ECONOMICS EXPERIMENTS ON INSURANCE, GENDER ROLES AND LABOR CONTRACT DESIGN

A Dissertation

by

PEILU ZHANG

Submitted to the Office of Graduate and Professional Studies of
Texas A&M University
in partial fulfillment of the requirements for the degree of

DOCTOR OF PHILOSOPHY

Chair of Committee,	Marco A. Palma
Committee Members,	Catherine C. Eckel
	Ragan Petrie
	Yu (Yvette) Zhang
Head of Department,	Mark L. Waller

May 2021

Major Subject: Agricultural Economics

ABSTRACT

From an individual perspective, one important way through which economics affects our daily lives is the framing of choices we make about our jobs, school, and what we have for dinner. People make thousands of decisions everyday, and economics deals with the optimization of those decisions in terms of individuals and society as a whole. Classical economic theory deals with the optimization of individuals' utility based on the rationality assumption. This dissertation studies individuals' decision-making from the perspective of behavioral economics which is primarily concerned with bounded rationality. I specifically focus on three different types of decision-making: insurance-purchasing, competition entry and employees' labor effort choice.

In chapter 1, I compare agricultural insurance purchasing decisions under three different insurance schemes: purely voluntary, purely compulsory and mixed insurance (i.e. insurance with a compulsory and a voluntary part). The question of what type of agricultural insurance has the highest social welfare has received a great deal of attention in theoretical studies. However, a general consensus is far from being reached. In addition, there is no empirical study comparing all three insurance schemes in previous literature. It is very difficult to find a natural experiment or use empirical data to compare all three insurance types simultaneously. In order to investigate this question, adverse selection and moral hazard are two main issues that need to be jointly addressed. In this chapter, I use the Balloon Analogue Risk Task (BART) as the assessment of risk-taking and insurance context to conduct an online experiment. I find adverse selection in purely voluntary insurance, but advantageous selection in mixed insurance. Moral hazard exists in all three types of insurance, but it is smaller in mixed insurance. The ancillary results suggest that under the combined effects of significant moral hazard and "no adverse selection" in purely compulsory insurance make it the insurance type with the lowest social earnings. Overall there is no crowding-out effect of the compulsory part on residual voluntary purchases in mixed insurance.

In chapter 2, I study the effects of social roles on willingness to compete, especially for women. It has been well documented that women often respond less favorably to competition than men even

when they have similar abilities. Economists are increasingly interested in investigating whether such gender differences in competitiveness may be useful for explaining persistent labor market differences. Competition aversion may explain why women are less likely to seek job promotions or to choose more lucrative and competitive fields [1]. Selecting out of certain labor markets is costly for society, especially when competent women are reluctant to compete for positions for which they are the best suited candidate. In this chapter, I ask whether social roles play an important role in individuals' willingness to compete by running a laboratory experiment. Subjects compete in two-person teams. In the treatment, one team member is randomly assigned the role of "breadwinner", and the other person is randomly assigned as the "supporter". There are no differences between the roles in terms of payment, power, or effort. The only difference is the framing of the gender roles reminiscent of western society. In the baseline, subjects compete without any role assignment. We find women's WTC increases by 36.5% when they are assigned as breadwinners compared to women in the baseline. The increase in WTC is mainly driven by high-ability women, and their expected earnings are 44.2% higher compared to high-ability women in the baseline. In the treatment, breadwinner's WTC is significantly higher than the supporter's WTC for both men and women, and the overall gender gap in WTC is not significant. We examine confidence, risk preferences, responsibility and social norms for competitiveness as potential mechanisms through which social roles affect WTC. We further test and replicate our laboratory results in an online experiment using roles prevalent in the workplace: "manager" and "assistant".

In chapter 3, I focus on the effects of promises on employees' effort choices. A strong employment relationship benefits the satisfaction, productivity and welfare of employees, and it is the key to the success of an organization. We experimentally examine the effects of using non-binding promises along with a claimed wage from the employee to boost the wage and effort level in a one-shot gift-exchange game. The "claim and promise" setup allowed us to test reciprocity in the "gift exchange" between employers and employees and the guilt-aversion theory in promise-keeping. We find that when the employer trusts the employee and provides at least the claimed wage, the employee reciprocates by keeping his promise or exerting an even higher level of effort. However,

when the employer offers a wage lower than the claimed wage, the employee retaliates by breaking the promise. The main results hold when employees must perform a real-effort task. In the real-effort paradigm, employers are less trusting, and there is less retaliation from employees due to the extra information about ability. In both stated and real effort paradigms, the wage, effort level, and final social payoffs are higher in the trust scenario of the “claim and promise" treatment compared to the baseline. We used pupil dilation and eye-tracking lookup patterns to help assess guilt and reciprocity dynamically and test the psychological game theory model.

DEDICATION

To my dear family.

ACKNOWLEDGMENTS

I would like to express my sincere gratitude to my advisor, Prof. Marco Palma, for his continuous support and encouragement throughout the course of my Ph.D. It was a great privilege to work and study under his guidance. Prof. Palma is not only the research mentor, but also the life mentor for me. His patience, motivation, and enthusiasm will always guide me move forward.

Next, I acknowledge my committee member, Prof. Catherine Eckel. Her work has deeply inspired me. I am extremely grateful for her suggestions on my research and the help she has offered me when I was in the job market. My sincere thanks also goes to Prof. Ragan Petrie, and Prof. Yu Zhang for their valuable comments on my research and help in my life.

I am sincerely grateful to my loving parents who have given huge support to me throughout. The love, patience, caring, understanding, and education to me from my parents are the biggest reason I can finish this dissertation and keep pursuing my academic dreams. I am very much thankful to my husband, Xin, a macroeconomist, the smartest man I ever know, who knows every single piece of my research. He is always the first person I discuss with whenever I either have a question or an idea. He has helped me a lot in solving challenging economics theories and technical questions. On top of that, his love and support lighted up the future during stressful times. My special thanks also goes to my best friend Ziwei. Her smile, encouragement, and company have brought so much happiness to my life. I couldn't have finished my Ph.D without all my beloved ones.

CONTRIBUTORS AND FUNDING SOURCES

Contributors

This work was supervised by a dissertation committee consisting of Professors Marco A. Palma and Yu (Yvette) Zhang from the Department of Agricultural Economics and Professors Catherine C. Eckel and Ragan Petrie from the Department of Economics.

All work for the dissertation was completed independently by the student.

Funding Sources

Chapter 3 was supported by the IFREE Small Grants Program. Graduate study was supported by a scholarship from Texas A&M University.

NOMENCLATURE

AOI	Areas of Interest
AS	Adverse Selection
BART	Balloon Analogue Risk Task
CRRA	Constant Relative Risk Aversion
C&P	Claim and Promise
MHL	Moral Hazard Level in the Last Balloon
PGT	Psychological Game Theory
WTC	Willingness to Compete

TABLE OF CONTENTS

LIST OF FIGURES

LIST OF TABLES

1. INTRODUCTION AND COMPARISONS OF INSURANCE TYPES: AN ONLINE EXPERIMENT[1]

1.1 Introduction

Agricultural production is inherently a business with risk and uncertainty.[2] Farmers face a variety of production and price uncertainty resulting from weather, pests, the market, and many other factors. Agricultural insurance plays an important role in protecting farmers against these risks and helping them make better investment decisions. Insurance is not only a topic of interest to farmers, but also to insurance providers and policy makers.[3] Agricultural insurance can be classified into two broad categories: Compulsory (government/public) and Voluntary (market/private). Producers can freely choose to purchase voluntary insurance or not, but they are unable to refuse compulsory insurance. For example, in Japan, crop insurance is compulsory for farmers who grow crops over three-fourths of an acre. In India and Philippines, farmers who access loans from banks and other financial institutions are required to buy crop insurance. Compulsory agricultural insurance is usually provided and required by the government. The other category —voluntary insurance— is normally provided by private insurance companies. Many kinds of agricultural insurance, such as flood insurance, livestock and weather insurance in many countries around the world belong to this category. In practice, the vast majority of compulsory insurance is partially compulsory. In this case, public insurance provides only partial coverage, and it allows for supplemental voluntary purchases. For example, flood insurance is compulsory to high-risk flood areas in the United States, but the compulsory requirement can be for partial coverage. Mixed insurance is also relevant in other domains outside of agriculture. For example, the U.S. Medicare program covers only half of all health expenditures for Americans aged 65 and older. In these cases, people can freely choose to buy additional health insurance from the private market to increase their coverage. In our article,

[1] Reprinted with permission from “Compulsory versus Voluntary Insurance: An Online Experiment" by Peilu Zhang and Marco A. Palma, 2021. American Journal of Agricultural Economics, 103.1, 106-125, Copyright [2021] by Wiley.

[2] This study was approved by the IRB.

[3] In the United States, Federal Crop Insurance represents 9% of the latest Farm Bill.

we refer to this type of insurance as mixed insurance.

Not surprisingly, the question of what type of agricultural insurance has the highest social welfare has received a great deal of attention in theoretical studies. However, a general consensus is far from being reached. In addition, there is no empirical study comparing all three insurance schemes in previous literature. It is very difficult to find a natural experiment or use empirical data to compare all three insurance types simultaneously. In order to investigate this question, adverse selection and moral hazard are two main issues that need to be jointly addressed. The objective of this paper is to conduct an experiment to compare the three insurance schemes in terms of both adverse selection and moral hazard. A theoretical model is constructed and tested using the experimental data.

In our experiment, we use the Balloon Analogue Risk Task (BART) as the assessment of risk-taking and economic context in which insurance options are set to protect participants from zero earnings due to explosions of balloons. BART, developed by [4], is a balloon-pumping task used to measure risk-taking behavior. By setting different balloons with or without insurance options, our experimental design allows for testing advantageous/adverse selection and moral hazard for each insurance scheme.

Adverse selection is a well-known phenomenon in the insurance market. It describes a situation where an individual's demand for insurance is positively correlated with the individual's riskiness. Due to asymmetric information, insurers are not able to distinguish high-risk individuals from low-risk individuals.[4] More recently, however, a growing empirical literature suggests a negative relationship between insurance demand and riskiness [10, 11, 12, 13], which conflicts with the implications of the standard economic model of insurance. One factor potentially causing this negative relationship is heterogeneity in risk aversion, which is absent in standard adverse selection models (for a review, see [14]). Higher-risk individuals are less likely to buy insurance because they are also less risk-averse. [15] first used the term "propitious selection" to describe

[4]Adverse selection was initiated by formal models in [5] and [6]. It has been found in different insurance markets. For example, [7], [8], [9] use empirical evidence to show adverse selection in crop insurance.

this phenomenon.[5] In this paper, balloons in BART without an insurance option allow us to measure risk preferences. We investigate adverse selection by comparing the level of risk aversion of individuals who purchased insurance to individuals who did not purchase insurance. We find that mixed insurance leads to advantageous (propitious) selection, as the partial compulsory part induces the least risk-averse individuals not to buy the additional (voluntary) insurance. We also find adverse selection still exists in purely voluntary insurance. Since we include risk preferences in the discussion of selection, adverse selection in our case is endogenous. Note that endogenous adverse selection is fairly common in agricultural production. For example, [20] show that farmers who are more willing to invest in risky production methods and technologies are more likely to purchase index insurance. In contrast, compulsory insurance is intended to prevent adverse or advantageous selection, as it forces all individuals to purchase insurance [21, 5].

Moral hazard is another factor that influences insurance welfare. Moral hazard arises when an individual engages in riskier behavior after purchasing insurance because the risks are transferred to the insurance company. Moral hazard has been extensively tested in crop insurance by investigating input (chemical/pesticide use) or output (yield) [22, 23, 24, 25, 26, 27].[6] In the present experiment, we compare insured subjects' performance in balloons with insurance to their performance without insurance to study moral hazard. We show moral hazard exists in all three types of insurance. Our BART set-up also provides a measure of the degree of moral hazard, and we find mixed insurance has a smaller degree of moral hazard.

Finally, we include adverse selection and the degree of moral hazard into the analysis of social earnings. Under the combined effects of "no adverse selection" and significant moral hazard, purely compulsory insurance has the lowest social earnings.[7] Based on our social earnings analysis, mixed insurance is the preferred insurance type.

[5][16] provide a theoretical model of advantageous selection. [17] points out that the "propitious selection" argument is appealing, however it has not gained much interest. [18] and [19] use a theoretical model and simulation techniques respectively, to suggest propitious selection cannot imply a final negative relationship after insurance at equilibrium.

[6]Moral hazard has also been studied outside of the agricultural field. For example, [28] and [29] evaluate moral hazard in health insurance by studying medical care expenses and health care services utilization.

[7]Social earnings are used as an indicator of social welfare in our experiment. Social earnings consists of three parts: the government, the insurer, and the consumer.

To the best our knowledge, our paper is the first to compare all three insurance types in terms of both adverse selection and moral hazard theoretically and experimentally. There has been considerable experimental work in insurance topics, including both laboratory and field experiments. Lab insurance experiments focus more on market design, such as testing prospect theory [30], and studying insurance purchase decisions and strategy. Most lab experiments use hypothetical situations. For example, [31] asked participants to imagine a hypothetical scenario where they own a house worth US $100,000, and then asked them to make insurance purchasing decisions. [32] studied the effects of microinsurance on farmers' sow production by conducting a randomized field experiment in China. [33] studied crop insurance in a pilot experiment in rural Ghana. [34] simulated a fertilizer purchase situation to study moral hazard in weather-index insurance. [35] takes advantage of a two-level randomized allocation of insurance to identify asymmetric information in crop insurance in the Philippines. Unlike previous field experiments focusing on one specific agricultural insurance market, our study compares the three broad agricultural insurance categories.

The rest of paper is organized as follows. Section 1.2 presents the theoretical model and hypotheses. Section 1.3 introduces the Balloon Analogue Risk Task. In section 3.3, we present the experimental design and procedures. The analysis and results are presented in section 2.4. Section 1.6 shows robustness checks to test the validity of BART, and section 3.7 concludes.

1.2 Theoretical Hypotheses

We first use backward induction to show the two properties implied by advantageous selection, and test them in voluntary and mixed insurance. We then solve the equilibrium insurance contract when advantageous selection exists. Finally, we explore the degree of moral hazard under the equilibrium insurance contract.

1.2.1 Properties of Advantageous Selection

Consider a simple model with two possible states of the world. In S_1 there is no accident, and individual i suffers no loss. In S_2, an accident occurs and i suffers a loss of C. We assume individuals only differ in risk aversion, and consider two levels of risk aversion, $\theta \in \{L, H\}$, which is not observable by insurance providers. Type L indicates less risk-averse individuals, and more risk-averse individuals are indicated by type H. All individuals have identical levels of initial wealth W. The probability of loss in S_2 for each individual is π_i. Individuals can take activities z_i to prevent the accident in S_2. The preventive activities z_i are assumed to be costly and unobservable by the insurers. Clearly, π_i is decreasing with z_i.

Following [18], we study insurance in a three-stage procedure. In the first stage, zero-profit insurance contracts are offered by the insurance company, which includes the per unit premium p (i.e., premium-indemnity ratio) and the coverage rate $\delta \in [0, 1]$. The total premium is $P = p * \delta$. In the second stage, individuals decide which contract to purchase. In the final stage, they decide how much preventive activities z_i to exert.

[18] make two properties underlying the propitious selection argument, and they focus on investigating whether propitious selection can explain a negative correlation between risk and insurance coverage after purchasing insurance. We focus on studying the existence of advantageous selection itself in different insurance types, and we examine the two properties of advantageous selection.

We solve the three-stage setup by backward induction, starting with the choice of preventive activities z_i. We use von-Neuman-Morgenstein's utility function u_θ to represent the individual preferences over lotteries, where u_H is obtained by an increasing concave transformation of u_L. The following equation shows the expected utility of a type θ individual who chooses the insurance contract (P, δ) and exerts preventive activities z.

$$U_\theta(z, P, \delta) = \pi(z)u_\theta(w - P - (1 - \delta)C) + (1 - \pi(z))u_\theta(w - P) - c(z) \tag{1.1}$$

where $c(z)$ is the cost of z. Thus, the optimal $z_\theta(P, \delta)$ of type θ is given by

$$z_\theta(P, \delta) = arg \max_{z>=0} U_\theta(z, P, \delta) \tag{1.2}$$

Since for a given level of δ, the premium will not change with z, insurance provides motives for individuals to decrease the level of z. The optimal z decreases with the coverage rate. Advantageous selection first assumes:

Property 1 (Regularity): For any insurance contract, a more risk-averse individual (H) exerts more preventive activities z_i compared to a less risk-averse individual (L).

We now look at the second stage to determine which insurance contract (P, δ) the individual will choose. Plugging the optimal z (2) in the utility function (1), we obtain the indirect utility function of type θ for an insurance contract. The choice of insurance contract depends on whether the marginal willingness to pay for insurance is monotone in risk-aversion, (i.e., the single-crossing property presented next). There are two simultaneous effects going in opposite directions, and we refer to them as the “preference effect" and the “risk effect". On the one hand, the more risk-averse individual H has a higher willingness to pay for insurance compared to the less risk-averse individual L (preference effect). On the other hand, the more risk-averse individual H exerts more preventive activities and has lower risk, resulting in a lower willingness to pay for insurance (risk effect). Table 1.1 summarizes these two opposing effects.

Table 1.1: Comparison of the Willingness-to-pay for Insurance (Reprinted from [2])

	WTP		WTP
Preference effect	More Risk-averse (low-risk)	$>$	Less Risk-averse (high-risk)
Risk effect	More Risk-averse (low-risk)	$<$	Less Risk-averse (high-risk)

The advantageous selection argument assumes that:

Property 2 (Single crossing): Given the optimal preventive activities z, for any insurance contract, a more risk-averse individual (H) has a higher marginal willingness to pay for insurance.

Property 2 assumes that the “preference effect" dominates the “risk effect" for any given insurance contract. To summarize, advantageous selection exists only when Property 1 and Property 2 hold. In Property 1, there are no opposite effects at play, and it is straightforward for it to hold for all three insurance schemes. In other words, advantageous selection exits when the “preference effect" dominates the “risk effect". Adverse selection exits when the “risk effect" dominates the “preference effect".

We now test the dominance of the two effects in purely voluntary and mixed insurance respectively. We follow [36]'s dual theory under which the expected utility is linear in wealth but non linear in probabilities. Now in equation (1), the utility of purchasing the insurance contract (P, δ) for an individual is

$$\begin{aligned} U(\pi, P, \delta; \alpha) &= (1+\alpha)\pi(w - P - (1-\delta)C) + (1 - (1+\alpha)\pi)(w - P) \\ &= w - P - (1+\alpha)(1-\delta)\pi C. \end{aligned} \tag{1.3}$$

where $\alpha >= 0$ captures the level of risk-aversion where a higher α represents higher risk-aversion. In other words, the individual overestimates the expected utility in state S_2 by a fraction α. With this utility function, the optimal preventive activity z is chosen by

$$\max_z = w - P - (1+\alpha)(1-\delta)\pi(z)C - c(z) \tag{1.4}$$

Following [18], we assume that $C = 1$, $c(z) = z^2$ and $\pi(z) = \frac{1-z}{2} \in [0, \frac{1}{2}]$ for $z \in [0, 1]$. Taking the first order condition of equation (4), we obtain the optimal preventive activity:

$$z_\alpha(\delta) = \frac{(1+\alpha)\delta - (1+\alpha)}{4} \tag{1.5}$$

We define $\theta = \frac{1+\alpha}{2}$ (with $\theta \in [\frac{1}{2}, 1]$), then z in equation (5) is:

$$z_\theta(\delta) = \frac{(1-\delta)\theta}{2} \in [0, \frac{1}{2}] \tag{1.6}$$

and the corresponding probability of an accident for type θ is:

$$\pi_\theta(\delta) = \frac{1}{2} - \frac{(1-\delta)\theta}{4} \in [\frac{1}{4}, \frac{1}{2}]. \tag{1.7}$$

Plugging equation (7) in the utility function (3), we have the induced utility function:

$$\begin{aligned} U_\theta(P,\delta) &= w - P - (1+\alpha)(1-\delta)\pi \\ &= w - P - \theta(1-\delta) + \frac{(1-\delta)^2}{2}\theta^2. \end{aligned} \tag{1.8}$$

Using the envelope theorem to take the derivative of P in equation (8) with respect to δ, we get the marginal willingness to pay for insurance:

$$\left.\frac{\partial_P}{\partial_\delta}\right|_{u_\theta} = \theta - (1-\delta)\theta^2 \tag{1.9}$$

In mixed insurance, a uniform compulsory insurance contract $(P^{'}, A)$ is provided before people make purchasing decisions. The utility of purchasing the residual voluntary insurance contract $(P^{''}, \delta_1)$ for an individual is:

$$\begin{aligned} U(\pi, P^{'} + P^{''}, \delta_1; \alpha) &= (1+\alpha)\pi(w - P^{'} - P^{''} - (1 - A - \delta_1)C) + (1-(1+\alpha)\pi)(w-P) \\ &= w - P^{'} - P^{''} - (1+\alpha)(1-A-\delta_1)\pi C. \end{aligned} \tag{1.10}$$

where $A + \delta_1 <= \delta$.

By doing the same process for purely voluntary insurance above, we get the marginal willingness to pay for the residual voluntary part:

$$\left.\frac{\partial_{P'+P''}}{\partial_{\delta_1}}\right|_{u_\theta} = \theta - (1 - A - \delta_1)\theta^2 \tag{1.11}$$

We compare equation (9) and (11) to test the difference in willingness to pay for insurance when we introduce mixed insurance:

$$\left.\frac{\partial_P}{\partial_\delta}\right|_{u_\theta} - \left.\frac{\partial_{P'+P''}}{\partial_{\delta_1}}\right|_{u_\theta} = (\delta - (A+\delta_1))\theta^2 \tag{1.12}$$

Thus, the resulting sign in equation (12) depends on $\delta - (A + \delta_1)$. Less risk averse individuals are less likely to buy the additional voluntary part in mixed insurance, and hence we have:

$$\begin{aligned} &\delta > (A + \delta_1) \text{for less risk-averse individuals} \\ &\delta = (A + \delta_1) \text{for more risk-averse individuals.} \end{aligned} \tag{1.13}$$

In other words, the partially compulsory insurance in mixed insurance reduces the willingness to pay for the additional voluntary insurance for less risk-averse individuals (i.e., crowding-out effect), but it has no effect on more risk-averse individuals. Thus, Property 2 is more likely to hold in mixed insurance. We then have the following hypothesis:

Hypothesis 1: Given the same total insurance contract, the “Preference effect" is more likely to dominate the “Risk effect" in mixed insurance compared to purely voluntary insurance, i.e., mixed insurance leads to advantageous selection.

1.2.2 Equilibrium Contracts

We solve the pooling equilibrium insurance contracts when advantageous selection exists. Consider the proportion of type L, less risk-averse individuals in the population is λ, and type H, more risk-averse individuals is $1 - \lambda$. In Figure 1.1, the two curves $P(L)$ and $P(H)$ represents the premium for each type. Due to moral hazard, the premium is an increasing and convex function of coverage. Property 1 ensures that for any given coverage, the premium for less risk-averse type $P(L)$, lies above it for more risk-averse type $P(H)$, except for $\delta \in \{0, 1\}$. When the insurer provides the same contract to both types L and H, the charged premium is $\lambda P(L) + (1 - \lambda P(H))$ which is represented by the dotted curve. The single-crossing property is depicted by $w(l)$ and $w(h)$. It is clear from the graph that a pooling equilibrium does not exist, as it is always possible for an insurer to propose a contract that attracts only the more risk-averse individual and makes a positive profit.

[18] show the existence of a separating equilibrium (see Appendix A.1). Under the separating equilibrium, the more risk-averse individuals are overinsured. The intuition is based on Property 2,

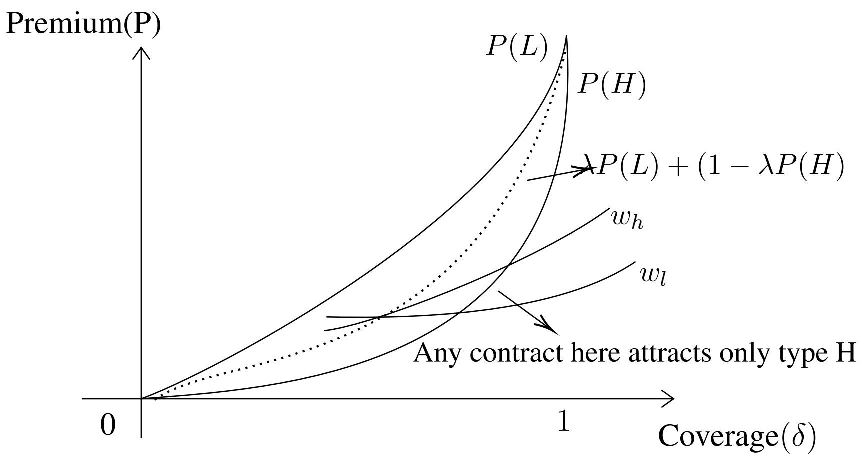


Figure 1.1: Non-existence of Pooling Equilibrium of Insurance Contract when Advantageous Selection Exists (Reprinted from [2])

the insurer needs to over-provide insurance to the more risk-averse individuals so that the insurer can separate them from the less risk-averse individuals. Thus, the final degree of moral hazard after insurance provision also depends on two opposite forces. On the one hand, more risk-averse individuals exert more preventive activities z; on the other hand, more risk-averse individuals purchase more insurance, and preventive activities decrease with insurance coverage. We denote these two effects as the “preference effect" and the “coverage effect" in Table 1.2.

Table 1.2: Comparison of the Degree of Preventive Activities (Reprinted from [2])

	Preventive Activity		Preventive Activity
Preference effect	more risk-averse (more coverage)	$>$	less risk-averse (less coverage)
Coverage effect	more risk-averse (more coverage)	$<$	less risk-averse (less coverage)

We derived the expected utility function for purely voluntary and mixed insurance separately (equations 3 & 10). We now derive the expected utility function for purely compulsory insurance:

$$U_c(\pi, P^*, \delta^*; \alpha) = (1+\alpha)\pi(w - P^* - (1-\delta^*)C) + (1-(1+\alpha)\pi)(w - P^*)$$
$$= w - P^* - (1+\alpha)(1-\delta^*)\pi C. \tag{1.14}$$

Purely compulsory insurance means identical purchase which is indicated by δ^* in equation (14). P^* is the identical premium. In order to do comparison, we set $\delta^* = \delta$ and $P^* = P$. We denote Z_v, Z_m and Z_c as the optimal preventative activity in purely voluntary, mixed, and purely compulsory insurance respectively (i.e, Z maximizing equations 3, 10, 14 respectively).

In mixed insurance, we expect people who choose not to buy the additional (voluntary) insurance are less risk-averse, and this means they have less coverage. Since they are less risk-averse, we can expect the coverage effect to dominate the preference effect. [18] argue this dominance as well. Thus, the final degree of moral hazard in mixed insurance is lower than in purely voluntary insurance, i.e, $Z_m < Z_v$. In purely compulsory insurance, both types of individuals are in the consumer pool. Compared to mixed insurance, less risk-averse individuals in purely compulsory insurance have more coverage, and the degree of moral hazard is higher, i.e., $Z_c > Z_m$. Compared to purely voluntary insurance, more type H (more risk-averse) individuals are in the consumer pool in purely compulsory insurance, and the degree of moral hazard is lower, i.e., $Z_c < Z_v$. We then have the following hypothesis.

Hypothesis 2: Given the same total insurance contract, mixed insurance has the lowest degree of moral hazard, and purely voluntary insurance has the highest degree of moral hazard.

We next describe the experiment to test our theoretical hypotheses.

1.3 The Balloon Analogue Risk Task (BART)

BART is a computerized measure of risk-taking behavior developed by [4]. In this task, participants are presented with a balloon and they receive a monetary reward for each successful pump. However, if the balloon explodes, they receive nothing. A higher number of pumps yields higher potential earnings, but it also represents a higher risk of explosion. Therefore, risk-taking is measured by the selected number of pumps, with more pumps indicating more risk-taking behavior. The participants know the balloon may explode at some point and that a higher number of pumps has a higher risk of explosion, but they do not know the actual probability function. The probability of explosion of a balloon is arranged by constructing an array of N numbers. In our experiment, the array of each balloon is 1-128.[8] Thus, the probability that a balloon will explode at the first pump is 1/128. The probability of explosion at the second pump is 1/127 if the balloon did not explode after the first pump; 1/126 at the third pump, and so on up until the 128th pump, at which the probability of explosion is 1. According to this arrangement, the expected earnings for each balloon are a bow-shaped function with a maximum at the 64th pump, which is also the expected explosion point.

The participants' decisions can be formalized as the choice of the lotteries:

$$L = \begin{cases} 0 & probability : k/128 = 1 - (128 - k)/128 \\ \gamma k & probability : (128 - k)/128 = (127/128 * 126/127 * ... * (128 - k)/(129 - k)) \end{cases}$$

where k is the number of pumps which in our experiment $\in [0, 128]$, while $\gamma > 0$ is a scale factor. The expected value of these lotteries is equal to $\gamma(128k - k^2)/128$. Assuming a constant relative risk aversion (CRRA) utility function $u(k) = k^r$, the BART allows for the estimation of the coefficient of risk aversion. The implied levels of *r* for every possible choice *k* are shown in Appendix Table A.1. The table implies a *risk-neutral* individual would choose $k^* = 64$. A *risk-seeking* participant whose utility function is convex would choose more than 64 pumps; a *risk-averse* participant with a concave utility function would choose less than 64 pumps. The insurance premium and coverage in our experiment are designed based on this algorithm of BART (explained in detail in the experimental design section).

[3] developed the Automatic version of BART, in which participants input their desired number of pumps into a box and the balloon is pumped automatically. If a balloon explodes before the

[8] In the original paper, the association of the BART with self-reported risk behavior in the real world occurred only with data from the balloon with a maximum number of 128 pumps [4]. For this reason, we choose a range of [0,128] pumps in our experiment.

indicated number of pumps is reached, participants lose all their earnings for that balloon. [3] show that this version does not change the validity of BART as an assessment of risk-taking. In order to observe risk-taking of successful pumps and explosions (i.e., avoid the data truncation), we use the Automatic Version of BART. In the original version of BART, risk-taking can only be observed for balloons that did not explode. Thus, risk-taking in our experiment is quantified by the selected number of pumps.

To prevent potential negative outcomes associated with risk-taking behavior in agricultural production, agricultural economists started to develop more reliable approaches to study behavior under risk and uncertainty back in the 20th Century [37, 38]. [39] is the first to conduct an incentivized gamble choice game with India farmers to test their risk preferences. Nowadays, there are several methods to elicit risk preferences that have been widely used and tested, including [40], [41], and [42], among others. In this paper, using BART as the assessment of risk-taking and insurance context in our experiment has three important advantages. First pumping a balloon is a simple task and a relatively under-studied context, and this task may incur “accidents" (explosions) so that we can incorporate insurance options. Second we can use the performance with and without insurance to test for both adverse selection and moral hazard. Third, BART also provides a good environment to study endogenous adverse/advantageous selection.[9]

1.4 Experimental Design

The experiment was conducted online using Amazon MTurk.[10] We use a between subject design with three treatments. The treatments are the three types of insurance: purely voluntary,

[9] The main criticism of BART is that the procedure is ambiguous as it does not provide objective probabilities to participants. [43] develop a new protocol based on BART to address the methodological issues in BART (BERT). In our experiment, although subjects do not know the objective probabilities, they know that a higher number of pumps yields higher earnings and also higher risk of explosions, and more importantly, we inform participants the maximum number of pumps of each balloon to reduce the ambiguity [44].

[10] To avoid limiting the subject pool to only students, we conduct our BART insurance experiment online using Amazon Mturk. We are aware that using Mturk comes at a cost such as no face-to-face instructions. However, our task, pumping a balloon is a simple task demanding low cognitive efforts. It sets up a context that is familiar and easy to grasp for participants [45]. In addition, [46] find that the data obtained from Mturk are at least as reliable as those obtained via traditional methods. [47] show Amazon MTurk workers can provide consistent and economically meaningful data. Our experiment actually further confirms it by showing the average number of pumps in BART is very similar to the original results in [4]. The average number of pumps without insurance options is 57 in our experiment, and the average number of pumps in the original paper is 59.

purely compulsory, and mixed insurance. Subjects signed a consent form and then proceeded to the BART task. In the BART section, subjects were informed that they will play with **30** sequential balloons which have a potential maximum of 128 pumps each; the reward for each successful pump is ¢1. The explosion points of each balloon were randomly drawn on the spot. Subjects were asked to indicate the number of pumps they want to select for each balloon. To avoid the effects of earnings from previous balloons on the performance of subsequent balloon, we randomly select three balloons at the end of the experiment to determine the participant's final payments.

Treatment 1: Purely Voluntary Insurance. For the **first** and **last** balloon, subjects were allowed to voluntarily buy insurance at a premium of ¢40 before pumping those two balloons. Subjects did not know this information until they played with those particular balloons. For the other **28** balloons, they played BART without the insurance option. The insurance in this treatment is voluntary. If the insured balloon explodes, participants receive ¢64. If the insured balloon does not explode, participants receive nothing from the insurance, and the cost is not refunded. In order to ensure understanding of the procedure and the insurance scheme, subjects had to correctly answer a quiz before proceeding to pump the balloons.

In our experiment, the maximum number of pumps for a balloon is 128, which implies the balloon will surely explode at the 128th pump. According to the BART algorithm, the optimal number of pumps in terms of expected rewards is 64. Thus, we set the insurance coverage equal to the actual earnings at the optimal pump, which is ¢64. The probability of a balloon exploding at the 64th pump is 1-(127/128*126/127*...*64/65)=1/2, and hence, the expected insurance benefits are 64*1/2=32. The actuarially fair insurance premium is ¢32. However, in real life, insurance is rarely actuarially fair, since insurance companies make profits, and also due to transaction costs, administration fees, moral hazard, adverse selection and risk premium.[11] Therefore, the premium is usually higher than the expected benefits from insurance. Thus, in our paper, for simplicity, we set the premium to ¢40.[12]

[11] Risk premium is the maximum amount a person will pay above actuarially fair premiums.

[12] Since in this paper, our objective is to compare different insurance types under the same coverage and premium instead of the effects of different premiums and coverage, for simplicity, we calculate the premium above based on a risk-neutral individual.

In our experiment, we only set the insurance option for the first and last balloon. We view participants' insurance purchasing decisions in the first balloon as a reflection of purchasing behavior before learning. The performance of the last balloon can be viewed as insurance purchasing behavior after learning. All the other balloons serve as within subject comparisons without insurance allocation, i.e., the measurement of risk-taking by BART.

Treatment 2: Purely Compulsory Insurance. This treatment is identical to the first treatment, except that insurance is compulsory, which means subjects were required to buy the insurance for the first and last balloon at a cost of ¢40.

Treatment 3: Mixed Insurance. This treatment is the same as the second treatment, except that the insurance is partially compulsory. This means the compulsory insurance only pays ¢32 to the subject if the insured balloon explodes, and the subject is allowed to buy additional voluntary insurance to obtain full coverage. The premium for the compulsory part is ¢20. The premium and coverage for the voluntary part are also ¢20 and ¢32 cents, respectively. The calculations of the premium and coverage are the same as in treatment 1, and we simply split the premium and coverage into two equal parts.

At the end of the BART task, all participants are asked to respond to the Domain-Specific Risk-Taking Scale (DOSPERT) questionnaire [48],[13] a Sensation Seeking Scale questionnaire [49], a gamble-choice task [41] and a demographic survey. The DOSPERT and Sensation Seeking Scale are used to check whether participants in different treatments have different original risk preferences. They are also used as robustness checks to evaluate whether the riskiness in BART is associated with psychological measures of risk-taking and self-reported risk behavior. The gamble choice, as one of the most widely used methods of eliciting risk preferences in laboratory experiments, serves the same purpose.

The experiment was computerized in Inquisit [50]. In total there were 305 subjects, with about 100 participants per treatment. Six subjects were excluded due to incomplete information. There were 13 subjects whose choice of pumps were less than the price of insurance when they chose to

[13]The DOSPERT scale is designed to elicit the domain-specific nature of risk preferences.

buy insurance, and we treat them as not understanding the insurance scheme and exclude them from the analysis. Hence, the final sample consists of 286 subjects. Table 1.3 summarizes the Insurance features for each treatment. The instructions, screenshots, questionnaires and demographic survey questions are available in the Appendix A.

Table 1.3: Insurance Features for each Treatment. (Reprinted from [2])

Treatments	First & Last balloon	Ins coverage (if explodes)	Ins premium	2-29th balloon	No. Subjects
Purely Voluntary	Vol insurance	¢64	¢40	Normal BART	93
Purely Compulsory	Com insurance	¢64	¢40	Normal BART	96
Mixed	Vol+Com	¢32+¢32	¢20+¢20	Normal BART	97

Based on the premium and coverage of insurance in our experimental design, we calculate the expected utility of buying (additional) insurance and not buying (additional) insurance assuming a CRRA utility function. We show the calculation process in Appendix A3. The intuition is that we have the coefficient of risk aversion r for each pump in BART, and then for each r, we find its optimal number of pumps after full/partial insurance. Then we can calculate the expected utility of buying or not buying (additional) insurance. In general, the results suggest that risk-averse individuals will choose to buy (additional) insurance, but risk-seeking individuals will choose not to buy (additional) insurance. This means our experimental design aligns with the economic theory section about the behavior of individuals with different risk preferences. Next, we show the results from our experimental data to test our two main Hypotheses.

1.5 Main Results

The main objective of this article is to compare the three insurance schemes. We first show the analysis of adverse selection and moral hazard corresponding to our hypotheses. We also show two ancillary results about social earnings and the overall effects of the compulsory part on the residual voluntary part in mixed insurance. In our experiment, the explosion points of each balloon were randomly drawn on the spot instead of fixed points, and this may cause different learning paths across treatments. Before showing the results, we test the distribution of explosion points and individuals' average number of pumps for each balloon (path-dependence) across treatments (see figures in Appendix A4). Kolmogorov-Smirnov tests suggest there is no difference in the distribution of explosion points or individuals' path-dependence ($p > 0.1$ for all comparisons). Balanced background characteristics across treatments are shown in Appendix A5.[14]

1.5.1 Analysis of Adverse Selection

Result 1 *1a.* Mixed insurance leads to advantageous selection, and adverse selection exists in purely voluntary insurance.

1b. The advantageous selection in mixed insurance is mainly contributed by the least risk-averse participants.

In purely voluntary insurance, the purchasing rate in the first balloon is 38.71%, and 59.14% in the last balloon. In mixed insurance, the purchasing rate for the additional insurance in the first balloon is 43.3%, and 59.79% in the last balloon. To study adverse selection, we explore the compositions of individuals with different risk-aversion among subjects who purchased and those who did not purchase insurance.

Figure 2.1 compares the average number of pumps in the middle 28 balloons of participants who purchased (additional) insurance in the first or last balloon (or both) to participants who did not purchase insurance at all. The results are consistent regardless of whether we consider the first and last balloon separately.

