more costly promises lead recipients to trust more frequently. However, there is no corresponding, significant difference in the trustworthiness of their senders. Further, when asked their beliefs explicitly, recipients do not believe that higher cost promises are more likely to be trustworthy. This challenges the usual interpretation of the receiver's choice in trust games with communication, wherein their choice is based solely on their own monetary payoff maximization given their beliefs about the sender's likely action. Their trust is, at least partially, other regarding. We conduct a second experiment to further test this result, using a game in which the sender's message cost has no possible influence on their decision. The results of this second experiment are inconclusive.

4.1 Introduction

Communication encourages cooperation in games well beyond what is expected solely from the players' monetary incentives. The increase in trust and cooperation is particularly pronounced when players communicate through free-form messages, where impact far exceeds prewritten messages with the same literal meaning (Ismayilov and Potters, 2017; Charness and Dufwenberg, 2010). This deviation from standard theory is well established but not completely understood.

Free-form communication introduces numerous possible mechanisms that may increase cooperation. By varying message content, senders may signal traits that correlate with trustworthiness, such as intelligence (Lange et al., 2016; Ruffle and Tobol, 2017). Senders may also select messages to credibly signal their intentions by reducing social distance between sender and receiver (Ismayilov and Potters, 2016), reducing ambiguity (Li et al., 2019), or increasing the magnitude of a potential lie (Geraldes et al., 2021). In this chapter, we explore a new potential explanation of the out-sized effect of free-form communication: effort cost.

Our interest in effort cost is motivated by an observation from Chapter 3's trust game with one-way messaging. The game includes the unique design feature that each message is sent to ten different receivers, who each decide whether to trust its sender. Examining the messages trusted by all/none of their recipients shows a stark contrast between the messages in each group. Compared to the least trusted messages, the most trusted messages are longer, more detailed, and contain fewer grammatical or conceptual errors. We interpret the difference between the most and least trusted messages as primarily differences in effort. Table 3.1 presents the most and least trusted messages.

We examine varying levels of effort cost as a potential signaling mechanism because, in the real-world, cheap talk is not necessarily cheap. Often, senders incur a great deal of cost to send a message which neither credibly signals their type nor changes the payoff

structure of the underlying interaction. In everyday economic interactions, people may employ a wide range of channels through which they send cheap talk messages about the quality of their offerings or their promised future actions. A single followup text message or numerous telephone calls from a salesperson to a prospective buyer may both transmit functionally identical relevant information to the buyer but involves a substantial difference in effort. Similarly, a politician running for office may choose between communicating their positions to constituents through bulk mail or through door-to-door visits. In these real-world environments, there are near limitless potential differences between message channels in addition to their effort costs.

In this chapter, we investigate effort as a mechanism for increasing the effectiveness of cheap talk communication. We conduct a trust game experiment with messaging between subjects, replacing free-form communication with the option to send a single, fixed message, promising to cooperate, with the cost of completing a real-effort task. We vary the level of required effort and examine the impact on the Recipients' trust and Senders' trustworthiness. The reminder of this section reviews relevant literature and our hypotheses. Section 2 presents the results of our primary experiment. In section 3, we present the results of a follow-up experiment motivated by the results of Experiment 1. Section 4 summarizes and concludes.

4.1.1 Background

We study communication in the context of a simplified trust game first used in Berg et al. (1995), depicted in Figure 4.1. The game involves two players, the Sender and Receiver. The Receiver begins with an endowment of experimental currency, which they may either divide evenly between both players (Out) or pass control to the Sender (In). If the Receiver passes, the total number of tokens increases and the Sender chooses to either divide the tokens evenly (Left) or keep the entire pool, leaving the Receiver with nothing (Right).

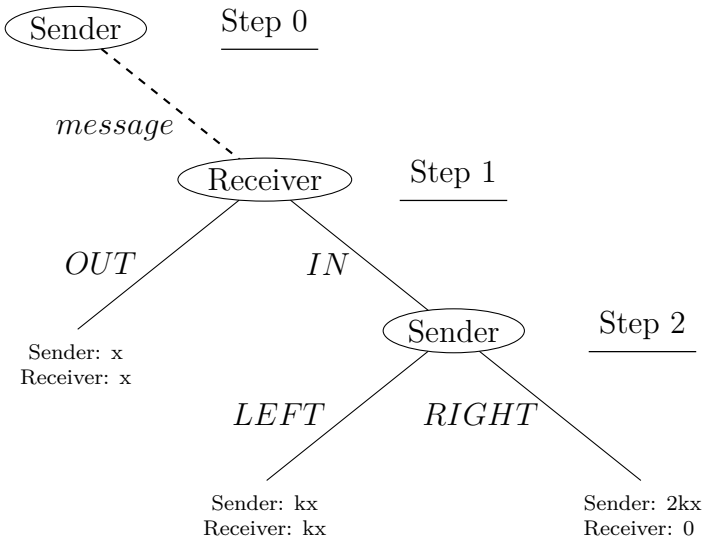


Figure 4.1: Game tree of a simple trust game.

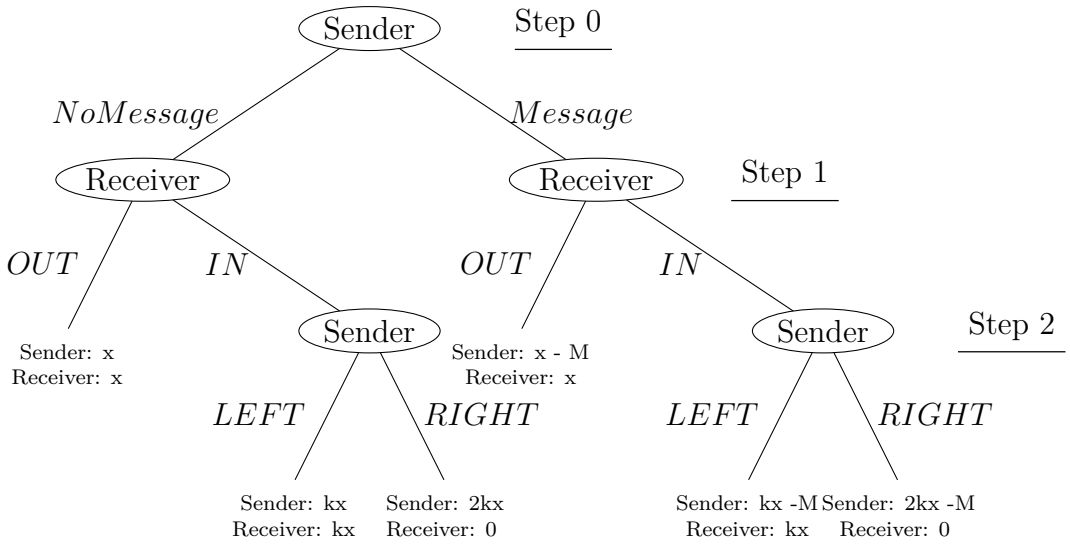
When players are concerned only for their own, material payoffs, the Subgame Prefect Equilibrium (SPE) for this game is trivial. As the sender's payoff for choosing Right is higher, the Receiver knows their payoff will be zero if they choose In and instead choose Out. Potential gains in surplus are undermined by the Sender's individual incentives. The well-established finding in numerous experiments is that both Senders and Receivers routinely deviate from the predicted equilibrium strategies, especially when allowed to communicate (Charness and Dufwenberg, 2006). With selfish preferences, cheap talk communication should not affect the outcome of the game as there is no impact on payoffs from these costless messages.

Augmenting the game tree to include a fixed message costs does not change the SPE prediction as the Sender's choice in step 2 remains Right. As such, the Receiver's equilibrium strategy is to choose Out in both cases. Because of this trivial similarity between the two sub-games, the SPE also predicts that the Senders would never choose to incur the communication cost as it only negatively affects their payoffs.

With the trust game described above in mind — where standard, backward induction predicts no potential impact of communication cost — we proceed to review the existing theoretical

models of cheap talk communication and related, costly extensions. These models do little to explain any direct impact of message cost, independent of other considerations.

Figure 4.2: Simple Trust Game with Message Cost



Augmenting the simple trust game with the introduction of a message cost does not change the SPE prediction.

Formal models of effective cheap talk communication originate with Crawford and Sobel (1982) (CS). The CS model features an uninformed Receiver whose payoffs depend on their action's agreement with the state of the world, and a Sender who is informed about the state of the world but whose payoffs make them upwardly biased. Depending on the degree of alignment between the two players' preferences (i.e. the size of the sender's bias), information exchange is possible even with solely cheap talk messages. Later extensions augment this model to include costs to the sender for misreporting information, whether those be moral aversion to lying, risk of being discovered, etc. (Kartik, 2005; Kartik et al., 2007). Inherently costly communication, regardless of the nature of the message sent, is typically modeled using the size of the message space (Hertel and Smith, 2013), the accuracy of the message (Dewatripont and Tirole, 2005), or the complexity of the message (Sobel, 2012).

