

Finding True Beliefs:
Applying Rank-Dependent Expected Utility Theory
to Proper Scoring Rules

A thesis presented
by

Pramod Thammaiah

To

Applied Mathematics

in partial fulfillment of the honors requirements

for the degree of

Bachelor of Arts

Harvard College

Cambridge, Massachusetts

April 1, 2011

Abstract

Proper scoring rules are designed to elicit truthful probability beliefs from expected-value maximizing agents. However, there is evidence that in certain contexts agents are not expected-value maximizers. Thus, we apply rank-dependent expected utility theory, a more general model of decision-making that incorporates probability weighting and non-linear utility functions, to the analysis of the quadratic scoring rule. Current literature provides a way to find an agent’s true beliefs in the case of two outcomes. We have two main theoretical contributions. First, we prove the existence of a unique optimal report and characterize its structure. Second, we use this characterization to find an agent’s true beliefs for any number of outcomes.

We demonstrate the feasibility of our methodology by conducting an experiment on Amazon Mechanical Turk, an online crowdsourcing marketplace. The empirical analysis leads to surprising results. There was no statistically significant difference in performance between the control and treatment groups. In aggregate, subjects were extremely close to bayesian beliefs and there was no evidence of bias. Additionally, there was a statistically significant decrease in performance of adjusted reports over unadjusted reports. We offer three potential explanations: subjects are expected-value maximizers for small gains, subjects did not understand the payment scheme, or the incentives were too small for them to be considered. Applying our methodology and findings to contexts with larger stakes and in-person experimentation are interesting avenues for further research.

Contents

1	Introduction	3
1.1	Motivation	3
1.2	Relevant Literature	4
1.3	Outline	6
2	Background Theory	7
2.1	Proper Scoring Rules	7
2.2	Rank-Dependent Expected Utility Theory	9
3	Solving for True Beliefs	14
3.1	Preliminaries	14
3.2	Ordering Property of the Optimal Report	15
3.3	Rank-Dependent Expected Utility Maximization	21
3.4	Finding the System of Linear Equations	23
3.5	Solving the System of Linear Equations	24
3.6	Finding True Beliefs from Decision Weights	25
3.7	The General Solution	26

<i>CONTENTS</i>	2
4 Experimental Design	27
4.1 Finding the Utility and Probability Weighting Functions	28
4.1.1 Theoretical Method	28
4.1.2 Practical Implementation	30
4.2 Eliciting Probability Reports	32
4.3 Amazon Mechanical Turk	35
5 Empirical Results	38
5.1 Utility Function	38
5.2 Probability Weighting Function	40
5.3 Assessing Performance	41
5.4 Inferring the Probability Weighting Function from Predictions	43
5.5 Distribution of Prediction Errors	45
6 Discussion	48
6.1 Potential Explanations	48
6.2 Future Work	50
6.3 Conclusion	50
7 Acknowledgements	52
A Instructions	53
Bibliography	63

Chapter 1

Introduction

There are many instances where it is useful to elicit the private information of a self-interested expert. One particular context is asking an expert for his beliefs about the probabilities of future outcomes. For example, a policy-maker may need to ask a meteorologist how likely an incoming hurricane will be a category 1, 2, 3, 4, or 5 hurricane. In order to elicit the expert's true beliefs, one could construct a payment mechanism that rewards accuracy. For expected-value maximizing experts, a *proper scoring rule* incentivizes the expert to report truthfully [14]. However, if the expert does not maximize expected value, then the given report may not be the expert's true beliefs. The central contribution of this thesis is a method to determine the true beliefs of an agent under rank-dependent expected utility theory.

1.1 Motivation

Proper scoring rules are an easy and incentive-compatible way to elicit quantitative and precise probability beliefs in both subjective and objective settings [15]. They were first proposed in [7] and have since been used in a wide variety of contexts including accounting [25], business [11], education [8], medicine [19], and politics [20], among others. Given their

application in a wide variety of contexts and potential for even broader application, it is important to consider deviations from the expected-value maximization assumption.