In purely voluntary insurance (Figure 1.2a), the line describing participants who purchased insurance lies above the line of those who did not purchase insurance. Using a two-tailed Mann-Whitney U-Test with 28 observations for each subgroup, we find the average number of pumps in the 2-29th balloons for those who "buy insurance" is significantly higher than those who "do not buy insurance" ($p < 0.001$). A Kolmogorov-Smirnov test shows that the two distributions

[14]The only exception is participants in the purely voluntary insurance treatment are older than participants in the mixed insurance treatment, but in the following estimations, there are no significant effects of age.

(a)

(b)

Figure 1.2: Adverse selection: Average number of pumps in the 2-29th balloons by insurance purchasing decision. (Reprinted from [2]) *Note:* The black line represents the 64th pump which theoretically maximizes the expected earnings. In our experiment, except for the first and last balloon, the average number of pumps per balloon per subject is 57 for purely voluntary and mixed insurance, and 59 for purely compulsory insurance. This result is similar to [3] who report an average of 59 pumps.

are statistically different ($p < 0.001$). Thus, the figure indicates the presence of adverse selection in purely voluntary insurance, as the insurance is more attractive to less risk-averse (higher-risk) individuals.

The results in mixed insurance are opposite (Figure 1.2b): the average number of pumps in the 2-29 balloons for those who chose to buy additional insurance is significantly lower than for those who did not choose to buy additional insurance at all (Mann-Whitney U-Test, $p = 0.002$.) In Figure 1.2b, the dash and solid lines swap positions, and the result of a Kolmogorov-Smirnov test (p=0.012) confirms that the two distributions are statistically different. Table 2.4 summarizes the average number of pumps in the 2-29th balloon by insurance purchasing decision for both treatments. In mixed insurance, when participants make decisions on purchasing the additional voluntary part, less risk-averse (higher-risk) individuals are less likely to buy additional insurance compared to more risk-averse (lower-risk) individuals, which leads to advantageous selection.

Participants went through 30 sequential balloons, and their performance may be erratic which could affect identification. We calculate the coefficient of variation (CV) of the number of pumps in 2-29th balloon for each subject. The CV is 0.297, 0.304, 0.267 for purely voluntary, mixed and purely compulsory insurance respectively. P-values from Mann-Whitney U-Tests suggest that the

Table 1.4: Average number of pumps in the 2-29th balloons by insurance decisions. (Reprinted from [2])

	Purely Voluntary	Mixed
Buy (additional)	59	55
	(1.9)	(1.9)
Not buy (additional)	52	58
	(2.6)	(3.2)

Notes: Standard errors are in parentheses.

CV did not differ across treatments ($p > 0.3$ for all comparisons).

We now explore the potential causes of advantageous selection in Mixed Insurance. Table 2.5 shows the average number of pumps in the 2-29th balloons by gender. In purely voluntary insurance, adverse selection exists for males and females. However, in mixed insurance, there is no difference in the number of pumps in the 2-29th balloon between women who buy and women who do not buy additional insurance. In addition, advantageous selection exists for males. Males who do not buy additional insurance in mixed insurance are the least risk-averse (highest-risk) participants in these two treatments.

Table 1.5: Gender differences in adverse selection by treatment. (Reprinted from [2])

	Purely Voluntary			Mixed		
	Buy	Not buy	P-value	Buy	Not buy	P-value
Male	61	53	0.001	58	63	0.002
	(2.7)	(3.2)		(2.7)	(3.4)	
Female	57	51	< 0.001	53	51	0.436
	(2.9)	(3.9)		(2.8)	(5.6)	

Notes: P-values are from two-sided Mann-Whitney U-Tests. Standard errors are in parentheses.

We first use “64 pumps" as the threshold to estimate a linear probability model (LPM) for subjects with different risk preferences. As shown in columns 1-2 of Table 1.6, mixed insurance (“Mixed") has no significant effects on insurance purchases of either risk-averse (< 64 pumps) or risk-seeking (> 64 pumps) participants, but the sign of “Mixed" is opposite for these two subgroups: positive for risk-averse subjects, and negative for risk-seeking subjects. We further estimate the LPM for the top 15% risk-seeking (high-risk) participants whose average numbers of pumps are higher than 70. The results in column 3 show that the mixed insurance scheme significantly decreases the likelihood of the risk-seeking participants to purchase insurance. In column 4, we control for the average number of pumps in 2-29th balloon (MeanPumps229), gender (Male), age, education, income and household size (HHsize), and the coefficient of mixed insurance is marginally significant.[15]

Table 1.6: The Effects of Mixed Insurance on Purchase Rate (Reprinted from [2])

	$<=64$ pumps (1)	>64 pumps (2)	top 15% pumps (3) ($>=70$ pumps)	top 15% pumps (4) ($>=70$ pumps)
Mixed	$0.094(0.081)$	$-0.189(0.130)$	$-0.335^{**}(0.156)$	$-0.318^{*}(0.179)$
MeanPumps229				$-0.001(0.010)$
Male				$0.099(0.182)$
Age				$0.010(0.008)$
Education				$-0.014(0.080)$
Income				$-0.005(0.073)$
HHsize				$-0.086(0.073)$
Constant	$0.592^{***}(0.057)$	$0.818^{***}(0.096)$	$0.923^{***}(0.118)$	$0.835(0.973)$
Control	No	No	No	Yes
Observations	141	49	30	30
R^2	0.01	0.04	0.14	0. 28

Notes: The dependent variable is the dummy variable for choosing to buy insurance in the first or last (or both) balloon. Standard errors are in parentheses. $*p < 10\%, **p < 5\%, ***p < 1\%$. All results are robust to using probit.

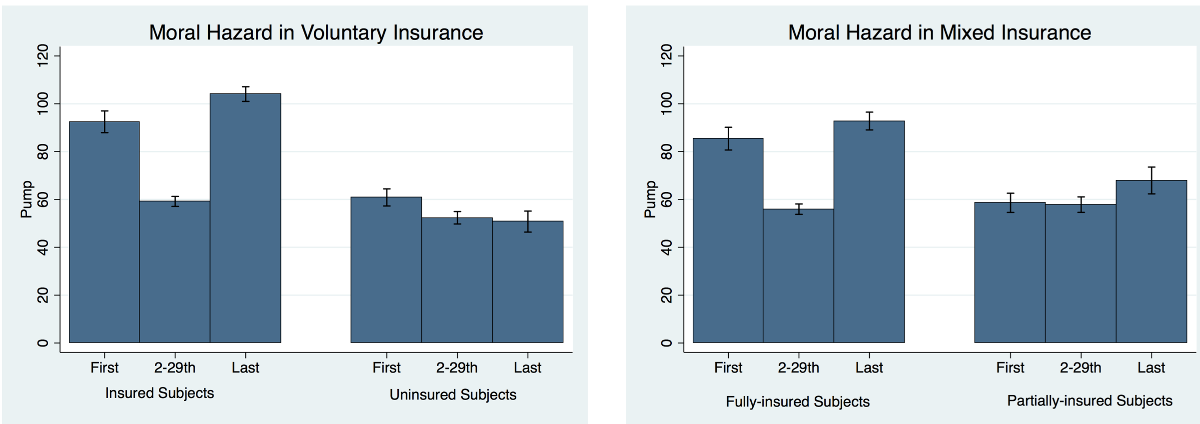
Thus far, we conclude that the elimination of adverse selection in mixed insurance is mainly contributed by the least risk-averse participants.[16] In mixed insurance, the partial compulsory part has a crowding-out effect on the voluntary part for less risk-averse individuals. This result is consistent with our theoretical Hypothesis 1.

[15] Education levels were defined as: EDU = 1 if Some High School or less, EDU = 2 if High School Diploma, EDU = 3 if Some College, EDU = 4 if 2 year/Associates Degree, EDU = 5 if 4 year/Bachelor's Degree, EDU = 6 if Some Graduate School, EDU = 7 if Graduate Degree; income levels were: INC = 1-10 corresponds to Less than $30,000, $30,000 - $39,999, $40,000 - $49,999, $50,000 - $59,999, $60,000 - $69,999, $70,000 - $79,999, $80,000 - $89,999, $90,000 - $99,999, $100,000 - $149,999, $150,000 or more respectively.

[16] As we mentioned in the introduction of BART, subjects are not aware of the accurate explosion probability of each pump, and hence we treat subjects with more pumps as less risk-averse, and include subjects whose numbers of pumps are higher than 70 into the group of least risk-averse participants.

1.5.2 Analysis of Moral Hazard

Result 2 There exists moral hazard in all three insurance types; but comparatively, mixed insurance has the lowest degree of moral hazard; purely voluntary insurance has the highest degree of moral hazard.



(a) Voluntary Insurance (b) Mixed Insurance

Figure 1.3: Moral Hazard: Comparison of the average number of pumps on different balloons in purely voluntary and mixed insurance. (Reprinted from [2])

Figure 1.3a depicts the average number of pumps on different balloons in purely voluntary insurance for insured and uninsured subjects separately. For insured subjects, we take the average number of pumps in the first ballon for subjects who purchased insurance in the first balloon (First), and do the same calculation for the last balloon (Last), and then take the average number of pumps in the 2-29th balloons for subjects who purchased insurance in the first or last or both balloons (2-29th).

The result shows that the average number of pumps in the 2-29th balloons is significantly lower than in the first and last balloon (Mann-Whitney U-tests, $p < 0.001$; $p < 0.001$). The number on the last balloon is higher than the first balloon ($p = 0.035$). The part of "uninsured subjects" of Figure 1.3a shows that there is no statistically significant difference in the average number of pumps between the 2-29th balloons and the first or last balloon for subjects who did not buy insurance at all ($p > 0.01$ for all comparisons). These results suggest moral hazard exists in purely voluntary insurance.

Figure 1.3b shows the results for mixed insurance. In mixed insurance, all subjects are insured (fully or partially), however, we find different results between subjects who purchased and those who did not purchase *additional* voluntary insurance. Using Mann-Whitney U-Tests we find a significant higher number of pumps in the first or last balloon compared to the 2-29th balloons ($p < 0.001$) for subjects who purchased additional insurance (i.e., fully-insured). There is no difference ($p > 0.1$ for each comparison) for subjects who did not buy additional insurance (i.e., partially-insured). This result suggests that moral hazard only exists for fully-insured subjects in mixed insurance.

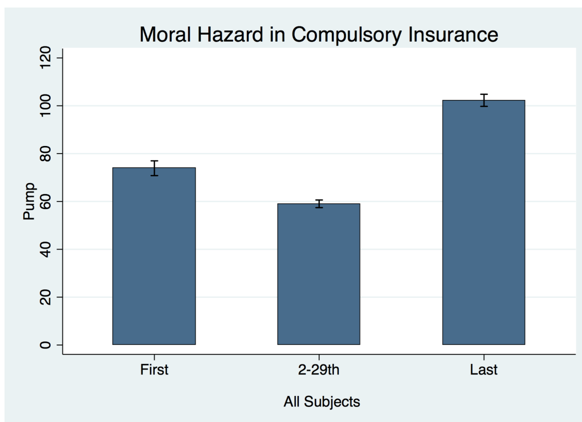


Figure 1.4: Moral Hazard: Comparison of the average number of pumps on different balloons in purely compulsory insurance. (Reprinted from [2])

Figure 2.6 shows the results for purely compulsory insurance. Every participant in the purely compulsory insurance treatment is required to buy full insurance. We find moral hazard exists in purely compulsory insurance. The average number of pumps in the first or last balloon is significantly higher than the average number of pumps in the 2-29th balloons (Mann-Whitney U-tests, p=0.001 for the first balloon; $p < 0.001$ for the last balloon).

BART allows us to further compare the degree of moral hazard for the three insurance schemes. The degree of moral hazard is quantified as the difference in the average number of pumps for insured subjects between the “first/last balloon" and the “2-29th balloons". Table 1.7 shows the average degree of moral hazard per insured subject for each treatment. We find that in the first balloon the degree of moral hazard in purely voluntary insurance (32.3) is the highest. We find no overall difference between mixed (15.2) and purely compulsory insurance (14.9), but individuals who were partially-insured in mixed insurance (5.1) have the lowest degree of moral hazard. In the last balloon, the degree of moral hazard in mixed insurance (26.0) is the lowest, which is mainly contributed by partially-insured individuals. There is no difference in the moral hazard level between purely voluntary (44.9) and compulsory insurance (43.3) in the last balloon.[17]

Table 1.7: Average degree of moral hazard per insured subject in the first or last balloon. (Reprinted from [2])

	Voluntary	Mixed (Fully-insured)	Mixed (Partially-insured)	Mixed	Compulsory
First	32.3	28.5	5.1	15.2	14.9
	(4.97)	(4.78)	(3.62)	(3.13)	(3.09)
Last	44.9	36.8	9.9	26.0	43.3
	(3.38)	(3.27)	(5.08)	(3.11)	(2.76)

Notes: Standard errors are in parentheses.

We estimate an OLS regression on the level of moral hazard in the first and last balloon. The independent variables are “Mixed" and “Compulsory" (Compulsory=0 if purely voluntary or mixed, Compulsory=1 if purely compulsory). We also control for age, education, gender, income and household size, but the coefficients for all these covariates are not significant. We show the results in Table 2.6. We estimate three regressions for the first and last balloons. The three regressions differ in the inclusion of subjects in mixed insurance: all the subjects in mixed insurance; only the fully-insured subjects; only the partially-insured subjects. The other insured subjects in the three regressions are the same: all the subjects in purely compulsory insurance and subjects who purchased insurance in purely voluntary insurance.

[17] P-values from Mann-Whitney *U*-Tests: V vs M(fully-insured): 0.471(F), 0.040(L); V vs M(partially-insured): < 0.001(F), < 0.001(L); V vs M: 0.002(F), < 0.001(L); V vs C: 0.004(F), 0.509(L); M(fully-insured) vs C: 0.015(F), 0.104(L); M(partially-insured) vs C: 0.086(F), < 0.001(L); M vs C: 0.839(F), < 0.001(L).

Table 1.8: OSL Regression on the Degree of Moral Hazard (Reprinted from [2])

	First Balloon: all subjects in M	First Balloon: M (fully-insured)	First Balloon: M(partially-insured)	Last Balloon: all subjects in M	Last Balloon: M (fully-insured)	Last Balloon: M(partially-insured)
Mixed	-17.169^{***}	-3.491	-27.305^{***}	-20.237^{***}	-10.001^{**}	-35.905^{***}
	(6.116)	(7.121)	(6.392)	(4.846)	(4.892)	(5.940)
Compulsory	-17.135^{***}	-16.885^{***}	-16.338^{***}	-2.559	-2.986	-2.254
	(6.070)	(6.067)	(5.803)	(4.796)	(4.364)	(4.759)
Constant	39.383^{***}	45.464^{***}	44.640^{***}	55.097^{***}	57.815^{***}	52.619^{***}
	(11.927)	(12.688)	(12.226)	(10.208)	(9.674)	(11.185)
Control	Yes	Yes	Yes	Yes	Yes	Yes
Observations	229	174	187	248	209	190
R^2	0.055	0.082	0.126	0.088	0.069	0.158

Notes: The dependent variable is the degree of moral hazard. Robust standard errors are in parentheses. $*p < 10\%, **p < 5\%, ***p < 1\%$.

Table 2.6 shows that in the first balloon, making insurance compulsory or mixed reduces moral hazard. The reduction of moral hazard in mixed insurance is mainly contributed by the partially-insured subjects. In the last balloon, the compulsory part has no effect, and mixed insurance reduces moral hazard. Both fully-insured and partially-insured subjects reduce the moral hazard in mixed insurance, but the magnitude of partially-insured subjects is larger, as shown in the last two columns of Table 2.6. Combining the results from the first and last balloon, we argue that mixed insurance has the lowest degree of moral hazard, and purely voluntary insurance has the highest degree of moral hazard. Recall in the analysis of adverse selection, we show subjects in mixed insurance who did not buy additional insurance are the least risk-averse subjects in our experiment. The results of moral hazard show that when the least risk-averse subjects are partially-insured, the “coverage effect" dominates the “preference effect", which is in line with our theoretical Hypothesis 2.

1.5.3 Ancillary results

Result 3 Purely compulsory insurance has the lowest social earnings.

We tested adverse selection and have the quantified moral hazard level, so we can include them into the regressions of social earnings. Social earnings in our experiment consists of three parts: the government, the insurer, and the consumer.[18] Table 1.9 shows the overall average social earnings,

[18] Some compulsory insurance in real life are paid by employers or by the insurer; for simplicity, we assume in our analysis that compulsory insurance is paid by the government.

the average net earnings from each subject for the insurer and government (henceforth, earnings of the insurer, government)[19], and also the average net earnings per consumer.

Table 1.9: Average Net Earnings (Dollars). (Reprinted from [2])

	First balloon				Last balloon			
	Society	Insurer	Consumer	Government	Society	Insurer	Consumer	Government
Purely Voluntary	0.25	-0.03	0.29	0	0.21	-0.07	0.28	0
	(0.03)	(0.02)	(0.02)	(omitted)	(0.04)	(0.02)	(0.02)	(omitted)
Purely Compulsory	0.24	0	0.22	0.03	0.11	0	0.26	-0.15
	(0.03)	(omitted)	(0.02)	(0.03)	(0.03)	(omitted)	(0.01)	(0.01)
Mixed	0.21	-0.02	0.22	0.002	0.18	-0.02	0.23	-0.02
	(0.03)	(0.01)	(0.02)	(0.02)	(0.03)	(0.01)	(0.02)	(0.02)

Notes: Standard errors are in parentheses.

We first combine the insurer and government into a broader "insurer" category. In the first balloon, there is no difference in overall social earnings or earnings of the insurer between any two insurance scheme. The consumer in purely voluntary insurance has higher earnings compared to mixed and compulsory insurance (Mann-Whitney U-tests, p=0.06 for both comparisons). In the last balloon, purely compulsory insurance has the lowest social earnings (p=0.003 (C vs V); p=0.02 (C vs M)), and there is no difference between purely voluntary and mixed insurance (p=0.447). We further find that the lower social earnings in purely compulsory insurance is due to the loss of earnings of the insurer/government ($p < 0.001$ for both comparisons). Earnings of the consumer in purely compulsory insurance are even higher than in mixed insurance (p=0.001).

The last balloon represents the purchasing behavior of participants after learning and the difference is in the last balloon, and hence we focus more on social earnings in the last balloon. We estimate an OLS regression for the overall social earnings in the last balloon. The independent variables are moral hazard level in the last balloon (MHL), and a dummy variable for adverse selection ($AS = 1$ if purely voluntary insurance, $AS = 0$ if mixed or purely compulsory insurance). We also control for gender, age, education, income and household size. The results shown in column (1) of Table 1.10 suggest that moral hazard has significantly negative effects on overall social

[19] For uninsured subjects, the earnings of the insurer and government are zero.

earnings; however the coefficient of adverse selection is not significant. We estimate regressions for earnings of the consumer and insurer. The results in columns (2) and (3) of Table 1.10 show that moral hazard reduces the earnings of the insurer, but it increases the earnings of consumers to a smaller degree. *Why does purely compulsory and voluntary insurance have different overall social earnings given that the degree of moral hazard in purely voluntary insurance is higher (Result 2)?* Column (2) shows that adverse selection has a marginal positive effect on earnings of the consumer. Since purely compulsory insurance has no adverse selection, its overall social earnings do not have the positive effects from adverse selection on the consumer part.

Table 1.10: OLS Regression on (Social) Earnings in Insurance (Dollars). (Reprinted from [2])

	Society (1)	Consumer (2)	Insurer & Government (3)
MHL	$-0.002^{***}(0.001)$	$0.001^{***}(0.000)$	$-0.002^{***}(0.000)$
AS	$0.044(0.04)$	$0.038^{*}(0.0191)$	$0.006(0.028)$
Male	$-0.019(0.038)$	$0.012(0.019)$	$-0.030(0.026)$
Age	$0.001(0.002)$	$-0.000(0.001)$	$0.001(0.001)$
Education	$0.010(0.013)$	$0.011^{*}(0.006)$	$-0.001(0.009)$
Income	$-0.008(0.007)$	$-0.004^{*}(0.004)$	$-0.001(0.069)$
HHsize	$0.006(0.013)$	$0.007^{*}(0.006)$	$-0.004(0.069)$
Constant	$0.167^{*}(0.099)$	$0.171^{**}(0.049)$	$0.272^{***}(0.098)$
Observations	286	286	286
R-squared	0.04	0.05	0.124

Notes: Standard errors are in parentheses. $*p < 10\%, **p < 5\%, ***p < 1\%$.

As mentioned in the introduction, the efficiency of each insurance type has no exact answer from economists. Our results suggest some implications of the combined effects of moral hazard and adverse selection on social welfare of different insurance types from the perspective of social earnings. When consumers in general can benefit from adverse selection which may depend on the distribution of the riskiness of consumers, the combined effects of significant moral hazard and “no adverse selection" makes purely compulsory insurance have less net social earnings than purely voluntary and mixed insurance.[20]

Result 4 When compulsory and voluntary insurance coexist, overall there is no crowding-out effect of the compulsory part on the residual voluntary purchases.

Since the majority of compulsory insurance is partial, investigating the effects of compulsory insurance on the residual voluntary part is also important to the insurance literature [51, 52, 53]. According to [54], compulsory and voluntary insurance are substitutes when they coexist. This means that there is a crowding-out effect of compulsory insurance on the residual voluntary market. Crowding-out effect indicates that increased government involvement in insurance markets substantially reduces demand for private insurance. [55] studied the U.S. Medicare program and found that Medicare does not have substantial effects on the coverage of the residual private insurance market. [56] reviewed the U.S. Medicaid program and showed that incomplete public insurance crowds out private insurance demand.[21] [57] find that crowding-out depends on the coverage rate of government insurance. In our “Mixed Insurance" treatment, subjects are first required to buy partially compulsory insurance, and then they are allowed to voluntarily purchase additional insurance. By comparing the purchasing rate in “Mixed Insurance" with “Purely Voluntary Insurance", our design also allows us to study the crowding-out effects of the compulsory part in “Mixed Insurance".[22]

Table 2.1 compares the proportion of subjects choosing to buy (additional) insurance in the first and last balloon by treatment. The last row shows the p-values from Mann-Whitney U-tests, which suggests there are no significant differences in the number of subjects choosing to buy insurance between purely voluntary and mixed insurance in either the first or last balloon. This means overall there is no crowding-out effects of the compulsory part in mixed insurance, which is also shown in our estimation results (Table 2.3).

In Table 2.3, we analyze a probit model on the purchase rate of the first and last balloon.

[20] The effect size in Table 1.10 is small, and this is because in many cases, the earnings of consumers are the losses of insurers. As a result, the final estimates are around zero.

[21] We did not find related studies in agricultural insurance.

[22] We assume if there is no partially compulsory part, the purchasing rate of the residual voluntary part in mixed insurance is similar to the purchasing rate of full voluntary insurance, since in mixed insurance, we divided the premium and coverage levels in full voluntary insurance at the same rate.

Table 1.11: Number of subjects choosing to buy insurance. (Reprinted from [2])

Insurance	First balloon	Last balloon
Purely Voluntary	38.7%	59.1%
Mixed	43.3%	59.8%
P-value	0.521	0.927

In column (3) and (4), we controlled for the average number of pumps in the 2-29th balloon (Meanpump2-29), age, gender, education, income and household size. All the coefficients of demographic covariates are not significant, and hence we do not report them in the table. The regression results suggest there is no effects of mixed insurance on purchase rate.

Table 1.12: The Crowding-out Effects of Mixed Insurance (Reprinted from [2])

	First (1)	Last (2)	First (3)	Last (4)
Mixed	0.118	0.017	0.125	0.042
	(0.184)	(0.184)	(0.191)	(0.191)
MeanPumps229			−0.001	−0.008
			(0.010)	(0.010)
Constant	-0.287^{**}	0.231^{*}	−1.015	−0.372
	(0.132)	(0.131)	(0.625)	(0.629)
Controls	No	No	Yes	Yes
Observations	190	190	190	190
Pseudo R^2	0.002	0.000	0.017	0.018

Notes: The dependent variable of the probit model is the dummy variable for choosing to buy insurance. Standard errors are in parentheses. $*p < 10\%, **p < 5\%, ***p < 1\%$

Recall in Result 1, we found that the compulsory part has a crowding-out effect on voluntary part for less risk-averse subjects. Combining these two results, we conclude that the compulsory part has a differential effect for people with different levels of risk aversion, with these effects offsetting each other.

1.6 Testing the Validity of BART as an Assessment of Risk-taking

BART has been previously shown to be correlated with self-reported risky behavior such as drug use and gambling [4]. We further test the validity of BART as an assessment of risk-taking in our experiment by comparing it with other measures of risk-taking. In particular, we use the Sensation Seeking Scale (SSS), DOSPERT[23] and a gamble-choice task (EG).[24] We consider Spearman rank correlations among risk-taking in BART and the other three risk measures.[25] Table 1.13 shows the Spearman's ρ in all the treatments. The results suggest that risk-taking behavior in BART is highly correlated with DOSPERT-investing and SSS-experience measures.[26] Risky behavior collected by the SSS-total scores and SSS-bor are positively correlated with BART at the 10% level. The correlation between risk-taking in BART and the gamble-choice is not statistically significant.[27] Self-reported risk preferences measures have been shown predictive of related behavior in the real world [60]. We conclude that risk-taking behavior collected in our experiment has some predictive power of risk-taking behavior in real life. This further confirms the validity of our analysis about adverse selection and moral hazard using the average number of pumps in BART for quantifying risk-taking behavior.

1.7 Concluding Remarks

The World Bank report about government support to agricultural insurance shows that agricultural insurance is voluntary in 78% of the surveyed countries [61]. About 13% of countries

[23]The SSS yields one total score and primary scales for: Disinhibition (SSS-DIS)-This scale represents the desire for social and sexual disinhibition as expressed in social drinking, partying, and a variety of sexual partners; Boredom Susceptibility (SSS-BOR)-This scale represents an aversion to repetition, routine, and dull people, and restlessness when things are unchanging; Thrill and Adventure Seeking (SSS-THR)-This scale contains items expressing a desire to engage in sports or other activities involving speed or danger; Experience Seeking (SSS-EXP)-This scale represents the seeking of experiences through the mind and senses, travel, and a nonconforming life-style. DOSPERT assesses risk taking in five content domains: financial decisions (separately for investing versus gambling), health/safety, recreational, ethical, and social decisions.

[24]In the incentivized gamble-choice task, participants were asked to choose one out of six lottery options with option 1 representing extreme risk aversion, and option 6 risk-loving. We code option 1 to 6 as 1, 2, 3, 4, 5, 6.

[25]We only use the average number of pumps in the 2-29th balloon as risk-taking in BART when testing for the correlations with other measures, as the pumps in the first and last balloon were affected by the insurance options.

[26]In our experiment, BART does not correlate with DOSPERT-all or DOSPERT-gamble, which is in line with the findings in [58]. However, in [58], they don't find correlations between BART and DOSPERT-investing.

[27]Risk preference evaluations have been shown not stable across elicitation techniques and context-dependent (see [59]).

Table 1.13: Spearman's ρ of the correlations among risk-taking in BART and the other three measures (Reprinted from [2])

	SSS-all	SSS-bor	SSS-dis	SSS-exp	SSS-thr		
Average number of pumps in 2-29th	.099*	.093*	.023	.156***	.062		
	Do-all	Do-ethics	Do-gamble	Do-invest	Do-health	Do-recreational	Do-social
Average number of pumps in 2-29th	.073	.033	.026	.139***	.039	.049	.077
	EG						
Average number of pumps in 2-29th	.033						

Notes: $*p < 10\%, **p < 5\%, ***p < 1\%$.

have compulsory insurance for either crop or livestock, and 11% have conditionally compulsory insurance for farmers who have investment loans. Most of the compulsory insurance is partially compulsory. In the literature, the comparisons of compulsory, voluntary and mixed insurance types are not well studied, making the overall efficiency of agricultural insurance scheme ambiguous. We use a simple experiment to compare the three insurance types in terms of both adverse selection and moral hazard.

By setting different balloons with or without insurance options respectively, we use BART as both the assessment of risk-taking behavior and insurance context. First, by comparing risk-taking behavior in balloons with no insurance options for people who buy and people who do not buy insurance, we find adverse selection in purely voluntary insurance but advantageous selection in mixed insurance. Specifically, the partial compulsory part in mixed insurance induces the less risk-averse (high-risk) individuals not to buy additional insurance.

Moral hazard exists in all three insurance types, but mixed insurance reduces the degree of moral hazard. Our results shed light on the combined effects of adverse selection and moral hazard on insurance schemes. We use social earnings as an indicator of social welfare, and find that purely compulsory insurance has the lowest social earnings.

The results about mixed insurance provide valuable insights, especially for developing countries where agricultural insurance is not yet well developed and very few schemes of insurance exist that implement different forms of mixed insurance. Further work about the optimal compulsory coverage rate in mixed insurance is necessary.

Our paper is the first to use BART to compare different insurance schemes. [62] modeled the Balloon Analog Insurance Task (BAIT) after BART to study protective risk management behavior. Further work —especially field evidence— is needed validate BART (and other risk instruments) to study insurance markets. We are also aware that different premium, coverage, subject pool and specific insurance markets need to be explored within each insurance scheme. We hope this paper serves as motivation to study the effects of different insurance parameters on insurance take up using incentivized mechanisms.

2. SOCIAL ROLES AND COMPETITIVENESS: MY WILLINGNESS TO COMPETE DEPENDS ON WHO I AM (SUPPOSED TO BE)[1]

2.1 Introduction

From executive boardrooms to battlegrounds women are making great strides to attain parity with male counterparts. The 2020 Global Gender Gap Report points out that many countries have achieved important milestones towards gender equality, but this process is very ponderous [63]. It is well documented in the economics literature that females are less competitive than males, even when they have similar ability [64, 65]. Economists have tried to explore whether such gender differences in competitiveness may be useful for explaining the persistent labor market gender gap [66].[2] If women are reluctant to compete, then they may also be less likely to seek job promotions or to choose more lucrative and competitive fields [1]. Selecting out of certain labor markets is costly for society, especially when competent women are reluctant to compete for positions for which they are the best suited candidate.

A direct way to close the gender gap in competitiveness is to assume that preferences for competition are malleable and induce women to become more competitive.[3] The other approach is to take preferences as given and to implement institutional changes to encourage women to compete, especially high ability women. In this paper, we use this second approach to evaluate how social roles affect individual preferences for competition. We do this by exogenously assigning roles/titles that convey specific social norms for competitiveness to individuals. We set up a laboratory experiment using conventional western society roles for men and women, namely “breadwinner" and “supporter". To make the roles more salient, we set up a group competition in pairs instead of individual competitions. We introduce in our design willingness to compete (WTC) scales (0-5)

[1]This study was approved by the IRB.

[2]The Global Gender Gap Report, documents a persistent structural gender gap in labor markets, such as in top managerial or CEO positions, STEM field studies, and in the acquisition of emerging skills.

[3][67] find that the gender gap in tournament entry reverses in a matrilineal society compared to patriarchal societies. [68] find that girls in selective single-sex schools are more likely to enter competitions against boys compared to girls from mixed-sex schools. Single-sex education has been proposed as an alternative to increase the competitiveness of women [69].

instead of a binary choice for entering the competition.[4] In the treatment, before subjects proceed to choose their WTC level for their groups, we exogenously and randomly assign the role of “breadwinner" or “supporter" to each group member. The two roles were selected based on the results from surveys within the same study population.[5] There are no differences between the two roles in terms of payment, power, or effort in our experiment. The only difference is the social role titles. There is no role assignment in the control condition.

In psychology, women and men belong to two separate social categories or roles [70], and they are associated with different behavioral prescriptions or social norms [71]. For example, men’s social roles promote competitiveness, while women’s social roles penalize it [72]. The two selected roles in our treatment also have opposite social norms for competitiveness, analogous to men and women in western society. Our main hypothesis is that the role of breadwinner (supporter) will reduce (increase) the social cost of entering the competition, and subsequently, we expect the WTC of men and women to change depending on their randomly assigned roles.

Consistent with previous literature, our results show a significant gender gap in WTC in the control group. The WTC for males is 40% higher than females in the baseline. However, there is no significant gender difference in WTC in the treatment. When women are randomly assigned as breadwinners, their WTC significantly increases compared to women in the baseline or women who are randomly assigned as supporters. The gender gap between treated female breadwinners and untreated males is not statistically significant. More importantly, the increase in women’s WTC is mainly contributed by high-ability women and they become better off, earning more than high-ability women in the baseline. Men are also affected by the role assignment; male supporters are less likely to enter the competition compared to male breadwinners. By examining the underlying mechanisms behind our results, we demonstrate that social norms for competitiveness of different social roles (including men and women) and responsibility for making the competition decision

[4] WTC for group competition carries responsibility, and our WTC scale design allows us to study the responsibility of each role. We provide more details in the experimental design section.

[5] Before the experiment, we conducted a survey with over 1800 responses to ask respondents to select corresponding labels to the primary-earner and the secondary-earner in a household. The details of the procedure are explained in the following section.

for the group explain the gender gap in WTC. We also show the role assignment affected only the responsibility of male supporters, and therefore we argue that the changes in WTC brought by the role assignment are mainly driven by the social norms for competitiveness associated with different roles. A theoretical model is developed to investigate how social norms for competitiveness affect WTC. We also estimate a difference-in-differences model, and find that women are more responsive to the social norms for competitiveness.

In order to get implications for the real workplace, we further conduct an online experiment using Amazon Mturk and randomly assign roles from labor markets: “manager" and “assistant" to the treatment group. We replicate the main results from our laboratory experiment. When women are assigned as the “manager", their WTC significantly increases, and the gender gap between male and female managers is not significant.

In the book “Switch: How to Change Things When Change Is Hard", [73] share an example of a manufacturing firm, Brasilata, and how their engine of success was fueled by the creation and adoption of a new worker role called “inventor". New employees were asked to sign an “innovation contract" when they joined the firm. The “inventor" role was made up, but “the program succeeded beyond any reasonable expectations". Employees submitted 134,846 ideas, an average of 145.2 ideas per worker. Our paper sheds light on the importance of job roles/titles in the workplace and provides potential mechanisms to understand how job roles/titles affect competitive behavior. [74] suggest job titles contribute to change the gender-related perceptions of a job. [75] show that job titles elicited more sex stereotyping than job descriptions. Another policy implication of our study relates to gender quotas [76]. Nearly 100 countries around the world use some type of electoral quota aimed at improving the underrepresentation of women in government [77]. Gender quotas encourage women to pursue a political or board role which violates traditional stereotypic beliefs about the sexes [78]. Our paper provides evidence of the potential effects of the role itself on females' competitive behavior.

We induce subjects with different roles which may be prone to an experimenter demand effect. However, the experimenter's expectations connected with the roles of breadwinner and supporter are exactly the social norms we are after. In real life, social norms are a type of social expectation, and in this regard, the experimenter demand effect belongs to the treatment effect we want to evaluate.[6] To the best of our knowledge, our paper is the first to study the effects of social roles on the gender gap in competition by exogenously and directly inducing social roles with distinctive social norms for competitiveness to each gender.[7] The most relevant papers to our research are [88], [89], [90] and [91]. [88] investigate the effects of social norms on gender differences in competitive choices using different birth cohorts of individuals who were exposed to different socio-economic institutions in China. They suggest that exposure to different institutions/norms during crucial developmental-ages significantly molds an individuals' willingness to compete. [89] studies the impacts of social identity on WTC, but she focuses on building group identity during the experiment and testing its effects on WTC. [90] suggest preferences for competition can be influenced by gender, family, and professional identities. [91] prime subjects with power, by asking subjects to recall and write down a personal situation in which they had control over other individuals. They find that when subjects are primed with a high-power situation, the gender gap in competitiveness vanishes. Our paper departs from previous work by directly and exogenously assigning social roles that grant individuals permission to violate prevalent gender stereotypes for competition.

The rest of the paper is organized as follows. In section 2.2, we introduce the concept of social norms, and explain how the two social norm roles were selected. In section 3.3, we present the experimental design and explain the WTC scale. The analysis and results are presented in section 2.4. Section 2.5 examines the potential mechanisms behind the results. Section 2.6 introduces the procedure and results of the online sessions, and section 3.7 concludes.

[6]We also adopt other prevalent practices in the experimental economics profession to mitigate potential experimenter demand effects [79]. First, the experiment is incentivized, and it is costly for subjects to deviate from their true competitive preferences. Second, the design is a between-subject design, and therefore the purpose of the experiment is not obvious for subjects since they are only exposed to one treatment environment [80, 81]. Third, we implement a second online experiment using Mturk with a higher degree of anonymity as a robustness check and we replicate the results.

[7]Previous studies relevant to other institutional changes include changing the gender composition of competitions [82, 83, 84], changing the competitors [85, 86], and developing a sponsorship program of competition entry [87], among others.

2.2 Relevant Literature and Social Role Label Selection

[92] introduced identity concepts from psychology and sociology into economic behavior. They pose that identity is a sense of belonging to a social category, and it prescribes how people within a category should behave. More precisely, the social norms linked to a social identity/role provide unwritten standards of how to appropriately behave among members of that particular category [71]. One important application of the social role model is gender, where man and woman constitute two separate social categories. Man and woman are associated with different behavioral prescriptions. For example, in patriarchal cultures, “men should not do domestic work at home and men should earn more money than their wives."[8] Social roles can influence economic outcomes because deviating from social norms generates a social cost. The social cost includes both internal (shame, guilt, loss of self-esteem, etc.) and external (disapproval, punishment, etc.) sanctions [71]. In psychology, this social cost is known as the “backlash effect" [94].[9] In this paper, we focus on how social roles affect willingness to compete. As noted in the role incongruity theory, men's social roles promote competitiveness, while women's social roles penalize it [72]. Specifically, women are expected to conform to the role of a supporting wife and to maternal duties. On the other hand, men are expected to be competitive in the workplace [95, 96].[10]

Misaligning with the established gender roles results in social costs. First, we show some examples of external sanctions. [97] show that marriage satisfaction is lower in households where the wife's income exceeds the husband's income; in fact, if this is the case, couples are more likely to divorce.[11] [99] show that men in the dating market prefer women who are less professionally ambitious. Men also tend to avoid highly educated female partners [100, 101, 102]. Job promotions increase the likelihood of divorce for women, but not for men [103]. [104] argue that social identity costs make single female students demonstrate less ambition in the presence of single males. Sec-

[8] [93] suggest these prescriptions exist within many households in the United States.

[9] [94] defines “backlash effect" as social and economic reprisals for behaving counter-stereotypically.

[10] Throughout the paper the discussions about social norms for competitiveness are based on patriarchal cultures, unless otherwise noted.

[11] [98] use a lab experiment with couples who live together and suggest that there are no differences in preference for work-division in couples.

ondly, it is hard to find direct empirical evidence of utility loss from internalized norm deviation as it is subjective. However, there is evidence that even when the actions are unobservable, women behave under the guides of gender roles. Using field evidence from voluntary-based requests sent by university email, [105] document that women are more likely to accept low-promotability tasks compared to men. They further show that such gender gap is not driven by preferences, but rather by the belief that women are supposed to be more likely to volunteer than men. Internalized social norms also affect females' performance. [106] find that female students perform worse in math tasks when they are in classrooms in which traditional masculinity norms are present; however, in the absence of these norms, there is no gender gap in performance.