In the most closely related to our current environment, costly communication can also take

the form of burned money, where senders may choose to incur a cost as a signal in addition to their independent choice of costless, cheap talk messages (Austen-Smith and Banks, 2000). With the presence of costly, burned money signals, more detailed transfer of information may be possible than in pure cheap talk settings, depending on the degree of agreement between the senders' and receivers' preferences, by allowing senders to credibly communicate certain states of the world both parties mutually prefer, provided that the Sender has a sufficiently strong preference towards those states. In experiments that allow for burned money communication, subjects behave largely in accordance with theoretical predictions, using the additional signal available to them to effectively coordinate (De Haan et al., 2015; Krol and Krol, 2020). Although we suppose that Senders of higher-effort messages are choosing to incur a cost, burned money models are not a plausible mechanism through which this may work. For burned money to explain increased trust from greater effort, Senders would have to value the opportunity to be trustworthy more than their own monetary payoffs.

All of the preceding theoretical models of cheap talk, costly or otherwise, include the key features of an underlying true state of the world and at least a partial alignment of preferences between the sender and receiver. Further, the misalignment of incentives between Sender and Receiver is common knowledge to both players and is explicit in the Sender's payoff functions, making this framework poorly suited for our environment. In a trust game, the state of the world (the Sender's aversion to lying) and thus the misalignment of preferences are both unknown. For effort cost to impact decisions in the above models, one would need to assume that there exists a type of Sender that values the opportunity to be honest more than their own material payoffs.

Existing experiments which add communication costs to otherwise cheap talk messaging focus on the presence of costs to establish a communication system. The communication mechanism is then available to all players, acting as a public good. Higher costs decrease the frequency of messages sent. However, the impact of this decrease in communication varies depending on the details of the structure of the game. In environments where subjects

have aligned incentives and are tasked with finding novel or creative solutions to tasks, higher message costs improve the efficiency of game outcomes by filtering out low-quality responses, leaving only well-informed players communicating their information (Charness et al., 2020; Grözinger et al., 2020).

However, in other games, communication costs introduce frictions, which harm subjects' ability to effectively coordinate. Message costs prevent subjects from coordinating on efficient equilibria in stag hunt games (Blume et al., 2017) and in other minimum effort games with a greater number of possible equilibria (Kriss et al., 2016; Fehr, 2017). Subjects in these games exchange fewer messages after the introduction of communication costs and contribute less. The impact of communication costs is highly dependent on the particularities of their context game. Communication costs may also prevent defection in repeated games, with the costs of renegotiating after defection leading to greater compliance between the parties involved (Andersson and Wengström, 2007).

A thorough review of the literature shows that discussions of links between message costs and recipient trust are scarce. The best experimental evidence linking message cost and meaning comes from literature examining costly apologies. Subjects in vignette experiments are more likely to evaluate apologies as genuine when the sender incurs an associated cost, regardless of whether that cost has any benefit to the recipient (Ohtsubo and Watanabe, 2009; Ohtsubo et al., 2012). These findings are supported by further studies using fMRI imaging, although interactions remain unincentivised (Ohtsubo et al., 2018).

In this chapter, we contribute to the literature on free-form, cheap talk communication, examining a new potential source of its disproportionate impact on behavior. Our game structure, with costs independent of the content of a given message, is novel relative to past experimental work, but models commonplace economic interactions. We find that increased message cost leads to increased trust by Recipients. However, the increase in trust is not accompanied by a corresponding increase in the likelihood that Senders will keep their

promises. Further, Recipients' stated beliefs suggest that they are aware that the message cost has no impact on the Senders' behavior.

4.1.2 Hypothesis

Given that real-world decision makers choose to exert costly effort in excess of what is required to convey the literal meaning of a message, they must expect benefit by altering the actions of the recipient. To change their action in response to a message, a self-interested Receiver must change their beliefs about the Sender. Therefore, the effort cost must change the utility the Sender receives should they break their promise. If the cost of effort is low, relative to the increased potential cost of lying, the Sender may increase their expected utility by incurring the effort cost.

Thus, we propose one of the reasons free-form communication is so effective at increasing cooperation and trust is that Senders can credibly signal their intentions by exerting greater effort. Those who send more costly messages are more likely to keep their promises. Thus, recipients of these promises, being aware of the higher rates of trustworthiness, trust higher-cost messages more frequently. If this proposed mechanism is true, we expect to observe three key pieces of evidence from our experimental data.

- Hypothesis 1: We expect the frequency of trust to be higher with higher message cost, conditional on a promise being sent. We test this against the null hypothesis that the trust rates are the same across all cost levels.
- Hypothesis 2: Because we propose that higher costs strengthen Senders' commitment to their promises, we expect corresponding, higher shares of Senders acting in accordance with their promises when the cost is high. The null hypothesis is that the rates are identical across all costs, conditional on promises.
- Hypothesis 3: We expect the perceived cost of effort by the Senders to be low relative

to the material payoffs of the game.

- Hypothesis 4: Finally, if greater trust is driven by stronger commitment by the Senders, the Receivers must believe this to be the case. We expect the stated beliefs of Receivers regarding the likelihood Senders will share the tokens to increase along with message cost. Receivers believe that Senders of high-cost promises are more likely to share the tokens than others. We test this against the null that the beliefs stated by Receivers about the Sender's actions are equal across all cost levels.

4.2 Experiment 1

4.2.1 Experiment 1: Methods

We conduct a trust game with one-way communication on Amazon Mechanical Turk (MTurk). The game proceeds in two steps between two players the Sender and the Receiver. In step 1, the Receiver begins with an initial endowment of 16 tokens which they may either split evenly between both players, ending the game, or pass to the Sender. If they pass the tokens, the total sum of tokens increases to 24 tokens and the Sender may either split the tokens evenly between the two or keep them all.

To examine the effect of varying levels of effort, we implement a modified communication mechanism that allows the effort required to compose a message to vary without altering its literal meaning. Instead of free-form text, Senders may send a single, prewritten message, "I will choose Left," or no message at all. If they choose to send the message, senders incur a real-effort cost which varies across treatments; declining to send a message is costless, bypassing the real-effort task altogether. Participants are aware of this cost prior to making the choice to send a message.

We impose this real-effort cost by requiring that Senders complete a series of slider tasks to

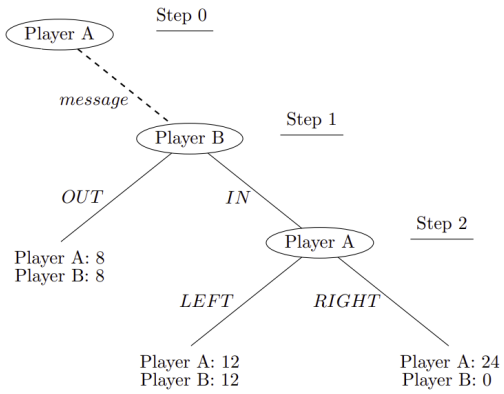


Figure 4.3: Trust game with payoffs used in experiment 1

send a message. Subjects must adjust a series of slider bars to match a randomly chosen target value ranging from 0 to 100. The cost of sending this message varies between our five treatments: 0, 1, 3, 8, or 20 correctly positioned sliders. We employ a between subject design with each participant encountering only one effort level treatment. Correctly positioning 20 sliders requires roughly 1 minute of real-effort for the average participant to complete. With an average total participation time of 8 minutes, this represents a substantial portion of the total effort put forth in their participation. All subjects must complete example slider tasks as part of their comprehension check, ensuring that they are familiar with the effort involved. By implementing this communication mechanism, we hold the information content constant, while varying the cost to the Sender. An image of the message interface is included in the Appendix, Figure B.1.

Each subject proceeds through the experiment in the same order. After reading the game instructions and successfully completing the comprehension check, each subject acts first as the Sender, making their message choice followed by their action choice, and then as the Receiver, choosing between In and Out. Each subject plays the game once as the Sender and five times as the Receiver. Subjects are paired in the order of their arrival, acting as Receiver for the five people who arrived immediately before and as Sender for the five

arriving immediately after. We run this game on Amazon Mechanical Turk (MTurk) using a game built on the oTree framework (Chen et al., 2016b). We restrict eligibility to include workers who have completed 1000 previous tasks with an approval rate of at least 99%, corresponding to roughly the top 75th percentile of MTurk workers (Robinson et al., 2019).

4.2.2 Experiment 1: Results

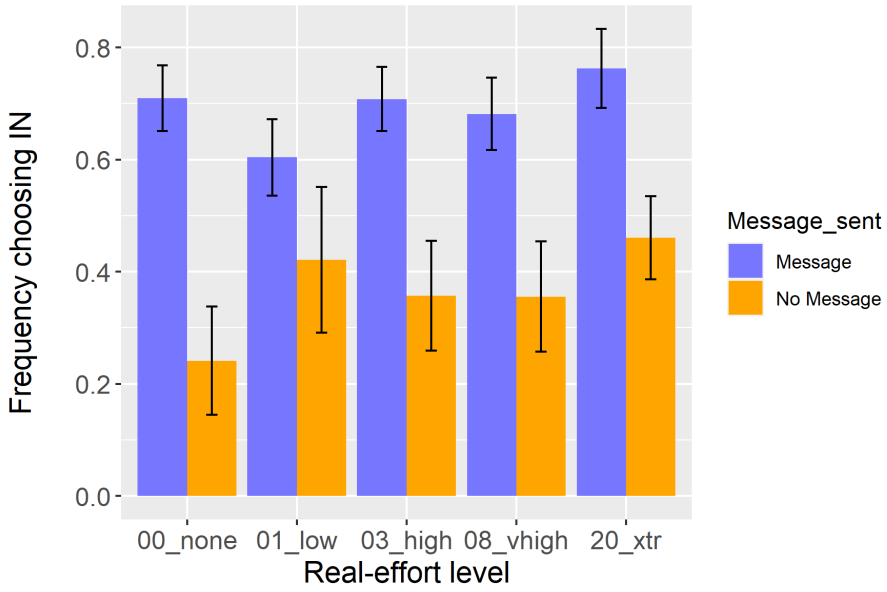
Our data consist of responses from 204 participants. The mean participation time and earnings were 8 minutes and $2.10 respectively. Figures 4.4 and 4.5 present the primary results across the five treatment groups.