The two deviations from this assumption that we consider are non-linear utilities and non-linear probability weighting. Both deviations are well-documented in the literature and there has been much work in establishing theories of decision-making to account for these phenomenon [16, 13, 21, 22]. The St. Petersburg paradox is an example of a violation of expected-value maximization where individuals choose a safe option rather than a more risky alternative with higher expected value [3]. Similarly, the Allais paradox is an important example of a violation of expected utility theory that prompted work into non-linear weighting of probabilities [13]. Individuals tend to overweight small probabilities and underweight large probabilities [22], offering a potential source of bias in the reports given by them.

The evidence against the expected-value model suggests that it is possible that agents incentivized by proper scoring rules are not revealing their true beliefs. Thus, finding a method that corrects for utility curvature and probability weighting, will enable one to improve the accuracy of the predictions.

1.2 Relevant Literature

There have been a few papers addressing deviations from the expected value assumption in proper scoring rules. However, those that have included probability weighting in their analysis have restricted themselves to only cases that have two outcomes.

An analysis of the effects on proper scoring rules of non-linear utility with expected utility maximization is undertaken in [24]. They consider a number of functional forms for the utility function and use numerical methods to determine the effects on the reports given. They find that risk-takers, agents with convex utilities, tend to overestimate the probability of the most likely event. Similarly, those with concave utilities, implying risk-averseness,

overestimate the probability of less likely events in order to “hedge” the payments. However, the paper does not consider the potential role of probability weighting.

The working paper [2] considers adjusting for risk-preferences and probability weighting through econometric methods. They illustrate how to estimate both risk attitudes and subjective probabilities using maximum likelihood methods under both expected utility and rank-dependent utility theory. They estimate risk attitudes from a task with objective probabilities as well as a task with subjective probabilities. Their work focuses on the quadratic and linear scoring rules.

The paper most closely related to our work is [15] which considers both non-linear utilities and probability weighting in the context of the quadratic scoring rule. They are able to characterize a solution for adjusting the report given by the agent for the case of two outcomes. We build off their findings and confirm their work in our general solution.

Additionally, [15] measures a “risk-correction curve” to correct subjective probability reports without assuming a complex model of decision-making. Instead, they observe reports for outcomes with objectively known probabilities and use these to calibrate the correction mechanism. Implicitly, they assume that there is some function $R(\cdot)$ that maps a probability belief p to the reported probability belief $R(p)$. They solicit measurements for $R(p)$ for many values of p in the following manner: take two ten-sided dice and let the first die represent the ten’s digit and the second die represent the one’s digit, and ask subjects to give reports on outcomes such as “The outcome of the roll with two 10-sided dice is in the range 01-25”. Thus, if they are given a report r in a setting with an unknown probability, they can infer the true belief of the agent by taking $R^{-1}(r)$. The advantage of this method is that it does not assume a specific decision-making model. However, its main limitation is that it cannot be easily generalized to more than two outcomes.

1.3 Outline

This thesis has the following format: chapter 2 covers the background theory needed to understand our main results. In particular, it will explain rank-dependent expected utility theory and proper scoring rules. Chapter 3 presents our main theoretical contributions and demonstrates how we derived a general solution for finding an agent's beliefs.

Next we explain our experiment and empirical results. Chapter 4 covers the experimental design. Chapter 5 presents the results of our experiment and the data analysis. Finally, 6 is the discussion of the implications of our work and suggestions for future research.

Chapter 2

Background Theory

This chapter formally explains the necessary theoretical background to understand the main results of this thesis. There are two sections, the first covering proper scoring rules and the second covering rank-dependent expected utility theory.

2.1 Proper Scoring Rules

For our purposes we only need to define proper scoring rules in the discrete probability space; a more thorough characterization of proper scoring rules can be found in [9, 14].

Suppose we have future event E that has n mutually exclusive and exhaustive outcomes denoted $1, 2, \dots, n$. An agent is asked to give a report $\vec{r} = (r_1, r_2, \dots, r_n)$ on the probabilities of each outcome occurring. We have that r_i corresponds to the probability that outcome i occurs and $\sum_{i=1}^n r_i = 1$. A *scoring rule* $s(\cdot)$ is a reward function that given a report \vec{r} will pay $s(\vec{r}, i)$ if i occurs. We restrict $s(\vec{r}, i)$ to $[-\infty, \infty)$ as defined in [18].