In order to test for the effects of social roles on WTC, we exogenously change the cost of social norm compliance for competitiveness socially prescribed to each gender by randomly assigning subjects in the treatment to two different titles/roles, namely “breadwinner" and “supporter". The two labels were selected based on two-rounds of surveys with the same study population before the lab experiment.[12] First, a total of 56 undergraduate students were recruited to provide open-ended social role labels corresponding to primary and secondary earners in a household. Specifically, the scripts asked:

In one word, how would you call someone who is the primary earner in a household:

In one word, how would you call someone who is the secondary earner in a household:

From these 112 responses, the three most frequently used labels for each role were selected for the second survey. The second survey consisted of the same two questions using a multiple choice format using the top three labels for each role in the first survey. In the first survey, for the primary earner, the top three labels were Breadwinner, Primary, and Provider. For the secondary earner, the top three labels were Supporter, Secondary, and Assistant. Then we conducted the second survey. A total of 3,794 responses were collected in the second survey with 1,897 responses to each question (40% males, 60% females). The second survey was a bulk email sent to all students on campus. The results show that 42% of respondents chose “breadwinner" for the primary earner

[12] We randomly selected 2 out of 100 participants to receive a $20 Amazon gift card.

label, and 55% chose “supporter" for the secondary earner label. Thus, the most popular labels from the two surveys, “breadwinner" and “supporter" were selected as the labels for our social roles treatment. Under the social role theory, a man is supposed to be the breadwinner of his family, while a woman is supposed to be the supporter [107].

2.3 Experimental Design

We implement a between subject design with a baseline control and a treatment group. Our experiment has two stages. The baseline and treatment group only differ in the second stage. The experiment was conducted at Texas A&M University. Subjects were students who were native English speakers recruited by bulk email. A total of 28 full sessions were conducted with eight subjects per session. Nine subjects experienced technical issues that affected their performance in the first stage and their subsequent choice in the second stage, and hence they were excluded from the analysis.[13] Ten subjects were not native English speakers and reported having difficulties understanding their assigned roles, so they were also excluded.[14] The final sample consists of 205 subjects. The experiment was computerized using Ztree [108], and had a duration of approximate 30 minutes.

2.3.1 Design

First Stage: In the first stage, subjects arrived to the experiment and they were seated in rows. Each session had 8 subjects. Following the original design in [64], subjects had five minutes to individually solve a real effort task under a noncompetitive piece-rate payment scheme. The real effort task consists of adding-up five two-digit numbers. Participants were not allowed to use a calculator, but they could use scratch paper. Participants received 10 cents for each correctly solved problem. Contrary to the original design in [64], we do not reveal the subjects’ performance in this stage until the end of the experiment. Since our focus is on the effects of social norms on WTC,

[13] It is possible that the other subjects in the sessions may be potentially affected by these instances. The results were robust to the exclusion of all the sessions.

[14] Although in the recruitment process, being a native English Speaker was required, there were still some participants who were not native English speakers who registered for the experiment. In order to have exactly 8 participants in each session, we allowed them to participate.

we try to minimize potential confounding effects of natural ability on competition entry decisions. The purpose of the first stage is to control for ability without a group or competition context.

Second Stage: After the first stage, each participant was randomly paired with another participant in the same room. We use team competition instead of individual competition following [109] and [110] so that we can assign different roles within a group. In order to identify the effects of different roles, the partner assignment was blind, and no information about partners was disclosed. Four groups were formed in each session.

In the second stage, each subject had 5 minutes to solve a task similar to the first stage. Before they proceed to the task, participants can choose between two possible payment schemes for their group: piece rate or tournament. In the piece rate scheme, each group receives 20 cents (10 cents per person) for each correctly answered question. In the tournament scheme, each group receives 80 cents for each correct answer only if they win, otherwise they receive nothing.[15] If a group enters the tournament, then they compete with all the other three groups in the session, regardless of the payment scheme chosen by the other three groups. This design feature removes potential concerns of participants competing only against those who choose the tournament scheme.

Each group member is asked to select the payment scheme for his group using a six point scale: {0, 1, 2, 3, 4, 5}, which is our willingness to compete measure. The probability of entering the tournament for each WTC scale is 0%, 20%, 40%, 60% 80% and 100% respectively. Individuals can choose (with certainty) the piece rate payment scheme by selecting 0 or the tournament by selecting 5. One of the two group members' WTC was randomly selected as the binding WTC for the group. The WTC selection is anonymous between the two group members, and participants do not know the payment scheme for their groups or their performance in this stage until the end of the experiment.[16]

[15]The expected payment under the piece rate payment scheme between the first and second stage is the same. In related literature, the usual payment for each correct answer is 50 cents. In our case, if we choose 50 cents in the first stage, then the payment for each correct answer would be $4 under the tournament payment scheme in the second stage. It has been shown that scaling up payments results in a significant increase in risk aversion [40, 111], and this impact might be heterogeneous by gender. To reduce the effects of the payment itself on competition entry, we lower our payment for the piece rate scheme to 10 cents in the first stage and 20 cents in the second stage.

[16]We want to test the effects of roles on performance, and minimize the confounding effects of the payment scheme on performance. We do not reveal the implemented payment scheme until the end of the experiment.

After selecting the payment scheme, each group proceeds to the second round real effort addition task. Subjects work separately on the task and then the payment for each group is determined by the combination of the number of correct answers of the two group members. The group earnings were equally divided between the two members. This design feature keeps subjects performance consistent and excludes the possibility of individuals competing on the basis of team performance.

Treatment Interventions: The difference between the treatment and the control group is the randomly assigned role to each group member in the second stage. For the treatment group in the second stage, each group member received the following information before answering the WTC scale question:

You were randomly selected to be the Breadwinner (Supporter) of the group, and your partner is the Supporter (Breadwinner) of the group.

In the control group, there is no such role assignment. As previously mentioned, in order to eliminate other factors that may affect competition entry, we inform subjects that the roles are *randomly assigned*, and subjects know their partner has been assigned the opposite role. Recall that the group payment is equally divided, and one of the two members' WTC is randomly selected as the binding WTC for the group. Given all these design features, the only difference between the two group members is the framing of the role.

Social Norms Belief Elicitation: We modified the method in [112] which is used to elicit social appropriateness to elicit social attitudes toward the competitiveness of each role in our experiment. Using a five-point-scale (1, 2, 3, 4, 5), subjects were asked to indicate how competitive they believe the following individual should be: a female supporter, a female breadwinner, a woman, a breadwinner, a male supporter, a supporter, a man, and a male breadwinner. For each question, if the participant's answer was the same as the most frequent answer in the session, they received a reward of ¢25. We do not reveal the outcome of each question until the end of the experiment in order to address possible hedging problems.

Performance Beliefs and Risk Preferences: We elicited participants' beliefs about their relative performance in the first task. We ask participants to guess whether they performed better than their second-stage partner. We also ask them to state their beliefs about their overall ranking in the session (1-8). Correct answers were incentivized with a ¢25 reward. The performance beliefs can be used to examine whether changes in WTC induced by the role assignment are derived from changes in confidence from the roles.

At the end of the experiment, all participants were asked to complete a incentivized risk gamble-choice task following [41] and a demographic survey. Participants received payments from the two tasks and the incentivized norm and belief elicitations, and the gamble choice. The average payment from the two tasks was $6.[17] Table 2.1 summarizes the experimental procedures and the number of subjects in each experimental condition. The instructions, questionnaires and surveys are available in the Appendix.

Table 2.1: Summary of Treatments and Number of Subjects.

	First stage (individually)	Second stage (group, WTC)	Female	Male	F-Breadwinner	M-Breadwinner	Total
Baseline Group	Piece rate	No roles	41	35	—	—	76
Treatment Group	Piece rate	Role assignment	73	56	38	26	129

2.3.2 WTC Scale

In our experimental design, we use a WTC scale instead of the conventional binary choice in previous literature. A more continuous measure of willingness to compete has been previously studied in the literature [113, 114, 115]. The results from previous studies suggest that the binary choice measure to some extent hides the intensity of preferences for competition. In this paper, we follow this line of work and focus on the responsibility required for making competition decisions for the group.

[17]The show-up fee was $5.

The competition in our experiment is a group competition, and in order to make our treatment (role assignment) more salient, we follow [110] and ask subjects to make WTC decisions for their groups instead of for themselves. In this regard, choosing to enter the competition or not is also a decision which carries responsibility. This is a particularly important feature in our design, since the role assignment may affect the WTC through changing the responsibility. We use the preference for randomization as a proxy for responsibility. Preference for randomization in the economics literature is a puzzle for decision theory as it may violate expected utility maximization [116]. As indicated in [117] and [118], one possible reason for preference for randomization is that participants may want to avoid the responsibility of making a choice, even when it results in suboptimal outcomes.[18] Thus, we design the WTC scale to investigate changes in responsibility for different roles.

Denote the piece-rate payment per correct answer as p_1, and the tournament payment as p_2. A WTC value of 1 in our scale, for example, indicates a 20% probability of entering the tournament, and an 80% probability of obtaining the piece-rate payment. Given the exante probability of winning the tournament in the design, the expected payment of each part of scale 1 is as follows:

$$scale1 = \begin{cases} 20\% tournament & \text{expected payment :20\% * (25\% * } p_2 \text{ + 75\% * 0)} \\ 80\%\text{piece rate} & \text{expected payment :80\% * } p_1 \end{cases}$$

For simplicity, we denote $p_1 = 1$, and $p_2 = 4p_1 = 4$. Each point in our scale represents a lottery with different probabilities assigned to different payoff levels: low payoff (0), middle payoff ($p_1 = 1$), and high payoff ($p_2 = 4$). For example, point 1 in the scale described above indicates 80% chance of earning 1, 5% chance of earning 4, and 20% chance of earning 0. The expected payoff is the same for each point in the scale, which equals to 1, but the variance varies

[18]In another bigger literature branch, preference for randomization is studied in relation to ambiguity/uncertainty aversion, where a lottery with 50%/50% probabilities (randomization) is compared to a lottery with unknown probabilities (ambiguity/uncertainty aversion) [119, 120, 121, 122]. We do not study this type of preference for randomization in current paper.

across the scale. Table 2.2 summarizes the lottery for each WTC scale.

Table 2.2: Indicated Lottery of Each WTC Scale

WTC Scale	Prob. of competition	Low Payoff (0)	Middle Payoff (1)	High Payoff (4)	Expected Payoff	Standard Deviation
0	0%	0	1	0	1	0
1	20%	15%	80%	5%	1	0.77
2	40%	30%	60%	10%	1	1.10
3	60%	45%	40%	15%	1	1.34
4	80%	60%	20%	20%	1	1.55
5	100%	75%	0	25%	1	1.73

From point 0 to point 5, the standard deviation increases. By assuming a constant relative risk aversion (CRRA) utility function ($\frac{x^{1-\gamma}}{1-\gamma}$) and based on expected utility maximization, risk-seeking individuals would choose point 5 in the scale, and risk-averse individuals would choose the piece rate (point 0 in the scale). Risk-neutral individuals would be indifferent.[19] Thus, choosing the middle points in the scale (1-4) reveals a preference for randomization. According to the corresponding tournament entry probability of each number, points 1 and 2 are the randomization attached to the decision of receiving a piece-rate payment (point 0 in the scale), and points 3 and 4 are the randomization attached to the decision of entering the competition (point 5 in the scale). The more the subject is willing to take responsibility for selecting the piece-rate payment, the lower his WTC; the more the subject is willing to take responsibility for entering the competition, the higher his WTC.

At the same time, the standard deviation increases with the WTC scale, which is in line with the risky-behavior nature of the tournament entry. Entering the tournament is a riskier decision compared to the fixed piece-rate scheme. In this regard, our WTC scale also makes the effects of risk preferences on tournament entry more salient. Females are more risk-averse than males [123], and therefore the WTC scale allows us to test whether our treatment effects are explained away by

[19]If we use the subjective probability of winning the tournament instead of 25%, the expected payments and standard deviations of the scales are monotone. Thus, we find the choices of individuals with different risk preferences do not change and are still 0 and 5. The only difference is that risk-neutral individuals would also choose either 0 or 5, since the expected payment is monotone.

gender differences in risk preferences. Based on the properties of our WTC scale, we argue that it is a combination of responsibility (for choosing for the group) and risk preferences.

2.4 Results

In this section, we first examine whether conditional on ability, men and women differ in their preferences for performing under a piece-rate *versus* a tournament scheme when competing in teams. We then examine whether exogenous role assignment changes their preferences for competition.

2.4.1 Willingness to Compete in the Baseline

Result 1. *In the baseline without role assignment, the WTC of women is lower than the WTC of men.*

In our experiment, the WTC was elicited during the second stage. Figure 2.1 shows the cumulative distribution of WTC by gender in the baseline. The average WTC of men is 3.11, and the average WTC of women is 2.22. Using a two-tailed Mann-Whitney U-Test, we find that the WTC of males is 40% higher than females ($p = 0.037$). Recall that, in the baseline, there is no role assignment. Participants have no information about their partners, except that the partner is another participant in the room.[20] We always have eight participants in each session. Our result suggests that men are more likely to choose a competitive tournament environment compared to women when competing in two-person teams.[21] Throughout the paper the reported test statistics refers to a two-tailed Mann-Whitney U-Test, unless otherwise noted.

Figure 2.2a provides detailed evidence of preference for randomization for both genders in the baseline. There are 68.29% females and 40.00% males choosing middle scales. There is a mass

[20] Gender composition of each session is observed by participants, and the percentage of each composition by session is shown in AppendixB.1. We tried to have the same number of males and females in each session, but we control for the differences in the following model estimations.

[21] Previous literature have explored gender differences in tournament entry using team competitions instead of individual competitions. [109] finds no gender gap in team competition entry. [110] find competing in two-person teams reduces the gender gap by two-thirds. In [109], the team was constructed by two subjects who had the same decision for tournament entry. Our team decision-making process is same to the process in [110], but the gender of the partner in [110] was revealed. Note that [110] use a binary decision to elicit preferences for competition which does not involve the responsibility for making the decision for the team in our paper.

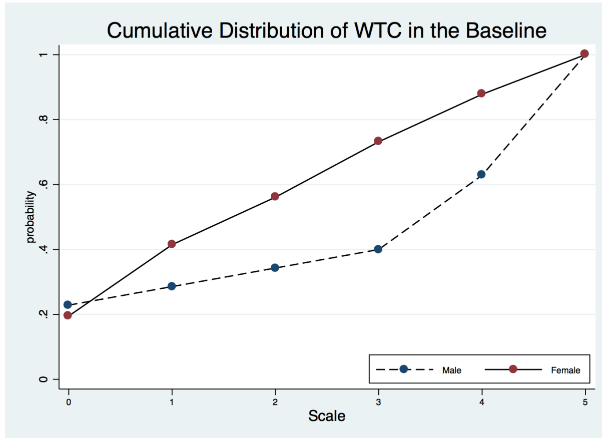
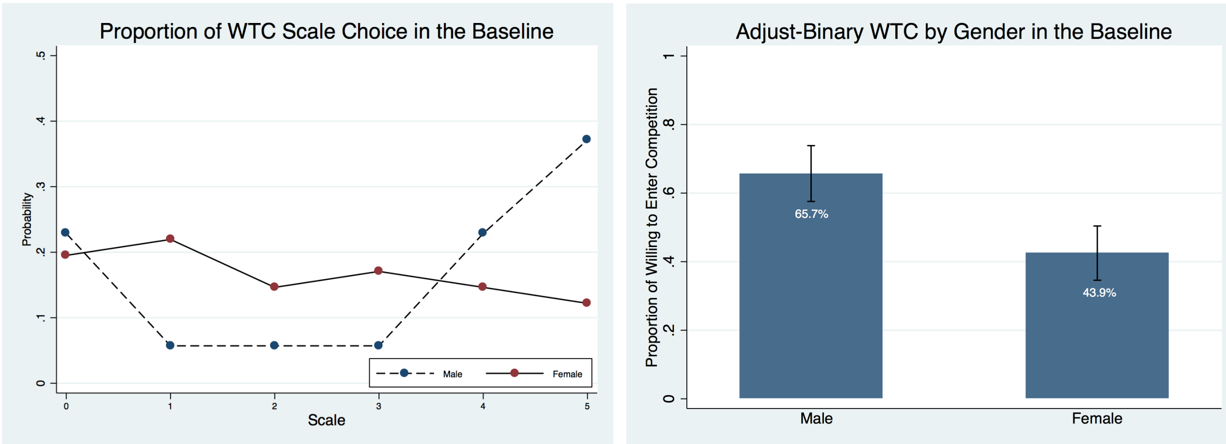


Figure 2.1: Cumulative Distribution of WTC in the Baseline

concentration for men in the $4-5$ points in the scale, and a mass concentration for women in the $1-3$ range. The distribution of WTC scale between men and women is significantly different (Kolmogorov-Smirnov test, $p = 0.031$). We further use the degree of randomization to quantify the level of responsibility for making the choice for the group. Based on the probability of entering the competition/piece-rate, the degree of randomization of point 4 in the scale is lower than the degree of randomization of point 3; and the degree of randomization of point 1 is lower than the degree of randomization of point 2. Points 0 and 5 in the scale have a zero degree of randomization. Thus, we set the responsibility measures equal to 0 for points 2 and 3; 1 for points 1 and 4; and 2 for points 0 and 5. In general, we find that men have lower preference for randomization and are more willing to take the competition choice responsibility compared to women ($p = 0.008$). This difference is mainly driven by the responsibility of entering the competition (i.e., points $3-5$ in the scale), and there is no significant difference in the responsibility of choosing the piece-rate (i.e., points $0-2$ in the scale) between man and women ($p = 0.021$ for points $3-5$, $p = 0.125$ for points $0-2$). Note that in our design the *lower* responsibility in points $0-2$ and *higher* responsibility in points $3-5$ indicate a higher WTC.

We further recoded the WTC scale to an adjusted-binary WTC choice with points 0-2 indicating choosing the piece-rate payment and corresponding to “0" in the binary choice, and points 3-5 indicating choosing to enter the competition and corresponding to “1" in the binary choice. The



(a) Proportions of WTC Scale Choice by Gender (b) Adjusted-Binary WTC (with standard error bars)

Figure 2.2: WTC in the Baseline

adjusted-binary WTC choice mutes the responsibility for making the choice for the group. Figure 2.2b shows there is a higher proportion of men choosing the scale $3-5$ compared to women (χ^2 test, $p = 0.056$). Thus, a higher proportion of choosing points $3-5$ (and also higher responsibility in points $3-5$) for men results in final higher WTC for men in the baseline.

2.4.1.1 Gender Differences in Performance in the Baseline

In order to rule out individual ability as a driver of WTC, subjects were not informed about their performance during the first stage until the end of the experiment. However, subjects may still form beliefs about their performance before choosing to enter the tournament or not. In the second stage, participants work in teams. Recall that participants do not know the payment scheme applied to their group until the end of the experiment. However, they can speculate about the payment scheme based on their own WTC, especially those with high WTC. Performance anticipation under a group context and a tournament environment may also affect subjects' decisions. Therefore, we examine whether gender differences in performance exist under the piece rate and tournament payment schemes in the baseline.

Table 2.3 summarizes the performance of men and women. In the first stage, the average number of correctly solved problems was 9.05 for women and 9.83 for men. The p-values in the last row show that there are no significant differences in performance between men and women. In the second stage, we first show the performance of all participants in order to evaluate whether the group context generates gender differences in performance. The difference in performance under the group context is not significant, which is consistent with [110]. On average, women correctly solve 9.34 problems, and men 9.40 ($p = 0.896$). Next, we compare the performance of individuals who actually end up being paid with a tournament payment scheme, and the difference is not significant ($p = 0.823$).

Previous literature (e.g., [124]) documents that competitions can boost performance; however using Wilcoxon signed-rank tests, we do not find differences in performance between the first and second stage for men or women (see p-values in the last column). This result is reasonable in our case, since performance and payment scheme information was not revealed until the end of the experiment. Given the similarity in performance of men and women, we argue that in the baseline, men have higher WTC than women of similar ability. Regression estimations and other potential underlying mechanisms will be explored later.

Table 2.3: Gender Difference in Performance in the Baseline

	First Stage 1 (all participants)	Second Stage 2 (all participants)	Second Stage 3 (only those with tournament pay)	p-value column 1 vs 2
Men	9.83 (0.49)	9.4(0.43)	9.08(0.55)	0.215
Women	9.05(0.52)	9.34(0.48)	9.12(0.64)	0.176
p-value	0.322	0.896	0.832	

Notes: Standard errors are in parentheses.

2.4.2 The Impact of Random Role Assignment on WTC in the Treatment Group

Result 2a. *Compared to the baseline, women's WTC significantly increases when they are assigned as breadwinners; there is no significant change in females' WTC when they are assigned as supporters. Males' WTC is not changed significantly by each role. Overall, there is no gender* *gap in WTC in the treatment group.*

2b. *The WTC of "supporters" is lower than the WTC of "breadwinners" for both men and women.*

In this section, we explore how the social role assignment impacts participants' WTC. Recall that in the treatment group, before choosing their WTC in the 0-5 scale, subjects were informed that they were randomly selected as the "breadwinner" or "supporter" of the group. We also informed participants that their partner was assigned the alternative role. Figure 2.3 shows that the gender gap in WTC decreases by 56% in the treatment group and it is no longer statistically significant (p-value = 0.123). We further analyze whether different roles have asymmetric effects by gender.

First, we compare the WTC of each role to the baseline. Figure 2.3 shows that the WTC of treated female breadwinners significantly increased compared to women in the baseline (p-value = 0.031).[22] More interestingly, there is no difference in the WTC between female breadwinners and males in the baseline (p-value = 0.373). The effect of the "breadwinner" role on males' WTC is not significant compared to the baseline (p-value = 0.285). We then compare the WTC of treated supporters to the baseline. There are no differences for both genders (p-value = 0.933 for women, p-value = 0.181 for men).[23] Although the difference in WTC between male supporters and males in the baseline is not significant, the gender gap in WTC between male supporters and females in the baseline is not significant (p-value = 0.436).

We further compare the WTC within the role assignment (*breadwinner* vs *supporter*). Figure 2.3 suggests that for both genders, breadwinners' WTC is significantly higher than the WTC of supporters ($p = 0.049$ for women, $p = 0.016$ for men). When we compare the WTC of each role across gender, we find that there is no gender gap in WTC between male and female "supporters" ($p = 0.513$); however male breadwinners are more willing to enter the competition than female breadwinners ($p = 0.032$).

Figure 2.4a shows that preference for randomization exists in the treatment for both genders.

[22] After we apply the multiple hypothesis testing in [125], the difference is marginal significant (both p-values with Thm3_1 and Remark3_7 = 0.053).

[23] Male supporters' WTC decreases, but due to lack of power, the difference is not significant.

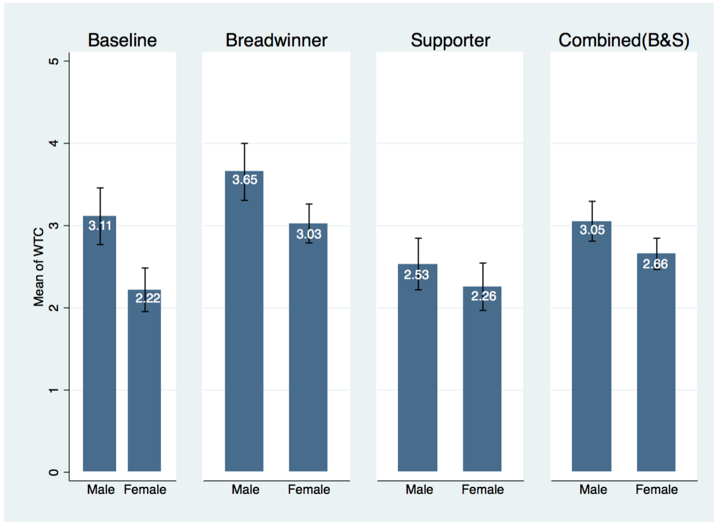
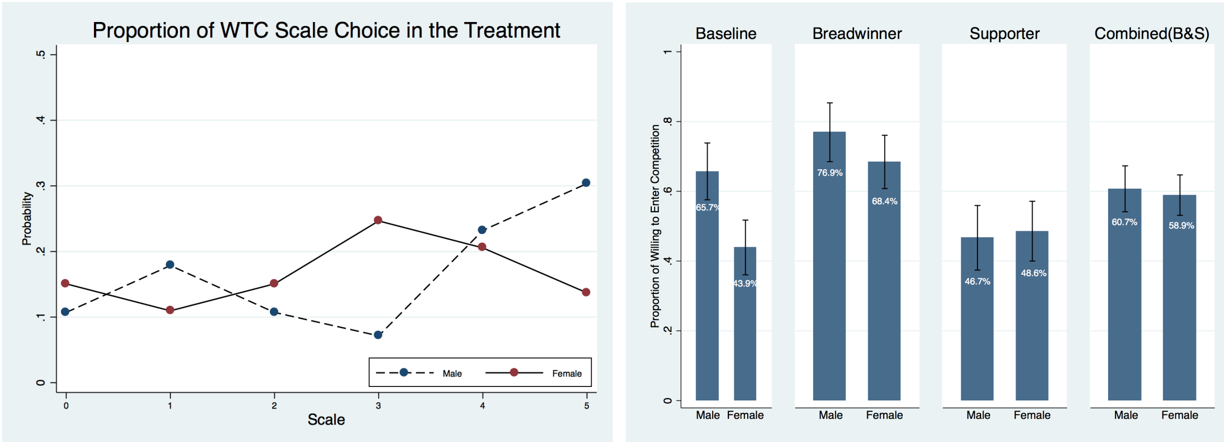


Figure 2.3: Comparisons of Mean of WTC Level (with standard error bars)

There are 71.23% females and 58.93% males choosing middle points in the scale, but the distribution of WTC scale between men and women in the treatment is not significantly different (Kolmogorov-Smirnov test, $p = 0.187$). Similar to the baseline, men in general have lower preference for randomization and are more willing to take the choice responsibility compared to women in the treatment ($p = 0.018$). We further mute the responsibility and recoded the WTC scale to an adjusted-binary WTC choice. Figure 2.4b shows that the WTC results using the adjusted-binary WTC choice are consist with the results using our complete WTC scale, except that there is no gender gap in WTC between female and male breadwinners (χ^2 test, p = 0.457).

Table 2.4 summarizes performance in the treatment and it shows that the results are similar to the baseline results. The *p*-values in the last row indicate that there are in general no gender differences in performance during the first or second stage. However, male supporters perform better than female supporters. Using Wilcoxon signed-rank tests, we evaluate the performance between the first and second stage for each role and gender. The *p*-values in columns suggest that the role assignment does not change the performance for men or women. The changes in WTC induced by the treatment (role assignment) are not likely due to changes in performance induced by the roles.



(a) Proportions of WTC Scale Choice by Gender

(b) Adjusted-Binary WTC in the Treatment

Figure 2.4: WTC in the Treatment

Table 2.4: Gender Difference in Performance in the Treatment Group

	First Stage (individual)	Second Stage (group)	Breadwinner 1st	Breadwinner 2nd	*p*-value	Supporter 1st	Supporter 2nd	*p*-value
Men	9.80	9.46	8.8	8.5	0.497	10.7	10.3	0.811
	(0.54)	(0.41)	(0.65)	(0.65)		(0.80)	(0.49)	
Women	8.86	8.97	8.8	9.3	0.286	8.9	8.6	0.441
	(0.39)	(0.34)	(0.58)	(0.46)		(0.52)	(0.51)	
P-value	0.202	0.163	0.763	0.377		0.031	0.010	

Notes: “1st" and “2nd" indicate first stage and second stage respectively. *p* values in columns are from comparisons between the first and second stage performance for each role and gender. Standard errors are in parentheses.

We estimate ordered probit models for the control and treatment groups. The dependent variable is WTC. The number of correct answers in the first stage is used to control for ability (Ability). We also control for confidence, risk preferences and the gender composition of each session. Confidence is indicated by converted participants’ beliefs of their relative rank in the session (1-8). Risk preferences are indicated by the choice of [41] incentivized risk gamble task, which has 6 choices (1-6). A higher number means more risk-tolerant behavior (RiskTolerance). Gender composition indicates the number of women in each session (No. Female). In Table 2.5, columns (1) and (2) show that in the baseline, after controlling for risk preferences, confidence and ability, the female coefficient is no longer significant.[24] In columns (3) and (4), we pool all data in the treatment. First, consistent with figure 2.3, there are no significant effects for gender. The positive and significant breadwinner coefficient shown in column (4) suggests that the "breadwinner" label makes people more willing to enter the competition compared to the "supporter" label. In column (5), we estimate a difference-in-differences model to investigate whether men and women respond differentially to the role assignment. The interaction term shows that there is no differential effect of assigned roles across gender. However, when we use the adjusted-binary WTC choice in column (6), we find women are more responsive to the role assignment. We explore the mechanisms behind these results in the following section. We also estimate a difference-in-differences model for "breadwinner" and "supporter" separately, and the results are similar (Table B.2).

2.5 Exploring the Mechanisms Behind the Changes in WTC

In this section, we explore the potential mechanisms behind the results presented before. We first ask whether the role assignment changes the confidence level or change the effects of risk aversion on WTC. We then test for the effects of social norms for competitiveness and examine changes in responsibility for making the competition choice for the group. Next, we separate subjects based on their performance in order to explore whether the role assignment has asymmetric effects on subjects with different ability.

2.5.1 Changes in Confidence and Risk Aversion

Recall that in the belief elicitation stage, we asked subjects to guess whether they believe they correctly solved more questions than their partner during the first stage. We also asked subjects to guess their performance rank in their session (1-8) during the first stage. Both questions were incentivized.

Table 2.6 shows the estimated performance rank in the first stage. The *p*-values in the last row

[24] In [64], the coefficient of gender stays significant after adding these controls. One reason for the difference in our case could be the WTC scale, which makes the effect of risk preferences on WTC more salient, and then such gender gap in our case is explained by risk preferences and confidence. [126] focus on experimental measurement error, and suggest that the gender gap in competition is well explained by risk attitudes and overconfidence once measurement error is taken into account. In this regard, our results align with [126].

Table 2.5: Regression Analysis of Ordered Probit Model of WTC.

	Baseline	Baseline	Treatment	Treatment	Treat & Base	
	(1)	(2)	(3)	(4)	(5)	(6) Binary-WTC
Female	-0.498^{**}	-0.299	-0.308	-0.277	-0.440^{*}	-0.463^{*}
	(0.249)	(0.264)	(0.187)	(0.213)	(0.253)	(0.259)
Breadwinner				0.401^{**}		
				(0.197)		
Treat					0.064	-0.064
					(0.235)	(0.218)
Treat*Female					0.309	0.681^{**}
					(0.312)	(0.346)
Ability		-0.012		-0.027	-0.022	-0.017
		(0.041)		(0.026)	(0.022)	(0.022)
Confidence		0.324^{***}		0.089	0.171^{***}	0.161^{***}
		(0.085)		(0.054)	(0.045)	0.054
RiskTolerance		0.149^{**}		0.211^{***}	0.211^{***}	0.268^{***}
		(0.076)		(0.063)	(0.047)	0.063
No. Female		-0.064		0.107	0.023	-0.001
		(0.118)		(0.088)	(0.068)	(0.053)
Observation	76	76	129	129	205	205
Pseudo R^2	0.015	0.094	0.006	0.065	0.059	0.122

Notes: Dependent variable in columns 1-5 is the WTC scale, and it is the adjusted-binary WTC choice in column 6. Standard errors in parentheses. Ability refers to performance in the 1st stage. Confidence refers number 1-8 with higher number indicating more confident. NO. female refers to the number of female in each session. $*p < 0.10, **p < 0.05, ***p < 0.01$

show that females are less confident than males, which is consistent with previous literature (see [65]). The *p*-values in the last three columns suggest the roles do not change the confidence level for neither men nor women. The beliefs for guessing the relative performance to their partners show similar results (see Appendix B.3), except that both males and females believe they perform better than their partners, and there is no gender difference. We further investigate whether the effect of confidence on the WTC change by role assignment using a difference-in-difference approach. We regress WTC on a dummy variable that takes the value of 1 if there is a role assignment and 0 for the control. The results are shown in columns (1) and (2) of Table 2.7. The results suggest confidence has a significant and positive effect on WTC for both “breadwinners" and “supporters", but the role assignment mitigates the impacts of confidence on WTC. Males are more confident than females, and therefore they are more likely to enter the competition. The role assignment then may reduce the gender gap in WTC by mitigating the effects of confidence on WTC.

Table 2.6: Estimated Performance Rank

	Baseline	Treat Treat	Breadwinner	Supporter	*p*-value Base vs BW	*p*-value Base vs SP	*p*-value BW vs SP
Men	2.54	2.84	2.81	2.87	0.683	0.431	0.839
Women	3.05	3.29	3.29	3.29	0.772	0.436	0.751
P-Value	0.035	0.087	0.298	0.148			

Notes: Subjects are asked to guess their performance ranks in a session (1-8). 1 means the best.

Table 2.7: Ordered Probit Estimates of Confidence and Risk on WTC.

	(1) Breadwinner & Base	(2) Supporter & Base	(3) Breadwinner & Base	(4) Supporter & Base
Treatment	-0.238	-0.733^{*}	0.735	-0.449
	(0.374)	(0.392)	(0.472)	(0.457)
Confidence	0.332^{***}	0.369^{***}		
	(0.081)	(0.082)		
Treat * Confidence	-0.240^{**}	-0.272^{**}		
	(0.107)	(0.113)		
RiskTolerance			0.157^{**}	0.174^{**}
			(0.074)	(0.074)
Treat * RiskTolerance			-0.061	0.158
			(0.108)	(0.121)
Observation	140	141	140	141
R-square	0.079	0.081	0.069	0.074

Notes: Dependent variable is WTC. Standard errors are in parentheses. All regressions control for gender, risk/confidence, performance in the first stage and gender composition in each session.

Risk preferences were measured using the [41] gamble choice task. We replicate the robust findings in the literature that women are more risk averse than men (p-value < 0.01). We further investigate whether the effects of risk preferences change by the role assignment and the results are presented in columns (3) and (4) of Table 2.7. In contrast to confidence, there are no significant interaction terms, suggesting that risk aversion has no differential impacts on the WTC of different groups.

2.5.2 Responsibility for Making the Choice for the Group

In our setup, the WTC scale also involves responsibility for making the choice for the group. The change in responsibility will affect the WTC level. Recall that we use the degree of randomization to set the responsibility score for each WTC scale. The responsibility score equals 0 for points 2 and 3; 1 for points 1 and 4; and 2 for points 0 and 5. Table 2.8 summarizes the responsibility scores for the baseline and treatment groups.

Table 2.8: Responsibility Scores

	Baseline			Breadwinner			Supporter		
	Total	Scale $0-2$	Scale $3-5$	Total	Scale $0-2$	Scale $3-5$	Total	Scale $0-2$	Scale $3-5$
Men	1.49	1.50	1.48	1.50	1.33	1.55	1.00	0.88	1.14
Women	1.00	1.09	0.89	0.82	0.75	0.85	0.971	1.17	0.76
P-value	0.008	0.125	0.021	0.001	0.135	0.003	0.884	0.306	0.191
Observation (M/F)	35/41	12/23	23/18	26/38	6/12	20/26	30/35	16/18	14/17

P-values are from two-tailed Mann-Whitney U-Tests.

Table 2.8 shows that in the baseline and with the role “breadwinner", men have a lower propensity for randomization, and are more willing to take the responsibility compared to women. The gender gap in responsibility mainly comes from a larger proportion of points $3-5$ in the scale by males. There is no gender difference in responsibility for the role of “supporter". Higher responsibility in the piece-rate (points $0-2$ in the scale) indicates a lower WTC, and higher responsibility in competition entry (points $3-5$ in the scale) indicates a higher WTC. The gender gap in responsibility partially explains the gender gap in WTC in the baseline and with the role “breadwinner", and no gender gap in WTC with the role of “supporter".

We further compare the responsibility between each role and the baseline within each gender. We find the role assignment only significantly changes male's responsibility score when they are assigned as the “supporter". Male supporters are less willing to take responsibility compared males in the baseline and male breadwinners ($p = 0.011$ for baseline, $p = 0.015$ for breadwinner), with the difference mainly driven by the responsibility in the $0-2$ points in the scale. Females' responsibility are not likely to be affected by the role assignment. The distributions of each WTC scale for men and women are shown in Appendix B.6.

2.5.3 The Effects of Social Norms

We modified the method in [112] to elicit social norms for competitiveness of each role and gender. Participants were asked to:

"Please indicate in a 1-5 scale how competitive you think a _____ should be. If your answer is the same as the most frequent answer in the experiment, you will receive a payment of 25 cents."

The social norms were elicited for the following: woman, female supporter, female breadwinner, breadwinner, man, male supporter, supporter, and male breadwinner in a household. Each subject was asked the whole set. Figure 2.5 shows the mean responses for the social norms of each role and gender in the treatment group.[25] Using Wilcoxon signed-rank tests, we find that most participants believe other participants believe *Females* should be less competitive than *Males* (p-value= 0.052). The social norms for the competitiveness of *Supporters* are lower than *Breadwinners* (p-value < 0.01). Adding the two labels to the gender closes the gender gap of social norms for competitiveness (p-value = 0.470 for Breadwinner, p-value = 0.132 for Supporter). We further find the social norms for competitiveness of *female and male supporters* are lower than females and males without any labels (p-value < 0.01 for both genders). In contrast, the social norms for competitiveness of *female and male breadwinners* are higher than women and men without labels (p-value < 0.01 for both genders). Interestingly, the results are consistent when considering the social norms reported by men and women separately (see Figure B.3).

Based on our main results shown before, we suggest the results using adjusted-binary WTC choice match the elicited social norms for competitiveness. We also developed a theoretical model to investigate how social norms for competitiveness affect WTC (see Appendix B.5). In the model, we argue the intrinsic self-image cost of entering the competition is higher for a woman compared

[25] In the control group, subjects had not seen the two labels until this section, and some subjects commented that they did not understand the label "Supporter" until they saw "Breadwinner" in the following question. Thus, we focus on the elicitation of social norms in the treatment. Including the social norms elicited in the control does not change the results (see Figure B.1).

to a man, and it is also higher for a supporter compared to a breadwinner. Assigning the role "breadwinner" to a woman reduces her social cost of entering the competition. Thus, we argue that the social norms for competitiveness can be a major mechanism through which the social roles affect WTC.