Across all five treatments, the prewritten promises increase cooperation between players. Receivers who receive the message promising to choose Left are significantly more likely to choose In compared to those whose partners decline to send the message, regardless of the real-effort cost. Likewise, Senders who send the message are also significantly more likely to choose Left across all treatments, supporting the Receivers' decision to entrust them with the tokens.

Our primary hypothesis is that greater real-effort cost to send a message will increase both trust (the likelihood a Receiver chooses In) and trustworthiness (the likelihood a Sender chooses Left). Senders bear an increased cost of violating a carefully crafted promise making it a more reliable commitment mechanism, which recipients correctly interpret. Our subjects' choices when acting as Receiver follow the expected pattern when there is real-effort cost required. Compared to the Low treatment (1 slider task), recipients of promises requiring High (3) and Extreme (20) levels of effort to send are significantly more likely to choose Left (p-values .011 and .001 respectively.) Furthermore, promises that required 20 slider tasks are trusted at the highest rate of all, with a significant difference between the Low and Very High treatments at the 5% level (p-values .001 and .049). Full tables p-values for pairwise comparisons are included in Appendix B.

Table 4.1 presents the results of OLS regressions regressing the frequency choosing In on the number of sliders required. When a message is sent and None treatments are excluded, we find a positive, statistically significant correlation between effort and trust. Estimates including None treatments are positive, but fail to meet the threshold of significant at the 5% level.

Figure 4.4: Receiver Choices by Required Effort Level



When Senders decline to send a message, there is no clear relationship between effort cost and trust. We find no statistically significant difference between the rate of trust receivers place in senders declining to send a message between the Low, High, and Very High effort levels. In extreme effort treatment, Receivers are significantly more likely to choose Left after receiving no message compared to None treatments (p-value = .001). However, there is no significant difference between the Extreme and the Low treatment. As shown in Table 4.1, we find no statistically significant correlation between the number of sliders required and the frequency of Receivers choosing In when no message is sent.

Having seen that Recipients' trust increases as the effort to send a promise to play Left increases, we now move to the Senders' choices between Right and Left. For our proposed

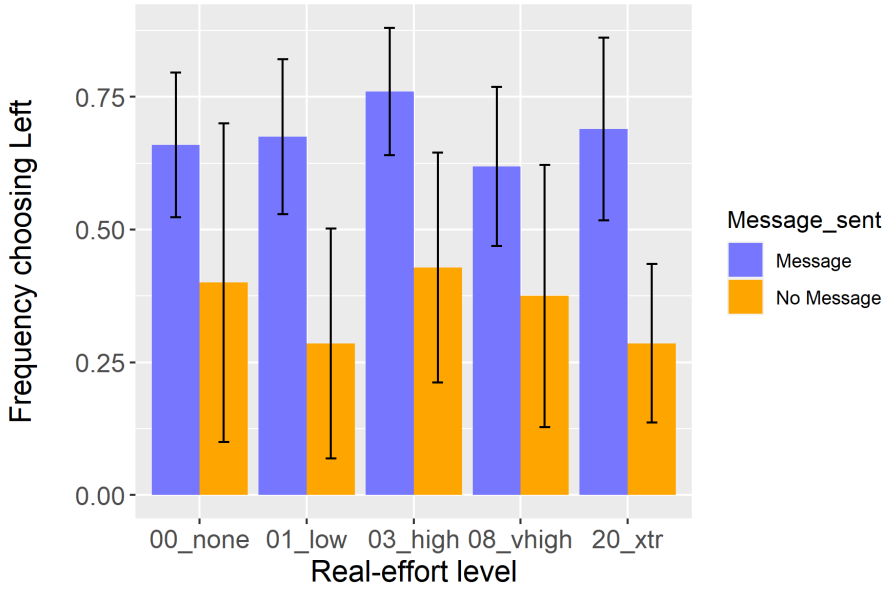
Table 4.1: Regression Estimates: Choices on Effort

	Trust		**Share**	
	Message (p-value)	No Message (p-value)	Message (p-value)	No Message (p-value)
Including No effort	0.0043‘ (0.057)	0.0063 (0.107)	-.0001 (0.891)	-.0043 (0.642)
Excluding No effort	0.0061* (0.014)	0.0043 (0.326)	-.0019 (0.732)	-.0039 (0.703)
Observations	1010/779	230/200	208/160	40/35

OLS coefficient estimates when regressing trust frequency (Receivers choosing Left) and Share frequency (Senders choosing Left) on the number of sliders required.
‘ : $p < .1, * : p < .05, ** : p < .01$

mechanism to hold, the increase in trust from Receivers must be supported by a corresponding increase in trustworthiness from the Senders. In three of the five treatments, Senders who choose to send the promise are significantly more likely to choose Left than those who do not. However, conditional on the message, or lack thereof, we do not find statistically significant differences in the Sender's Left/Right between any of the five effort level treatments.

Figure 4.5: Sender Actions by Required Effort Level

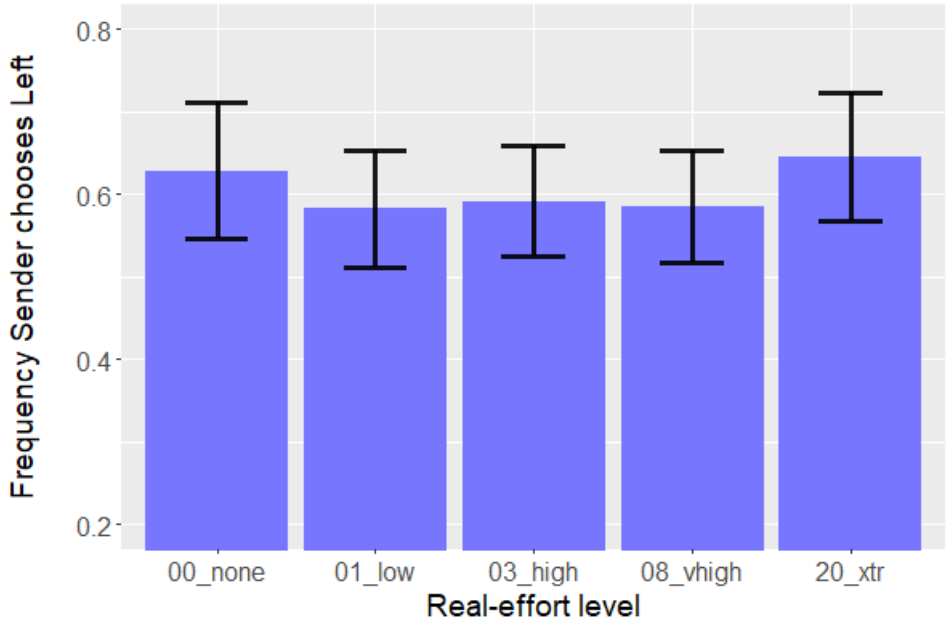


As effort cost increases, trust increases without evidence of a corresponding increase in

Senders' trustworthiness. In our proposed mechanism, we expect that increased trust is driven by the accurate belief that higher effort message will be more trustworthy as it is more costly to violate. We now examine subjects' stated beliefs about Senders' actions, given that a message was sent. If their beliefs about the likelihood that a message sender will choose Left increases along with trust, they believe that the message cost reinforces commitment but are mistaken. Figure 4.6 presents subjects' beliefs by effort level.

We find no evidence of significant differences in subjects' beliefs between any of our five treatments. Beliefs agree with our results regarding the Senders' choices. Subjects do not believe that higher effort promises lead to a higher frequency of their Senders choosing Left.

Figure 4.6: First Order Receiver Beliefs Following a Promise



Subjects' average estimates of the frequency choosing Left under each treatment. 95% confidence intervals are plotted in black

Finally, at the end of the experiment, we ask the subjects their hypothetical willingness to pay for the opportunity to send a message without completing the slider task. Across all treatments, the median response was zero tokens, indicating a complete unwillingness to

incur monetary losses to reduce real-effort expenditure. This supports Hypothesis 3 that the cost of effort is small relative to the material payoffs.

4.2.3 Experiment 1: Discussion

At first glance, Experiment 1 yields puzzling results which suggest a contradiction between Receivers' beliefs and actions. Consistent with our proposed mechanism, a higher level of effort required to send a promise is indeed associated with a higher likelihood that its recipient will choose In, passing the tokens to their partner. However, this change in action is not supported by a corresponding change in beliefs. When asked directly to provide their estimate of the likelihood that a message's sender will choose Left, subjects' mean responses do not change between levels of effort. What makes this apparent misalignment between stated beliefs and actions more notable is that these beliefs are correct. Examining subjects' choices in the Sender role, we see no significant differences in Left/Right choices between any of our treatments.