Suppose that the agent has beliefs $\vec{p} = (p_1, p_2, \dots, p_n)$ such that he believes outcome i occurs with probability p_i . Then $s(\cdot)$ is a *proper scoring rule* if the expected value of the payment is maximized, from the agent's perspective, when $\vec{r} = \vec{p}$. A *strictly proper scoring*

rule is a scoring rule whose expected value is uniquely maximized with $\vec{r} = \vec{p}$. [17, 4] describe a number of strictly proper scoring rules such as the quadratic, spherical, and logarithmic scoring rules.

For the rest of this thesis we will focus on the quadratic scoring rule because it is a commonly used proper scoring rule [15]. Additionally, unlike the logarithmic scoring rule, there are finite upper and lower bounds for the payments, which is desirable in an experimental setting. The quadratic scoring rule is defined as follows:

$$s(\vec{r}, i) = 2r_i - \sum_{k=1}^n r_k^2$$

The payments fall in the interval $[-1, 1]$. However, in an experimental setting we want to avoid negative payments because it is difficult to take away money from subjects relative to giving them money. Additionally, we would have to account for loss aversion effects in our analysis [13, 22]. Thus, we make use of the fact that strictly proper scoring rules remain strictly proper under linear transformation [4] and get the following, more general form:

$$s(\vec{r}, i) = b \left(a + 2r_i - \sum_{k=1}^n r_k^2 \right)$$

This scoring rule has a maximum payoff of $b(a + 1)$ when $r_i = 1$ and outcome i occurs. It has a minimum payoff of $b(a - 1)$ when $r_i = 0$, $r_j = 1$ for some $i \neq j$, and outcome i occurs. For illustrative purposes, we show how under expected-value maximization, the quadratic scoring rule results in $\vec{r} = \vec{p}$ as desired. The maximization problem is defined below:

$$\max_{\vec{r}} \sum_{j=1}^n p_j s(\vec{r}, j) \quad \text{s.t.} \quad \sum_{j=1}^n r_j = 1$$

Note that a maximum exists because the feasible set is convex and the objective function is bounded. We can solve using first-order conditions because we have a strictly concave objective function (shown in section 3.2). We proceed by using the lagrangian and substituting

in the quadratic scoring rule:

$$\max_{\vec{r}} \quad ba \sum_{j=1}^n p_j + \sum_{j=1}^n 2bp_j r_j - b \left(\sum_{j=1}^n r_j^2 \right) \sum_{j=1}^n p_j + \lambda \left(1 - \sum_{j=1}^n r_j \right)$$

Note that $\sum_{j=1}^n p_j = 1$. If we take the first-order condition with respect to r_j we get the following:

$$2bp_j - 2br_j^* - \lambda = 0$$

We get $\lambda = 0$ by summing the FOC over all j and noting that $\sum_{j=1}^n p_j = \sum_{j=1}^n r_j^* = 1$. Thus, the optimal report of the agent is $r_j^* = p_j$ for $j = 1, 2, \dots, n$ as desired.

2.2 Rank-Dependent Expected Utility Theory

Expected utility theory is the standard model in economics for understanding decision-making agents in uncertain situations [21]. It is popular for its convenient mathematical properties. However, there are many empirical violations of its basic axioms. The independence axiom states that given lotteries L_1, L_2, L_3 with $L_1 \succeq L_2$, then for all $p \in [0, 1]$ we have $[p, L_1; 1 - p; L_3] \succeq [p, L_2; 1 - p; L_3]$. It is particularly controversial because it implies linear weighting of probabilities over different outcomes.

We present an example of the Allais paradox found in [13] as a violation of the independence axiom and a motivation for considering non-linear probability weighting. Consider the following two sets of binary choices:

- Choice A: \$3000 with 100% probability
- Choice B: \$4000 with 80% probability

The majority of people preferred A to B. Now consider the second choice:

- Choice C: \$3000 with 25% probability

- Choice D: \$4000 with 20% probability

In this case, the majority of people preferred D to C. To see how these preferences violate the independence axiom let $L_1 = [1, \$3000]$, $L_2 = [0.8, \$4000; 0.2, \$0]$, and $L_3 = [1, \$0]$. We have that $L_1 \succeq L_2$ from A being preferred over B, but then have $[0.25, L_2; 0.75, L_3] \succeq [0.25, L_1; 0.75, L_3]$.