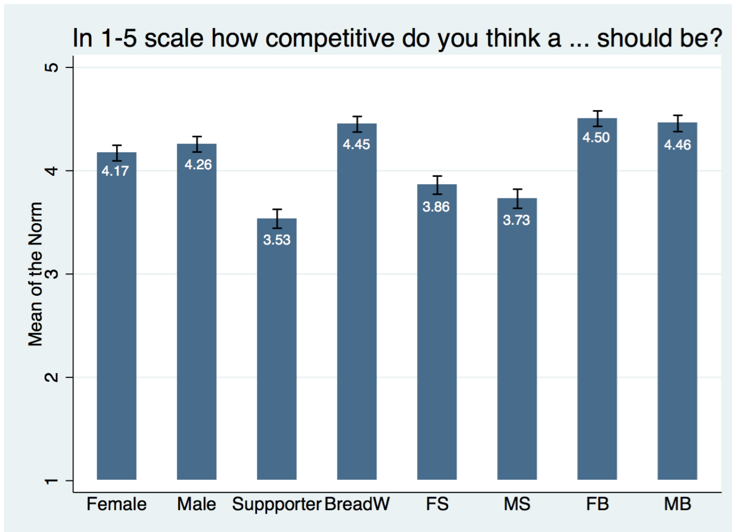


Figure 2.5: Social Norms for Competitiveness (with SE)

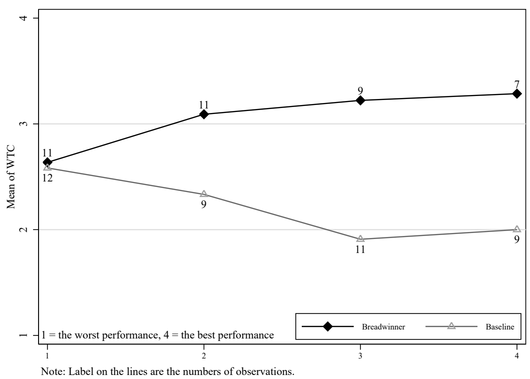
To sum up, social norms for competitiveness of different social roles (including men and women) and the responsibility for making the choice for the group explain the gender gap in WTC in our experimental set-up (both in the baseline and treatment groups). The changes in WTC brought by the role assignment mainly come from the social norms associated with the social roles, and based on results in Table 2.5, women are more responsive to such social norms.

2.5.4 WTC by Performance

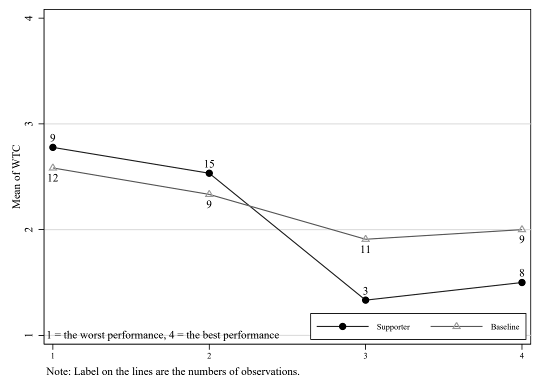
In this section, we are interested in whether the role assignment differentially affects the WTC of subjects with different ability.

Figure 2.6 shows the mean WTC for women (panels a and b) and man (panels c and d) conditional on their performance quartile during the first stage.[26] Panel a of Figure 2.6 shows that the

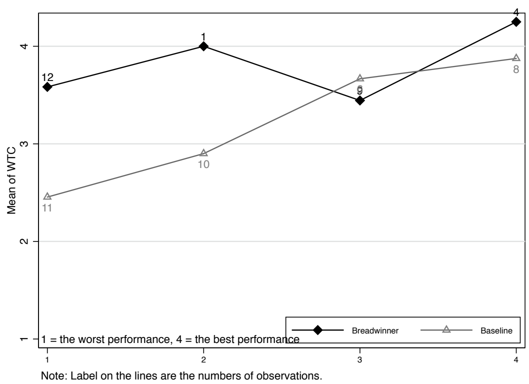
[26]We use performance in the first stage before participants make the tournament entry decision, and there is no



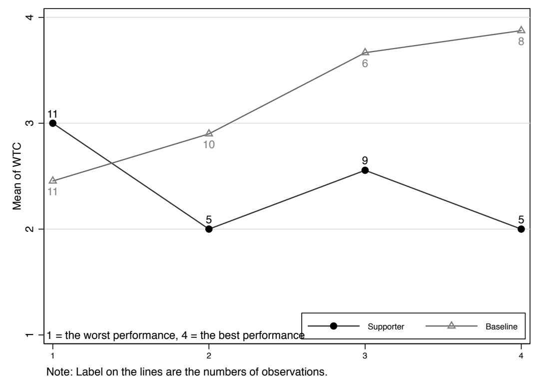
(a) Female (breadwinner and baseline)



(b) Female (supporter and baseline)



(c) Male (breadwinner and baseline)



(d) Male (supporter and baseline)

Figure 2.6: WTC by Performance Quantiles

increase in WTC of female breadwinners is mainly contributed by high-performing women. The difference in WTC between the baseline and female breadwinners in the first two best quantiles (3 and 4) is significant (p-value = 0.040). In panel b of Figure 2.6, we do not find similar patterns of changes when comparing women in the baseline and women supporters. We also show the WTC by performance quantile for males (Figure 6-c and 6-d). We do not find similar results for male breadwinners, but we find that the WTC of male supporters in the top two performance-quantiles is lower than males in the baseline within the same quantiles (p-value = 0.048). This means the decrease in WTC of male supporters is also mainly contributed by high-performing men. Based on our analysis about social norms for competitiveness, one possible explanation is high-performing subjects are more likely to obey social norms.

group pairing effect in the first stage.

We finally compare the expected earnings in the second stage for females in different performance quantiles. When we calculated the expected earnings, we hypothesized that subjects compete individually with 8 subjects per session. The piece-rate payment is 0.1 per correct answer, and the tournament-payment is 0.8 per correct answer. We then use subjects' performance in the first stage (ability), their ranks in the first stage and their WTC to calculate the expected earnings in the second stage. Results in Table 2.9 show the top performing-quantiles women are significantly better off when they are assigned as breadwinners in the treatment compared to the baseline. In fact, they earn 44.2% more by entering the competition. For males, we do not find expected earning differences between the control and treatment by performance quantiles (see TableB.4). We also show actual earnings by quantiles for men and women, and the results are robust (see Appendix B.7). Highest-ability women with the "breadwinner" title earn twice the actual amount of money than highest-ability women in the control.

Table 2.9: Women's Expected Earnings in the Second Stage by Performance Quartile ($)

		Baseline	Treat	Breadwinner	Supporter	*p*-value Base vs Treat	*p*-value Base vs BW	*p*-value Base vs S
Worst	1-2	0.94(0.17)	1.10(0.13)	1.03(0.19)	1.16(0.19)	0.516	0.836	0.369
Best	3-4	3.62(0.66)	4.32(0.53)	5.22(0.71)	3.00(0.63)	0.236	0.041	0.710

Notes: 1= the worst performance; 4 = the best performance. Standard errors are in parentheses.

2.6 Online Experiment

By using the roles of "breadwinner" and "supporter", we focus on the traditional family roles of males and females in western society. We further conducted an online experiment using Amazon Mturk with a more general population.[27] In the online sessions, we implement a between subjects design with a baseline and a treatment group. However, the treatment roles are "manager" and

[27] In the online sessions, we ask people not to use a calculator. It is still possible people still used the calculator; however, the number of correct answers solved in the online experiment is very low. Although the absolute numbers might be affected, the relative comparisons should not be affected in our experiment.

“assistant". We use these two roles for two reasons. First, we want to evaluate whether the social norm effects in the lab still hold with different titles. Second, we focus on roles existing in the actual workplace. The online experiment was computerized with Otree [127]. The experimental design for the online experiment is identical to the lab experiment. The only difference is the treatment roles and related incentivized questions in social norm elicitation and questionnaire parts. We scaled down the payments on Mturk, and the average payment of the two tasks was $1.5. Table B.7 summarizes the experimental procedures and the number of subjects in the online sessions.

We conduct the same analysis for the online experiment, and obtain similar results (Figure 2.7). There is a gender gap in WTC in the baseline (p-value = 0.044), and in aggregate there is no gender gap in the treatment group (p-value = 0.966). When females are randomly assigned as managers, their WTC increases significantly compared to females in the baseline (p-value = 0.043) and female assistants (p-value = 0.035). The gender gap disappears between female managers and untreated males in the baseline (p-value = 0.882). When males are assigned as assistants, their WTC is lower than the WTC of males in the baseline (p-value = 0.069). Male assistants also have lower WTC compared to male managers, but the difference is not significant (p-value = 0.113). When we compare WTC within the treatment, there is no gender gap for both managers and assistants (p-value = 0.898 for manager, p-value = 0.748 for assistant).

We further check the performance in the two stages, confidence and risk preferences. We also elicit the social norms for the new titles. The results are congruent with what we find in the lab experiment (please see the Appendix for more details). The online experiment provides robustness of the effects of social role titles on the WTC we found in the laboratory experiment.

2.7 Concluding Remarks

PricewaterhouseCoopers (PwC) recently surveyed more than 3,600 professional women (aged 28-40) to learn about their career development experiences and aspirations.[28] The survey results show that 82% of women are confident in their ability to fulfill their career aspirations and 73%

[28] PricewaterhouseCoopers (doing business as PwC) is one of the Big Four auditors in the world. The report of the survey can be found at www.pwc.com/timetotalk.

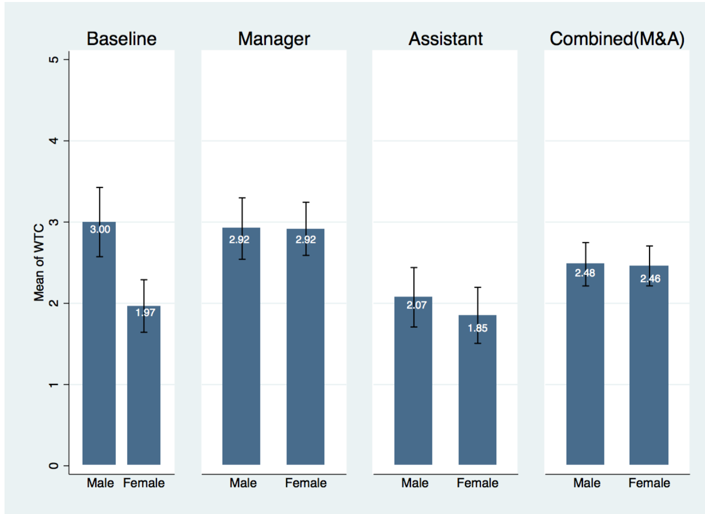


Figure 2.7: Comparisons of Mean of WTC in Online Experiment (with standard error bars)

are actively seeking career advancement opportunities. However, 42% feel nervous about how starting a family might impact their careers. Affected by traditional social norms, women are expected to spend more time in child care and household activities than men [128]. As a result, reducing the limits of social norms for competitiveness on women becomes important. Especially for high-ability women, since it is costly for them, their families and society if they do not pursue professional positions for which they are the best suited candidate. PwC's global chairman Bob Moritz states "It must go hand in hand with efforts to mitigate any unconscious biases and gender stereotypes that have traditionally impacted career success and progression in workplaces around the world".

However, gender norms remain extremely powerful because they are highly visible, biologically and culturally rooted. Changing norms is almost impossible during a short period of time, but the results of our experiment show that another social role added to females with different prescriptive social norms for competitiveness can be promoted to mitigate the effects of traditional gender norms and change their competitive behavior. We use a simple experiment and provide evidence that females' willingness to compete is limited by social norms. We show females' WTC significantly increases when they are *randomly assigned* the role of "breadwinner" or "manager". Although the role is randomly assigned, and there are no changes in responsibility or power bestowed on this role, it still gives women a permission “nudge" to break the glass ceiling of traditional gender roles for competition. This result is further reinforced by examinations of the underlying mechanism behind the results. Becoming the “breadwinner" or “manager" mitigates the impacts of confidence on competition entry and reduces the social cost of entering the competition for women. According to our social norm elicitation, under the title of “breadwinner" or “manager", it becomes socially permissible for women to be competitive.

Importantly, the increase in female breadwinners' WTC is mainly contributed by high-ability women, and they earn 44.2% more money compared to high-ability women in the baseline. The effects of the social role nudge in our experiment imply that restructuring institutions, such that assigning leadership encouraging titles to women, has the potential to reduce gender disparities in economic and labor market outcomes.

In our experiment, for simplicity and clean identification, we do not disclose any information about partners. Future work may investigate how subjects change their WTC when the gender of the other group member is disclosed to them, or when the role is assigned based on performance. It is also possible that social roles may have heterogeneous effects on women of different ages and women in different cultures, which can be another avenue for future research.

3. BARGAINING WITH PROMISES: THE EFFECTS OF PROMISES ON EMPLOYEES' EFFORT CHOICE[1]

3.1 Introduction

From politicians running for office to couples swearing eternal love, promises are universal in human culture. In labor markets, promise-based management plays a crucial role in fostering employees' engagement in the workplace, and the positive correlation between employees' engagement and organizational performance is proven [129, 130]. Well-crafted, thoughtful promises from employees can build trust and productivity [131]. In the workplace, the procedure of signing a contract or reaching an agreement where the employee makes a promise to the employer typically involves negotiation. During negotiation, the employee's promise usually comes with a request or claim from the employee for increased salary, a promotion, or flexible working schedules. The employer then chooses whether to trust the employee, and sequentially, the employee allocates effort. In this paper, we follow these sequential interactions between the employer and employee that include a "claim and promise" option in the labor contract to test the effectiveness of promise-based management in a laboratory experiment.

The labor contract between employers and employees is incomplete in that the employee has the residual ability to adjust her/his performance since performance is not always observable or easily enforceable after implementing the contract. If the employee is entirely self-interested, s/he would exert minimal effort. However, numerous scholars have documented a "gift exchange" between employers and employees where employees respond to generous wage levels (i.e., the gift) by exerting above minimal effort [132, 133, 134, 135]. [136] first introduced the gift-exchange game to study labor contract relationships in a laboratory experiment. Since then, several studies have tested different methods to obtain Pareto improvements in a gift-exchange game to provide innovation in the workplace to harness motivation and increase employees' performance [137,

[1]This study was funded by the IFREE Small Grants Program, and preregistered at AEA RCT Registry: AEARCTR-0003533. This study was approved by the IRB.

138, 139].[2] In the current paper, we focus on employees' promises in the workplace. In particular, we explore the effects of using a non-binding promise along with a claimed wage ("claim and promise") from the employee on boosting the labor contract efficiency in a one-shot gift-exchange game using both stated and real effort paradigms.

We implemented a two by two experimental design varying the promise (no "claim and promise" vs. "claim and promise") and effort paradigm (stated-effort vs. real-effort). We started with the stated-effort paradigm. In the baseline, subjects play a modified two-stage gift-exchange game. During the first stage, the employer chooses a wage level and announces a non-binding desired effort level. In the second stage, the employee chooses the effort level with a fixed cost for each level of effort. In the treatment group, the employee chooses whether to make a "claim and promise" or not before the employer selects the wage level. The employee initiated the "claim and promise" using a brief limited message to the employer: "If the firm gives me the wage w, I will choose the effort level e". Both the claimed wage and promised effort are non-binding. Adding this step increases the number of interactions between the employer and the employee and allows us to follow the negotiation process. The difference between the actual wage and the claimed wage and the difference between the desired effort and the promised effort may affect reciprocity between the employer and the employee.

Our results suggest that most employees (92.7%) chose to make a "claim and promise" during the first stage in the stated-effort paradigm. When the employer trusted the employee and provided a wage greater than or equal to the claimed wage, the employee *reciprocated* by keeping his promise or exerting an even higher level of effort. Both the actual wage and effort level increased, in this case, compared to the baseline. The wage level increased by 15.1%, and the effort level increased by 28.3%, which resulted in 15.3% higher final social payoffs. However, when the

[2] [137] analyze the consequences of control in a principal-agent game where the principal can control the agent by implementing a minimum performance constraint before choosing an activity. They find that most agents reduce their performance in response to the controlling decision. [138] found that delegation can improve effort levels and firm profits by conducting an experiment in which an employer chooses for each worker whether to assign the wage or to allow the worker to choose her/his wage (i.e. "delegate" the wage choice). [139] found that social comparisons concerning wages and decision rights affect workers' performance when a firm can choose workers' wages or let them choose their own.

employer offered a wage lower than the claimed wage, the employee *retaliated* by breaking his promise. The employer tried to negotiate with the employee desiring a different effort level than the promised effort based on the difference between the actual wage he offered and the wage the employee claimed. However, the employee cared more about the trust of the employer and ignored the negotiation. The employee retaliated whenever the actual wage was lower than the claimed wage.

To obtain more implications for a real workplace environment, we also ran the baseline and “claim and promise" treatment using a real-effort task. In the real-effort task, subjects added up sets of five two-digit numbers for three minutes. Compared to the stated-effort paradigm, we added an incentivized practice stage after subjects read the instruction and before they started the game. The employee was informed about his performance and the median performance of participants in their sessions (i.e., relative performance) before starting the game. The employer was only informed about the median performance of the session. We used the performance during the practice stage as a proxy for ability.

We replicated the main results obtained in the stated-effort paradigm in terms of direction, but the magnitude of the results was much larger in the real-effort paradigm. When the employer chose to trust the employee, the wage level increased 55.9%, the effort level increased 35.8%, and the final social payoffs increased 30.0%. There were differences between the stated and real-effort conditions. We found lower trust rates of the employer and a lower degree of retaliation of the employee in the real-effort paradigm. Due to information about ability, the employer could infer the probability of promise-keeping by comparing the promised effort and the median performance during the practice stage. Thus, the employer was more likely to choose not to trust the employee when the promised effort was higher than the median performance during the practice stage. The employee also knew whether his promised effort was higher than her/his performance during the practice stage. Consequently, when the employer provided a lower wage with a lower desired effort, the employee chose not to retaliate as much as in the stated-effort paradigm. Based on these results, we argue that the negotiation between the employer and the employee in the real-effort paradigm is successful.

One of the leading theories in the literature regarding why people keep their promises is the belief-based *"guilt-aversion"* model from psychological game theory (PGT)[140, 141, 142].[3] In contrast to traditional game theory, PGT models emotions directly into the utility function [146]. The "guilt-aversion" model assumes that people care about others' expectations. Promises change others' expectations, and subsequently, people feel guilty if they break promises. Scholars from various disciplines have explored guilt-aversion to explain other prosocial behavior [147, 148, 149, 150]. Most relevant to our work is [151]'s test of guilt-aversion as a mechanism to explain the "gift exchange". Based on PGT, *reciprocity* is another potential explanation of the "gift exchange" [152, 153]. Our theoretical analysis focused on stated-effort, applied PGT, and included both reciprocity and guilt in the utility function. The "claim and promise" treatment served to change the reciprocity component through the "claimed wage" and change the guilt component through the "promised effort". We showed simulated results of the PGT models for both the baseline and the treatment. Another contribution of this paper is that we used choice-process data (pupil dilation and eye-tracking lookup patterns) to test subjects' feeling of guilt and reciprocity dynamically and then provide direct evidence for PGT models.[4]

The rest of the paper is presented in the sections to follow. Section 3.2 offers a review of the literature. Section 3.3 explains the experimental design. Section 3.4 constructs the theoretical model and reports the simulation results. Section 3.5 presents the main results of our experiment. Section 3.6 provides a discussion of choice-process data, followed by the conclusions in Section 3.7.

[3]In noting the "commitment-based theory" which does not belong to PGT, [143] argued people have an innate preference for promise keeping per se. The discussion on which theory poses stronger data demands related to promise-keeping has not yet reached a consensus in the literature [144, 145]. In this paper, we focus on PGT and the guilt-aversion theory.

[4]Each computer station included a screen-based eye-tracker and a high-resolution camera, which captured facial expressions (see Figure C.1).

3.2 Literature Review

Scholars have intensively studied promises theoretically [154] and within different economic games, such as in the (hidden action) trust game, the dictator game [140, 144, 155, 156, 157, 145, 158], and the mistress game [159]. Limited literature exists on promises in experimental labor economics. To investigate the effects of cheap-talk evaluations on effort choice in a one-shot gift exchange game, [160] provided an “evaluation" option to the employer after an employee’s choice of effort. [161] evaluated promises in group decision-making using a gift-exchange game. The contract used in [162] bears the closest resemblance to the one we used in our study. In the [162] design, the treatments differed in two factors: (1) who proposes the contract (the firm or the worker) and (2) whether the proposed contract includes a non-binding specification of the worker’s effort level. Their key finding was that the worker’s actual effort choice was highest when the worker proposed the contract, and the contract included the worker’s promised effort level. Our paper departs from [162] in several ways. Their study focused on whether the tendency to keep agreements depended on who proposed the agreement; our paper focused on the sequential negotiation process between employers and employees where employees made promises to employers with claimed/proposed wages, and then employers paid wages with desired effort levels. In [162], the worker’s proposal could be either accepted or rejected by the firm. If the firm rejected the worker’s proposal, the game ended, making the proposed wage binding. In our design, the employer could choose any wage level they wanted (not necessarily the claimed wage) after reviewing the employee’s “claim and promise," which made the wage non-binding. We believe the employer’s actual wage is a crucial factor reflecting the employer’s trust and affecting reciprocity in the negotiation process. [162] measured the feeling of guilt using a self-reported conscious affect-3 (TOSCA-3) score at the end of the experiment [163]. TOSCA-3 tests the intrinsic characteristics of personality, which is likely not affected by the decisions during a laboratory session. Using biometric equipment permitted tracking changes in feelings of guilt throughout the experiment and permitted its use to test PGT models. In addition, we implemented a real-effort paradigm to include ability in the promise setup.

3.3 Experimental Design

A total of 314 undergraduate and graduate students from Texas A&M University participated in this experiment. Each session involved 6-14 participants. The experiment was computerized using Ztree [108], and lasted approximately 30 minutes. Each participant earned approximately $25 (including a $5 show-up fee). The experiment instructions are available in the Appendix.

Upon arriving at the laboratory, participants were randomly assigned to a computer station. Each computer station was equipped with an eye-tracker and a high-resolution camera to capture participants' facial expressions. In each session, participants were randomly assigned as an employer or an employee, and each employer was randomly paired with an employee. Participants only knew that they were paired with another participant in the room. Partners received no information about each other. To avoid reputation effects, each treatment lasted only one round (see [164] and [162]). Our two by two design consisted of a baseline and the "Claim & Promise" treatment (hereafter C&P) for stated and real effort paradigms. Each treatment was a modified version of [136]'s gift-exchange game.

3.3.1 Stated Effort

Stated-Effort Baseline: Subjects played the gift-exchange game in two stages:

Stage 1: a) The employer chose a wage level, w for the employee;

b) The employer announced a non-binding effort level, $\hat{e}$, that he wanted the employee to do for the chosen wage.

Stage 2: The employee chose the effort level, e, after reviewing the wage level set by the employer.

The combination of wage and effort level determined the monetary payoffs for the employer and the employee as follows:

$$\begin{aligned} \pi_F &= (120 - w) * e \\ \pi_W &= w - c(e) - 20 \end{aligned} \tag{3.1}$$

where $c(e)$ was the cost of effort, a function increasing in e. Wages were integers between 20 and 120.[5] We used tokens in the experiment, and each token equaled $0.50. Table 3.1 shows the feasible effort levels and the cost of effort.

Table 3.1: Effort-Cost Table

Effort Level	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Cost	0	1	2	4	6	8	10	12	15	18

Stated-Effort C&P: The only difference between the baseline and the C&P treatment was an additional stage before the employer chose the wage level. The first stage became:

Stage 1: The employee decided whether to make a “claim and promise" or not:

“If the firm gives me the wage=_____, I will choose the effort level=_____". This claim and promise were non-binding.

The next two stages were identical to the baseline.

3.3.2 Real Effort

In the real-effort paradigm, subjects performed a real effort task that consisted of adding up sets of five two-digit numbers for three minutes. Then, the effort level was the number of correct answers to the additions, s.

Real-Effort Baseline: After subjects read the instructions and before they started the game, they participated in a practice stage. Each subject was given three minutes to do the task and received $0.25 for each correct answer. Employees learned about their performance and the median number of correct sums for the session after finishing the practice task.[6] To avoid the potential confounding effects of employers' ability on their wage decisions, employers were only informed

[5]We used the redemption value of 120. It was possible for the employee to take the employer's entire profit. Previous studies suggest that, except for the difference in the level of the firms' earnings, the results were very similar using a redemption value of 120 or 240 [165, 138].

[6]The program assigns the roles initially, but we show the assigned roles to participants after the practice stage.

about the median number of correct sums for the session. They did not know their performance until the end of the experiment. Performance in the practice stage served as a proxy for subjects' ability. Note that providing information about the median performance was to inform the employer about the employee pool's performance level and to inform the employee about their relative performance. In real work environments, the employee is usually aware of his ability. After the practice stage, participants proceeded to the following sequential stages.

Stage 1: The employer chose the wage level, w; and announced a non-binding number of correct additions $\hat{s}$ he wanted the employee to solve.

Stage 2: After reviewing the employer's wage level, the employee proceeded to the addition task.

The number of correct answers was updated on the computer screen after each question so that the employees could decide to continue solving more questions or to stop exerting effort at any point. Note that there was no fixed cost function for each subject in the real-effort paradigm, and the cost was their real effort. The payoff functions were:

$$\begin{aligned} \pi_F &= (120 - w) * (s/10) \\ \pi_W &= w - 20 \end{aligned} \tag{3.2}$$

Real-Effort C&P: The only difference from the baseline was that there was an additional stage before the employer chose the wage level after the practice stage. The first stage then became:

Stage 1: The employee decided whether to make a "claim and promise" or not:

"If the firm gives me the wage=_____, I will solve _____ correct sums". This claim and promise were non-binding.

After the one-shot gift-exchange game in the stated and real effort paradigm, we elicited the employer's beliefs. We asked the employer to guess the effort level the employee provided. A payment of $1 incentivized correct answers. Finally, subjects proceeded to the TOSCA-3 questionnaire and a demographic survey. We used the same five scenarios as [162] from TOSCA-3 in our experiment. Each scenario's answers were on a scale from 0 to 4, with "0" indicating no feeling of guilt while “4" indicated strong feelings of guilt. Table 3.2 summarizes the number of observations in each treatment and also the number of employees who chose to make a “claim and promise."

Table 3.2: Number of subjects in each treatment group.

	Total No. (Pair of Employer & Employee)	No. of Promise
Stated-Effort Baseline	76 (38)	-
Stated-Effort C&P	82 (41)	38 (92.7%)
Real-Effort Baseline	78 (39)	-
Real-Effort C&P	82 (41)	35 (85.4%)

3.3.3 Eye-tracking and Pupil Size

We used eye-tracking to record visual attention and pupil size during the entire experiment. The eye-tracker recorded where participants were looking at the computer screen. It provided data including, but not limited to, pupil dilation, gaze location, time spent on a stimulus (fixation), the time length of fixations, and the number of revisits. The measurement rate was 60 Hz, which means each metric was measured 60 times per second. Pupil dilation has been previously used in the economics literature to measure emotional arousal. More relevant to our experiment, researchers [166, 167, 168] have linked pupil dilation to guilt.[7]

Each participant was required to read the consent form that included disclosure information about the biometric tools before agreeing to participate in the experiment. Before starting the experiment, we performed a nine-point eye-tracking calibration for each subject to ensure high-quality data. We successfully recorded eye-tracking data for 147 subjects in the stated-effort paradigm and 143 subjects in the real-effort paradigm.

[7] We have 11 Tobii X2-60s and 5 Tobii Spectrum devices. All 16 stations also use iMotions AFFDEX for facial expressions. We evaluated facial expressions for evidence of guilt. However, we find subjects are neutral most of the time during the experiment (80.4% of the entire experiment).

3.4 Theoretical Model

3.4.1 Model Setup

Our theoretical model focused on stated-effort, which was the fixed cost function and did not involve ability. In the baseline, subjects played the two-stage gift-exchange game above. In the treatment, we added a “claim and promise" before the first stage. The “claim and promise" was non-binding as cheap-talk, and therefore based on traditional game theory, the sub-game perfect Nash Equilibrium (SPNE) is the same in both the baseline and the treatment group, with the employer choosing the minimum wage and the employee choosing the minimum effort. However, the literature documents higher wages and effort levels and a “gift exchange" between employers and employees [132]. Thus, we applied psychological game theory (PGT) to our gift-exchange setting.

In PGT, contrary to traditional game theory, utility is not solely motivated by material payoffs but also by belief-dependent motivations (see a survey, [146]). Reciprocity and emotions are two examples of belief-dependent motivations. Reciprocity means people respond to kindness with kindness and be unkind when treated unkindly. Reciprocity was the intuition to explain the wage-effort exchange when [132] first documented gift exchange in labor markets.[8] [153] later brought in *expectation* to gift exchange and present reciprocity as the employee’s response to the distance between the expected and actual wage.

Emotion is another motivation captured by PGT to study behavior. Among the emotion category, *guilt* is one of the most widely used emotions to explain cooperative behavior in different economics games, including cooperation in the gift-exchange game [141, 151]. Individuals feel guilty if they do not meet others’ expectations, and hence, guilt-aversion has also been used explain promise-keeping [140]. In previous literature, guilt-aversion is tested by eliciting second-order beliefs [140].[9] However, [172] pointed out that second-order belief elicitation may cause “false consensus," as players might believe others have thoughts similar to their own when asked

[8] [169] first incorporated reciprocity into game theory, and [152] expanded it to extensive game forms.

[9] In previous studies, guilt-aversion has also been tested through second-order beliefs elicitation to explain other prosocial behavior [170, 149, 171]. The results of whether second-order beliefs can explain behavior are mixed. [172] showed matched recipients’ expectations directly to the donors, and they suggested the correlation between behavior and expectations was close to zero.

to state their second-order beliefs. The “induced belief" is the solution to “false consensus" issues suggested by [172]. In the current design, we used the “desired effort" $\hat{e}$ as a proxy for the induced belief to study guilt-aversion.

We tested both reciprocity and guilt-aversion to explain the gift exchange in our design. We set the baseline and the C&P treatment model separately; the C&P treatment changed the reciprocity component in the model through the “claimed wage" and the guilt component through the “promised effort."

In the baseline, there was no claimed wage and promised effort, and the worker's utility depended on actual wage(aw), desired effort(de), and actual effort(ae). The following function consisting of three parts was how we modeled the worker's utility.

$$\begin{aligned} U_w(ae; aw, de) = \underbrace{\pi_w(ae, aw)}_{\text{material payoff}} + \underbrace{\lambda_R * (aw - 20) * \pi_f(ae, aw)}_{\text{reciprocity}} \\ - \underbrace{\lambda_G * T_G * max\{E[\pi_f; de] - \pi_f(ae), 0\}}_{\text{guilt}} \end{aligned} \tag{3.3}$$

where $\pi_w(ae, aw)$ is the material payoff of the worker, and $\pi_w(ae, aw) = aw - c(ae) - 20$ in our design. λ_R is the sensitivity parameter of reciprocity, and $\lambda_R \geq 0$; $aw - 20$ (the distance between the actual wage and minimum wage) represents the kindness received from the firm, and π_f is the payoff of the employer ($\pi_f = (120 - aw) * ae$). The entire *reciprocity* part indicates that the more the worker cares about reciprocity and the more kindness he receives from the firm, the more weight s/he puts on the firm's payoff.

The third part of the utility function modeled guilt. λ_G is the sensitivity parameter of guilt, and $\lambda_G \geq 0$. Since the desired effort (de) indicates the firm's expectation about the effort level and guilt is from letting down others, we modeled guilt to be generated by the distance between the firm's expected payoff by the desired effort and the firm's actual payoff by the actual effort. When $ae \geq de$ (i.e., $\pi_f \geq E(\pi_f)$), there is no guilt, and hence we set $max\{E[\pi_f; de] - \pi_f(ae), 0\}$. T_G is the trigger of guilt. [146] argued that the feeling of guilt should be “reasonable". For example, if the firm paid the minimum wage with the maximum desired effort in our experiment, then there should be no guilt from employees if they did not exert the desired effort. Thus, we added T_G to the model. We further set $T_G = \frac{aw-20}{de}$, and this means the higher the wage rate by the firm based on desired effort, the higher the level of guilt for the worker. Finally, based on previous literature, people derive utility from reciprocity and disutility from guilt, and therefore in our model, reciprocity increased utility, and guilt reduced the worker's utility.

In the C&P treatment, the worker's utility also depended on promised effort(pe) and claimed wage(cw):

$$\begin{aligned} U_w(ae; aw, cw, de, pe) = \underbrace{\pi_w(ae, aw)}_{\text{material payoff}} + \underbrace{\lambda_R * (aw - cw) * \pi_f(ae, aw)}_{\text{reciprocity}} \\ - \underbrace{\lambda_G * T_{G2} * max\{min(E[\pi_f; pe], E[\pi_f; de]) - \pi_f(ae), 0\}}_{\text{guilt}} \end{aligned} \tag{3.4}$$

where $(aw - cw)$ captures reciprocity in the treatment. If $aw >= cw$, then it means the worker received kindness from the firm; if $aw < cw$, then it means the worker was treated unkindly by the firm. T_{G2} is the trigger of guilt in the treatment, and we set $T_{G2} = \frac{aw-20}{de} * \frac{cw-20}{pe}$, which indicates that the level of guilt also depends on the claimed wage and promised effort. In the treatment, we assumed the smaller of promised-effort and desired-effort drove the worker's guilt. Thus, instead of using the distance between desired effort and actual effort to represent guilt, we used $[min(E[\pi_f; pe], E[\pi_f; de]) - \pi_f(ae)]$.

3.4.2 Simulation Results

We first put the material payoffs and the trigger of guilt into the utility functions. Then the worker's utility in the baseline becomes:ă

$$\begin{aligned} U_w(ae; aw, de) = [aw - c(ae) - 20] + \lambda_R * (aw - 20) * [(120 - aw) * ae] \\ -\lambda_G * \frac{aw - 20}{de} * max\{(120 - aw) * de - (120 - aw) * ae, 0\} \end{aligned} \tag{3.5}$$

The work's utility in the treatment becomes:

$$\begin{aligned} U_w(ae; aw, cw, de, pe) = [aw - c(ae) - 20] + \lambda_R * (aw - cw) * [(120 - aw) * ae] \\ -\lambda_G * (\frac{aw - 20}{de} * \frac{cw - 20}{pe}) * max\{min[(120 - aw) * pe, (120 - aw) * de] \\ -(120 - aw) * ae, 0\} \end{aligned} \tag{3.6}$$

We then ran simulations using backward induction, assuming λ_R and λ_G follow uniform distributions: $\lambda_R \sim U(0, a)$, $\lambda_G \sim U(0, b)$. To run simulations, we also assumed the firm knew the worker's type λ_R and λ_G. To make the magnitude reasonable, we took square roots of the reciprocity and the trigger of guilt (T_G and T_{G2}) when we ran simulations.

Table 3.3 shows the simulated contract outcomes based on different maximum values of λ_R and λ_G. The results suggest that when the model does not include reciprocity and guilt (i.e., $\lambda_R, \lambda_G \sim U(0, 0)$), employers would always choose the minimum wage and employees would always choose minimum effort. The wage and effort levels increase as employees' sensitivity for reciprocity and guilt increases. When employees' sensitivity for reciprocity is sufficiently large, employees increase effort levels even when the received wage does not change. The ρ values by Spearman tests show the correlation between wage and effort. Our simulation results support the "gift exchange," but the positive correlation between wage and effort declines when employees' sensitivity for reciprocity and guilt increases as suggested by the ρ values. This is because employees would always choose high levels of effort with large λ_R and λ_G. Finally, when we compared the C&P treatment with the baseline, we found the wage and effort levels were higher in the C&P treatment when λ_R and λ_G were small. We next used data from our experiment to test for treatment effects in the following section.

3.5 Results

This section presents and discusses our findings regarding how "claim & promise" affects contract outcomes. We presented stated and real effort paradigms separately and also compared these two paradigms. We used the choice-process data to explore guilt and reciprocity in the next section.

Table 3.3: Simulated Contract Outcomes.

	$\lambda_R, \lambda_G \sim U(0,0)$			$U(0,0.01)$			$U(0,0.02)$			$U(0,0.04)$			$U(0,0.06)$		
	Wage	Effort	ρ	Wage	Effort	ρ	Wage	Effort	ρ	Wage	Effort	ρ	Wage	Effort	ρ
Baseline	20	0.1	-	22.40	0.11	0.99	27.86	0.22	0.73	33.02	0.47	0.61	33.25	0.68	0.34
C&P Treatment	20	0.1	-	45.80	0.26	0.61	44.76	0.36	0.63	43.72	0.48	0.43	42.24	0.48	0.33

Notes: ρ in columns are from Spearman tests.

3.5.1 Stated Effort

Result 1a: The employee reciprocated to the employer by keeping his promise when the actual wage was greater than or equal to the claimed wage; the employee retaliated against the employer by breaking his promise and choosing a lower effort than the desired effort when the actual wage was lower than the claimed wage.

Result 1b: Both wage and effort levels increased significantly when the actual wage was greater than or equal to the claimed wage compared to the baseline.

In the baseline, we first replicated the “gift exchange" between the employer and the employee from previous studies [132]. The employee’s actual chosen effort was positively correlated with the employer’s paid wage (Spearman’s test, $\rho = 0.431$, $p = 0.009$), which was also in-line with our simulated results based on the PGT model. However, the actual effort (e) was significantly lower than the employer’s desired effort (e_d) in the baseline (Wilcoxon signed-rank test, $p = 0.004$), as shown in Table 3.4.[10]

In the C&P treatment, the positive correlation between the effort and wage level still held (Spearman’s test, $\rho = 0.329$, $p = 0.036$). When we compared the C&P treatment to the baseline in general, we found employers paid a higher wage to employees; however, employees chose the same level of effort as in the baseline, which was lower than their promised effort (e_p) (Wilcoxon signed-rank test, $p = 0.014$) and the desired effort (e_d) ($p < 0.001$). There was no significant change in

[10] In four sessions, the rate of participation slipped. To keep an even number of participants, we asked one of the graduate students working in the lab to be the last participant. After we removed these four observations, 38 employers and 36 employees remained in the baseline study; 39 employers and 41 employees remained in the treatment. Since only three employees chose not to make a “claim and promise" in the treatment, we did not report the outcome for them. Including the three pairs of subjects does not change our results regarding the overall comparison between the baseline and treatment.

overall social payoffs.[11] Based on the overall comparison, we see employees broke their promises. Employees did not respond to employers' higher wage by choosing a higher effort level in the C&P treatment. The experiment's primary purpose was to understand how the "claim and promise" affected the employers' choices of wages, and subsequently, employees' effort. Employees might choose to break or keep their promises based on the difference between the actual wage (*aw*) and the claimed wage (*cw*). Panel B of Table 3.4 shows the treatment outcomes by separate cases.

Table 3.4: Summary statistics of contract outcomes in the stated-effort paradigm.