These patterns of behavior – increasing trust despite Receivers' beliefs and Senders' actions remaining unchanged – implies that our proposed mechanism through which free-form communication has such an exceptional impact is incorrect. The amount of effort required to send a promise does not change the reliability of its content. However, the recipients of these messages do change their behavior as effort cost increases. Furthermore, Receivers increase their trust in their partners in spite of their stated beliefs. Receivers know that Senders of more costly messages are no more likely to choose Left if given the opportunity.

This mismatch between the Receivers' actions and the payoff maximizing response implied by their beliefs raises important questions about our interpretation of the trust game. Although we have, up to this point, treated a Receiver's choice to play In as synonymous with trust, subjects' beliefs raise doubts about the propriety of this interpretation. Although the consequence of this action is to entrust the outcome of the game to one's partner, our evi-

dence suggests that this is not entirely driven by the belief that their partners will keep their promise. The recipients of higher-effort promises appear to be motivated to choose Left in response for reasons other than beliefs about the Sender's action.

If Receivers truly choose In at higher rates without believing that there is a corresponding increase in trustworthiness, then there is a significant gap in the current understanding of communication in trust games. Current models assume that Receivers trust promises (choose In) because the choice maximizes their expected monetary payoffs based on their first order beliefs about Senders' choices given that they sent a promise. Promises increase the likelihood of trustworthiness on the part of the Sender either through guilt aversion based on the Sender's second order beliefs (Charness and Dufwenberg, 2006; Khalmetski, 2016; Ederer and Stremitzer, 2017; Di Bartolomeo et al., 2023) or through the Sender's aversion to the inherent inconsistency involved in telling a lie (Vanberg, 2008) — Expectation based and Commitment based explanations. Our results appear to show that other factors influence Receivers' In choices. Receivers stated first order beliefs about Sender's In/Out decisions show no change as effort changes yet Receivers choices do change. Factors other than expected payoff influence Receivers' choices.

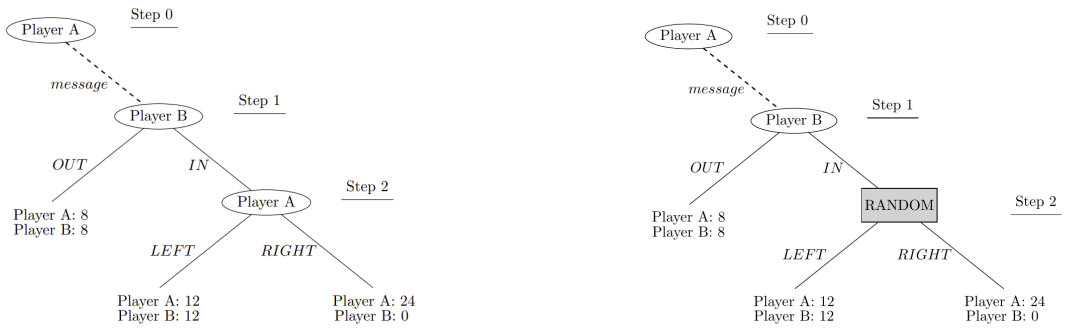
To investigate this further, we conduct a second experiment to focus on these other factors, removing the possibility that recipients can infer any information relevant to their monetary payoff from the communication mechanism by introducing a chance that the Sender's decision is out of their control. In doing so, we can isolate any effect caused by the changing levels of effort required of message Senders independent of any possible signal about the Sender's Left/Right choice. When the Sender's decision is made by a random device, Receivers' expected payoffs are fixed and known. Any changes in Receivers' behavior as effort changes cannot be explained using existing models of trust.

4.3 Experiment 2

4.3.1 Experiment 2: Methods

In light of the results of Experiment 1, we conduct a second experiment to further explore the apparent contradiction between Receiver's trust and Senders' trustworthiness as effort cost increases. To do this, we conduct a modified version of the previous trust game with communication, introducing conditions where a Sender's action choice is taken out of their control after they have sent their message.

As before, we recruit subjects from MTurk participate in a trust game with restricted communication. After completing a comprehension check, participants proceed through the game acting first the Sender and then as the Receiver. In the sender role, the task remains unchanged from Experiment 1. Senders first choose whether or not to send a single fixed message,"I will choose Left". If they choose to send the message, they incur real-effort cost which varies across two treatments. The treatments; Low and High effort require 1 and 8 slider tasks, respectively.



(a) Without random Sender choice

(b) With random Sender choice

Figure 4.7: Game trees with and without random choice as seen by Receivers

After the sender has made their message and action choices, they then play five rounds of the game as the Receiver. As before, the Receiver will receive the message or lack thereof, knowing the effort cost required to send it. However, we introduce the chance that the Sender's action choice is disregarded and determined randomly, with no influence from the Sender. If the Left/Right choice is random, Receivers are informed clearly that the Sender has no bearing on the outcome prior to making their decision between In and Out.

This mechanism is similar to one used in Ederer and Stremitzer (2017) to exogenously vary Receivers' expectations by introducing an unreliable mechanism which alters the level of control the Sender has over their own choice. We use this mechanism to compare varying levels of effort when expectations are constant and unambiguous. When nature chooses a random mechanism after the message but before the Receiver's choice, any changes variation in trust rates cannot be attributed to either commitment based or expectations based explanations.

4.3.2 Experiment 2: Hypotheses

The typical interpretation of a Receiver's In choice is that they trust the Sender to share the tokens. The Receiver's trust is based on maximizing their own expected payoff given their beliefs about the Sender. If this interpretation is a full account of the Receivers' considerations, effort cost cannot plausibly affect the Receiver's choice when the Sender does not control the choice between Left and Right.

In the cases where the In/Out decision is made by chance rather than by the Sender, there is no possibility that the Receiver may infer any information about their expected payoff from a message. Thus, changing the effort required to send the message should also have no impact on its recipient's behavior.

However, the results from Experiment 1 suggest that the different levels of effort do not

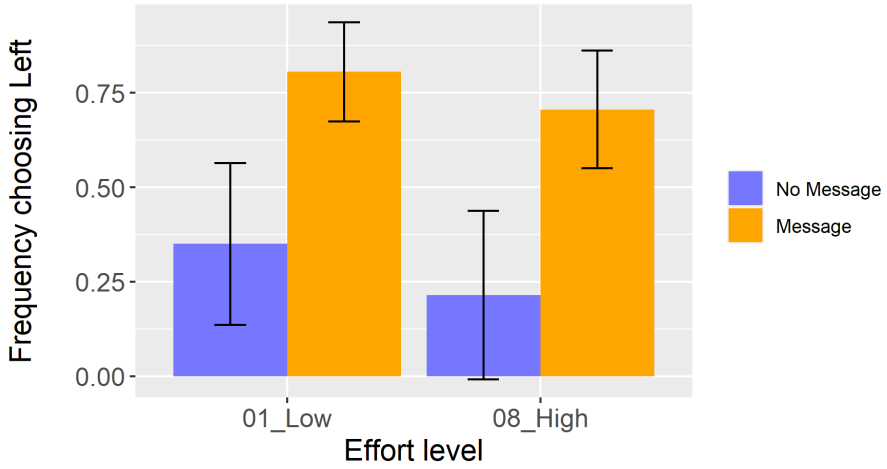
change the Receivers' estimations of the likelihood the Sender will share the tokens. If there truly is no change in beliefs, it implies that Receivers view their choice to pass the tokens as a fixed lottery, regardless of the effort cost. As such, we expect the pattern of increased trust with effort to remain even when the Sender's intentions are irrelevant.

We test this against the null that the mean likelihood that a recipient will choose In is equal across all treatments when the In/Out choice is removed from the Sender.

4.3.3 Experiment 2: Results

Experiment 2 was conducted on MTurk during the Spring of 2024. Our data consist of choices made by 103 participants. Average participation time was less than 10 minutes and typical earnings were $2.00. The procedures and hypotheses for Experiment 2 were preregistered with the Open Science Foundation (https://osf.io/vt63c)

Figure 4.8: Sender Choices by Required Effort



95% confidence intervals are plotted in black

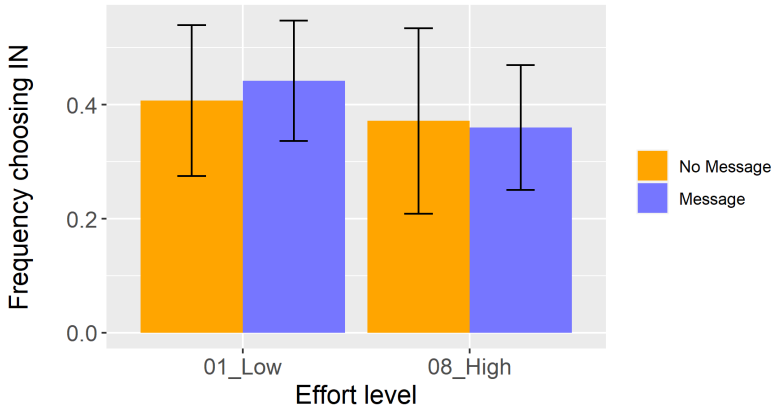
Figure 4.8 presents data relevant to Senders' choices. Consistent with previous literature and our own results in Experiment 1, Senders who choose to send the promise are more

likely to choose Left than those who send no message across both effort levels. Again, as in experiment 1, we find no statistically significant difference in Senders' Left/Right choices between treatments; effort is not associated with greater trustworthiness.

The design of Experiment 2 assumes that subjects' decisions will be unaffected unless the Senders' choices are replaced with a random mechanism. As the Sender's decisions all occur prior to the random mechanism, we expect them to be similar to Experiment 1. The results for Senders' choices are consistent with this assumption.

We then test our primary hypothesis for Experiment 2: Receivers will choose In at higher rates for higher effort promises, regardless of whether their partner's decision is made randomly or by a human. If Receivers have other regarding preferences towards Senders, Receivers should choose In more frequently in response to high effort promises when the Left/Right choice is made randomly. Figure 4.9 presents Receiver choices when they are informed the Sender's choice will be random.

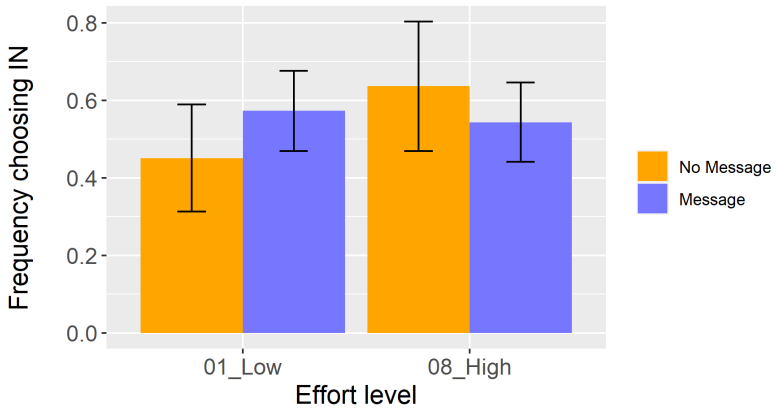
Figure 4.9: Receiver Responses when Sender Choices are Random



95% confidence intervals are plotted in black

The choices made by Receivers in our experiment do not support our hypothesis. We find no statistically significant difference between the frequency of Receivers choosing In after receiving the promise between treatments. Further, we do not find a statistically significant

Figure 4.10: Receiver Response when Senders Retain Control



95% confidence intervals are plotted in black

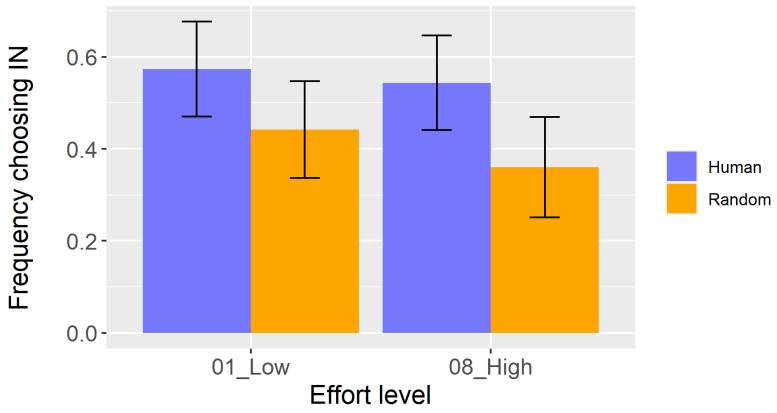
difference in the frequency choosing In between between treatments when no message is sent or between message/no message cases within the same treatment.

Our design for Experiment 2 assumes that when the Sender's choice is made by a human, Receivers' decisions should follow the patterns observed in Experiment 1: higher In frequencies in response to promises and higher In frequencies following promises in higher effort treatments. Receivers know when making their In/Out decision that their human partner will make the Left/Right decision. The information and structure of the game are identical to Experiment 1. Figure 4.10 presents Receivers' choices when their human partner makes the Left/Right decision.

Unlike in Experiment 1, we do not observe that promises increase the frequency of Receivers choosing In. There is no statistically significant difference between message/no message cases in either treatment. Further, we do not observe an increase in the frequency choosing In as effort cost increases. These two results are not consistent with our findings in Experiment 1 although the game structure and available information are the same for the Receivers making these choices.

Figures 4.11 and 4.12 present data comparing Receivers' responses to messages (or lack thereof) when the Left/Right decision is made by a human versus when it is made randomly.

Figure 4.11: Receiver Choices Following a Message

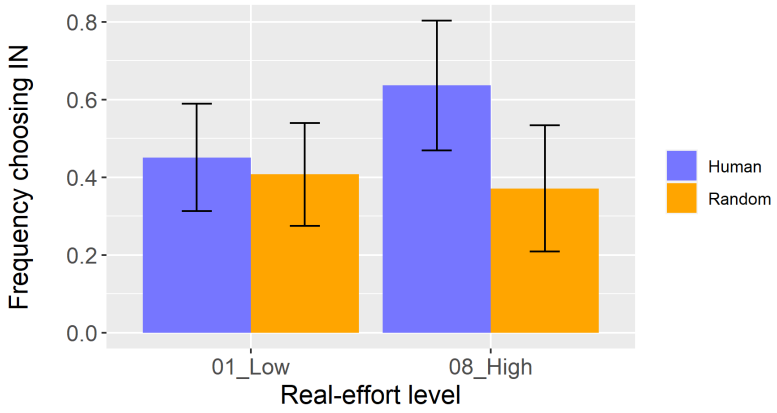


95% confidence intervals are plotted in black

Receivers of promises in Experiment 2 choose In less often when the Left/Right decision is random instead of human controlled. However, this difference is not statistically significant in either the high or low effort treatments.

The same is true for Receivers who receive no message (Figure 4.12). The proportion of Receivers choosing In is lower in random choice cases versus human choice cases, but this difference is not statistically significant.

Figure 4.12: Receiver Choices Following No Message



95% confidence intervals are plotted in black

4.3.4 Experiment 2: Discussion

The results of Experiment 2 do not support the hypothesis developed in response to Experiment 1. In Experiment 1, 'trust' rates in response to promises increase as the effort required to send the promise increases. There was no corresponding statistically significant increase in Senders' trustworthiness nor was there an increase in the Receivers' first order beliefs regarding Senders' actions. We designed Experiment 2 to test whether we can observe an increase in 'trust' in an environment where the Senders' actions and the Receivers' first-order beliefs cannot possibly be related to the Senders' message choice or effort level. We create this environment by introducing the possibility that a Sender's choice will be replaced with a coin-flip after they have made their message choice and exerted the necessary effort but before the Recipient chooses In or Out.

Our data in Experiment 2 do not support our hypothesis. When the In/Out choice is made randomly, Receivers are not significantly more likely to choose In in response to a promise when the effort cost is higher. Our hypothesis predicts that Receivers will more often 'trust' Senders who incurred a greater cost to send their promise independent of any information they may infer about the Sender's reliability.

However, other results from Experiment 2 indicate that there may be a flaw in the design of our decision task. In cases when the Left/Right choice is not random — the game proceeding, in theory, identically to the game in Experiment 1 — we do not replicate the results of Experiment 1. Senders who send promises are still more trustworthy than those who do not send a message. However, Receivers do not respond accordingly. When the Left/Right choice is made by their human partner, Receivers are not significantly more likely to choose In following a promise versus no message. This failure to replicate the promise effect, present in both Experiment 1 and numerous prior studies, raises doubts about the validity of our results. Introducing the potential random Left/Right choice may have made the task more difficult for subjects to comprehend in their brief participation time or introduced ambiguity

about when the random mechanism is determined or to whom payoff is assigned. Experiment 2 does not support the hypothesis developed following Experiment 1, but it is uncertain whether this a true reflection of participants' decision making or the result of our game's design.

4.4 Conclusion

In this chapter, we present the results of two experiments designed to investigate why free-form communication has an outsized effect on cooperation in economic decision making. We isolate one factor out of the limitless possible signals contained in free-form messages — real-effort cost — and implement a communication mechanism in a trust game which allows us to vary the level of effort required to send a message while leaving the message content unchanged. Senders may send a fixed message promising to share the payoff equally or send no message at all. Sending the message requires the Sender to exert real-effort in the form of a slider task which we vary across treatments.

We find that Receivers trust higher effort promises significantly more often than lower-effort promises. Across all effort levels, Receivers trust promises more than the absence of a message, consistent with existing literature. This increase in trust with effort is not complemented by a statistically significant increase in their Senders' trustworthiness. One is tempted to attribute this increase in trust without increasing trustworthiness to an error in the Receivers' judgement, inferring greater trustworthiness falsely from higher effort cost. However, the stated first order beliefs of the Receivers do not support this. When asked to estimate the frequency senders of promises choose Left for varying levels of effort, there is no statistically significant difference between treatments.

In total, these three findings (effort increasing trust but not trustworthiness or beliefs about trustworthiness) suggest that Receivers' In/Out decisions are influenced by more than their

own payoff maximization. If higher effort cost truly does not change Receivers' beliefs but it increases the frequency with which they play In, then Receivers' choice to play In cannot solely be called 'trust.' The evidence available to us suggests that the choice to 'trust' may partially be motivated by the Receivers' consideration of the Senders' expectations.

At present, the evidence for this result is somewhat weak and based on interpreting the lack of a significant differences in beliefs as evidence that they are unchanging. To address this weakness, we conduct a second experiment, modifying the procedures in Experiment 1 so that Receivers' beliefs about Senders' actions cannot possibly affect their decisions if they are solely payoff-maximizing. Following the Senders transmitting a promise and incurring the associated effort cost, their Left/Right choice is determined randomly instead of by the Sender's choice. Receivers know that the Left/Right choice is independent of their paired Sender's decision and thus their message cannot have any information relevant to the Receiver's payoff maximization. This design creates a condition under which we are certain Receivers' beliefs are constant.