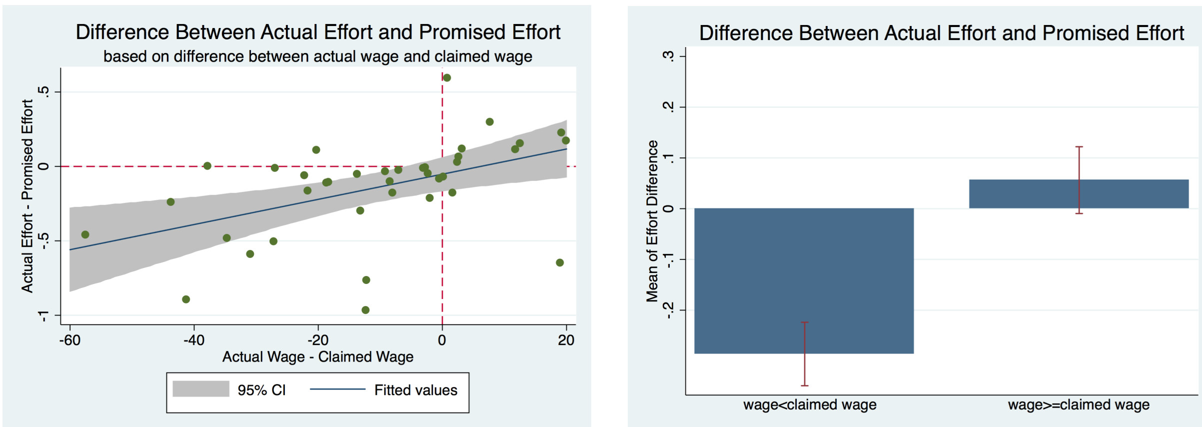
	Obs.(f/w)	w	e	e_d	e_p	w_c	π_f	π_w	Overall Payoffs
A									
Baseline	38/36	62.79 (3.12)	0.53 (0.04)	0.67 (0.03)			28.23 (2.57)	35.86 (3.04)	63.77 (3.12)
C&P Treatment	36/38	72.61 (3.20)	0.54 (0.04)	0.65 (0.03)	0.68 (0.03)	83.42 (3.45)	23.17 (2.19)	45.47 (2.88)	69.37 (2.18)
p-value (Base vs Treat)		0.031	0.922	0.469			0.209	0.033	0.201
B									
C&P Treatment ($aw >= cw$)	15/16	74.00 (5.78)	0.68 (0.07)	0.69 (0.06)	0.63 (0.06)	68.75 (5.25)	28.73 (3.80)	44.13 (5.09)	73.50 (2.82)
C&P Treatment ($aw < cw$)	21/22	71.62 (3.74)	0.43 (0.05)	0.63 (0.04)	0.72 (0.04)	94.09 (3.01)	19.19 (2.32)	46.45 (3.42)	66.36 (3.03)
p-value ($aw >= cw$ vs $aw < cw$)		0.686	0.007	0.529	0.209	< 0.001	0.024	0.906	0.047
p-value ($aw >= cw$ vs baseline)		0.045	0.065	0.818			0.533	0.190	0.032

Notes: p-values are from two sided Mann-Whitney U tests. In each cell, the numbers refer to mean, with standard errors in parentheses. We report wage of employer w, actual effort of employee e, desired effort of employer e_d, promised effort of employee e_p, claimed wage of employee w_c, employer payoffs in tokens π_f and employee payoffs in tokens π_w. In part B, we report these variables separately based on the difference between actual wage (aw) and claimed wage (cw) in the treatment. Note that we do not include the three pairs of employers and employees who chose NOT to make a "claim & promise." The exchange rate in our experiment is $1 token = \$0.5$.

Panel B of Table 3.4 suggests that when the employer trusted the employee and provided at least the claimed wage, the employee reciprocated by keeping the promised effort. Notably, there was no significant difference between the actual and promised effort (0.68 vs 0.63, Wilcoxon signed-rank test, $p = 0.103$). However, when the employer offered a wage lower than the claimed wage, the employee responded by breaking his promise, and actual effort was significantly lower than the promised effort (0.43 vs 0.72, Wilcoxon signed-rank test, $p < 0.001$). Detailed distributions where the "actual effort - promised effort" was positively correlated with "actual wage - claimed wage" (Spearman's test, $\rho = 0.621$, $p < 0.001$) comprise Figure 3.1a. This distribution suggests a conditional reciprocity in our experiment. We also showed the average distance between the actual and promised effort based on the wage difference in Figure 3.1b. The difference in the

[11] The changes in the payoffs of employers and employees strongly depend on the specific function forms. Therefore we focused on comparing the overall social payoffs rather than the payoffs for employers or employees.

(advantageous/disadvantageous) distance between these two scenarios was statistically significant (Mann-Whitney U test, $p < 0.001$).



(a) Distribution

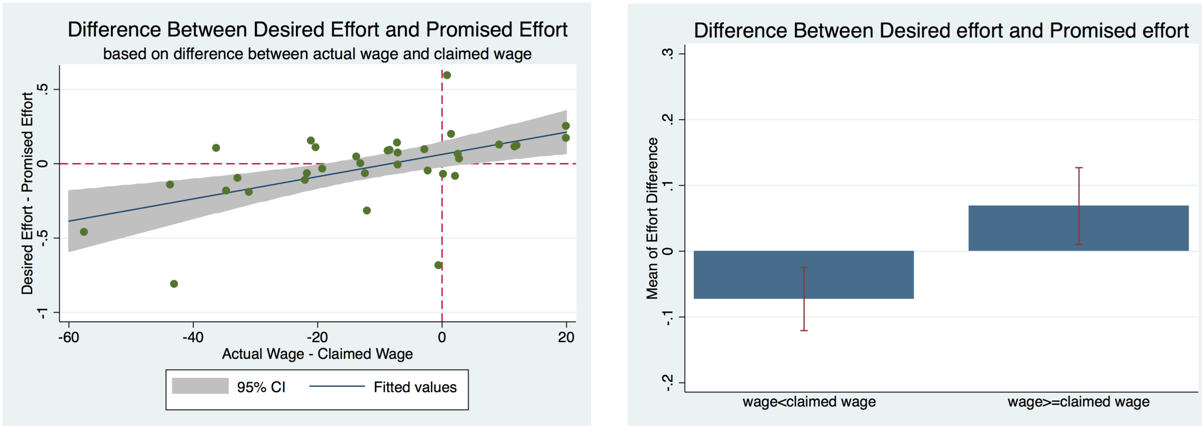
(b) Mean with Standard Error Bars

Figure 3.1: Difference Between Actual Effort and Promised Effort (Stated Effort)

The employer's actual wage and the desired effort were similar in these two scenarios, as shown by the *p*-values in the second to last row of Table 3.4. The reason why the actual wage was lower than the claimed wage was that the claimed wage was too high to be accepted by the employer (94.09 in $aw < cw$ vs. 68.75 in $aw >= cw$). The employer tried to negotiate with the employee desiring a different effort level from the promised effort when he chose to pay a different wage from the claimed wage. Figure 3.2a suggests that “desired effort - promised effort" positively correlated with “actual wage - claimed wage" (Spearman's test, $\rho = 0.622$, $p < 0.001$).[12] Figure 3.2b suggests the mean distance between the desired and promised effort when $aw < cw$ was negative. However, it was positive when $aw >= cw$. The difference in the distance between these two scenarios was significant (Mann-Whitney U test, $p = 0.016$).[13]

[12]The correlation between the desired effort and the actual wage is also positive and significant (Spearman's test, $\rho = 0.481$, $p = 0.003$).

[13]Using Wilcoxon signed-rank tests, we find when the employer chooses to pay an equal or higher wage than the claimed wage (i.e., $aw >= cw$), he desires higher effort than the promised effort (0.69 vs 0.63, $p = 0.039$). When $aw < cw$, there is no significant difference between the desired and promised effort (0.72 vs 0.63, $p = 0.126$).



(a) Distribution

(b) Mean with Standard Error Bars

Figure 3.2: Difference Between Desired Effort and Promised Effort (Stated Effort)

Since the employer desired different effort levels based on different wage levels he paid to the employee, we compared the actual and desired effort (ae vs. de) based on the wage difference to further explore the *conditional reciprocity* in our experiment. In the case of "$aw >= cw$", we found the employee tried to reach the desired effort, which was higher than his promised effort, and there was no significant difference between actual effort and desired effort (Wilcoxon signed-rank test, $p = 0.928$). When $aw < cw$, 38% of employers desired the same effort to the promised effort; 42% and 19% of employers desired a lower and higher effort than the promised effort, respectively. Table 3.4 suggests that, in general, the actual effort was significantly lower than the desired effort in the case of "$aw < cw$" (Wilcoxon signed-rank test, $p < 0.001$). We further found that even for the 42% of employers whose desired effort was lower than the promised effort, the actual effort was significantly lower than the desired effort ($p = 0.012$). In the real workplace, there are negotiations between the employer and the employee for salary, work schedule, working conditions, etc. For example, in collective bargaining, the negotiation process occurs between the representative of employees and the employer. Our results provide insights on such negotiations in labor contract procedures. Once the actual wage is lower than the claimed wage, the employee retaliates, even though the employer compromises on the desired level of effort.

The overall social payoffs were highest in the case of "$aw >= cw$" compared to the case of

"$aw < cw$" or the baseline. The *p*-values in the last row of Table 3.4 show that both the wage level and effort level were significantly higher than in the baseline when the employer chose to trust the employee ("$aw >= cw$"). This suggested our promise setup, in the case of "$aw >= cw$", enhanced the efficiency of the labor contract design. When "$aw < cw$", there was no significant difference in the overall social payoffs between the treatment and the baseline. Finally, we estimated an ordered probit model of actual effort; Table 3.5 shows the results. Columns (1) and (2) support the results above that our C&P treatment significantly increased the effort level when we compared the scenario "$aw >= cw$" to the baseline, and the result held after we controlled for wages, desired effort, and socio-demographic variables. In columns (3) and (4), we compared the scenario "$aw < cw$" to the baseline. After including control variables, we found that employees significantly reduced their effort levels when employers chose not to trust them. In column (5), we estimated a regression only for the C&P treatment group, and the results suggested when the employer paid greater than or equal to the claimed wage, the employee significantly increased her/his effort.

Table 3.5: Ordered Probit Estimations of Actual Effort (Stated Effort).

	Baseline vs Treat ($aw >= cw$)		Baseline vs Treat ($aw < cw$)		Treat
	(1)	(2)	(3)	(4)	(5)
Wage		0.020**		0.021**	0.003
		(0.009)		(0.009)	(0.015)
C&P Treat	0.644**	0.564**	-0.410	−0.683***	
	(0.314)	(0.285)	(0.264)	(0.245)	
Desired effort		0.930		0.801	2.992
		(1.090)		(1.256)	(2.043)
Promised effort					-0.106
					(1.584)
$AW >= CW$ (vs $aw < cw$)					1.455***
					(0.464)
Control	No	Yes	No	Yes	Yes
Observation	52	52	58	58	38
Pseudo R square	0.018	0.078	0.009	0.050	0.165

Notes: Dependent variable is the actual effort level. We control for age, gender, race and income. Robust Standard errors are in parentheses. $*p < 0.10, **p < 0.05, ***p < 0.01$

3.5.2 Real effort

Result 2: The employee reciprocated to the employer by correctly solving more questions than promised when the actual wage was greater than or equal to the claimed wage. When the actual wage was lower than the claimed wage, there was no significant retaliation from the employee in the real-effort paradigm.

In the real-effort paradigm, the effort level was the number of correct answers the employee solved. We conducted the baseline and the “C&P" treatment with a real-effort task. Recall that we added an incentivized practice stage, and we used subjects’ performance in the practice stage as a proxy for their ability. Table 3.6 summarizes the outcomes of the contract with real effort.[14]

Table 3.6: Summary statistics of contract outcome in the real-effort paradigm.

	Obs.(f/w)	Ability(employee)	w	e	e_d	e_p	w_c	π_f	π_w	Overall Payoffs
A										
Baseline	38/38	6.66	54.92	5.89	7.05			37.89	34.92	72.81
		(0.38)	(2.92)	(0.34)	(0.30)			(2.76)	(2.92)	(2.52)
Treatment	35/33	6.76	65.14	6.88	7.20	7.27	89.39	37.74	45.00	83.08
		(0.50)	(3.43)	(0.49)	(0.28)	(0.42)	(3.67)	(3.96)	(3.63)	(3.26)
p-value (Base vs Treat)		0.968	0.018	0.096	0.669			0.539	0.026	0.046
B										
Treatment ($aw >= cw$)	8/8	7.88	85.63	8.00	7.63	6.75	69.38	29.00	65.63	94.63
		(1.46)	(8.42)	(1.12)	(0.82)	(0.88)	(5.38)	(7.03)	(8.42)	(4.19)
Treatment ($aw < cw$)	27/25	6.40	59.07	6.52	7.07	7.44	95.80	40.33	38.40	79.38
		(0.47)	(2.85)	(0.53)	(0.28)	(0.49)	(3.75)	(4.64)	(3.01)	(3.84)
p-value ($aw >= cw$ vs $aw < cw$)		0.213	0.007	0.196	0.748	0.495	0.003	0.312	0.007	0.019
p-value ($aw >= cw$ vs baseline)		0.095	0.002	0.013				0.212	0.002	< 0.001

Notes: p-values are from two sided Mann-Whitney U tests. In each cell, the numbers refer to mean, with standard errors in parentheses. The variables are the same as in the stated-effort paradigm, except that the effort is the number of correct answers in the real-effort paradigm. We also report the “Ability" which is the number of correct answers in the practice stage of the employee. In part B, we report these variables separately based on the difference between actual wage (aw) and claimed wage (cw) in the treatment.

In the baseline with the real-effort task, we did not find a significant correlation between the number of correct answers solved by the employee and the employer’s wage (Spearman’s test,

[14] In one session, two subjects were talking to each other during the experiment. We figured out they were friends after the experiment, and hence we removed their data. In two sessions, the participation rate was not full. To keep an even number of participants, we asked one of the graduate students working in the lab to be the last participant. The graduate student workers’ data are excluded. The final tests included 38 employers and 38 employees in the baseline and 41 employers and 39 employees in the treatment. Six employees chose not to make a “claim and promise" in the treatment; we did not report the outcome for them. Including the six subjects does not change our results regarding the overall comparison between the baseline and the “C&P" treatment.

$\rho = 0.053$, $p = 0.753$). This indicates there is no significant "gift exchange" in the baseline. The actual effort was significantly lower than the desired effort in the baseline (Wilcoxon signed-rank test, $p = 0.018$). More interestingly, the employee's actual effort was also lower than his ability (Wilcoxon signed-rank test, $p = 0.075$). This means that in the baseline, the employee exerted less effort when earning money for the employer. In other words, the employee did not try his/her best to work for the employer.

In panel A of Table 3.6, we first showed no difference in employees' ability between the baseline and the C&P treatment.[15] Then, we replicated the outcome of the *wage* level shown in the stated-effort paradigm. The employer paid a higher wage in the C&P treatment compared to the baseline. In contrast to stated-effort, the employee responded to the higher wage by solving more correct treatment answers in the real-effort paradigm. Also, there was no significant difference between the actual effort and the promised effort (Wilcoxon signed-rank test, $p = 0.602$), nor between the actual effort and the desired effort in the treatment ($p = 0.443$). The difference between the actual effort and ability in the C&P treatment was not significant ($p = 0.673$). This suggests that the employee exerted his/her full ability to work for the employer. Based on the overall comparison in the real-effort paradigm, the overall social payoffs in the C&P treatment significantly increased compared to the baseline.

In panel B of Table 3.6, we separated the subjects who chose to make a "claim and promise" in the treatment into two possible scenarios: "$aw >= cw$" and "$aw < cw$".[16] There were 8 out of 35 employers (22.9%) choosing to trust employees and pay at least the claimed wage. The trust rate was lower than in the stated-effort paradigm. Recall that, after the practice stage, the employer knew the median number of correct answers in the session, and hence the employer knew the

[15]There is no difference in employers' ability between the treatment and the baseline (Mann-Whitney *U* tests, $p = 0.375$).

[16]There are six subjects choosing not to make a "claim and promise" in the treatment. The six employees' performance in the practice stage (4.83) was less than or equal to the median performance. The difference in the ability between employees who chose not to make a "claim and promise" and other employees in our experiment is marginally significant (two sided Mann-Whitney *U* test, $p = 0.088$). When the employee chose not to make a "claim and promise," the employer paid a lower wage. Consequently, the employee offered a lower effort, and the overall payoffs were lower than the case with a "claim and promise." Due to the lack of power, only the difference in the overall payoffs is significant ($p = 0.059$).

employee pool's median ability. Figure 3.3 shows the difference between the actual wage and the claimed wage based on the distance between the promised effort and median performance. It suggested the trust decreases as "promised effort - median performance" increased (Spearman test, $\rho = -0.305, p = 0.074$). This, in turn, suggested the employer inferred the probability of promise-keeping/breaking by comparing the median performance and the promised effort. This explained why the trust rate was lower in the real-effort paradigm.

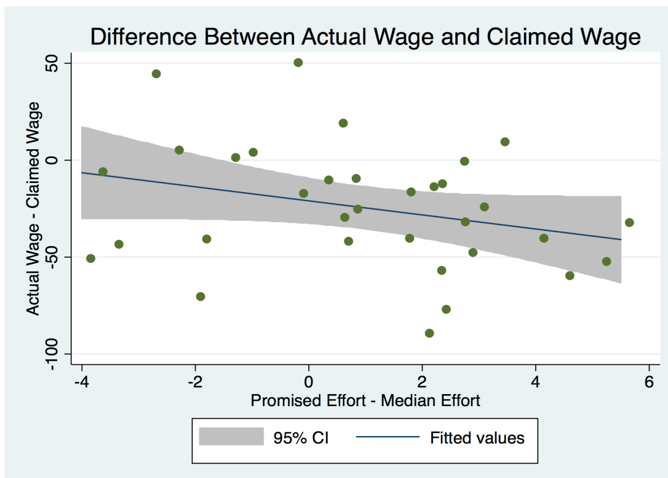
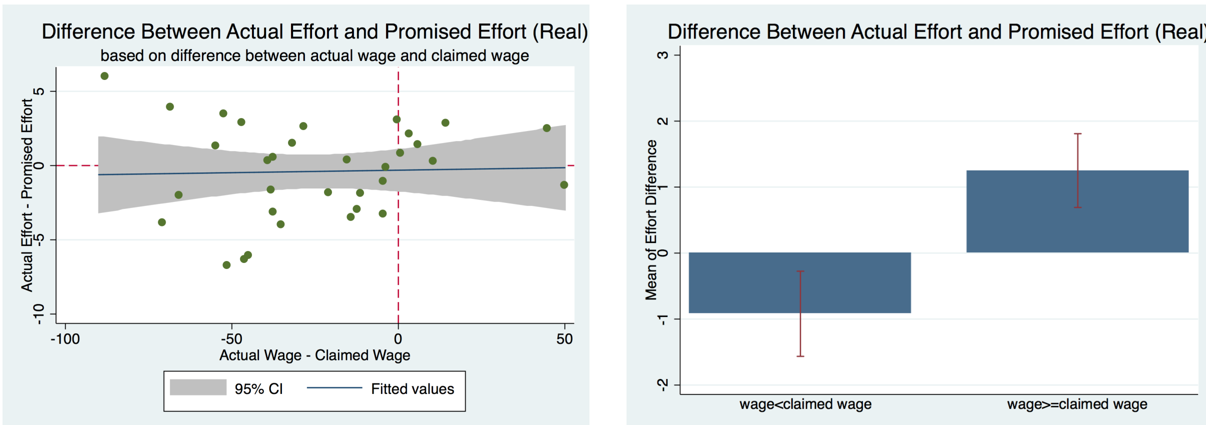


Figure 3.3: Difference Between Actual Wage & Claimed Wage

We further analyzed the employee's reaction in a similar way to the stated-effort paradigm. We compared the distance between the actual and promised effort, the distance between the desired and promised effort, and the distance between the actual and desired effort in the cases of "$aw >= cw$" and "$aw < cw$" separately. First, we found that, on average, the employee reciprocated to the employer by solving more questions correctly than s/he promised when "$aw >= cw$" (Wilcoxon signed-rank test, $p = 0.072$). In contrast to the stated-effort paradigm, in the case of "$aw < cw$", there was no difference between the actual and promised effort (Wilcoxon signed-rank test, $p = 0.184$). The correlation between (actual effort - promised effort) and (actual wage - claimed wage) in Figure 3.4a was not statistically significant (Spearman test, $\rho = -0.102, p = 0.570$), but the magnitude of the difference in Figure 3.4b was significant ($p = 0.042$). We concluded that the employee's conditional reciprocity to the employer held in the real-effort paradigm, but the negative side of reciprocity (i.e., retaliation) was not significant.



(a) Distribution

(b) Mean with Standard Error Bars

Figure 3.4: Difference Between Actual Effort and Promised Effort (Real)

In the stated-effort paradigm, we found the employer tried to negotiate with the employee in desiring different effort levels from the promised effort when he provided different wages from the claimed wages. We did not find such significant negotiation strategy in the real-effort paradigm. The correlation between "desired effort - promised effort" and "actual wage - claimed wage" in real-effort was not significant (Spearman test, $\rho = 0.132$, $p = 0.450$). The average "desired effort - promised effort" was 0.88 in the case of "$aw >= cw$", and -0.30 in the case of "$aw < cw$", but the difference in the distance was not significant (Mann-Whitney U test test, $p = 0.127$). Instead, we found employers in real-effort more likely to use *ability* information to negotiate with employees. Employers desired effort levels based on the median performance (Spearman test, $\rho = 0.485$, $p < 0.001$), and they also paid wages based on the median performance ($\rho = 0.217$, $p = 0.058$).

We further compared the actual effort with the desired effort. First, we replicated the positive reciprocity in the case of "$aw >= cw$". The employee tried to solve the desired number of correct answers when trusted. There was no significant difference between the actual and desired effort

(Wilcoxon signed-rank test, $p = 0.309$). Unlike what we found in stated-effort, there was no significant retaliation in the case of "$aw < cw$" in the real-effort paradigm, and employees tried to reach the desired effort by employers. There was no significant difference between the actual and desired effort when "$aw < cw$" (Wilcoxon signed-rank test, $p = 0.188$). The difference in the distance ("actual effort - desired effort") between these two scenarios was not significant (0.37 vs -0.55, $p = 0.162$). These results suggested that the negotiation between employers and employees based on employees' real-effort ability was more successful than the stated-effort. We elaborated on the comparison between the two effort paradigms in the following section.

We estimated an ordered probit regression of actual effort with the real-effort task. Table 3.7 displays the results. Columns (1) and (2) show that in the real-effort paradigm, compared to the baseline, the "C&P" setup significantly raised the number of correct answers after controlling for wage, employee's ability, desired effort, and socio-demographic variables. Columns (3)-(6) suggest the C&P treatment had a positive effect on effort when "$aw >= cw$", with no negative effects on effort when "$aw < cw$", compared to the baseline. In column (7), we compared the case "$aw >= cw$" to "$aw < cw$" within the treatment group. The effect of "$aw >= cw$" was marginally significant, and this was due to the result of no significant retaliation found when "$aw < cw$".

3.5.3 Summary of the Results of Stated-effort versus Real-effort

Result 3a:: Due to the consideration of ability or not, there were two major differences between the stated and real-effort paradigms. There was significant "gift-exchange" in stated-effort, but not in real-effort; the degree of retaliation from the employee in real-effort was smaller compared to stated effort when "$aw < cw$". When "$aw >= cw$", the positive reciprocity from the employee was similar in these two effort paradigms.

Result 3b: The employer desired higher effort than his stated-effort expectation, but there was no difference between desired and expected effort in the real-effort paradigm.

The behavior of stated effort versus real effort has been a controversial topic in previous experimental labor market studies. [173] ssummarized and extensively compared these two paradigms. It is much easier to control the cost function with stated effort. The real-effort paradigm has the

Table 3.7: Ordered Probit Estimations of Actual Effort (Real Effort).

	Baseline vs Treat		Baseline vs Treat ($aw >= cw$)		Baseline vs Treat ($aw < cw$)		Treat
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
Wage		-0.007		-0.015		-0.006	-0.017
		(0.008)		(0.010)		(0.009)	(0.012)
Ability		0.248***		0.396***		0.233***	0.174**
		(0.065)		(0.069)		(0.081)	(0.074)
C&P Treat	0.388	0.503**	0.874**	1.060**	0.262	0.417	
	(0.238)	(0.261)	(0.412)	(0.526)	(0.257)	(0.323)	
Desired effort		0.167**		0.243**		0.173**	0.079
		(0.075)		(0.096)		(0.081)	(0.118)
Promised effort							-0.024
							(0.134)
$AW >= CW$ (vs $aw < cw$)							0.728*
							(0.428)
Control	No	Yes	No	Yes	No	Yes	Yes
Observation	71	71	46	46	63	63	33
Pseudo R square	0.008	0.093	0.024	0.178	0.004	0.075	0.101

Notes: Dependent variable is the actual number of correct answers solved by the employee. Robust standard errors are in parentheses. $*p < 0.10, **p < 0.05, ***p < 0.01$. “Ability" is indicated by the performance of the employee in the practice stage.

confounding effects of ability; however, it is ubiquitous behavior in the real world. Few studies compare these two effort paradigms with parallel experimental treatments [174, 175]. Scholars have not reached a consensus between the two methodologies in different situations. In our paper, the treatment in the stated and real-effort paradigm was parallel, except that in real-effort, we added the practice stage, which revealed information about ability in both the baseline and “C&P" treatment. This made the difference between stated-effort and real-effort more salient in our experiment.

We first compared the baseline results without “claim and promise" between these two effort paradigms. As we showed in the main results above, we found significant “gift exchange" in stated-effort ($\rho = 0.431$, $p = 0.009$), but not in real-effort ($\rho = 0.053$, $p = 0.753$). The regression results of real-effort in Table 3.7 also suggested that wage level did not significantly affect effort level. However, the effort level strongly depended on ability. This potentially explained that the insignificant “gift exchange" in the real-effort paradigm was due to limitations on ability. Note that in the regression results (Table3.7 & Table3.5), the coefficients of wage were negative in real-effort, but there were no unexpected negative coefficients of wages in stated-effort. To some extent, these results verified the advantage of a clear cost function in the stated-effort paradigm suggested by [173].[17]

We next compared the C&P treatment between stated and real-effort. Based on our main results above, the positive reciprocity in the case of "$aw >= cw$" was similar in these two effort paradigms, but the retaliation in the case of "$aw < cw$" was different. The retaliation in stated-effort was significant, but it was not significant in real-effort. We further calculated the degree of positive reciprocity and retaliation, which were measured by the absolute distance between the actual effort and promised effort in the case of "$aw >= cw$", and the absolute distance between the actual effort and desired effort in the case of "$aw < cw$", respectively. We found no significant difference in the degree of positive reciprocity between these two effort paradigms (0.13 vs 0.06, Mann-Whitney *U* test, *p*-value=0.234), but the degree of retaliation in real-effort was significantly lower than stated-effort (0.06 vs 0.21, *p*-value=0.046).

In the real-effort paradigm, employees were aware of their ability and the session's median performance (relative ability) before they started the game. In the case of "$aw < cw$", we found the employee's promised effort was significantly higher than his ability (Wilcoxon signed-rank test, $p = 0.050$), and the employee's claimed wage was not based on his ability (Spearman test, $\rho = -0.030, p = 0.869$) nor based on relative ability ($\rho = 0.034, p = 0.830$). The employee might feel guilty about the inflated promised effort and claimed wage (see more details in the following section). Thus, when "$aw < cw$" and the employer paid wage and desired effort based on ability, the employee might choose not to retaliate. Such could be one potential explanation for the lack of retaliation in the real-effort paradigm when "$aw < cw$".

For the employer, recall that we elicited employers' incentivized beliefs about the actual effort level at the end of the experiment. In both the stated and real-effort paradigms, the wage level was significantly correlated with employers' beliefs (Spearman test, $\rho = 0.571, p < 0.001$ for baseline and $p = 0.011$ for treatment in the stated-effort; $\rho = 0.320, p = 0.051$ for baseline and $p < 0.001$ for treatment in the real-effort). The p-values in Table 3.8 show that in stated-effort, the employer desired higher effort than his belief, but there was no difference between desired effort and beliefs

[17] Since the effort level with a real-effort task depends on the provided work time duration, we cannot directly compare the effort level and the wage rate between the two effort paradigms.

in the real-effort task. In real-effort, the employer knew the median performance, and he used this information to build beliefs and desired effort. The correlation between the desired effort and the median performance was positive and significant (Spearman test, $\rho = 0.724$, $p < 0.001$). We thus argue that since there was no involvement about ability in stated-effort, employers tried to desire higher effort than his expectation.[18]

Table 3.8: Comparison between Stated-Effort and Real-Effort

	Baseline			C&P Treatment		
	Desired Effort	Belief (employer)	P-value	Desired Effort	Belief (employer)	P-value
Stated Effort	0.67 (0.03)	0.49 (0.03)	< 0.001	0.65 (0.03)	0.51 (0.04)	< 0.001
Real Effort	7.05 (0.30)	6.62 (0.30)	0.123	7.20 (0.28)	6.73 (0.30)	0.112

Notes: The p-values in the column are from Wilcoxon signed-rank tests. In each cell, the numbers refer to mean, with standard errors in parentheses.

Thus far, we concluded that our "C&P" treatment generated a similar direction of behavior (conditional reciprocity) for stated-effort and real-effort paradigms. However, the quantitative outcome was quite different. In the real-effort paradigm, due to the involvement of ability, the employer paid a wage based on the median performance of the employee, and the employee did not retaliate against the employer. In this regard, the negotiation between the employer and employee was more successful in the real-effort paradigm. In the real workplace, the employee had private information about his/her ability and his/her relative ability, and this information could influence the feeling of guilt and, subsequently, her/his effort choice. However, if the work did not require high ability, the employee would choose to retaliate when the employer did not trust his promise.

[18] We further find with a similar average effort level, the wage rate was significantly higher in stated-effort than in real-effort (Mann-Whitney U test, $p = 0.009$). Note that when we compared the effort level, we divided the effort level in the real-effort by 10. The average effort level in the stated-effort was 0.53, and it was 0.59 in the real-effort. This suggests the employer paid a wage premium for the higher desired effort.

3.6 Choice-process Data

3.6.1 Guilt

[140] suggested that promises could drive behavior because people cared about others' expectations, and they might feel guilty when their behavior let down others' expectations. [140] arrived at this conclusion using a guilt aversion model, and they tested the model using second-order beliefs. [162] measured the feeling of guilt and tested the correlation between guilt and promise-keeping using the self-reported TOSCA-3 questionnaire. In our experiment, we used desired effort as the induced-belief to include guilt in our model. We further used choice-process data from biometric tools to directly measure guilt, and we focused on pupil dilation responses [176]. We also incorporated the TOSCA-3 questionnaire at the end of the experiment. Since we only had the promise setup in the treatment, we focused on guilt for the C&P treatment. In the C&P treatment, we successfully recorded 71 subjects in stated-effort (36 employees, 35 employers) and 72 subjects in real-effort (36 employees, 36 employers).

Researchers have previously used pupil dilation responses to measure *guilt* [177, 166, 167, 178, 179]; most of them focused on the guilt in deception.[19] The deception literature suggests pupils dilate when people feel guilty. We followed this literature to assume pupils dilate when people feel guilty in our promise setup.

We first measured the pupil size of each TOSCA-3 question screen. At the end of the experiment, subjects responded to five questions from the TOSCA-3 questionnaire. The answers about guilt were on a scale from 0 to 4, with 0 indicating no guilt feelings while 4 indicating strong guilt feelings.[20] Overall, we found (combining stated and real effort), the correlation between pupil size and TOSCA-3 scores were positive but not significant (Spearman test, $\rho = 0.026$, $p = 0.498$).[21] However, such correlation in the real-effort for employees are positive and marginally significant

[19] Pupils dilate for various reasons, including emotion, arousal, cognitive difficulty, and pain [180, 181, 168, 182].

[20] We followed [162] and chose the questions related to guilt as in Question B and randomly picked one of the other questions as Question A.

[21] To calculate pupil size, we first took an average between the pupil size of subjects' left and right eye (in millimeters) on each timestamp (60 times per second), and then we calculated the average pupil size per screen (average of all timestamps spent on that screen).

(Spearman test, $\rho = 0.142$, $p = 0.062$). One possible explanation is that three of the five TOSCA-3 questions related to scenarios of employees in the workplace, and the role assignment in our experiment could have had a priming effect, especially for the “employee" in the real-effort paradigm. This result, to some extent, verified our assumption that pupils dilate when people feel guilty.[22]

We next studied pupil dilation for the main experimental decision screens. In both stated and real effort, the screen in which the employee chose his actual effort after reviewing the employer’s wage was the most important screen to study guilt feelings in our promise setup, hereafter, the “effort choice" screen. In the stated-effort paradigm, the “effort choice" screen included information about the actual wage paid by the employer, the desired effort by the employer, a review of his “claim and promise," and also the employee’s actual effort choice options. In the real-effort paradigm, the only difference from stated-effort was that subjects proceeded to the real task after reviewing the information instead of choosing the effort level. In order to obtain accurate choice-process data on each information, we first defined non-overlapping Areas of Interest (AOIs) in the “effort choice" screen (see Appendix C.2). AOI is a tool to select regions of a displayed stimulus and to extract eye-tracking metrics specifically for those regions.[23] We then calculated the average pupil dilation for all the AOIs.

In the C&P treatment, employees’ guilt feeling depended on the wage paid by the employer. Thus, we explored pupil dilation on the “effort choice" screen when “$aw >= cw$" and “$aw < cw$" separately. In stated-effort, we found that when “$aw >= cw$", the correlation between the pupil size and promise-keeping behavior (i.e., “actual effort - promised effort") was negative and marginally significant (Spearman test, $\rho = -0.231$, $p = 0.082$). We did not find such correlation when “$aw < cw$". This suggests that when the employer chose to trust the employee, the larger

[22][162] found a significant positive correlation between guilt scores and effort chosen by the employee in a one-shot gift exchange game. [162] also suggested guilt scores positively correlate with promise-keeping behavior by one-tailed Spearman tests (i.e., negative correlation between TOSCA-3 scores and “promised effort - actual effort"). However, in our experiment, we did not find such a correlation in either stated or real effort. We argue that during the experiment, participants’ feelings of guilt were changeable, and especially in our case, the feeling of guilt also depended on the wage paid by the employer.

[23]Note that we use fixation AOI instead of gaze AOI. Gaze points show on what the eyes were focusing. For example, in our experiment, eye tracker collected data with a sampling rate of 60 Hz, and we had 60 individual gaze points per millisecond. If a series of gaze points were very close -in time and/or space - this gaze cluster constituted a fixation, denoting lock of the eyes on an object for some time. Fixations were excellent measures of visual attention.

the distance between the actual and promised effort the lower the employee's guilt.

In real-effort, employers' trusting rate was low, and hence we focused on the case when "$aw < cw$". Since the employee in the real-effort paradigm proceeded to the real effort task after reviewing the information instead of choosing the effort level, we could not use the actual effort to study guilt. Instead, we used promised effort, desired effort, and ability. We found that the correlation between employees' pupil dilation on the "effort choice" screen and inflated promises (i.e., "promised effort - median performance") was positive and significant ($\rho = 0.32$, $p = 0.031$), and the correlation between pupil dilation and "desired effort - ability" was negative ($\rho = -0.26$, $p = 0.055$).[24] We argue that when the employee made a promise that was greater than his ability, and the desired effort was lower than his ability, s/he experienced guilt. This result supports our explanation about the smaller degree of retaliation in the real-effort paradigm when $aw < cw$.

3.6.2 Reciprocity

We used eye-tracking lookup patterns to provide evidence for reciprocity.[25] Lookup indicates care and attention, more care generated higher lookup rates. Based on the definition of reciprocity, where people cared for others who cared about them, we applied lookup patterns as a measurement of reciprocity. We followed [167] to use the fixation duration to present lookup patterns.

In the state-effort paradigm, we first showed the information employees cared in the "effort choice" screen. In Table 3.9, the *fixation duration* and *revisits* showed that employees cared most about the employer's wage. More interestingly, the second information about which employees cared most was the desired effort by employers, which did not affect their payoffs, and on average, employees looked up the "desired effort" for 1.09 seconds with four revisits (Appendix C.4 shows the distribution of lookup by heatmaps).[26] Regarding reciprocity, we found that in the case when

[24] The "effort choice" screen did not list "ability" and "median performance," but subjects may have had memories about the performance.

[25] [167] used lookup patterns to infer the feeling of guilt. However, [167] found high lookup generated more deception, which was inconsistent with their guilt hypothesis that "the more one cares about others' payoffs and looks at them, the less one should deceive."

[26] Following the protocol of [167], we excluded fixations shorter than 50 msec. However, these observations did not change the results. Note that similar to [167], participants in our experiment made frequent rapid fixations, and the time per lookup was low.

“$aw >= cw$", employees looked up the information from employers (i.e., “actual wage" and “desired effort") for a longer time when they received a higher wage from employers (Spearman test, $\rho = 0.468$, $p = 0.021$). We did not find a correlation in the case of “$aw < cw$" ($\rho = -0.209$, $p = 0.295$).

Table 3.9: Average Lookup Patterns (Stated Effort)

	Claimed Wage	Promised Effort	Actual Wage	Desired Effort	Actual Effort
Time to first fixation (in seconds)	16.09	24.11	16.31	23.99	15.24
Fixation duration (in seconds)	0.73	0.83	1.19	1.09	0.47
Revisits (No.)	2	3	4	4	9

Notes: Since subjects need to make choice on “Actual Effort", subjects have the most revisits on it. The order (from left to right) of the information in the table is the same as listed in the screen (from up to down).

In the real-effort paradigm, we found similar lookup patterns as in stated-effort, shown in Table C.1. Employees cared about employers' actual wage and desired effort based on the time employees look at them. The only difference is that employees cared more about their promised effort in real-effort compared to stated-effort. We calculated the “promised effort to desired effort" fixation duration ratio, and we found employees looked at the “promised effort" for a longer time in real-effort (0.96) compared to stated-effort (0.76). This result suggests that due to ability in real-effort, employees cared more about their promises. Since we had low trusting rates in real-effort, we did not find a significant correlation between lookup time durations and wages when “$aw >= cw$" nor the case of “$aw < cw$" (Spearman test, $\rho = 0.185$, $p = 0.208$).

3.7 Concluding Remarks

We conducted a laboratory experiment to study the effects of non-binding promises on labor contracts using a one-shot gift exchange game. Our treatment consisted of a “claim and promise" before the conventional gift exchange game. The employee could choose whether to make a “claim and promise" or not before the employer selected the wage level. The non-binding message of “claim and promise" was designed to provide more sequential interactions between the employer and the employee.

We tested the baseline and treatment in both stated-effort and real-effort, and we found similar qualitative results in these two effort paradigms. Our empirical analyses suggested that when the employer chose to pay a wage greater than or equal to the claimed wage, the employee reciprocated by keeping his promise, and in this case, the wage level, effort level, and final social payoff were the highest. When the employer chose a lower wage than the claimed wage, the employee retaliated and broke his promise in stated-effort. However, in real-effort, due to the provision of ability information, the retaliation was not significant. We found evidence that the C&P treatment generated more interactions between employers and employees. The employer treated the C&P treatment as a negotiation stage when paying a different wage than the claimed wage and desired a different effort level than the promised effort. In real-effort, employers were more likely to negotiate with employees based on the employees' ability.