The results of Experiment 2 do not support our interpretation of the Experiment 1 results. When the Senders' action is chosen at random, Receivers are not significantly more likely to choose In following a more costly promise. However, we also do not observe a statistically significant increase in trust when the Senders make the Left/Right decision as they do in Experiment 1. Further, we do not find a significant increase in trust rates following promises. This is puzzling given that the effect of promises seen both in the existing literature and in Experiment 1 in this chapter. The lack of impact of promises on their recipients suggests that introducing the potential random mechanism had unintended effects on our subjects. We cannot determine whether Experiment 2's results are a true reflection of subjects' decisions given the information or if they are an artifact of misunderstanding about the random mechanism. In future work, we will investigate these results further, using a redesigned task to ensure clarity and comprehension.

Despite the inconclusive results of Experiment 2, the work presented in this chapter still reveals novel aspects of communication between economic agents. Receivers trust higher-effort messages more frequently, holding the message content constant. This finding may explain part of the outsized effectiveness of free-form communication compared to fixed, prewritten signals. A Sender may craft a message to display effort, increasing their expected payoff in excess of the cost they incur. Determining the reason for this — whether other-regarding towards Senders or payoff-maximizing based on Senders' choices — requires further investigation.

Bibliography

Andersson, O. and Wengström, E. (2007). Do antitrust laws facilitate collusion? experimental evidence on costly communication in duopolies. *Scandinavian Journal of Economics*, 109(2):321–339.

Andreoni, J. and Rao, J. M. (2011). The power of asking: How communication affects selfishness, empathy, and altruism. *Journal of public economics*, 95(7-8):513–520.

Andres, M., Bruttel, L., and Friedrichsen, J. (2023). How communication makes the difference between a cartel and tacit collusion: A machine learning approach. *European Economic Review*, 152:104331.

Asparouhova, E., Hertzel, M., and Lemmon, M. (2009). Inference from streaks in random outcomes: Experimental evidence on beliefs in regime shifting and the law of small numbers. *Management Science*, 55(11):1766–1782.

Austen-Smith, D. and Banks, J. S. (2000). Cheap talk and burned money. *Journal of Economic Theory*, 91(1):1–16.

Bar-Hillel, M. and Wagenaar, W. A. (1991). The perception of randomness. *Advances in applied mathematics*, 12(4):428–454.

Ben-Ner, A., Putterman, L., and Ren, T. (2007). Lavish returns on cheap talk: Non-binding communication in a trust experiment. *Available at SSRN 1013582*.

Benjamin, D. J., Moore, D. A., and Rabin, M. (2017). Biased beliefs about random samples: Evidence from two integrated experiments. Technical report, National Bureau of Economic Research.

Berg, J., Dickhaut, J., and McCabe, K. (1995). Trust, reciprocity, and social history. *Games and economic behavior*, 10(1):122–142.

Blume, A., Kriss, P. H., and Weber, R. A. (2017). Pre-play communication with forgone costly messages: experimental evidence on forward induction. *Experimental Economics*, 20:368–395.

Bou, S., Brandts, J., Cayón, M., and Guillén, P. (2016). The price of luck: paying for the hot hand of others. *Journal of the Economic Science Association*, 2:60–72.

Buchan, N. R., Johnson, E. J., and Croson, R. T. (2006). Let's get personal: An international examination of the influence of communication, culture and social distance on other regarding preferences. *Journal of Economic Behavior & Organization*, 60(3):373–398.

Burns, B. D. and Corpus, B. (2004). Randomness and inductions from streaks:"gambler's fallacy" versus" hot hand ". *Psychonomic bulletin & review*, 11:179–184.

Bursztyn, L., Egorov, G., Haaland, I., Rao, A., and Roth, C. (2023). Justifying dissent. *The Quarterly Journal of Economics*, 138(3):1403–1451.

Camerer, C. F., Nave, G., and Smith, A. (2019). Dynamic unstructured bargaining with private information: theory, experiment, and outcome prediction via machine learning. *Management Science*, 65(4):1867–1890.

Charness, G., Cooper, D. J., and Grossman, Z. (2020). Silence is golden: team problem solving and communication costs. *Experimental Economics*, 23:668–693.

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

Charness, G. and Dufwenberg, M. (2010). Bare promises: An experiment. *Economics letters*, 107(2):281–283.

Charness, G. and Gneezy, U. (2008). What's in a name? anonymity and social distance in dictator and ultimatum games. *Journal of Economic Behavior & Organization*, 68(1):29–35.

Chen, D. L., Moskowitz, T. J., and Shue, K. (2016a). Decision making under the gambler's fallacy: Evidence from asylum judges, loan officers, and baseball umpires. *The Quarterly Journal of Economics*, 131(3):1181–1242.

Chen, D. L., Schonger, M., and Wickens, C. (2016b). otree—an open-source platform for laboratory, online, and field experiments. *Journal of Behavioral and Experimental Finance*, 9:88–97.

Chen, H.-L., Jegadeesh, N., and Wermers, R. (2000). The value of active mutual fund management: An examination of the stockholdings and trades of fund managers. *Journal of Financial and quantitative Analysis*, 35(3):343–368.

Chen, J. and Houser, D. (2017). Promises and lies: can observers detect deception in written messages. *Experimental Economics*, 20:396–419.

Clotfelter, C. T. and Cook, P. J. (1993). The "gambler's fallacy" in lottery play. *Management Science*, 39(12):1521–1525.

Cooper, D. J. and Kühn, K.-U. (2016). Communication and cooperation: A methodological study. *Southern economic journal*, 82(4):1167–1185.

Crawford, V. (1998). A survey of experiments on communication via cheap talk. *Journal of Economic theory*, 78(2):286–298.

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

Criscuolo, P., Dahlander, L., Grohsjean, T., and Salter, A. (2021). The sequence effect

in panel decisions: Evidence from the evaluation of research and development projects. *Organization Science*, 32(4):987–1008.

De Haan, T., Offerman, T., and Sloof, R. (2015). Money talks? an experimental investigation of cheap talk and burned money. *International Economic Review*, 56(4):1385–1426.

Devetag, G., Di Guida, S., and Polonio, L. (2016). An eye-tracking study of feature-based choice in one-shot games. *Experimental Economics*, 19:177–201.

Dewatripont, M. and Tirole, J. (2005). Modes of communication. *Journal of political economy*, 113(6):1217–1238.

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

Ederer, F. and Stremitzer, A. (2017). Promises and expectations. *Games and Economic Behavior*, 106:161–178.

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

Ellingsen, T., Östling, R., and Wengström, E. (2018). How does communication affect beliefs in one-shot games with complete information? *Games and Economic Behavior*, 107:153–181.

Fehr, D. (2017). Costly communication and learning from failure in organizational coordination. *European Economic Review*, 93:106–122.

Fellbaum, C. (2010). Wordnet. In *Theory and applications of ontology: computer applications*, pages 231–243. Springer.

Fiedler, M. and Haruvy, E. (2009). The lab versus the virtual lab and virtual field—an experimental investigation of trust games with communication. *Journal of Economic Behavior & Organization*, 72(2):716–724.

Fonseca, M. A. and Normann, H.-T. (2012). Explicit vs. tacit collusion—the impact of communication in oligopoly experiments. *European Economic Review*, 56(8):1759–1772.

Frey, B. J. and Dueck, D. (2007). Clustering by passing messages between data points. *science*, 315(5814):972–976.

Frydman, C. and Nave, G. (2017). Extrapolative beliefs in perceptual and economic decisions: Evidence of a common mechanism. *Management Science*, 63(7):2340–2352.

Fudenberg, D., Kleinberg, J., Liang, A., and Mullainathan, S. (2022). Measuring the completeness of economic models. *Journal of Political Economy*, 130(4):956–990.

Fudenberg, D. and Liang, A. (2019). Predicting and understanding initial play. *American Economic Review*, 109(12):4112–4141.

Geraldes, D., Heinicke, F., and Kim, D. G. (2021). Big and small lies. *Journal of Behavioral and Experimental Economics*, 91:101666.

Gilovich, T., Vallone, R., and Tversky, A. (1985). The hot hand in basketball: On the misperception of random sequences. *Cognitive psychology*, 17(3):295–314.

Green, B. and Zwiebel, J. (2018). The hot-hand fallacy: Cognitive mistakes or equilibrium adjustments? evidence from major league baseball. *Management Science*, 64(11):5315–5348.

Grimm, S., Kocher, M. G., Krawczyk, M., and Le Lec, F. (2021). Sharing or gambling? on risk attitudes in social contexts. *Experimental Economics*, 24:1075–1104.

Grözinger, N., Irlenbusch, B., Laske, K., and Schröder, M. (2020). Innovation and communication media in virtual teams–an experimental study. *Journal of Economic Behavior & Organization*, 180:201–218.

Guryan, J. and Kearney, M. S. (2008). Gambling at lucky stores: Empirical evidence from state lottery sales. *American Economic Review*, 98(1):458–473.

Hanaki, N. and Ozkes, A. I. (2022). Strategic environment effect and communication. *Experimental Economics*, pages 1–34.

Hertel, J. and Smith, J. (2013). Not so cheap talk: Costly and discrete communication. *Theory and decision*, 75:267–291.

Horton, J. J., Rand, D. G., and Zeckhauser, R. J. (2011). The online laboratory: Conducting experiments in a real labor market. *Experimental economics*, 14:399–425.

Huber, J., Kirchler, M., and Stöckl, T. (2010). The hot hand belief and the gambler's fallacy in investment decisions under risk. *Theory and decision*, 68:445–462.

Ismayilov, H. and Potters, J. (2016). Why do promises affect trustworthiness, or do they? *Experimental Economics*, 19:382–393.

Ismayilov, H. and Potters, J. (2017). Elicited vs. voluntary promises. *Journal of Economic Psychology*, 62:295–312.

Jetter, M. and Walker, J. K. (2015). Game, set, and match: Do women and men perform differently in competitive situations? *Journal of Economic Behavior & Organization*, 119:96–108.

Jiao, P. (2017). Belief in mean reversion and the disposition effect: An experimental test. *Journal of Behavioral Finance*, 18(1):29–44.

Kartik, N. (2005). Information transmission with almost-cheap talk. *Available at SSRN 817508*.

Kartik, N., Ottaviani, M., and Squintani, F. (2007). Credulity, lies, and costly talk. *Journal of Economic theory*, 134(1):93–116.

Khalmetski, K. (2016). Testing guilt aversion with an exogenous shift in beliefs. *Games and Economic Behavior*, 97:110–119.

Kriss, P. H., Blume, A., and Weber, R. A. (2016). Coordination with decentralized costly communication. *Journal of Economic Behavior & Organization*, 130:225–241.

Krol, M. and Krol, M. E. (2020). On the strategic value of 'shooting yourself in the foot': an experimental study of burning money. *International Journal of Game Theory*, 49(1):23–45.

Lange, B. P., Hennighausen, C., Brill, M., and Schwab, F. (2016). Only cheap talk after all? new experimental psychological findings on the role of verbal proficiency in mate choice. *Psychology of Language and Communication*, 20(1):1–22.

Li, C., Turmunkh, U., and Wakker, P. P. (2019). Trust as a decision under ambiguity. *Experimental Economics*, 22:51–75.

Lien, J. W. and Yuan, J. (2015). Selling to biased believers: strategies of online lottery ticket vendors. *Economic Inquiry*, 53(3):1506–1521.

Livingston, J. A. (2012). The hot hand and the cold hand in professional golf. *Journal of Economic Behavior & Organization*, 81(1):172–184.

Loh, R. K. and Warachka, M. (2012). Streaks in earnings surprises and the cross-section of stock returns. *Management Science*, 58(7):1305–1321.

Loper, E. and Bird, S. (2002). Nltk: The natural language toolkit. *arXiv preprint cs/0205028*.

Lundquist, T., Ellingsen, T., Gribbe, E., and Johannesson, M. (2009). The aversion to lying. *Journal of Economic Behavior & Organization*, 70(1-2):81–92.

Marcus, M., Santorini, B., and Marcinkiewicz, M. A. (1993). Building a large annotated corpus of english: The penn treebank.

Mihalcea, R., Corley, C., Strapparava, C., et al. (2006). Corpus-based and knowledge-based measures of text semantic similarity. In *Aaai*, volume 6, pages 775–780.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Miller, J. B. and Sanjurjo, A. (2019). A cold shower for the hot hand fallacy: Robust evidence that belief in the hot hand is justified.

Misirlisoy, E. and Haggard, P. (2014). Asymmetric predictability and cognitive competition in football penalty shootouts. *Current Biology*, 24(16):1918–1922.

Mohlin, E. and Johannesson, M. (2008). Communication: Content or relationship? *Journal of Economic Behavior & Organization*, 65(3-4):409–419.

Ohtsubo, Y., Matsunaga, M., Tanaka, H., Suzuki, K., Kobayashi, F., Shibata, E., Hori, R., Umemura, T., and Ohira, H. (2018). Costly apologies communicate conciliatory intention: an fmri study on forgiveness in response to costly apologies. *Evolution and Human Behavior*, 39(2):249–256.

Ohtsubo, Y. and Watanabe, E. (2009). Do sincere apologies need to be costly? test of a costly signaling model of apology. *Evolution and Human Behavior*, 30(2):114–123.

Ohtsubo, Y., Watanabe, E., Kim, J., Kulas, J. T., Muluk, H., Nazar, G., Wang, F., and Zhang, J. (2012). Are costly apologies universally perceived as being sincere? a test of the costly apology-perceived sincerity relationship in seven countries. *Journal of Evolutionary Psychology*, 10(4):187–204.

Ötting, M., Langrock, R., Deutscher, C., and Leos-Barajas, V. (2020). The hot hand in professional darts. *Journal of the Royal Statistical Society Series A: Statistics in Society*, 183(2):565–580.

Penczynski, S. P. (2019). Using machine learning for communication classification. *Experimental Economics*, 22(4):1002–1029.

Polonio, L. and Coricelli, G. (2019). Testing the level of consistency between choices and beliefs in games using eye-tracking. *Games and Economic Behavior*, 113:566–586.

Powdthavee, N. and Riyanto, Y. E. (2015). Would you pay for transparently useless advice? a test of boundaries of beliefs in the folly of predictions. *Review of Economics and Statistics*, 97(2):257–272.

Rabin, M. (2002). Inference by believers in the law of small numbers. *The Quarterly Journal of Economics*, 117(3):775–816.

Rabin, M. and Vayanos, D. (2010). The gambler's and hot-hand fallacies: Theory and applications. *The Review of Economic Studies*, 77(2):730–778.

Rao, K. and Hastie, R. (2023). Predicting outcomes in a sequence of binary events: Belief updating and gambler's fallacy reasoning. *Cognitive Science*, 47(1):e13211.

Rehurek, R. and Sojka, P. (2011). Gensim–python framework for vector space modelling. *NLP Centre, Faculty of Informatics, Masaryk University, Brno, Czech Republic*, 3(2):2.

Robinson, J., Rosenzweig, C., Moss, A. J., and Litman, L. (2019). Tapped out or barely tapped? recommendations for how to harness the vast and largely unused potential of the mechanical turk participant pool. *PloS one*, 14(12):e0226394.

Roney, C. J. and Trick, L. M. (2009). Sympathetic magic and perceptions of randomness: The hot hand versus the gambler's fallacy. *Thinking & reasoning*, 15(2):197–210.

Ruffle, B. J. and Tobol, Y. (2017). Clever enough to tell the truth. *Experimental Economics*, 20:130–155.

Snowberg, E. and Yariv, L. (2021). Testing the waters: Behavior across participant pools. *American Economic Review*, 111(2):687–719.

Sobel, J. (2012). Complexity versus conflict in communication. In *2012 46th Annual Conference on Information Sciences and Systems (CISS)*, pages 1–6. IEEE.

Sparck Jones, K. (1972). A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21.

Sundali, J. and Croson, R. (2006). Biases in casino betting: The hot hand and the gambler's fallacy. *Judgment and Decision making*, 1(1):1–12.

Tebbe, E. and Wegener, B. (2022). Is natural language processing the cheap charlie of analyzing cheap talk? a horse race between classifiers on experimental communication data. *Journal of Behavioral and Experimental Economics*, 96:101808.

Terrell, D. (1994). A test of the gambler's fallacy: Evidence from pari-mutuel games. *Journal of risk and uncertainty*, 8:309–317.

Terrell, D. (1998). Biases in assessments of probabilities: New evidence from greyhound races. *Journal of Risk and Uncertainty*, 17:151–167.

Tetlock, P. C. (2007). Giving content to investor sentiment: The role of media in the stock market. *The Journal of finance*, 62(3):1139–1168.

Tversky, A. and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological bulletin*, 76(2):105.

Vanberg, C. (2008). Why do people keep their promises? an experimental test of two explanations 1. *Econometrica*, 76(6):1467–1480.

Weber, M. and Camerer, C. F. (1998). The disposition effect in securities trading: An experimental analysis. *Journal of Economic Behavior & Organization*, 33(2):167–184.

Wilson, T., Wiebe, J., and Hoffmann, P. (2005). Recognizing contextual polarity in phrase-level sentiment analysis. In *Proceedings of human language technology conference and conference on empirical methods in natural language processing*, pages 347–354.

Woods, D., Abdallah, M., Bagchi, S., Sundaram, S., and Cason, T. (2022). Network defense and behavioral biases: an experimental study. *Experimental Economics*, 25(1):254–286.

Appendices

Appendix A

Appendix for Chapter 3

A.1 Affinity Propagation Algorithm Description

Affinity Propagation begins with a similarity matrix $\mathbf{S}_{n\times n}$ and iteratively updates two matrices: $\mathbf{R}_{n\times n}$, the responsibility matrix and $\mathbf{A}_{n\times n}$, the availability matrix until stable clusters emerge. Both $\mathbf{A}$ and $\mathbf{R}$ initialize with all zero elements. The algorithm proceeds as follows:

1) $\mathbf{R}$ updates following:

$$r(i,k) \leftarrow s(i,k) - \max_{\forall k \neq k'} \{a(i,k') + s(i,k')\}$$

2) $\mathbf{A}$ updates following:

$$a(i,k) \leftarrow \min\{0, r(k,k) + \sum_{i' \notin i,k} \max(0, r(i',k))\}$$

$$a(k,k) \leftarrow \sum_{i' \neq k} \max(0, r(i',k))$$

Final clusters assignments are given by the criteria matrix $\mathbf{C} = \mathbf{A} + \mathbf{S}$; element i chooses k as its exemplar if the highest value in row i in $\mathbf{C}$ is in column k. This process repeats, updating $\mathbf{A}$ and $\mathbf{R}$ until cluster assignments remain unchanged between iterations.