We included reciprocity and guilt in PGT models to explain the gift exchange in our design, and we further provided direct evidence for the PGT model using choice-process data. Pupil dilation showed evidence for employees' guilt when employers trusted the employees, and they broke their promises. Eye-tracking lookup patterns suggested employees cared about the wage and the desired effort by the employer. The higher the wage received in the stated-effort paradigm, the more the employee cared about the employer's wage and desired effort. This result provided evidence for reciprocity. We also used choice-process data to suggest that one potential reason why the retaliation in real-effort was not significant was that subjects felt guilty when they over-promised (i.e., their promised effort was higher than their abilities).

Our set-up closely followed the negotiation process between employers and employees in real workplace environments. Our findings have important implications for the use of Promised-based Management to motivate employees in the workplace. For work that does not require high ability (represented by stated-effort in our experiment), employees care more about employers' trust when they make promises. For work that requires high ability (represented by real-effort), negotiation between employers and employees base on ability information boosts the efficiency in the workplace. We did not test for heterogeneous treatment effects on different subgroups. Future work might profitably focus on the promise-keeping of different groups (e.g., gender, performance) in the workplace, especially in a field labor experiment.

4. CONCLUSIONS AND FUTURE STUDY

In this dissertation, I focus on three types of economics decision-making: insurance purchasing, competition entry and employees' effort choice. Under the framework of behavioral economics, I run three different economics experiments to study these economics decision-making. The results provide important applications for the design of insurance schemes, gender equality in the labor market and employment relationship in the workplace, respectively.

In chapter 1, I compare the three insurance schemes (compulsory, voluntary and mixed insurance) in terms of both adverse selection and moral hazard. The results about mixed insurance provide valuable insights, especially for developing countries where agricultural insurance is not yet well developed and very few schemes of insurance exist that implement different forms of mixed insurance. The test of the validity of BART as an instrument to study insurance, and the design of compulsory coverage rate in mixed insurance could be interesting future studies.

In chapter 2, I show the costs of gender roles for females to enter competitions. I encourage more women, especially those with high abilities, to enter the competition by randomly assigning them a different role/title with different social norms for competitiveness. The randomly assigned role (breadwinner) gives women a permission “nudge" to break the glass ceiling of traditional gender roles for competition. The results have emphasized the importance of pushing institutional changes to help women to break the limits of traditional gender norms in the workplace. In potential future studies, I would be interested to see the effects of gender norms on women's decision in the workplace for women with different ages and from different cultures.

In chapter 3, I show the effects of promises from the employee on the employer's wage choice and the employee's effort choice. The findings have important implications for the use of Promised-based Management to motivate employees in the workplace. I allow the employee to make a promise before the employer decides the wage level. This step adds more interactions between employees and employers, and it increases the efficiency in the workplace. More field evidence of the effects of employees' promises in the workplace is needed in the future study.

REFERENCES

[1] H. Ibarra, N. M. Carter, C. Silva, *et al.*, “Why men still get more promotions than women,” *Harvard Business Review*, vol. 88, no. 9, pp. 80–85, 2010.

[2] P. Zhang and M. A. Palma, “Compulsory versus voluntary insurance: An online experiment,” *American Journal of Agricultural Economics*, vol. 103, no. 1, pp. 106–125, 2021.

[3] T. J. Pleskac, T. S. Wallsten, P. Wang, and C. Lejuez, “Development of an automatic response mode to improve the clinical utility of sequential risk-taking tasks.,” *Experimental and clinical psychopharmacology*, vol. 16, no. 6, p. 555, 2008.

[4] C. W. Lejuez, J. P. Read, C. W. Kahler, J. B. Richards, S. E. Ramsey, G. L. Stuart, D. R. Strong, and R. A. Brown, “Evaluation of a behavioral measure of risk taking: the balloon analogue risk task (bart).,” *Journal of Experimental Psychology: Applied*, vol. 8, no. 2, p. 75, 2002.

[5] G. A. Akerlof, “The market for" lemons": Quality uncertainty and the market mechanism,” *The quarterly journal of economics*, pp. 488–500, 1970.

[6] M. Rothschild and J. Stiglitz, “Equilibrium in competitive insurance markets: An essay on the economics of imperfect information,” in *Uncertainty in economics*, pp. 257–280, Elsevier, 1978.

[7] J. R. Skees and M. R. Reed, “Rate making for farm-level crop insurance: implications for adverse selection,” *American Journal of Agricultural Economics*, vol. 68, no. 3, pp. 653–659, 1986.

[8] R. E. Just, L. Calvin, and J. Quiggin, “Adverse selection in crop insurance: Actuarial and asymmetric information incentives,” *American Journal of Agricultural Economics*, vol. 81, no. 4, pp. 834–849, 1999.

[9] J. D. Woodard, A. D. Pavlista, G. D. Schnitkey, P. A. Burgener, and K. A. Ward, "Government insurance program design, incentive effects, and technology adoption: the case of skip-row crop insurance," *American Journal of Agricultural Economics*, vol. 94, no. 4, pp. 823–837, 2012.

[10] P.-A. Chiappori and B. Salanie, "Testing for asymmetric information in insurance markets," *Journal of political Economy*, vol. 108, no. 1, pp. 56–78, 2000.

[11] A. Finkelstein and K. McGarry, "Multiple dimensions of private information: evidence from the long-term care insurance market," *American Economic Review*, vol. 96, no. 4, pp. 938–958, 2006.

[12] D. M. Cutler, A. Finkelstein, and K. McGarry, "Preference heterogeneity and insurance markets: Explaining a puzzle of insurance," *American Economic Review*, vol. 98, no. 2, pp. 157–62, 2008.

[13] P. Hudson, W. W. Botzen, J. Czajkowski, and H. Kreibich, "Risk selection and moral hazard in natural disaster insurance markets: Empirical evidence from germany and the united states," *Philadelphia: Risk Management and Decision Processes Center, The Wharton School, University of Pennsylvania*, 2014.

[14] A. Cohen and P. Siegelman, "Testing for adverse selection in insurance markets," *Journal of Risk and insurance*, vol. 77, no. 1, pp. 39–84, 2010.

[15] D. Hemenway *et al.*, "Propitious selection," *Quarterly Journal of Economics*, vol. 105, no. 4, pp. 1063–1069, 1990.

[16] D. De Meza and D. C. Webb, "Advantageous selection in insurance markets," *RAND Journal of Economics*, pp. 249–262, 2001.

[17] P. Siegelman, "Adverse selection in insurance markets: an exaggerated threat," *Yale LJ*, vol. 113, p. 1223, 2003.

[18] P. De Donder and J. Hindriks, "Adverse selection, moral hazard and propitious selection," *Journal of Risk and Uncertainty*, vol. 38, no. 1, pp. 73–86, 2009.

[19] T. Karagyozova and P. Siegelman, “Can propitious selection stabilize insurance markets?,” *Journal of Insurance Issues*, pp. 121–158, 2012.

[20] A. M. Mobarak and M. R. Rosenzweig, “Informal risk sharing, index insurance, and risk taking in developing countries,” *American Economic Review*, vol. 103, no. 3, pp. 375–80, 2013.

[21] K. J. Arrow, “Uncertainty and the welfare economics of medical care,” *American Economic Review*, vol. L111, no. 5, pp. 941–973, 1963.

[22] J. K. Horowitz and E. Lichtenberg, “Insurance, moral hazard, and chemical use in agriculture,” *American Journal of Agricultural Economics*, vol. 75, no. 4, pp. 926–935, 1993.

[23] V. H. Smith and B. K. Goodwin, “Crop insurance, moral hazard, and agricultural chemical use,” *American Journal of Agricultural Economics*, vol. 78, no. 2, pp. 428–438, 1996.

[24] K. H. Coble, T. O. Knight, R. D. Pope, and J. R. Williams, “An expected-indemnity approach to the measurement of moral hazard in crop insurance,” *American journal of agricultural economics*, vol. 79, no. 1, pp. 216–226, 1997.

[25] C. G. Turvey, M. Hoy, and Z. Islam, “The role of ex ante regulations in addressing problems of moral hazard in agricultural insurance,” *Agricultural Finance Review*, 2002.

[26] M. J. Roberts, N. Key, and E. O’Donoghue, “Estimating the extent of moral hazard in crop insurance using administrative data,” *Review of Agricultural Economics*, vol. 28, no. 3, pp. 381–390, 2006.

[27] J. He, X. Zheng, R. M. Rejesus, and J. M. Yorobe Jr, “Moral hazard and adverse selection effects of cost-of-production crop insurance: evidence from the philippines,” *Australian Journal of Agricultural and Resource Economics*, vol. 63, no. 1, pp. 166–197, 2019.

[28] M. V. Pauly, “The economics of moral hazard: comment,” *The American Economic Review*, vol. 58, no. 3, pp. 531–537, 1968.

[29] C. Sapelli and B. Vial, "Self-selection and moral hazard in chilean health insurance," *Journal of health economics*, vol. 22, no. 3, pp. 459–476, 2003.

[30] P. J. Schoemaker and H. C. Kunreuther, "An experimental study of insurance decisions," *Journal of Risk and Insurance*, pp. 603–618, 1979.

[31] H. Kunreuther and M. Pauly, "Role of deliberative thinking and emotions in insurance decisionmaking: An experimental study," tech. rep., Working Paper 2014-03. Risk Management and Decision Processes Center, Wharton School, University of Pennsylvania, 2014.

[32] H. Cai, Y. Chen, H. Fang, and L.-A. Zhou, "The effect of microinsurance on economic activities: evidence from a randomized field experiment," *Review of Economics and Statistics*, vol. 97, no. 2, pp. 287–300, 2015.

[33] D. Karlan, E. Kutsoati, M. McMillan, and C. Udry, "Crop price indemnified loans for farmers: A pilot experiment in rural ghana," *Journal of Risk and Insurance*, vol. 78, no. 1, pp. 37–55, 2011.

[34] R. V. Hill and A. Viceisza, "A field experiment on the impact of weather shocks and insurance on risky investment," *Experimental Economics*, vol. 15, no. 2, pp. 341–371, 2012.

[35] S. Gunnsteinsson, "Experimental identification of asymmetric information: Evidence on crop insurance in the philippines," *Journal of Development Economics*, vol. 144, p. 102414, 2020.

[36] M. E. Yaari, "The dual theory of choice under risk," *Econometrica: Journal of the Econometric Society*, pp. 95–115, 1987.

[37] R. R. Officer and A. N. Halter, "Utility analysis in a practical setting," *American Journal of Agricultural Economics*, vol. 50, no. 2, pp. 257–277, 1968.

[38] J. L. Dillon and P. L. Scandizzo, "Risk attitudes of subsistence farmers in northeast brazil: A sampling approach," *American Journal of Agricultural Economics*, vol. 60, no. 3, pp. 425–435, 1978.

[39] H. P. Binswanger, “Attitudes toward risk: Experimental measurement in rural india,” *American journal of agricultural economics*, vol. 62, no. 3, pp. 395–407, 1980.

[40] C. A. Holt and S. K. Laury, “Risk aversion and incentive effects,” *American economic review*, vol. 92, no. 5, pp. 1644–1655, 2002.

[41] C. C. Eckel and P. J. Grossman, “Forecasting risk attitudes: An experimental study using actual and forecast gamble choices,” *Journal of Economic Behavior & Organization*, vol. 68, no. 1, pp. 1–17, 2008.

[42] P. Crosetto and A. Filippin, “A theoretical and experimental appraisal of four risk elicitation methods,” *Experimental Economics*, vol. 19, no. 3, pp. 613–641, 2016.

[43] K. Fairley, J. M. Parelman, M. Jones, and R. M. Carter, “Risky health choices and the balloon economic risk protocol,” *Journal of Economic Psychology*, vol. 73, pp. 15–33, 2019.

[44] P. Crosetto and A. Filippin, “The “bomb" risk elicitation task,” *Journal of Risk and Uncertainty*, vol. 47, no. 1, pp. 31–65, 2013.

[45] G. Charness, U. Gneezy, and A. Imas, “Experimental methods: Eliciting risk preferences,” *Journal of Economic Behavior & Organization*, vol. 87, pp. 43–51, 2013.

[46] M. Buhrmester, T. Kwang, and S. D. Gosling, “Amazon’s mechanical turk: A new source of inexpensive, yet high-quality, data?,” *Perspectives on psychological science*, vol. 6, no. 1, pp. 3–5, 2011.

[47] D. Johnson and J. Ryan, “Amazon mechanical turk workers can provide consistent and economically meaningful data,” 2018.

[48] A.-R. Blais and E. U. Weber, “A domain-specific risk-taking (dospert) scale for adult populations,” *Judgment and Decision making*, vol. 1, no. 1, 2006.

[49] M. Zuckerman, E. A. Kolin, L. Price, and I. Zoob, “Development of a sensation-seeking scale.,” *Journal of consulting psychology*, vol. 28, no. 6, p. 477, 1964.

[50] Inquisit, “Computer software,” *Retrieved from http://www.millisecond.com.*, 2016.

[51] E. Briys, Y. Kahane, and Y. Kroll, "Voluntary insurance coverage, compulsory insurance, and risky-riskless portfolio opportunities," *Journal of Risk and Insurance*, pp. 713–722, 1988.

[52] J. Hindriks, "Public versus private insurance with dual theory: a political economy argument," *The GENEVA Papers on Risk and Insurance-Theory*, vol. 26, no. 3, pp. 225–241, 2001.

[53] R. Chetty and E. Saez, "Optimal taxation and social insurance with endogenous private insurance," *American Economic Journal: Economic Policy*, vol. 2, no. 2, pp. 85–116, 2010.

[54] I. Ehrlich and G. S. Becker, "Market insurance, self-insurance, and self-protection," *Journal of political Economy*, vol. 80, no. 4, pp. 623–648, 1972.

[55] A. Finkelstein, "The interaction of partial public insurance programs and residual private insurance markets: evidence from the us medicare program," *Journal of Health Economics*, vol. 23, no. 1, pp. 1–24, 2004.

[56] J. R. Brown and A. Finkelstein, "The interaction of public and private insurance: Medicaid and the long-term care insurance market," *The American Economic Review*, vol. 98, no. 3, pp. 1083–1102, 2008.

[57] K. Sakai, M. Okura, *et al.*, "An economic analysis of compulsory and voluntary insurance," *International Journal of Academic Research in Accounting, Finance and Management Sciences*, vol. 2, no. 2, pp. 1–8, 2012.

[58] P. Crosetto and A. Filippin, "A theoretical and experimental appraisal of five risk elicitation methods. jena economic research paper 2013-009," 2013.

[59] A. Reynaud and S. Couture, "Stability of risk preference measures: results from a field experiment on french farmers," *Theory and decision*, vol. 73, no. 2, pp. 203–221, 2012.

[60] J. P. Beauchamp, D. Cesarini, and M. Johannesson, "The psychometric and empirical properties of measures of risk preferences," *Journal of Risk and uncertainty*, vol. 54, no. 3, pp. 203–237, 2017.

[61] O. Mahul and C. J. Stutley, *Government support to agricultural insurance: challenges and options for developing countries*. The World Bank, 2010.

[62] B. G. Essex, C. W. Lejuez, R. Y. Qian, K. Bernstein, and D. H. Zald, "The balloon analog insurance task (bait): A behavioral measure of protective risk management," *PloS one*, vol. 6, no. 6, 2011.

[63] W. E. Forum, "The global gender gap report," World Economic Forum Geneva, 2020.

[64] M. Niederle and L. Vesterlund, "Do women shy away from competition? do men compete too much?," *The quarterly journal of economics*, vol. 122, no. 3, pp. 1067–1101, 2007.

[65] M. Niederle and L. Vesterlund, "Gender and competition," *Annu. Rev. Econ.*, vol. 3, no. 1, pp. 601–630, 2011.

[66] T. Buser, M. Niederle, and H. Oosterbeek, "Gender, competitiveness, and career choices," *The Quarterly Journal of Economics*, vol. 129, no. 3, pp. 1409–1447, 2014.

[67] U. Gneezy, K. L. Leonard, and J. A. List, "Gender differences in competition: Evidence from a matrilineal and a patriarchal society," *Econometrica*, vol. 77, no. 5, pp. 1637–1664, 2009.

[68] A. L. Booth and P. Nolen, "Gender differences in competition: the role of single-sex education," tech. rep., CEPR Discussion Paper 7214, 2009.

[69] R. G. Fryer Jr and S. D. Levitt, "An empirical analysis of the gender gap in mathematics," *American Economic Journal: Applied Economics*, vol. 2, no. 2, pp. 210–40, 2010.

[70] D. B. Kandel, M. Davies, and V. H. Raveis, "The stressfulness of daily social roles for women: Marital, occupational and household roles," *Journal of health and social behavior*, pp. 64–78, 1985.

[71] J. Elster, "Social norms and economic theory," *Journal of economic perspectives*, vol. 3, no. 4, pp. 99–117, 1989.

[72] A. H. Eagly and S. J. Karau, "Role congruity theory of prejudice toward female leaders.," *Psychological review*, vol. 109, no. 3, p. 573, 2002.

[73] D. Heath and C. Heath, *Switch: How to change things when change is hard*. Random House, 2011.

[74] J. N. Cleveland and L. A. Smith, "The effects of job title and task composition on job and incumbent perceptions 1," *Journal of Applied Social Psychology*, vol. 19, no. 9, pp. 744–757, 1989.

[75] J. C. Jessell and L. Beymer, "The effects of job title vs. job description on occupational sex typing," *Sex Roles*, vol. 27, no. 1-2, pp. 73–83, 1992.

[76] R. Pande and D. Ford, "Gender quotas and female leadership," 2012.

[77] D. Dahlerup, *Women, quotas and politics*. Routledge, 2013.

[78] M. L. Krook, "Gender quotas, norms, and politics," *Politics & Gender*, vol. 2, no. 1, p. 110, 2006.

[79] J. de Quidt, L. Vesterlund, and A. J. Wilson, "Experimenter demand effects," in *Handbook of Research Methods and Applications in Experimental Economics* (A. Schram and A. Ule, eds.), ch. 20, pp. 384–399, Edward Elgar Pub, 2019.

[80] D. J. Zizzo, "Experimenter demand effects in economic experiments," *Experimental Economics*, vol. 13, no. 1, pp. 75–98, 2010.

[81] G. Charness, U. Gneezy, and M. A. Kuhn, "Experimental methods: Between-subject and within-subject design," *Journal of Economic Behavior & Organization*, vol. 81, no. 1, pp. 1–8, 2012.

[82] M. Niederle, C. Segal, and L. Vesterlund, "How costly is diversity? affirmative action in light of gender differences in competitiveness," *Management Science*, vol. 59, no. 1, pp. 1–16, 2013.

[83] M. Sutter and D. Rützler, "Gender differences in competition emerge early in life," 2010.

[84] N. Datta Gupta, A. Poulsen, and M. C. Villeval, "Gender matching and competitiveness: Experimental evidence," *Economic Inquiry*, vol. 51, no. 1, pp. 816–835, 2013.

[85] N. D. Große and G. Riener, "Explaining gender differences in competitiveness: Gender-task stereotypes," tech. rep., Jena economic research papers, 2010.

[86] C. L. Apicella, E. E. Demiral, and J. Mollerstrom, "No gender difference in willingness to compete when competing against self," *American Economic Review*, vol. 107, no. 5, pp. 136–40, 2017.

[87] N. R. Baldiga and K. B. Coffman, "Laboratory evidence on the effects of sponsorship on the competitive preferences of men and women," *Management Science*, vol. 64, no. 2, pp. 888–901, 2018.

[88] A. Booth, E. Fan, X. Meng, and D. Zhang, "Gender differences in willingness to compete: The role of culture and institutions," *The Economic Journal*, vol. 129, no. 618, pp. 734–764, 2019.

[89] M.-P. Dargnies, "Social identity and competitiveness," *Available at SSRN 1815003*, 2011.

[90] C. B. Cadsby, M. Servátka, and F. Song, "How competitive are female professionals? a tale of identity conflict," *Journal of Economic Behavior & Organization*, vol. 92, pp. 284–303, 2013.

[91] L. Balafoutas, H. Fornwagner, and M. Sutter, "Closing the gender gap in competitiveness through priming," *Nature communications*, vol. 9, no. 1, p. 4359, 2018.

[92] G. A. Akerlof and R. E. Kranton, "Economics and identity," *The Quarterly Journal of Economics*, vol. 115, no. 3, pp. 715–753, 2000.

[93] A. Hochschild and A. Machung, *The second shift: Working families and the revolution at home*. Penguin, 2012.

[94] L. A. Rudman, "Self-promotion as a risk factor for women: the costs and benefits of counterstereotypical impression management.," *Journal of personality and social psychology*, vol. 74, no. 3, p. 629, 1998.

[95] A. M. Blackstone, "Gender roles and society," *Human Ecology: An Encyclopedia of Children, Families, Communities, and Environments*, pp. 335–338, 2003.

[96] A. Oakley, *Sex, gender and society*. Routledge, 2016.

[97] M. Bertrand, E. Kamenica, and J. Pan, "Gender identity and relative income within households," *The Quarterly Journal of Economics*, vol. 130, no. 2, pp. 571–614, 2015.

[98] F. Cochard, H. Couprie, and A. Hopfensitz, "What if women earned more than their spouses? an experimental investigation of work-division in couples," *Experimental Economics*, vol. 21, no. 1, pp. 50–71, 2018.

[99] R. Fisman, S. S. Iyengar, E. Kamenica, and I. Simonson, "Gender differences in mate selection: Evidence from a speed dating experiment," *The Quarterly Journal of Economics*, vol. 121, no. 2, pp. 673–697, 2006.

[100] S. L. Brown and B. P. Lewis, "Relational dominance and mate-selection criteria: Evidence that males attend to female dominance," *Evolution and Human Behavior*, vol. 25, no. 6, pp. 406–415, 2004.

[101] T. Greitemeyer, "What do men and women want in a partner? are educated partners always more desirable?," *Journal of Experimental Social Psychology*, vol. 43, no. 2, pp. 180–194, 2007.

[102] G. J. Hitsch, A. Hortaçsu, and D. Ariely, "Matching and sorting in online dating," *American Economic Review*, vol. 100, no. 1, pp. 130–63, 2010.

[103] O. Folke and J. Rickne, "All the single ladies: Job promotions and the durability of marriage," *American Economic Journal: Applied Economics*, vol. 12, no. 1, pp. 260–87, 2020.

[104] L. Bursztyn, T. Fujiwara, and A. Pallais, "'acting wife': Marriage market incentives and labor market investments," *American Economic Review*, vol. 107, no. 11, pp. 3288–3319, 2017.

[105] L. Babcock, M. P. Recalde, L. Vesterlund, and L. Weingart, "Gender differences in accepting and receiving requests for tasks with low promotability," *American Economic Review*, vol. 107, no. 3, pp. 714–47, 2017.

[106] Z. Salikutluk and S. Heyne, "Do gender roles and norms affect performance in maths? the impact of adolescents' and their peers' gender conceptions on maths grades," *European Sociological Review*, vol. 33, no. 3, pp. 368–381, 2017.

[107] A. H. Eagly and M. E. Kite, "Are stereotypes of nationalities applied to both women and men?," *Journal of Personality and Social Psychology*, vol. 53, no. 3, p. 451, 1987.

[108] U. Fischbacher, "z-tree: Zurich toolbox for ready-made economic experiments," *Experimental economics*, vol. 10, no. 2, pp. 171–178, 2007.

[109] M.-P. Dargnies, "Men too sometimes shy away from competition: The case of team competition," *Management Science*, vol. 58, no. 11, pp. 1982–2000, 2012.

[110] A. Healy and J. Pate, "Can teams help to close the gender competition gap?," *The Economic Journal*, vol. 121, no. 555, pp. 1192–1204, 2011.

[111] C. A. Holt and S. K. Laury, "Risk aversion and incentive effects: New data without order effects," *American Economic Review*, vol. 95, no. 3, pp. 902–912, 2005.

[112] E. L. Krupka and R. A. Weber, "Identifying social norms using coordination games: Why does dictator game sharing vary?," *Journal of the European Economic Association*, vol. 11, no. 3, pp. 495–524, 2013.

[113] R. Petrie and C. Segal, "Gender differences in competitiveness: The role of prizes," *GMU working paper in economics*, 2015.

[114] J. Ifcher and H. Zarghamee, “Pricing competition: a new laboratory measure of gender differences in the willingness to compete,” *Experimental Economics*, vol. 19, no. 3, pp. 642–662, 2016.

[115] S. Saccardo, A. Pietrasz, and U. Gneezy, “On the size of the gender difference in competitiveness,” *Management Science*, vol. 64, no. 4, pp. 1541–1554, 2017.

[116] D. J. Bennett, *Randomness*. Harvard University Press, 2009.

[117] K. K. Li, “Preference towards control in risk taking: Control, no control, or randomize?,” *Journal of Risk and Uncertainty*, vol. 43, no. 1, pp. 39–63, 2011.

[118] N. Dwenger, D. Kübler, and G. Weizsäcker, “Preference for randomization: Empirical and experimental evidence,” tech. rep., SFB 649 Discussion Paper, 2013.

[119] F. J. Anscombe, R. J. Aumann, *et al.*, “A definition of subjective probability,” *Annals of mathematical statistics*, vol. 34, no. 1, pp. 199–205, 1963.

[120] J. Eichberger and D. Kelsey, “Uncertainty aversion and preference for randomisation,” *Journal of Economic Theory*, vol. 71, no. 1, pp. 31–43, 1996.

[121] P. Klibanoff, “Stochastically independent randomization and uncertainty aversion,” *Economic Theory*, vol. 18, no. 3, pp. 605–620, 2001.

[122] K. Saito, “Preferences for flexibility and randomization under uncertainty,” *American Economic Review*, vol. 105, no. 3, pp. 1246–71, 2015.

[123] C. C. Eckel and P. J. Grossman, “Men, women and risk aversion: Experimental evidence,” *Handbook of experimental economics results*, vol. 1, pp. 1061–1073, 2008.

[124] U. Gneezy, M. Niederle, and A. Rustichini, “Performance in competitive environments: Gender differences,” *The Quarterly Journal of Economics*, vol. 118, no. 3, pp. 1049–1074, 2003.

[125] J. A. List, A. M. Shaikh, and Y. Xu, “Multiple hypothesis testing in experimental economics,” *Experimental Economics*, pp. 1–21, 2016.

[126] B. Gillen, E. Snowberg, and L. Yariv, “Experimenting with measurement error: Techniques with applications to the caltech cohort study,” *Journal of Political Economy*, vol. 127, no. 4, pp. 1826–1863, 2019.

[127] D. L. Chen, M. Schonger, and C. Wickens, “otree-an open-source platform for laboratory, online, and field experiments,” *Journal of Behavioral and Experimental Finance*, vol. 9, pp. 88–97, 2016.

[128] G. A. Akerlof, “A theory of social custom, of which unemployment may be one consequence,” *The quarterly journal of economics*, vol. 94, no. 4, pp. 749–775, 1980.

[129] S. Markos and M. S. Sridevi, “Employee engagement: The key to improving performance,” *International journal of business and management*, vol. 5, no. 12, p. 89, 2010.

[130] V. Kumar and A. Pansari, “Measuring the benefits of employee engagement,” *MIT Sloan Management Review*, vol. 56, no. 4, p. 67, 2015.

[131] D. Sull and C. Spinosa, “Promise-based management: The essence of execution,” *Harvard business review*, vol. 85, pp. 78–86, 141, 05 2007.

[132] G. A. Akerlof, “Labor contracts as partial gift exchange,” *The quarterly journal of economics*, vol. 97, no. 4, pp. 543–569, 1982.

[133] G. A. Akerlof and J. L. Yellen, “The fair wage-effort hypothesis and unemployment,” *The Quarterly Journal of Economics*, vol. 105, no. 2, pp. 255–283, 1990.

[134] E. Fehr, U. Fischbacher, and E. Tougareva, “Do high stakes and competition undermine fairness? evidence from russia,” *Evidence from Russia (July 2002)*, 2002.

[135] U. Gneezy and J. A. List, “Putting behavioral economics to work: Testing for gift exchange in labor markets using field experiments,” *Econometrica*, vol. 74, no. 5, pp. 1365–1384, 2006.

[136] E. Fehr, G. Kirchsteiger, and A. Riedl, "Does fairness prevent market clearing? an experimental investigation," *The quarterly journal of economics*, vol. 108, no. 2, pp. 437–459, 1993.

[137] A. Falk and M. Kosfeld, "The hidden costs of control," *American Economic Review*, vol. 96, no. 5, pp. 1611–1630, 2006.

[138] G. Charness, R. Cobo-Reyes, N. Jiménez, J. A. Lacomba, and F. Lagos, "The hidden advantage of delegation: Pareto improvements in a gift exchange game," *American Economic Review*, vol. 102, no. 5, pp. 2358–79, 2012.

[139] G. Charness, R. Cobo-Reyes, J. A. Lacomba, F. Lagos, and J. M. Pérez, "Social comparisons in wage delegation: Experimental evidence," *Experimental Economics*, vol. 19, no. 2, pp. 433–459, 2016.

[140] G. Charness and M. Dufwenberg, "Promises and partnership," *Econometrica*, vol. 74, no. 6, pp. 1579–1601, 2006.

[141] P. Battigalli and M. Dufwenberg, "Guilt in games," *American Economic Review*, vol. 97, no. 2, pp. 170–176, 2007.

[142] A. Beck, R. Kerschbamer, J. Qiu, and M. Sutter, "Shaping beliefs in experimental markets for expert services: Guilt aversion and the impact of promises and money-burning options," *Games and Economic Behavior*, vol. 81, pp. 145–164, 2013.

[143] C. Vanberg, "Why do people keep their promises? an experimental test of two explanations 1," *Econometrica*, vol. 76, no. 6, pp. 1467–1480, 2008.

[144] G. Charness and M. Dufwenberg, "Bare promises: An experiment," *Economics letters*, vol. 107, no. 2, pp. 281–283, 2010.

[145] G. Di Bartolomeo, M. Dufwenberg, S. Papa, and F. Passarelli, "Promises, expectations & causation," *Games and Economic Behavior*, vol. 113, pp. 137–146, 2019.

[146] P. Battigalli and M. Dufwenberg, "Belief-dependent motivations and psychological game theory," 2020.

[147] M. Dufwenberg, S. Gächter, and H. Hennig-Schmidt, "The framing of games and the psychology of play," *Games and Economic Behavior*, vol. 73, no. 2, pp. 459–478, 2011.

[148] K. Khalmetski, A. Ockenfels, and P. Werner, "Surprising gifts: Theory and laboratory evidence," *Journal of Economic Theory*, vol. 159, pp. 163–208, 2015.

[149] C. Bellemare, A. Sebald, and S. Suetens, "A note on testing guilt aversion," *Games and Economic Behavior*, vol. 102, pp. 233–239, 2017.

[150] S. Dhami, M. Wei, and A. al Nowaihi, "Public goods games and psychological utility: Theory and evidence," *Journal of Economic Behavior & Organization*, vol. 167, pp. 361–390, 2019.

[151] S. Dhami, M. Wei, and A. al Nowaihi, "Gift exchange models: Theory and evidence," *working paper*, 2020.

[152] M. Dufwenberg and G. Kirchsteiger, "A theory of sequential reciprocity," *Games and economic behavior*, vol. 47, no. 2, pp. 268–298, 2004.

[153] U. Malmendier and K. M. Schmidt, "You owe me," *American Economic Review*, vol. 107, no. 2, pp. 493–526, 2017.

[154] J. D. Carrillo and M. Dewatripont, "Promises, promises...," *The Economic Journal*, vol. 118, no. 531, pp. 1453–1473, 2008.

[155] C. Deck, M. Servátka, and S. Tucker, "An examination of the effect of messages on cooperation under double-blind and single-blind payoff procedures," *Experimental Economics*, vol. 16, no. 4, pp. 597–607, 2013.

[156] F. Ederer and A. Stremitzer, "Promises and expectations," *Games and Economic Behavior*, vol. 106, pp. 161–178, 2017.

[157] M. Dufwenberg, F. Li, and A. Smith, "Promises and punishment," *Unpublished manuscript*, 2018.

[158] S. Schwartz, E. Spires, and R. Young, "Why do people keep their promises? a further investigation," *Experimental Economics*, vol. 22, no. 2, pp. 530–551, 2019.

[159] J. Chen and D. Houser, "Broken promises and hidden partnerships: An experiment," *Journal of Economic Behavior & Organization*, vol. 157, pp. 754–774, 2019.

[160] N. Du and Q. Shahriar, "Cheap-talk evaluations in contract design," *Journal of behavioral and experimental economics*, vol. 77, pp. 78–87, 2018.

[161] V. Faralla, G. Borà, A. Innocenti, and M. Novarese, "Promises in group decision making," *Research in Economics*, 2019.

[162] G. Charness, N. Du, C.-L. Yang, and L. Yao, "Promises in contract design," *European Economic Review*, vol. 64, pp. 194–208, 2013.

[163] J. Tangney, R. Dearing, P. Wagner, and R. Gramzow, "The test of self-conscious affect-3 (tosca-3): George mason university," *Fairfax, VA*, 2000.

[164] P. J. Healy, "Group reputations, stereotypes, and cooperation in a repeated labor market," *American Economic Review*, vol. 97, no. 5, pp. 1751–1773, 2007.

[165] G. Charness and M. O. Jackson, "The role of responsibility in strategic risk-taking," *Journal of Economic Behavior & Organization*, vol. 69, no. 3, pp. 241–247, 2009.

[166] A. K. Webb, C. R. Honts, J. C. Kircher, P. Bernhardt, and A. E. Cook, "Effectiveness of pupil diameter in a probable-lie comparison question test for deception," *Legal and Criminological Psychology*, vol. 14, no. 2, pp. 279–292, 2009.

[167] J. T.-y. Wang, M. Spezio, and C. F. Camerer, "Pinocchio's pupil: using eyetracking and pupil dilation to understand truth telling and deception in sender-receiver games," *American Economic Review*, vol. 100, no. 3, pp. 984–1007, 2010.

[168] M. M. Bradley, L. Miccoli, M. A. Escrig, and P. J. Lang, "The pupil as a measure of emotional arousal and autonomic activation," *Psychophysiology*, vol. 45, no. 4, pp. 602–607, 2008.

[169] M. Rabin, "Incorporating fairness into game theory and economics," *The American economic review*, pp. 1281–1302, 1993.

[170] K. Khalmetski, "Testing guilt aversion with an exogenous shift in beliefs," *Games and Economic Behavior*, vol. 97, pp. 110–119, 2016.

[171] E. Cartwright, "A survey of belief-based guilt aversion in trust and dictator games," *Journal of Economic Behavior & Organization*, vol. 167, pp. 430–444, 2019.

[172] T. Ellingsen, M. Johannesson, S. Tjøtta, and G. Torsvik, "Testing guilt aversion," *Games and Economic Behavior*, vol. 68, no. 1, pp. 95–107, 2010.

[173] G. Charness, U. Gneezy, and A. Henderson, "Experimental methods: Measuring effort in economics experiments," *Journal of Economic Behavior & Organization*, vol. 149, pp. 74–87, 2018.

[174] A. Brüggen and M. Strobel, "Real effort versus chosen effort in experiments," *Economics Letters*, vol. 96, no. 2, pp. 232–236, 2007.

[175] G. Dutcher, T. Salmon, and K. J. Saral, "Is real effort more real?," *Available at SSRN 2701793*, 2015.

[176] J. T.-y. Wang, "Pupil dilation and eye tracking," *A handbook of process tracing methods for decision research: A critical review and user's guide*, pp. 185–204, 2011.

[177] M. Bradley and M. P. Janisse, "Accuracy demonstrations, threat, and the detection of deception: Cardiovascular, electrodermal, and pupillary measures," *Psychophysiology*, vol. 18, no. 3, pp. 307–315, 1981.

[178] G. Hochman, A. Glöckner, S. Fiedler, and S. Ayal, "“i can see it in your eyes": Biased processing and increased arousal in dishonest responses,” *Journal of Behavioral Decision Making*, vol. 29, no. 2-3, pp. 322–335, 2016.

[179] R. Hepach, A. Vaish, and M. Tomasello, “Children’s intrinsic motivation to provide help themselves after accidentally harming others,” *Child development*, vol. 88, no. 4, pp. 1251–1264, 2017.

[180] E. H. Hess, “Attitude and pupil size,” *Scientific american*, vol. 212, no. 4, pp. 46–55, 1965.

[181] J. Beatty, “Task-evoked pupillary responses, processing load, and the structure of processing resources.,” *Psychological bulletin*, vol. 91, no. 2, p. 276, 1982.

[182] M. Schulte-Mecklenbeck, A. Kühberger, and J. G. Johnson, *A handbook of process tracing methods for decision research: A critical review and user’s guide*. Psychology Press, 2011.

APPENDIX A

A.1 The Existence of A Separating Equilibrium

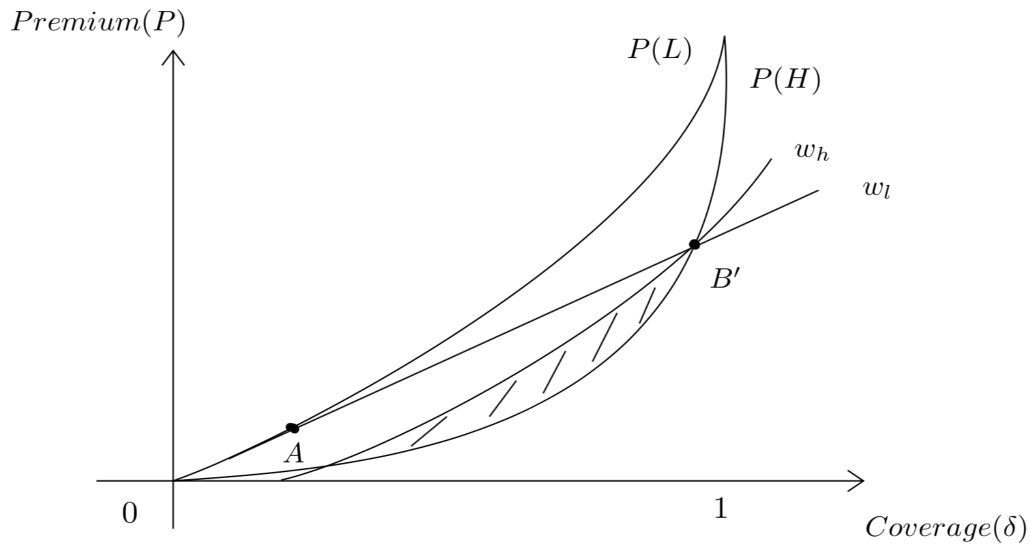


Figure A.1: Existence of a separating equilibrium of insurance contract. (Reprinted from [2])

In the separating equilibrium, point A represents the insurance contract for less risk-averse individuals (L). In order to separate more risk-averse individuals (H) from less risk-averse (L) individuals, the insurer needs to over-provide insurance to more risk-averse individuals (H), since all feasible contracts preferred by more risk-averse individuals in the shaded area are also preferred by less risk-averse individuals.