Intuitively, the Affinity Propagation algorithm considers, first, each element's choice for its own exemplar out of the other elements in the set, $\mathbf{R}$. After this, these preferences are

updated to account for the overall appropriateness of a given message to serve as an exemplar in the final cluster assignments. if a given message is chosen as a most preferred exemplar by relatively few other messages, it will incur a penalty reflecting its overall unsuitability to act as an exemplar.

A.2 Full Results from Table 3.4: Bag-of-words Model

Table A.1: Bag of Words Model: Full

Tokens	Trust	T-stat	Trustworthiness	T-stat
i	0.003	(0.212)	0.063	(1.411)
choose	-0.023	(-1.512)	-0.099*	(-2.259)
in	0.068***	(3.545)	0.036	(0.662)
left	0.060**	(3.278)	0.083	(1.608)
you	0.004	(0.217)	0.076	(1.618)
to	0.029*	(2.069)	0.088*	(2.187)
will	-0.005	(-0.282)	0.081	(1.773)
be	-0.012	(-0.847)	-0.077	(-1.942)
the	0.028	(1.761)	-0.069	(-1.517)
both	-0.013	(-0.635)	0.042	(0.764)
we	0.047*	(2.202)	0.138*	(2.306)
and	0.031*	(1.998)	0.046	(1.051)
a	-0.009	(-0.529)	-0.064	(-1.326)
go	-0.013	(-0.732)	-0.114*	(-2.326)
if	0.004	(0.188)	-0.001	(-0.022)
get	0.033	(1.689)	-0.023	(-0.430)
of	0.016	(0.754)	-0.020	(-0.333)
for	0.008	(0.398)	0.106*	(1.974)
let	0.011	(0.487)	-0.116	(-1.815)
so	-0.026	(-1.196)	-0.108	(-1.726)
u	0.048*	(2.007)	0.003	(0.043)
player	-0.023	(-0.863)	0.060	(0.794)
's	0.001	(0.038)	0.159*	(2.363)
that	0.042*	(2.043)	0.043	(0.742)
can	0.046*	(2.084)	-0.143*	(-2.275)
make	0.066**	(3.015)	0.010	(0.165)
out	-0.101***	(-4.933)	0.087	(1.509)
hi	-0.001	(-0.055)	0.047	(0.866)
hello	0.029	(1.464)	0.069	(1.243)
12	0.041	(1.664)	-0.007	(-0.102)
play	-0.022	(-1.007)	0.028	(0.462)
do	-0.021	(-0.947)	-0.062	(-1.009)
token	0.018	(0.720)	-0.074	(-1.063)
'm	0.010	(0.411)	0.009	(0.137)
fair	0.052*	(2.316)	0.113	(1.793)
this	0.063**	(2.730)	0.099	(1.510)
it	-0.046	(-1.856)	-0.120	(-1.708)
good	0.004	(0.189)	-0.039	(-0.615)
'll	-0.028	(-1.127)	0.046	(0.655)
best	0.028	(1.170)	0.005	(0.076)
please	0.013	(0.549)	0.070	(1.065)
more	0.030	(1.181)	-0.139	(-1.924)
maximize	0.067*	(2.247)	-0.148	(-1.692)
point	-0.031	(-1.285)	0.025	(0.357)
our	-0.007	(-0.198)	-0.004	(-0.042)
each	-0.015	(-0.545)	0.008	(0.100)
most	-0.023	(-0.808)	0.121	(1.532)
game	-0.050	(-1.763)	-0.082	(-1.000)
all	-0.033	(-1.191)	-0.024	(-0.307)
b	0.016	(0.452)	-0.056	(-0.575)
always	0.053*	(2.068)	0.136	(1.869)
money	0.008	(0.263)	-0.047	(-0.570)
pick	0.021	(0.719)	0.008	(0.096)
with	0.006	(0.229)	0.000	(0.001)
have	0.053	(1.788)	0.088	(1.040)
right	-0.130***	(-4.614)	-0.224**	(-2.767)

Each coefficient is associated with a dummy variable indicating a message includes the relevant word. Control variables for messages' literal meaning are omitted from the table. $* : p < .05, ** : p < .01, * * * : p < .001$

Appendix B

Appendix for Chapter 4

B.1 Pairwise Comparisons for Experiment 1

Table B.1: Receiver comparison without message

	01_Low	03_High	08_vHigh	20_xtr
00_None	0.58	.144	.142	.013
01_Low		.721	.727	.343
03_High			.504	.134
08_vHigh				.127

P-values comparing effort levels when no message is sent. P-values are for one-sided t-tests examining whether the mean of the column treatment exceeds the mean of the row treatment.

Table B.2: Receiver comparison with message

	01_Low	03_High	08_vHigh	20_xtr
00_None	.989	.519	.736	.129
01_Low		.011	.052	.001
03_High			.723	.118
08_vHigh				.049

P-values comparing effort levels when a message is sent. P-values are for one-sided t-tests examining whether the mean of the column treatment excedes the mean of the row treatment.

B.2 Communication Mechanism

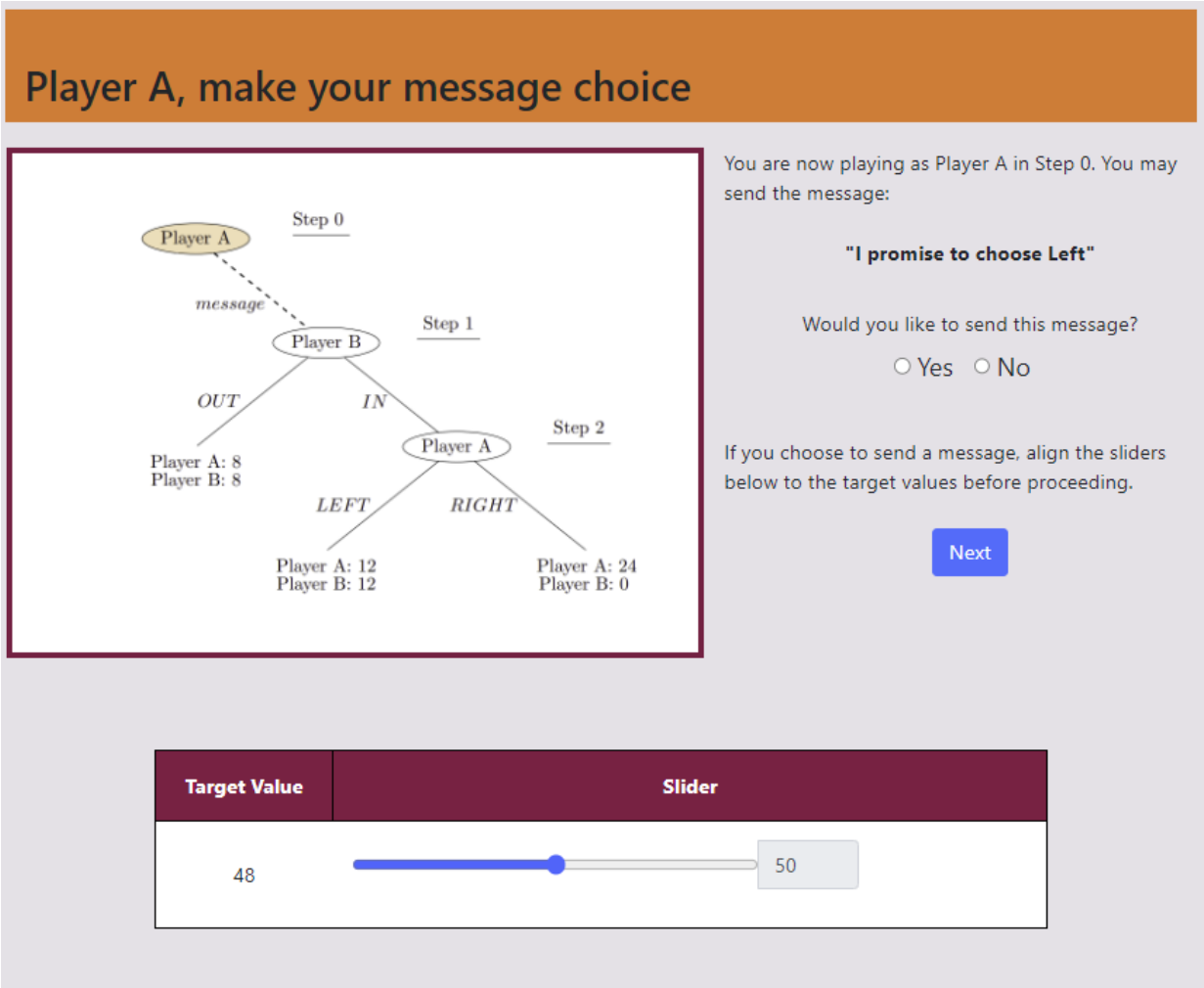


Figure B.1: Senders' interface in both Experiment 1 and 2