A.2 Estimates of *r* for the BART, assuming CRRA $u(k) = k^r$

Assuming a constant relative risk aversion (CRRA) utility function $u(k) = k^r$, the BART allows for the estimation of the coefficient of risk aversion. The implied levels of *r* for every possible choice *k*, $k \in [1, 128]$ are shown in Appendix Table A.1.

Table A.1: Estimates of r for the BART, assuming CRRA $u(k) = k^r$ (Reprinted from [2])

K	r	**K**	r	**K**	r
1	$0 \leq r \leq 0.011$	**44**	$0.515 \leq r \leq 0.532$	**87**	$2.085 \leq r \leq 2.161$
2	$0.012 \leq r \leq 0.019$	**45**	$0.533 \leq r \leq 0.551$	**88**	$2.162 \leq r \leq 2.240$
3	$0.020 \leq r \leq 0.027$	**46**	$0.552 \leq r \leq 0.570$	**89**	$2.241 \leq r \leq 2.324$
4	$0.028 \leq r \leq 0.036$	**47**	$0.571 \leq r \leq 0.590$	**90**	$2.325 \leq r \leq 2.413$
5	$0.037 \leq r \leq 0.044$	**48**	$0.591 \leq r \leq 0.610$	**91**	$2.414 \leq r \leq 2.506$
6	$0.045 \leq r \leq 0.053$	**49**	$0.611 \leq r \leq 0.630$	**92**	$2.507 \leq r \leq 2.605$
7	$0.054 \leq r \leq 0.062$	**50**	$0.631 \leq r \leq 0.651$	**93**	$2.606 \leq r \leq 2.710$
8	$0.063 \leq r \leq 0.071$	**51**	$0.652 \leq r \leq 0.673$	**94**	$2.711 \leq r \leq 2.821$
9	$0.072 \leq r \leq 0.080$	**52**	$0.674 \leq r \leq 0.695$	**95**	$2.822 \leq r \leq 2.938$
10	$0.081 \leq r \leq 0.089$	**53**	$0.696 \leq r \leq 0.718$	**96**	$2.939 \leq r \leq 3.063$
11	$0.090 \leq r \leq 0.098$	**54**	$0.719 \leq r \leq 0.741$	**97**	$3.064 \leq r \leq 3.196$
12	$0.099 \leq r \leq 0.108$	**55**	$0.742 \leq r \leq 0.765$	**98**	$3.197 \leq r \leq 3.339$
13	$0.109 \leq r \leq 0.117$	**56**	$0.766 \leq r \leq 0.790$	**99**	$3.340 \leq r \leq 3.491$
14	$0.118 \leq r \leq 0.127$	**57**	$0.791 \leq r \leq 0.815$	**100**	$3.492 \leq r \leq 3.654$
15	$0.128 \leq r \leq 0.137$	**58**	$0.816 \leq r \leq 0.841$	**101**	$3.655 \leq r \leq 3.830$
16	$0.138 \leq r \leq 0.147$	**59**	$0.842 \leq r \leq 0.868$	**102**	$3.831 \leq r \leq 4.02$
17	$0.148 \leq r \leq 0.158$	**60**	$0.869 \leq r \leq 0.896$	**103**	$4.021 \leq r \leq 4.225$
18	$0.159 \leq r \leq 0.168$	**61**	$0.897 \leq r \leq 0.924$	**104**	$4.226 \leq r \leq 4.447$
19	$0.169 \leq r \leq 0.179$	**62**	$0.925 \leq r \leq 0.954$	**105**	$4.448 \leq r \leq 4.689$
20	$0.180 \leq r \leq 0.190$	**63**	$0.955 \leq r \leq 0.984$	**106**	$4.690 \leq r \leq 4.954$
21	$0.191 \leq r \leq 0.201$	**64**	$0.985 \leq r \leq 1.015$	**107**	$4.955 \leq r \leq 5.224$
22	$0.202 \leq r \leq 0.213$	**65**	$1.016 \leq r \leq 1.048$	**108**	$5.225 \leq r \leq 5.565$
23	$0.214 \leq r \leq 0.224$	**66**	$1.049 \leq r \leq 1.081$	**109**	$5.566 \leq r \leq 5.920$
24	$0.225 \leq r \leq 0.236$	**67**	$1.082 \leq r \leq 1.115$	**110**	$5.921 \leq r \leq 6.315$
25	$0.237 \leq r \leq 0.248$	**68**	$1.116 \leq r \leq 1.151$	**111**	$6.316 \leq r \leq 6.759$
26	$0.249 \leq r \leq 0.261$	**69**	$1.152 \leq r \leq 1.188$	**112**	$6.760 \leq r \leq 7.260$
27	$0.262 \leq r \leq 0.273$	**70**	$1.189 \leq r \leq 1.226$	**113**	$7.261 \leq r \leq 7.830$
28	$0.274 \leq r \leq 0.286$	**71**	$1.227 \leq r \leq 1.265$	**114**	$7.831 \leq r \leq 8.485$
29	$0.287 \leq r \leq 0.299$	**72**	$1.266 \leq r \leq 1.306$	**115**	$8.486 \leq r \leq 9.244$
30	$0.300 \leq r \leq 0.312$	**73**	$1.307 \leq r \leq 1.348$	**116**	$9.245 \leq r \leq 10.136$
31	$0.313 \leq r \leq 0.326$	**74**	$1.349 \leq r \leq 1.392$	**117**	$10.137 \leq r \leq 11.198$
32	$0.327 \leq r \leq 0.340$	**75**	$1.393 \leq r \leq 1.438$	**118**	$11.199 \leq r \leq 12.485$
33	$0.341 \leq r \leq 0.354$	**76**	$1.439 \leq r \leq 1.485$	**119**	$12.486 \leq r \leq 14.074$
34	$0.355 \leq r \leq 0.368$	**77**	$1.486 \leq r \leq 1.534$	**120**	$14.075 \leq r \leq 16.090$
35	$0.369 \leq r \leq 0.383$	**78**	$1.535 \leq r \leq 1.585$	**121**	$16.091 \leq r \leq 18.729$
36	$0.384 \leq r \leq 0.398$	**79**	$1.586 \leq r \leq 1.639$	**122**	$18.730 \leq r \leq 22.334$
37	$0.399 \leq r \leq 0.414$	**80**	$1.640 \leq r \leq 1.694$	**123**	$23.335 \leq r \leq 27.558$
38	$0.415 \leq r \leq 0.430$	**81**	$1.695 \leq r \leq 1.752$	**124**	$27.559 \leq r \leq 35.816$
39	$0.431 \leq r \leq 0.446$	**82**	$1.753 \leq r \leq 1.813$	**125**	$35.817 \leq r \leq 50.885$
40	$0.447 \leq r \leq 0.462$	**83**	$1.814 \leq r \leq 1.876$	**126**	$50.886 \leq r \leq 87.682$
41	$0.463 \leq r \leq 0.479$	**84**	$1.877 \leq r \leq 1.942$	**127**	$87.683 \leq r \leq 146.285$
42	$0.480 \leq r \leq 0.497$	**85**	$1.943 \leq r \leq 2.011$	**128**	$r \geq 146.286$
43	$0.498 \leq r \leq 0.514$	**86**	$2.012 \leq r \leq 2.084$		

A.3 Empirical Predictions

Based on the insurance premium and coverage in our experiment and the explosion probability of each pump in BART, the expected utility of *buying* insurance is:

$$EU(\text{buying (additional) insurance}) = U(24)k_f/128 + U(k_f - 40)(128 - k_f)/128, \quad k_f \in (40, 128] \tag{A.1}$$

and the expected utility function of *not buying* insurance in *purely voluntary insurance* normalizing $U(0) = 0$ is:

$$EU(\text{not buying insurance}) = U(k)(128 - k)/128, \quad k \in (0, 128] \tag{A.2}$$

and the expected utility function of *not buying additional* insurance in *mixed insurance* is:

$$EU(\text{not buying additional insurance}) = U(12)k_p/128 + U(k_p - 20)(128 - k_p)/128, \quad k_p \in (20, 128] \tag{A.3}$$

Since insurance has an incentive effect which changes the lotteries, we use k, k_f, k_p to indicate the choice of pumps without insurance, with full insurance and with partial insurance respectively. Recall that by assuming a CRRA utility function, we get the levels of *r* for every possible choice of pumps without insurance (Appendix Table A.1). We first calculate the optimal choice of pumps for each *r* after insurance, i.e., k_f and k_p (Appendix). Then we can calculate each part of the following equations based on a CRRA:

$$\begin{gathered} EU[\text{buying insurance } (k_f \in (40, 128])] - EU[\text{not buying insurance } (k \in (0, 128])] : \\ [U(24)k_f/128 + U(k_f - 40)(128 - k_f)/128] - [U(k)(128 - k)/128] \end{gathered} \tag{A.4}$$

$$EU(\text{buying additional insurance}) - EU(\text{not buying additional insurance } (k_p \in (20, 128])) :$$

$$[U(24)k_f/128 + U(k_f - 40)(128 - k_f)/128] - [U(12)k_p/128 + U(k_p - 20)(128 - k_p)/128] \tag{A.5}$$

Tables A.2 and A.3 show results of equation (A4) and (A5) around the risk-neutral subjects. In purely voluntary insurance, the difference between the expected utility of buying and not buying insurance is negative for risk-loving individuals, zero for risk-neutral, and positive for risk-averse individuals. In mixed insurance, since we simply split the premium and coverage of purely voluntary insurance into two equal parts as the premium and coverage of the mixed insurance, the difference between the expected utility of buying and not buying additional insurance is positive for a small group of risk-loving individuals. Note that when we calculate equations A1, A2 and A3 for each pump and the coefficient r, we find that the difference in the magnitude of expected utility is large across pumps. Thus, we do not use the comparisons of the utility differences across pumps in A.2 and A.3 to predict adverse/advantageous selection.

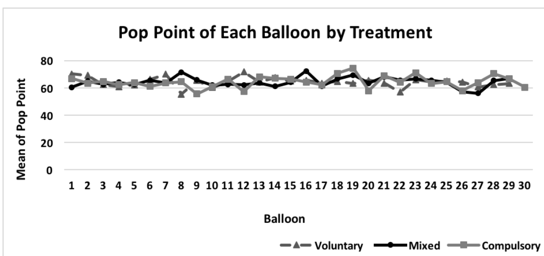
Table A.2: EU[buying insurance ($k_f \in (40, 128]$)] $- EU$[not buying insurance ($k \in (0, 128]$)] (Reprinted from [2])

K	$EU(Buy) - EU(NotBuy)$	K	$EU(Buy) - EU(NotBuy)$	K	$EU(Buy) - EU(NotBuy)$
55	1.737	**64**	0	**73**	-24.397
56	1.733	**65**	-0.706	**74**	-32.935
57	1.706	**66**	-1.643	**75**	-44.336
58	1.651	**67**	-2.976	**76**	-59.617
59	1.558	**68**	-4.497	**77**	-80.247
60	1.415	**69**	-6.627	**78**	-108.309
61	1.209	**70**	-9.427	**79**	-146.755
62	0.922	**71**	-13.115	**80**	-199.828
63	0.529	**72**	-17.972	$>= 81$	$<= -273.871$

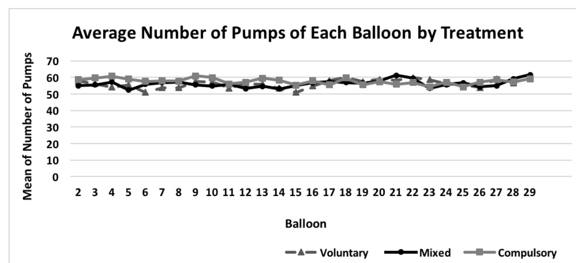
Table A.3: EU(buying additional insurance) $-$ EU(not buying additional insurance) (Reprinted from [2])

K	$EU(Buy) - EU(NotBuy)$	K	$EU(Buy) - EU(NotBuy)$	K	$EU(Buy) - EU(NotBuy)$
55	1.122	**64**	2	**73**	-0.242
56	1.207	**65**	2.080	**74**	-1.695
57	1.297	**66**	2.132	**75**	-3.821
58	1.393	**67**	2.152	**76**	-6.889
59	1.492	**68**	2.114	**77**	-11.331
60	1.595	**69**	1.999	**78**	-17.715
61	1.699	**70**	1.761	**79**	-26.954
62	1.804	**71**	1.362	**80**	-40.245
63	1.908	**72**	0.721	$>= 81$	$<= -59.51294892$

A.4 Distribution of Explosion Points and Learning Paths



(a) Explosion Points Distribution by Treatment



(b) Leaning Paths by Treatment

Figure A.2: Distribution of Explosion Points and Learning Paths (Reprinted from [2])

A.5 Background characteristics and measures of risk-taking

Table A.4: Background characteristics and measures of risk-taking across treatments. (Reprinted from [2])

	Purely Voluntary	Mixed	Purely Compulsory	Purely Voluntary v. Mixed	Purely Voluntary v. Compulsory	Purely Compulsory v. Mixed
Gamble Choice(mean)	3.4	3.4	3.4	0.775	0.969	0.708
Pumps in 2-29th(mean)	56.85	56.45	59.02	0.778	0.413	0.476
SSS-all(mean)	15.39	17.23	16.84	0.073	0.183	0.800
DOSPERT-all(mean)	90.94	91.23	94.53	0.556	0.202	0.384
Age (mean)	38	35	37	0.021	0.294	0.242
Gender	50.5% (F) 49.5% (M)	47.5% (F) 52.5% (M)	46.7% (F) 53.3% (M)	0.669	0.616	0.939
Household(mean)	2.6	2.7	2.8	0.544	0.205	0.513
Education(median)	2 year/ Associates Degree	2 year/ Associates Degree	2 year/ Associates Degree	0.025	0.427	0.194
Income(mean)	40,000-49,999	40,000-49,999	40,000-49,999	0.499	0.215	0.596

Notes: The last three columns show the p-values from two-sided Mann-Whitney *U*-Tests.

A.6 Experimental Instructions

(1). General Instructions

Now you will be presented with 30 balloons in the computer screen.

You have to decide how many times you want to pump each balloon. For every successful pump you will earn money. However, the explosion point for each balloon is random. The maximum possible number of pumps for each balloon is 128. The explosion point is random and it can be anywhere in the range from the first (1st) to the last (128th) pump.

For each balloon, you will be asked to select how many times you want to pump it up. You get a MONETARY reward of $0.01 for every successful pump. HOWEVER, if a balloon explodes before it reaches the number of pumps you indicated, you earn $0.00 for that balloon.

After each trial, a new balloon will appear.

For SOME balloons, you have an opportunity to buy an Insurance to protect yourself against the risk of an explosion for that particular balloon. Please make your decisions carefully. (if Purely Voluntary or Mixed Treatment)

For SOME balloons, you are required to buy an Insurance to protect yourself against the risk of an explosion for that particular balloon. (if Purely Compulsory Treatment)

At the end of the experiment, 3 balloons will be RANDOMLY SELECTED, and you will be paid the amount of money earned for these three balloons.

(2). Summary

* You write the number of times you want to pump up each balloon in a provided textbox.

* Remember: each balloon can be pumped up to 128 times (it will surely pop at 128th pump).

* Each balloon is then pumped up until a) that number is reached or b) it pops. Whatever occurs first.

* If it does not explode, you make $0.01 for each pump.

* If it does explode, you will not make any money on that balloon.

* There are a total of 30 balloons.

* Only some balloons have the opportunity to purchase insurance. (if Purely Voluntary or Mixed Treatment)

For some balloons, you are REQUIRED to buy an Insurance. (if Purely Compulsory Treatment)

* At the end, you will be paid the exact amount you earned on THREE randomly selected balloons.

Continue when you are ready to start.

(3a).Insurance (Purely Voluntary)

On the following balloon, you have an opportunity to buy an insurance to protect yourself against the risk of explosion. The price of the insurance is $0.40.

If the balloon does explode, the insurance will pay you $0.64; if the balloon does not explode, the insurance will pay you nothing, and the cost is not refunded. However, you will keep the earnings you make in that balloon if it is selected at the end of the experiment.

Remember: each balloon can be pumped up to 128 times;

The insurance is only valid for this balloon.

Before proceeding to make your choices, you have to correctly answer the following three questions.

(1). If you choose to buy the insurance, and you pump 128 times, then how much would you earn for this balloon?

A. $0 B. $64 C. $24 D. $40

(2). If you choose NOT to buy the insurance, and you pump 64 times, and the balloon does not explode, then how much would you earn for this balloon?

A. $64 B. $24 C. $0 D. $40

(3). If you choose to buy the insurance, and you pump 70 times, and the balloon does not explode, then how much would you earn for this balloon?

A. $70 B. $30 C. $0 D. $40

<u>(3b).Insurance (Mixed)</u>

On the following balloon, you are REQUIRED to buy an insurance to protect yourself against the risk of explosion. The price of the insurance is $0.20, and it is compulsory.

If the balloon does explode, the insurance will pay you $0.32; if the balloon does not explode, the insurance will pay you nothing, and the cost is not refunded. However, you will keep the earnings you make in that balloon if it is selected at the end of the experiment.

Besides the compulsory insurance, you have an opportunity to buy another insurance VOLUNTARILY. The price of this insurance is also $0.20, and it is voluntary.

If the balloon does explode, the insurance will also pay you $0.32; if the balloon does not explode, the insurance will pay you nothing, and the cost is not refunded. However, you will keep the earnings you make in that balloon if it is selected at the end of the experiment.

Remember: each balloon can be pumped up to 128 times (it will surely pop at 128th pump);

if the balloon does not explode, you make $0.01 for each pump;

the insurance is only valid for this balloon.

Before proceeding to make your choices, you have to correctly answer the following three questions.

(1).If you choose NOT to buy the voluntary insurance, and you pump 128 times, then how much would you earn for this balloon?

A. $0.0 B. $0.64 C. $0.24 D. $0.12

(2). If you choose NOT to buy the voluntary insurance, and you pump 64 times, and the balloon does not explode, then how much would you earn for this balloon?

A. $0.64 B. $0.24 C. $0.0 D. $0.44

(3). If you choose to BUY the voluntary insurance, and you pump 70 times, and the balloon explodes, then how much would you earn for this balloon?

A. $0.12 B. $0.24 C. $0.58 D. $0.30

(3c).Insurance (Purely Compulsory)

On the following balloon, you are REQUIRED to buy an insurance to protect yourself against the risk of explosion. The price of the insurance is $0.40, and it is compulsory.

If the balloon does explode, the insurance will pay you $0.64; if the balloon does not explode, the insurance will pay you nothing, and the cost is not refunded. However, you will keep the earnings you make in that balloon if it is selected at the end of the experiment.

Remember: each balloon can be pumped up to 128 times (it will surely pop at 128th pump);

if the balloon does not explode, you make $0.01 for each pump;

the insurance is only valid for this balloon.

Before proceeding to pump the balloon, you have to correctly answer the following two questions.

(1).After you buy the insurance, you pump 128 times, and then how much would you earn for this balloon?

A. $0.0 B. $0.64 C. $0.24 D. $0.40

(2).After you buy the insurance, you pump 70 times, and the balloon does not explode. How much would you earn for this balloon?

A. $0.70 B. $0.30 C. $0.0 D. $0.40

(4a). Insurance choice (Purely voluntary)

Now please indicate your decision by Clicking the options below.

Yes, I buy the insurance at a cost of $0.40.

No, I do not buy the insurance.

(4b). Insurance choice (Mixed)

Now please indicate your decision by Clicking the options below.

Yes, I buy the additional insurance at a cost of $0.20.

No, I do not buy the additional insurance.

(5). BART

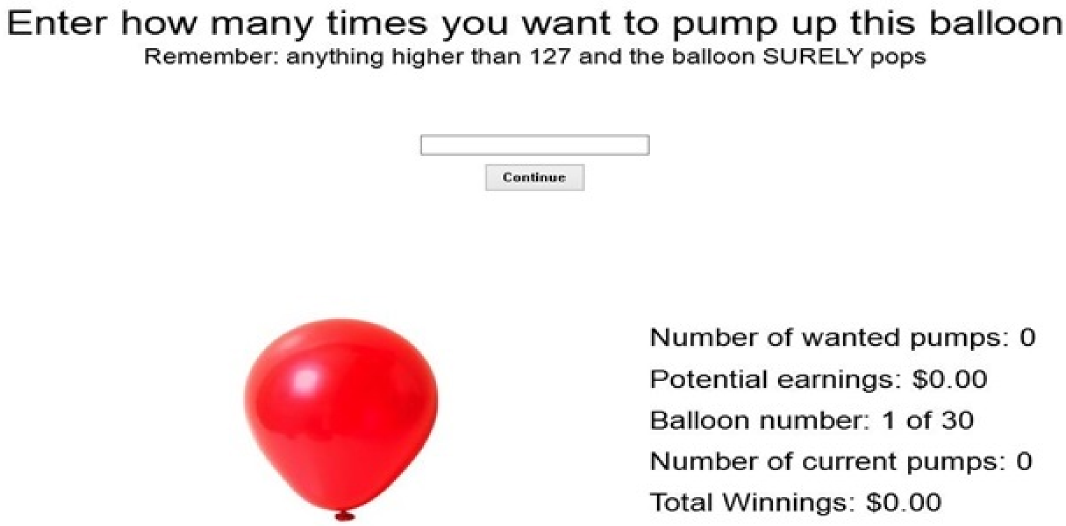


Figure A.3: Screenshot of BART. (Reprinted from [2])

(6). Earnings in BART

Congratulations!

Your earnings on 1-30 balloons are(),(),()...respectively, by randomly selecting three of them, your payment of this part is $().

The balloon task is now complete. Please Click continue to go to the next part.

(7). DOSPERT (Figure A.4)

For each of the following statements, please indicate the likelihood that you would engage in the described activity or behavior if you were to find yourself in that situation. Provide a rating from Extremely Unlikely to Extremely Likely.

(1). Admitting that your tastes are different from those of a friend.

(2). Going camping in the wilderness.

(3). Betting a day's income at the horse races.

(4). Investing 10% of your annual income in a moderate growth mutual fund.

(5). Drinking heavily at a social function.

(6). Taking some questionable deductions on your income tax return.

(7). Disagreeing with an authority figure on a major issue.

(8). Betting a day's income at a high-stake poker game.

(9). Having an affair with a married man/woman.

(10). Passing off somebody else's work as your own.

(11). Going down a ski run that is beyond your ability.

(12). Investing 5% of your annual income in a very speculative stock.

(13). Going whitewater rafting at high water in the spring.

(14). Betting a day's income on the outcome of a sporting event.

(15). Engaging in unprotected sex.

(16). Revealing a friend's secret to someone else.

(17). Driving a car without wearing a seat belt.

(18). Investing 10% of your annual income in a new business venture.

(19). Taking a skydiving class.

(20). Riding a motorcycle without a helmet.

(21). Choosing a career that you truly enjoy over a more prestigious one.

(22). Speaking your mind about an unpopular issue in a meeting at work.

(23). Sunbathing without sunscreen.

(24). Bungee jumping off a tall bridge.

(25). Piloting a small plane.

(26). Walking home alone at night in an unsafe area of town.

(27). Moving to a city far away from your extended family.

(28). Starting a new career in your mid-thirties.

(29). Leaving your young children alone at home while running an errand.

(30). Not returning a wallet you found that contains $200.

(8). Sensation Seeking Scale (Figure A.5)

Each of the items below contains two choices, A and B. Please click the letter of the choice which most describes your likes or the way you feel. In some cases you may find items in which both choices describe your likes or feelings. Please choose the one which better describes your likes or feelings. In some cases you may find items in which you do not like either choice. In these cases mark the choice you dislike least. Do not leave any items blank.

In this part, there are not right or wrong answers. Be frank and give your honest appraisal of yourself.

(1) A. I like “wild" uninhibited parties

B. I prefer quiet parties with good conversation

(2) A. There are some movies I enjoy seeing a second or even a third time

B. I can’t stand watching a movie that I’ve seen before

(3) A. I often wish I could be a mountain climber

B. I can’t understand people who risk their necks climbing mountains

(4) A. I dislike all body odors

B. I like some for the earthly body smells

(5) A. I get bored seeing the same old faces

B. I like to comfortable familiarity of everyday friends

(6) A. I like to explore a strange city or section of town by myself, even if it means getting lost

B. I prefer a guide when I am in a place I don’t know well

(7) A. I dislike people who do or say things just to shock or upset others

B. When you can predict almost everything a person will do and say he or she must be a bore

(8) A. I usually don't enjoy a movie or play where I can predict what will happen in advance

B. I don't mind watching a movie or a play where I can predict what will happen in advance

(9) A. I have tried marijuana or would like to

B. I would never smoke marijuana

(10) A. I would not like to try any drug which might produce strange and dangerous effects on me

B. I would like to try some of the new drugs that produce hallucinations

(11) A. A sensible person avoids activities that are dangerous

B. I sometimes like to do things that are a little frightening

(12) A. I dislike "swingers" (people who are uninhibited and free about sex)

B. I enjoy the company of real "swingers"

(13) A. I find that stimulants make me uncomfortable

B. I often like to get high (drinking liquor or smoking marijuana)

(14) A. I like to try new foods that I have never tasted before

B. I order the dishes with which I am familiar, so as to avoid disappointment and unpleasantness

(15) A. I enjoy looking at home movies or travel slides

B. Looking at someone's home movies or travel slides bores me tremendously

(16) A. I would like to take up the sport of water skiing

B. I would not like to take up water skiing

(17) A. I would like to try surf boarding

B. I would not like to try surf boarding

(18) A. I would like to take off on a trip with no preplanned or definite routes, or timetable

B. When I go on a trip I like to plan my route and timetable fairly carefully

(19) A. I prefer the "down to earth" kinds of people as friends

B. I would like to make friends in some of the "far out" groups like artists or "punks"

(20) A. I would not like to learn to fly an airplane

B. I would like to learn to fly an airplane

(21) A. I prefer the surface of the water to the depths

B. I would like to go scuba diving

(22) A. I would like to meet some persons who are homosexual (men or women)

B. I stay away from anyone I suspect of being “gay or lesbian"

(23) A. I would like to try parachute jumping

B. I would never want to try jumping out of a plane with or without a parachute

(24) A. I prefer friends who are excitingly unpredictable

B. I prefer friends who are reliable and predictable

(25) A. I am not interested in experience for its own sake

B. I like to have new and exciting experiences and sensations even if they are a little frightening, unconventional, or illegal

(26) A. The essence of good art is in its clarity, symmetry of form and harmony of colors

B. I often find beauty in the “clashing" colors and irregular forms of modern paintings

(27) A. I enjoy spending time in the familiar surroundings of home

B. I get very restless if I have to stay around home for any length of time

(28) A. I like to dive off the high board

B. I don’t like the feeling I get standing on the high board (or I don’t go near it at all)

(29) A. I like to date members of the opposite sex who are physically exciting

B. I like to date members of the opposite sex who share my values

(30) A. Heavy drinking usually ruins a party because some people get loud and boisterous

B. Keeping the drinks full is the key to a good party

(31) A. The worst social sin is to be rude

B. The worst social sin is to be a bore

(32) A. A person should have considerable sexual experience before marriage

B. It’s better if two married persons begin their sexual experience with each other

(33) A. Even if I had the money I would not care to associate with flight rich persons like those in the “jet set"

B. I could conceive of myself seeking pleasures around the world with the “jet set"

(34) A. I like people who are sharp and witty even if they do sometimes insult others

B. I dislike people who have their fun at the expense of hurting the feelings of others

(35) A. There is altogether too much portrayal of sex in movies

B. I enjoy watching many of the “sexy" scenes in movies

(36) A. I feel best after taking a couple of drinks

B. Something is wrong with people who need liquor to feel good

(37) A. People should dress according to some standard of taste, neatness, and style

B. People should dress in individual ways even if the effects are sometimes strange

(38) A. Sailing long distances in small sailing crafts is foolhardy

B. I would like to sail a long distance in a small but seaworthy sailing craft

(39) A. I have no patience with dull or boring persons

B. I find something interesting in almost every person I talk to

(40) A. Skiing down a high mountain slope is a good way to end up on crutches

B. I think I would enjoy the sensations of skiing very fast down a high mountain slope

(9). Gamble Choice

See Figure A.6: screenshot of gamble-choice task.

(10). Demographic survey

Please answer the following survey questions.

(1) Please enter your age in years.

(2) Please indicate the HIGHEST level of education you have completed.

Some High School or less

High School Diploma

Some College

2 year/Associates Degree

4 year/Bachelor's Degree

Some Graduate School

Graduate Degree

(3) Including yourself, how many people live in your household?

(4) Please indicate your gender.

Male

Female

(5) Please indicate your race.

Asian/ Pacific Islander

African American

Caucasian/ White

Native American/ Indigenous

Hispanic

Other (Please list below)

(6) Please indicate your household yearly income for 2016. (Include all forms of income, including salary, interest and dividend payments, tips, scholarship support, student loans, parental support, and allowance)

Less than $30,000

$30,000 - $39,999

$40,000 - $49,999

$50,000 - $59,999

$60,000 - $69,999

$70,000 - $79,999

$80,000 - $89,999

$90,000 - $99,999

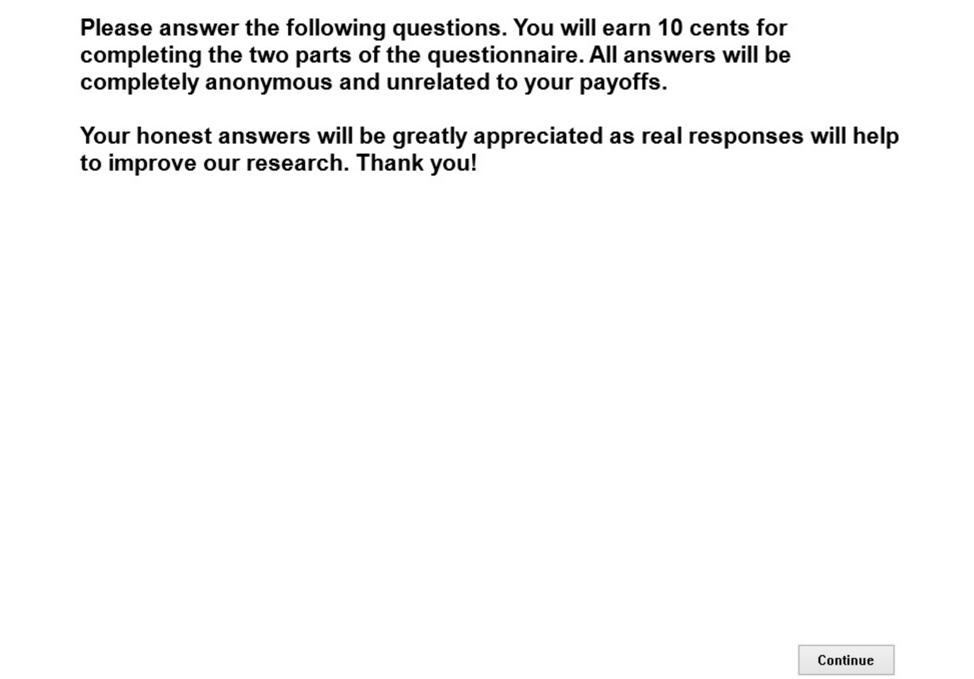
$100,000 - $149,999

$150,000 or more

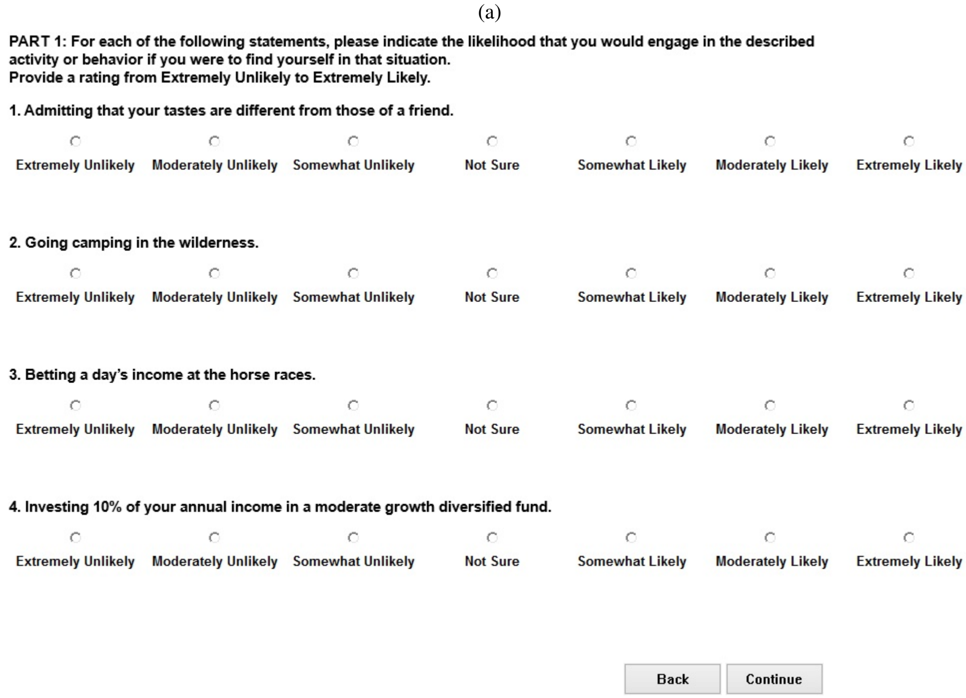
(7) Do you think the Insurance Option in the balloon task is clear?

Yes.

No.



(a)



(b)

Figure A.4: Screenshot of DOSPERT (Reprinted from [2])

PART 2: Each of the items below contains two choices, A and B. Please click the letter of the choice which most describes your likes or the way you feel. In some cases you may find items in which both choices describe your likes or feelings. Please choose the one which better describes your likes or feelings. In some cases you may find items in which you do not like either choice. In these cases mark the choice you dislike least. Do not leave any items blank.

In this part, there are not right or wrong answers. Be frank and give your honest appraisal of yourself.

1.

- A. I like “wild” uninhibited parties.
- B. I prefer quiet parties with good conversation.

2.

- A. There are some movies I enjoy seeing a second or even a third time.
- B. I can’t stand watching a movie I’ve seen before.

3.

- A. I often wish I could be a mountain climber.
- B. I can’t understand people who risk their necks climbing mountains.

4.

- A. I dislike all body odors.
- B. I like some of the earthy body smells.

Continue

Figure A.5: Screenshot of SSS. (Reprinted from [2])

In this part, please make a choice among the following six gamble choices. Each choice has two events A and B, and each event's chance of occurring is 50%. After you make your choice, the system will randomly choose an event and depending on your choice, you will get the corresponding payoff for that event.
If you select a gamble with a negative payoff for outcome B, negative payoffs will be deducted from your payment in previous parts of the experiment.

Gamble choice	The event	Probability	Payoff (cents)
1.	A	50%	10
	B	50%	10
2.	A	50%	18
	B	50%	6
3.	A	50%	26
	B	50%	2
4.	A	50%	34
	B	50%	-2
5.	A	50%	42
	B	50%	-6
6.	A	50%	44
	B	50%	-8

Continue

Figure A.6: Screenshot of gamble-choice task. (Reprinted from [2])

APPENDIX B

B.1 Gender composition in each session.

Table B.1: No. of Females by Session

	Freq. (Treatment)	Percent	Freq. (Baseline)	Percent
2	1	6.3	-	-
3	2	12.5	3	25
4	4	25	4	33.3
5	5	31.2	3	25
6	4	25	2	16.7
Total	16	100	12	100

B.2 Men and Women Respond Equally to Role Assignment

In column (1), we exclude the data of “supporter", and the dummy variable *treatment* takes the value of 1 if the treatment assignment is *breadwinner* and 0 for the *baseline*. In column (2), the “breadwinner" data is excluded, and *treatment* is 1 if the role is *supporter* and 0 for the *baseline*. In column (3), we only use the data of treatment group, and *treatment* is 1 if the role is *breadwinner* and it is 0 if the role is *supporter*. The coefficients of all the interaction terms are not significant, which means men and women respond equally to each role assignment.

Table B.2: Ordered Probit Regressions for the Difference-in-Differences Estimates of WTC.

	(1) Breadwinner & Base	(2) Supporter & Base	(3) Breadwinner & Supporter
Treatment	0.449	-0.121	0.586**
	(0.291)	(0.273)	(0.301)
Female	-0.389	-0.414	-0.131
	(0.256)	(0.257)	(0.278)
Treated * Female	0.069	0.400	-0.314
	(0.374)	(0.364)	(0.384)
Confidence	0.197***	0.233***	0.089
	(0.054)	(0.058)	(0.054)
Risk Tolerance	0.130**	0.236***	0.211***
	(0.055)	(0.059)	(0.063)
Observation	140	141	129
R-square	0.069	0.072	0.067

Notes: Dependent variable is WTC. Standard errors are in parentheses. All regressions also control for ability and gender composition, but none of them are significant. All results are robust to using OLS instead of ordered probit.

B.3 Confidence

Table B.3: Proportion of subjects who believe they performed better than their partner (confidence)

	Baseline	Breadwinner	Supporter	*p*-value Base vs BW	*p*-value Base vs SP	*p*-value BW vs SP
Men	80.0%(0.07)	73.1%(0.09)	83.3%(0.07)	0.528	0.732	0.355
Women	78.1%(0.07)	78.9%(0.07)	71.4%(0.08)	0.923	0.509	0.460
P-Value	0.836	0.589	0.260			

Notes: Standard errors are in parentheses.

B.4 Elicited Social Norms

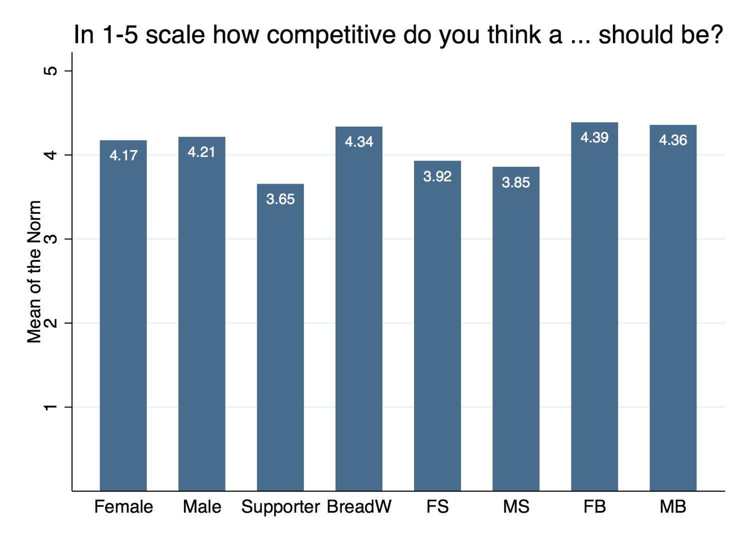
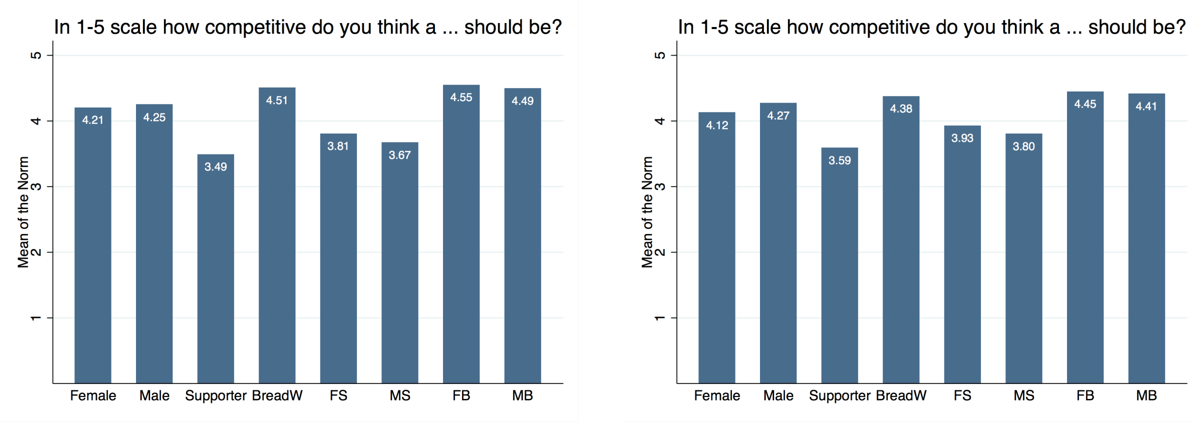


Figure B.1: Elicited Social Norms for Competitiveness (including baseline & treatment group)



(a) Female in Treatment

(b) Male in Treatment

Figure B.2: Elicited Social Norms for Competitiveness by Gender

B.5 Theoretical Model of Social Norms

In this section we develop a simple social norm model based on [92], and apply it to a "competition entry" environment. Consider social categories C, and each person j is mapped into one of the categories, c_j. Each person is aware of his/her own category and others' categories. In addition, each person may be mapped into multiple categories under different situations. For example, a person j can be both a woman and a professional. P indicates the prescribed behavior by social norms for each category C. We first propose the following additive utility function, which depends on an individual j's identity, as well as his/her actions and the actions of others:

$$U_j = U_j(a_j, a_{-j}, I_j) \tag{B.1}$$

where I_j is j's role or self-image, and a_j is the action taken by j. The action taken by others is indicated by a_{-j}. We further extend the self-image term by breaking it into two components:

$$I_j = I_{ji} + I_{je} \tag{B.2}$$

where I_{ji} is the internalized self-image, which captures the positive feeling or negative feeling a person j experiences from an action. I_{je} is the external self-image, and it captures the rewards or punishments from others after they observe j's action. We present I_{ji} and I_{je} separately:

$$I_{ji} = I_{ji}(a_j; c_j, \epsilon_j, P)$$
$$I_{je} = I_{ji}(a_j, a_{-j}; c_j, \epsilon_j, P)$$

First, we evaluate I_{ji}. An individual j's intrinsic role depends on his/her social category c_j, as well as the extent to which his/her own actions match the prescribed behavior indicated by P. The other given characteristics of j are captured by ϵ_j. Role also depends on the distance between ϵ_j and the ideal characteristics in category c_j, indicated by P. A person j's intrinsic social status within a category is given by the function $I_{ji}(\cdot)$. When the action is observable to others, the gains or losses of j's self-image further depend on how others interpret his/her action, which is revealed by a_{-j} in I_{je}. Based on this framework, the gains or losses associated with person j's self-image represent the increases or decreases in utility U_j driven by I_j.

We apply the above model to our experimental setting, to derive our hypotheses. Since in our experiment, the competitive action is unobservable because the WTC is private information, we normalize $I_{je} \sim 0$, and focus on I_{ji}. We also assume a person's identity is additive when he/she is mapped into multiple categories with different "P" which play in opposite directions.

Hypothesis 1: *The WTC of women is lower than the WTC of men.*

The current experiment studies gender and competitiveness, which can be formalized as follows. There is a set of categories C, and it consists of {Female, Male}. If a person j is male, then his social category is c_{jm}; if a person j is female, then her social category is c_{jw}. Let us assume that an individual j can choose one of two types of actions $a_j \in \{0, 1\}$ in terms of "competition entry". If an individual j enters the competition, then $a_j = 1$, otherwise $a_j = 0$. P describes the appropriate behavior for "competition entry" for each gender. Based on P, it is socially appropriate for males (c_{jm}) to choose $a_{jm} = 1$, but inappropriate for females (c_{jw}) to choose $a_{jw} = 1$.[1]

[1] For competitiveness, males have higher social status than females [72].

A female experiences a loss of I_{ji} when she enters the competition, but a male experiences a gain of I_{ji} when he enters the competition, which decreases and increases their respective U_j. Thus, consistent with previous literature, we hypothesize that in the control, the WTC of females will be lower than the WTC of males.

Hypothesis 2: *The WTC of supporters is lower than the WTC of breadwinners.*

We now study the social categories of family roles. The C set consists of {Supporter, Breadwinner}, and action (a_j) still represents “competition entry". P describes the appropriate behavior for “competition entry" for each role. According to P, a breadwinner (c_{jb}) has higher social status in competitiveness than a supporter (c_{js}). Therefore, we expect that the WTC of breadwinners will be higher than the WTC of supporters.

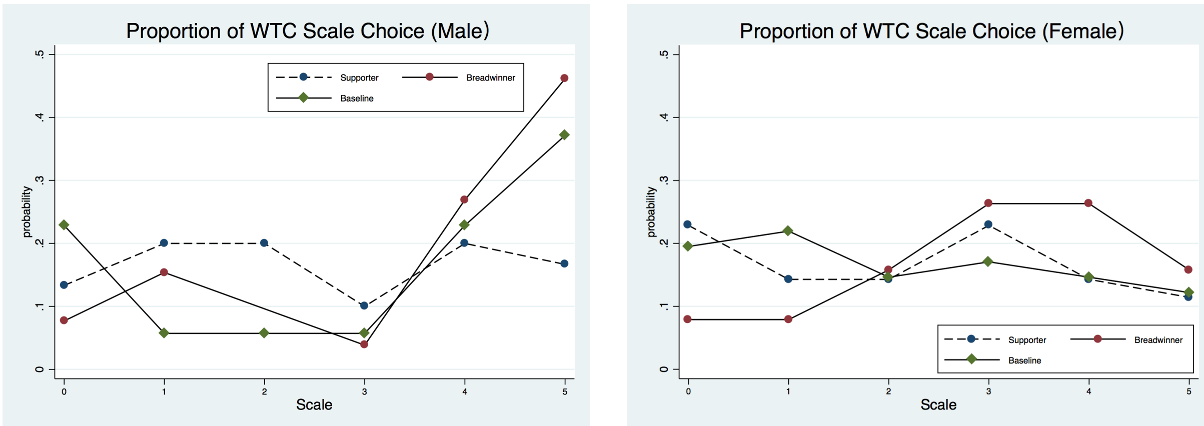
Hypothesis 3: *The “breadwinner" role for women increases their WTC, while the role of “supporter" for men decreases WTC.*

When a person j is assigned a role, he/she is mapped into two categories, gender and role. If a woman is assigned as the breadwinner, her dual category is denoted as $c_{jwb} = c_{jw} + c_{jb}$. If a man is assigned as the supporter, his dual category is denoted as $c_{jms} = c_{jm} + c_{js}$. Category gender and category role have separate P for “competition entry", which is the same as stated above. A female breadwinner's (c_{jwb}) intrinsic self-image experienced from entering competition comes from two parts: woman (I_{jiw}) and breadwinner (I_{jib}). We present her self-image as follows:

$$I'_{ji} = I_{jiw} + I_{jib}. \tag{B.3}$$

Since the role of “breadwinner" rewards competitiveness according to its P, then $I'_{ji} > I_{ji}$ for a woman. This means the role of “breadwinner" increases I_{ji} when a woman enters the competition. Therefore, a female breadwinner's WTC is higher than the WTC of a woman without this role. Applying the same analysis to a male supporter (c_{jms}), we expect the role of “supporter" to decrease the WTC of a man.

B.6 Distributions of each WTC scale



(a) Male (b) Female

Figure B.3: Distributions of each WTC scale

B.7 Earnings in the Second Stage by Performance Quartile ($)

Table B.4: Men's Expected Earnings in the Second Stage by Performance Quartile ($)

		Baseline	Treat	Breadwinner	Supporter	*p*-value Base vs Treat	*p*-value Base vs BW	*p*-value Base vs S
Worst	1-2	1.68(0.31)	1.73(0.24)	1.12(0.28)	1.58(0.32)	0.556	0.234	0.951
Best	3-4	6.74(0.99)	6.05(0.85)	5.46(0.98)	5.50(1.02)	0.852	0.396	0.334

Notes: 1= the worst performance; 4 = the best performance. Standard errors are in parentheses.

Table B.5: Women's Actual Earnings in the Second Stage by Performance Quartile ($)

		Baseline	Treat	Breadwinner	Supporter	*p*-value Base vs Treat	*p*-value Base vs BW	*p*-value Base vs S
Worst	1	1.09(0.69)	1.22(0.57)	1.11(0.83)	1.36(0.91)	0.845	0.821	0.714
	2	1.72(0.85)	1.35(0.52)	1.55(0.85)	1.2(0.62)	0.439	0.684	0.416
	3	1.93(0.94)	1.71(0.81)	1.51(1.09)	2.2(0.12)	0.794	0.227	0.298
Best	4	4.1(1.37)	5.35(1.04)	8.4(0.28)	2.73(1.36)	0.741	0.079	0.301

Notes: 1= the worst performance; 4 = the best performance. Standard errors are in parentheses.

Table B.6: Men's Actual Earnings in the Second Stage by Performance Quartile ($)

		Baseline	Treat	Breadwinner	Supporter	*p*-value Base vs Treat	*p*-value Base vs BW	*p*-value Base vs S
Worst	1	0.38(0.26)	1.42(0.63)	0.56(0.22)	2.8(1.2)	0.171	0.345	0.101
	2	1.63(0.81)	2.45(0.92)	2.1(.)	2.82(1.77)	0.573	0.243	0.332
	3	1.85(1.50)	1.11(0.36)	1.86(1.08)	2.01(1.06)	0.522	0.595	0.443
Best	4	2.95(1.25)	3.68(1.14)	1.68(0.56)	5.04(2.07)	0.8441	0.665	0.651

Notes: 1= the worst performance; 4 = the best performance. Standard errors are in parentheses.

B.8 The Online Experiment

Table B.7: Summary of Treatments and Number of Subjects (online experiment).

	First stage (individually)	Second stage (group, WTC)	Female	Male	F-Manager	M-Manager	Total
Baseline Group	Piece rate	No roles	30	22	—	—	52
Treatment Group	Piece rate	Role assignment	63	52	36	25	115

Table B.8: Gender Difference in Performance in the Baseline

	First Stage 1	Second Stage 2(all participants)	*p*-value column 1 vs 2
Men	6.82	7.6	0.502
Women	4.87	5.4	0.597
p-value	0.18	0.094	

Table B.9: Gender Difference in Performance in the Treatment Group

	First Stage 1(individual)	Second Stage 2(group)	Manager 3	Assistant 4	P-value column 1 vs 3	P-value column 1 vs 4
Men	4.60	5.88	6.72	5.11	0.009	0.144
Women	4.84	5.68	5.61	5.77	0.569	0.042
P-value	0.505	0.611				

Table B.10: Proportion of subjects who believe they performed better than their partner (confidence)

	Baseline	Manager	Assistant	*p*-value Base vs MA	*p*-value Base vs AS	*p*-value MA vs SP
Men	86.4%	64%	66.7%	0.083	0.115	0.842
Women	40%	52.8%	51.9%	0.304	0.374	0.942
P-Value	0.001	0.387	0.272			

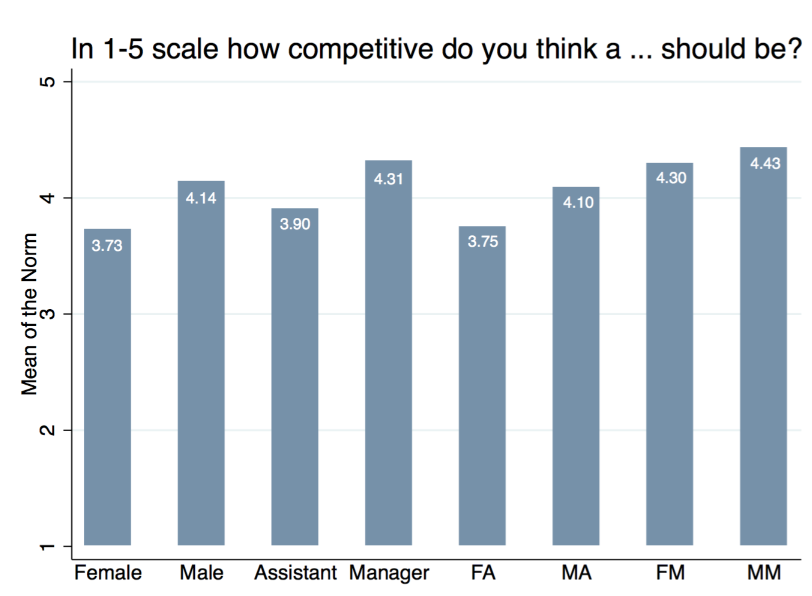


Figure B.4: Elicited Social Norms for Competitiveness in Treatment (online)

B.9 Instructions

WELCOME

In the experiment today, you will be asked to complete two different tasks. None of these will take more than 5 minutes. At the end of the experiment we will pay you based on your performance in these two tasks. The method we use to determine your earnings varies across tasks. Before each task we will describe in detail how your payment is determined.

Your total earnings from the experiment are the sum of your payment from the two tasks.

New page:

Task 1

For Task 1 you will be asked to solve a set of problems independently. You will be given 5 minutes to answer these problems. For calculation questions, you cannot use a calculator, however you are welcome to write the numbers down and make use of the provided scratch paper. Your answers to the problems are anonymous.

You get 10 cents per problem you solve correctly. Your payment does not decrease if you provide an incorrect answer to a problem. We refer to this payment as the piece rate payment. You will not be informed of how many questions you correctly solve and your corresponding payment in this task until the end of the experiment.

Please do not talk with one another for the duration of the experiment. If you have any questions, please raise your hand.

New page:

Task 2

For task 2 you will be randomly paired with a partner (one of the participants in this room, but you will not be told who she/he is), and you two will work as a group anonymously. Four groups in total will be formed in this session. As in Task 1 each group will be given 5 minutes to solve a similar set of problems. The sum of answers that the two members correctly solve will be the final number of correct answers of the group. The final payment of each group in this task will be equally split between the two group members.

New page:

Your payment in this task will also depend on your choice of a payment scheme: piece rate or tournament. If you choose *piece rate*, your group will be paid in the same way as Task 1, but with 20 cents per problem you solve correctly.

If you choose *tournament*, your group will be paid based on the performance of your group relative to that of the other three groups, regardless of the payment scheme chosen by the other groups. For instance, if the number of problems solved correctly in your group is the largest among all the four groups, your group will receive **80** cents per correct answer, four times the payment from the piece rate; otherwise your group will receive no payment for this task. If there are ties, all these winner groups will receive 80 cents per correct answer.

The choices and answers from you and your partner will be kept anonymous. You will not be informed of the corresponding payment of your group in this task until the end of the experiment.

New page:

Now, you were randomly selected to be the *BreadWinner* of your group, and your partner is the *Supporter*. (only for treatment)

Please indicate in 5-scale how much you want to choose the tournament payment scheme for your group. Your partner will also be asked this same question.

Either your selection or your partner's selection will be selected to be the final selection of your group by equal chance.

You (your partner) will not know the choice made by your partner (you) in this task. You will be informed of the final payment scheme of your group and your corresponding payment in this task at the end of the experiment.

New page:

Please indicate in 5-scale how much you want to choose the tournament payment scheme for your group. If you choose scale 1, and your choice is selected as your group's choice, then the probability of entering the tournament will be 20% (i.e., 80% of probability to use piece rate); if you choose scale 2, the probability of entering the tournament will be 40%. Scale 5 means entering tournament for sure.

Scale	Description
0	Piece rate
1	Entering the Tournament with probability 20%.
2	Entering the Tournament with probability 40%.
3	Entering the Tournament with probability 60%.
4	Entering the Tournament with probability 80%.
5	Entering the Tournament for sure.

Your choice is: 0, 1, 2, 3, 4, 5

New page:

Please indicate in the five-scale how competitive do you think a **woman** should be if she is the **supporter** in her household? If your answer is the same as the most frequent answer for this question in today's experiment, then you will get extra 25 cents.

1 (Not competitive at all), 2, 3, 4, 5 (Very competitive)

New page:

Please indicate in the five-scale how competitive do you think a **woman** should be if she is the **breadwinner** in her household? If your answer is the same as the most frequent answer for this question in today's experiment, then you will get extra 25 cents.

1 (Not competitive at all), 2, 3, 4, 5 (Very competitive)

New page:

Please indicate in the five-scale how competitive do you think a **woman** should be? If your answer is the same as the most frequent answer for this question in today's experiment, then you will get extra 25 cents.

1 (Not competitive at all), 2, 3, 4, 5 (Very competitive)

New page:

Please indicate in the five-scale how competitive do you think a **breadwinner** of a household should be? If your answer is the same as the most frequent answer for this question in today's experiment, then you will get extra 25 cents.

1 (Not competitive at all), 2, 3, 4, 5 (Very competitive)

New page:

Please indicate in the five-scale how competitive do you think a **supporter** of a household should be? If your answer is the same as the most frequent answer for this question in today's experiment, then you will get extra 25 cents.

1 (Not competitive at all), 2, 3, 4, 5 (Very competitive)

New page:

Please indicate in the five-scale how competitive do you think a **man** should be if he is the **supporter** in his household? If your answer is the same as the most frequent answer for this question in today's experiment, then you will get extra 25 cents.

1 (Not competitive at all), 2, 3, 4, 5 (Very competitive)

New page:

Please indicate in the five-scale how competitive do you think a **man** should be if he is the **breadwinner** in his household? If your answer is the same as the most frequent answer for this question in today's experiment, then you will get extra 25 cents.

1 (Not competitive at all), 2, 3, 4, 5 (Very competitive)

New page:

Please indicate in the five-scale how competitive do you think a **man** should be? If your answer is the same as the most frequent answer for this question in today's experiment, then you will get extra 25 cents.

1 (Not competitive at all), 2, 3, 4, 5 (Very competitive)

New page:

Do you think that your performance in Task 1 is better than your partner's performance? If your guess is correct, then you will get extra 25 cents.

Yes

No

New page:

Please estimate the rank of your performance relative to the rest of the participants in Task 1? For example, if you check 1, you consider yourself as better than all other seven participants. If your guess is the same as your rank, then you will get extra 25 cents.

1, 2, 3, 4, 5, 6, 7, 8

New page:

If you can choose your role in a household, what is your choice?

Breadwinner

Supporter

New page:

Please make a choice among the following six gamble choices. Each choice has two events A and B, and each event's chance of occurring is 50%. After you make your choice, the system will randomly choose an event for that choice, you will get the corresponding payoff for that event. Please select the gamble that you choose.

New page:

Do you think men or women generally do better in the "math problem" task that you just did?

Demographic survey

Please answer the following survey questions.

(1) Please enter your age in years.

(2) Please indicate the HIGHEST level of education you have completed.

Some High School or less

High School Diploma

Some College

2 year/Associates Degree

4 year/Bachelor's Degree

Some Graduate School

Graduate Degree

(3) Including yourself, how many people live in your household?

(4) Please indicate your gender.

Male

Female

(5) Please indicate your race.

Asian/ Pacific Islander

African American

Caucasian/ White

Native American/ Indigenous

Hispanic

Other (Please list below)

(6) Please indicate your household yearly income for 2016. (Include all forms of income, including salary, interest and dividend payments, tips, scholarship support, student loans, parental support, and allowance)

Less than $30,000

$30,000 - $39,999

$40,000 - $49,999

$50,000 - $59,999

$60,000 - $69,999

$70,000 - $79,999

$80,000 - $89,999

$90,000 - $99,999

$100,000 - $149,999

$150,000 or more

(7) Your major?

B.10 Word Selection Survey

Survey1

In one word, how would you call someone who is the primary earner in a household: _____

In one word, how would you call someone who is the secondary earner in a household: _____

Survey 2

1. In one word, how would you call someone who is the primary earner in a household:

Breadwinner

Primary

Provider

Other (please list) _____

2. In one word, how would you call someone who is the secondary earner in a household:

Secondary

Supporter

Assistant

Other (please list) _____

APPENDIX C

C.1 The set-up of each station.

Figure C.1: The set-up of each station.

C.2 AOIs in screen.

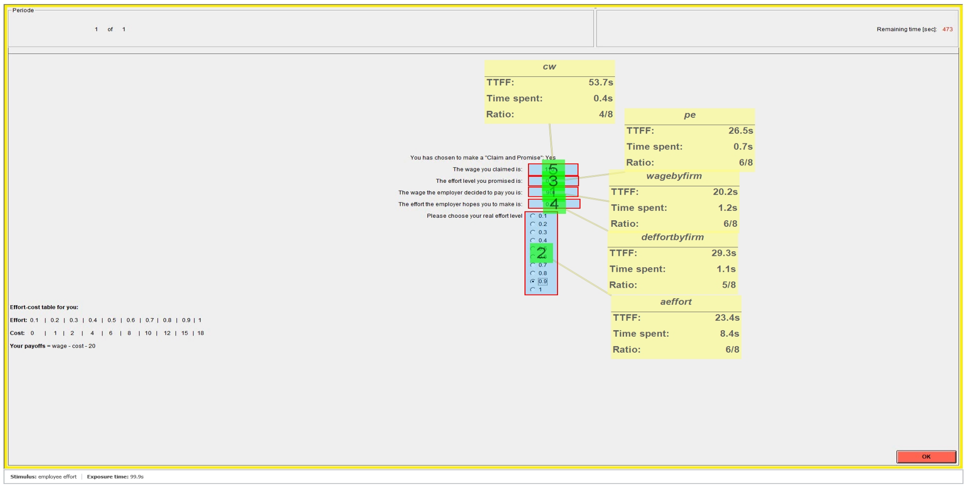


Figure C.2: AOIs in the “Effort Choice" Screen (stated)

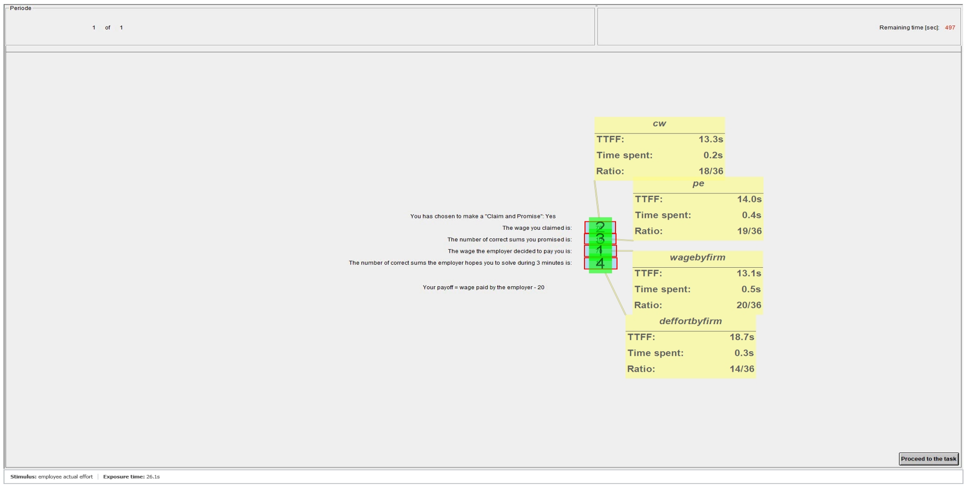
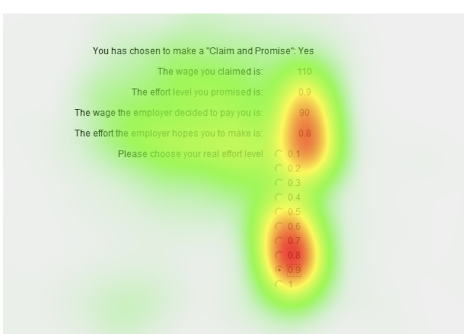


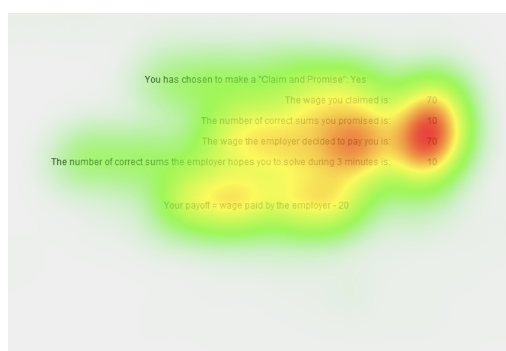
Figure C.3: AOIs in the “Effort Choice" Screen (real)

C.3 Heatmaps.

The Heatmaps is generated within iMotion, and it is an aggregated visualization and analysis of eye tracking data. The heatmaps helps identify what areas of the stimuli was focused on for the longest amount of time by all the participants, and what areas or elements may go missed entirely. **Red** indicates relatively more attention and **green** is relatively less attention.



(a) Stated Effort



(b) Real Effort

Figure C.4: Heatmap

C.4 Lookup patterns in the real-effort paradigm.

Table C.1: Average Lookup Patterns (Real Effort)

	Claimed Wage	Promised Effort	Actual Wage	Desired Effort
Time to first fixation (in seconds)	3.74	6.91	5.73	12.96
Fixation duration (in seconds)	0.45	0.73	0.91	0.76
Revisits (No.)	2	3	3	2

*Notes:*The order (from left to right) of the information in the table is the same as listed in the screen (from up to down).

C.5 Instructions

Stated-effort Baseline

Welcome! Thank you for participating in this study.

You will be making REAL decisions involving REAL money. So, please read the instructions carefully. Feel free to ask any questions by raising your hand. Please do not speak to other participants during the experiment. You can use the draft paper we provide, but you are NOT allowed to use the calculator.

The payoffs in the experiment are represented by Experimental Points. Your payment in the experiment will be converted to Dollars at a rate of **1 point = $0.5**. To ensure anonymity and confidentiality you were given a random number decal at the beginning of the experiment. You will need this number to collect your payment.

There are two types of participants: **employer** and **employee**.

At the beginning of the experiment the computer will randomly choose with equal chance whether you will be an employer or an employee. Each person selected to be an employer will be randomly paired with another person in the room who has been selected as an employee.

The identity of the participants will remain anonymous. You will never know the identity of the person you were paired with and the other person will never know your identity.

The experiment consists of **two** stages. The employer pays a wage w to the employee. After receiving the wage w, the employee will choose an effort level e for the employer. The details are as follows.

Stage 1 (employer): a) The employer chooses the wage level, w for his employee;

b) The employer also announces a non-binding effort level, $\hat{e}$, that he wants the employee to do.

Stage 2 (employee): The employee has to choose his actual effort level, e, after he reviews the wage level set by the employer.

Note: The final payment of both parties will be decided by the employee's actual decision on effort e, and will NOT be decided by the proposed effort $\hat{e}$.

The combination of wage and actual effort level determine monetary payoffs for the employer and the employee:

The cost of each real e:

Table C.2: Effort-Cost Table

Effort Level	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Cost	0	1	2	4	6	8	10	12	15	18

Calculation of payoffs:

Payoff of the employer = $(120 - w) * e$

The payoff of the employer equals: (120 - the wage he paid to his employee) * the actual effort level selected by the employee.

Payoff of the employee = $w - c(e) - 20$

The payoff of the employee equals: the wage received from the employer - the cost of the actual effort level he selected - 20.

Thus, the higher the effort level provided by the employee and the lower the wage, the larger the *employer's* payoffs. The lower the effort level provided by the employee and the higher the wage, the larger the *employee's* payoffs.

Example:

If the salary is 50 and the actual effort level is 0.5, then the payoffs would be:

Employer's payoffs = $(120 - 50) * 0.5$ = 35 points = $17.5

Employee's payoffs = $50 - 6 - 20$ = 24 points = $12

Questionnaire: Just to be sure that you understand the instructions you have to solve a simple questionnaire. When everyone in the room has answered the questionnaire correctly we will start the experiment.

The wage for an employee is 100, and the employee's actual effort level is 0.3, then the payoffs would be:

Employer's payoff (experimental points):

Employee's payoff (experimental points):

<u>Stated-effort “C&P" Treatment</u>

Welcome! Thank you for participating in this study.

You will be making REAL decisions involving REAL money. So, please read the instructions carefully. Feel free to ask any questions by raising your hand. Please do not speak to other participants during the experiment. You can use the draft paper we provide, but you are NOT allowed to use the calculator.

The payoffs in the experiment are represented by Experimental Points. Your payment in the experiment will be converted to Dollars at a rate of **1 point = $0.5**. To ensure anonymity and confidentiality you were given a random number decal at the beginning of the experiment. You will need this number to collect your payment.

There are two types of participants: **employer** and **employee**.

At the beginning of the experiment the computer will randomly choose with equal chance whether you will be an employer or an employee. Each person selected to be an employer will be randomly paired with another person in the room who has been selected as an employee.

The identity of the participants will remain anonymous. You will never know the identity of the person you were paired with and the other person will never know your identity.

The experiment consists of **three** stages. The employee chooses to make a “claim and promise" or not. The employer pays a wage w to the employee. After receiving the wage w, the employee will choose an effort level e for the employer. The details are as follows.

Stage 1 (employee): The employee decides whether to make a “claim and promise" or not:

“If the firm gives me the wage=_____, I will choose the effort level=_____". This claim and promise is non-binding.

Stage 2 (employer): a)The employer chooses the wage level, w for his employee;

b) The employer also announces a non-binding effort level, $\hat{e}$, that he wants the employee to do.

Stage 3 (employee): The employee has to choose his effort level, e, after he reviews the wage level set by the employer.

Note: The final payment of both parties will be decided by the employee's actual decision on effort e, and will NOT be decided by the promised or proposed effort.

The combination of wage and actual effort level determine monetary payoffs for the employer and the employee:

The cost of each real e:

Table C.3: Effort-Cost Table

Effort Level	0.1	0.2	0.3	0.4	0.5	0.6	0.7	0.8	0.9	1
Cost	0	1	2	4	6	8	10	12	15	18

Calculation of payoffs:

Payoff of the employer = $(120 - w) * e$

The payoff of the employer equals: (120 - the wage he paid to his employee) * the actual effort level selected by the employee.

Payoff of the employee = $w - c(e) - 20$

The payoff of the employee equals: the wage received from the employer - the cost of the actual effort level he selected - 20.

Thus, the higher the effort level provided by the employee and the lower the wage, the larger the *employer's* payoffs. The lower the effort level provided by the employee and the higher the wage, the larger the *employee's* payoffs.

Example:

If the actual wage is 50 and the actual effort level is 0.5, then the payoffs would be:

Employer's payoffs = $(120 - 50) * 0.5$ = 35 points = $17.5

Employee's payoffs = $50 - 6 - 20$ = 24 points = $12

Questionnaire: Just to be sure that you understand the instructions you have to solve a simple questionnaire. When everyone in the room has answered the questionnaire correctly we will start the experiment.

The employee promised that if the wage is 60, he will select the effort level of 6. The actual wage for an employee is 100 and the proposed effort level from the employer is 0.8, and actual effort level selected by the employee is 0.3, then the payoffs would be:

Employer's payoff (experimental points):

Employee's payoff (experimental points):

Real-effort Baseline

Welcome! Thank you for participating in this study.

You will be making REAL decisions involving REAL money. So, please read the instructions carefully. Feel free to ask any questions by raising your hand. Please do not speak to other participants during the experiment. You can use the draft paper we provide, but you are NOT allowed to use the calculator.

The payoffs in the experiment are represented by Experimental Points. Your payment in the experiment will be converted to Dollars at a rate of **1 point = $0.5**. To ensure anonymity and confidentiality you were given a random number decal at the beginning of the experiment. You will need this number to collect your payment.

There are two types of participants: **employer** and **employee**.

At the beginning of the experiment the computer will randomly choose with equal chance whether you will be an employer or an employee. Each person selected to be an employer will be randomly paired with another person in the room who has been selected as an employee.

The identity of the participants will remain anonymous. You will never know the identity of the person you were paired with and the other person will never know your identity.

The experiment consists of **two** stages. The employer pays a wage w to his/her employee, and the employee needs to do an addition task (e.g., 23+34+72+47+97) for the employer. The number of correct sums solved by the employee determines the payoffs of the employer. The details are as follows.

Stage 1 (employer): a) The employer chooses a wage level using an integer number between 20 and 120 (w) for the employee.

b) The employer also proposes a number of correct sums (s^*), that he/she wants the employee to solve.

Stage 2 (employer): After reviewing the wage level set by the employer, the employee has 3 minutes to do the task.

Note: The final payment of the employer will be decided by the employee's actual number of correct sums (s), and will **NOT** be decided by the proposed number of correct sums (s^*).

Calculation of payoffs:

Payoff of the employer = $(120 - w) * s/10$

The payoff of the employer equals: (120 - the wage he paid to his employee) * the actual number of correct sums solved by the employee/10.

Payoff of the employee = $w - 20$

The payoff of the employee equals: the wage received from the employer - 20.

Thus, the more the questions solved correctly by the employee and the lower the wage, the larger the *employer's* payoffs. The the higher the wage, the larger the *employee's* payoffs.

Example:

If the salary is 50 and actual number of correct sums is 5, then the payoffs would be:

Employer's payoffs = $(120 - 50) * (5/10)$ = 35 points = $17.5

Employee's payoffs = $50 - 20$ = 30 points = $15

Questionnaire:

Just to be sure that you understand the instructions you have to solve a simple questionnaire. When everyone in the room has answered the questionnaire correctly we will start the experiment.

The wage for an employee is 100 and the employer wants the employee to solve 8 number of correct sums. The actual number of correct sums solved by the employee is 3, then the payoffs would be:

Employer's payoff (experimental points):

Employee's payoff (experimental points):

Practice Stage:

Before the start of experiment, you will have the opportunity to perform the addition task so you get familiar with it.

You have **3** minutes to complete the task. You need to correctly solve adding-up sets of five two-digit numbers (e.g., 12+23+35+45+54). You will earn $0.50 for each correct answer, and $0 for each incorrect answer.

Real-effort "C&P" Treatment

Welcome! Thank you for participating in this study.

You will be making REAL decisions involving REAL money. So, please read the instructions carefully. Feel free to ask any questions by raising your hand. Please do not speak to other participants during the experiment. You can use the draft paper we provide, but you are NOT allowed to use the calculator.

The payoffs in the experiment are represented by Experimental Points. Your payment in the experiment will be converted to Dollars at a rate of **1 point = $0.5**. To ensure anonymity and confidentiality you were given a random number decal at the beginning of the experiment. You will need this number to collect your payment.

There are two types of participants: **employer** and **employee**.

At the beginning of the experiment the computer will randomly choose with equal chance whether you will be an employer or an employee. Each person selected to be an employer will be randomly paired with another person in the room who has been selected as an employee.

The identity of the participants will remain anonymous. You will never know the identity of the person you were paired with and the other person will never know your identity.

The experiment consists of **three** stages. The employee chooses whether to make a "claim and promise" or not. The employer pays a wage w to his/her employee, and the employee needs to do an addition task (e.g., 23+34+72+47+97) for the employer. The number of correct sums solved by the employee determines the payoffs of the employer. The details are as follows.

Stage 1 (employee): The employee chooses whether to make the following "claim and promise" or not:

"If the employer offers me the wage_____, I will solve this number of correct sums:_____."

Stage 2 (employer): a) The employer chooses a wage level using an integer number between 20 and 120 (w) for the employee.

b) The employer also proposes a number of correct sums (s^*), that he/she wants the employee to solve.

Stage 3 (employer): After reviewing the wage level set by the employer, the employee has 3 minutes to do the task.

Note: The final payment of the employer will be decided by the employee's actual number of correct sums (s), and will **NOT** be decided by promised or proposed number of correct sums.

Calculation of payoffs:

Payoff of the employer = $(120 - w) * s/10$

The payoff of the employer equals: (120 - the wage he paid to his employee) * the actual number of correct sums solved by the employee/10.

Payoff of the employee = $w - 20$

The payoff of the employee equals: the wage received from the employer - 20.

Thus, the more the questions solved correctly by the employee and the lower the wage, the larger the employer's payoffs. The the higher the wage, the larger the employee's payoffs.

Example:

If the actual wage is 50 and actual number of correct sums is 5, then the payoffs would be:

Employer's payoffs = $(120 - 50) * (5/10)$ = 35 points = $17.5

Employee's payoffs = $50 - 20$ = 30 points = $15

Questionnaire:

Just to be sure that you understand the instructions you have to solve a simple questionnaire. When everyone in the room has answered the questionnaire correctly we will start the experiment.

The employee promised that if the wage is 60, he will provide 6 correct sums. The actual wage for the employee is 100 and the employer wants the employee to solve 8 number of correct sums. The actual number of correct sums solved by the employee is 3, then the payoffs would be:

Employer's payoff (experimental points):

Employee's payoff (experimental points):

<u>Practice Stage</u>

Before the start of experiment, you will have the opportunity to perform the addition task so you get familiar with it.

You have **3** minutes to complete the task. You need to correctly solve adding-up sets of five two-digit numbers (e.g., 12+23+35+45+54). You will earn $0.50 for each correct answer, and $0 for each incorrect answer.

End of the Practice Stage (employee)

You have completed the practice task. The number of correct answer you solved is:

The median number of correct sums in this session is:

Please click the button below to proceed to the main part of the experiment.

End of the Practice Stage (employer)

You have completed the practice task.

The median number of correct sums in this session is:

Please click the button below to proceed to the main part of the experiment.

Role Assignment

The computer has selected you to be an **employee**.

The computer has selected you to be an **employer**.

TOSCA-3 (subjects need to make choice in the scale No(0)—Yes(4) for each A & B)

1. You make a mistake at work and find out a coworker is blamed for the error.

A. You would think: "life is not fair."

B. You would feel unhappy and eager to correct the situation.

2. You have recently moved away from your family, and everyone has been very helpful. A few times you needed to borrow money, but you paid it back as soon as you could.

A. You would be proud that you repaid your debts.

B. You would return the favors as quickly as you could.

3. You are driving down the road, and you hit a small animal.

A. You would feel: "well, it was an accident."

B. You'd feel bad you hadn't been more alert driving down the road.

4. You and a group of coworkers worked very hard on a project. Your boss singles you out for a bonus because the project was such a success.

A. You would feel your hard work had paid off.

B. You would feel you should not accept it.

5. You make a big mistake on an important project at work. People were depending on you, and your boss criticized you.

A. You would think: “Well, nobody’s perfect."

B. You would think: “I should have recognized the problem and done a better job."