To resolve this paradox we adopt non-linear probability weighting and a relaxed version of the independence axiom described in [16, 21]. The probability weighting function w satisfies the following properties [5]:

1. $w : [0, 1] \rightarrow [0, 1]$
2. w is strictly increasing. Thus, it has a well-defined inverse.
3. $w(0) = 0$ and $w(1) = 1$

Empirically, it has been seen that for small p , $w(p) > p$ and for large p , $w(p) < p$ [22]. The following graph illustrates the curvature of the probability weighting function:

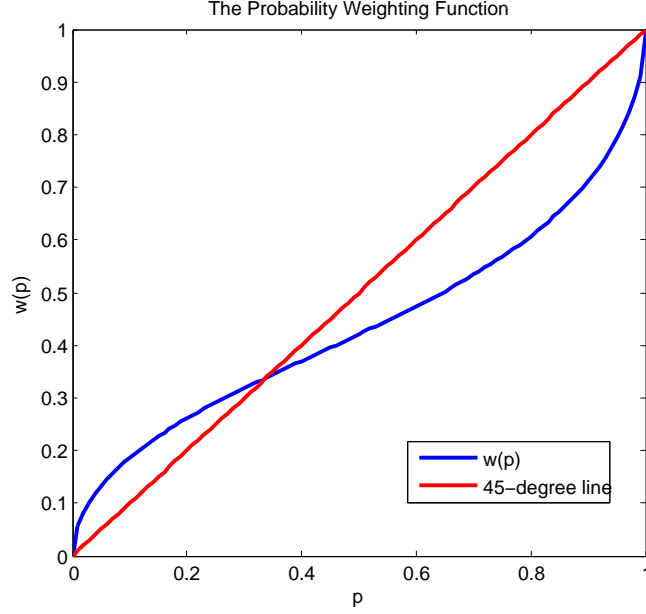


Figure 2.1: A common functional form ($w(p) = \frac{p^\gamma}{[p^\gamma + (1-p)^\gamma]^{\frac{1}{\gamma}}}$ with $\gamma = 0.61$) used in [22]

There are many decision-theories that incorporate probability weighting and are discussed in [21]. According to [21], rank-dependent expected utility (RDEU) theory, as proposed by [16], is one of the better candidates for replacing expected utility theory. Two papers, [2, 15], that also explore deviations from expected value maximization in the context of proper scoring rules, use rank-dependent expected utility theory as the decision-making model. We limit our description to the discrete version of the theory needed for our purposes.

This formulation is based in large part of the explanation given in [5]. A defining characteristic of RDEU is that the order of preferences over outcomes impacts the decision-weights assigned to outcomes. Additionally, decision weights for an outcome do not depend solely on the probability of that outcome occurring but the entire probability distribution. Suppose we have a set of outcomes $\vec{x} = (x_1, x_2, \dots, x_n)$ and an associated probability

distribution $\vec{p} = (p_1, p_2, \dots, p_n)$ where outcome x_i occurs with probability p_i , $\sum_{i=1}^m p_m = 1$, and $x_1 \succeq x_2 \succeq \dots \succeq x_m$. Then the agent maximizes his RDEU:

$$RDEU[p_1, x_1; \dots; p_m, x_m] = \sum_{i=1}^m \pi_i u(x_i) \quad (2.1)$$

where the decision-weights are defined as follows:

$$\pi_i = w\left(\sum_{j=1}^i p_j\right) - w\left(\sum_{j=1}^{i-1} p_j\right)$$

Note that $\pi_1 = w(p_1) - w(0) = w(p_1)$. Additionally, we can see that $\sum_{i=1}^n \pi_i = (w(1) - w(\sum_{j=1}^{n-1} p_j)) + w(\sum_{j=1}^{n-1} p_j) - w(\sum_{j=1}^{n-2} p_j) + \dots + (w(p_1) - w(0)) = w(1) - w(0) = 1$.

We now show how rank-dependent expected utility theory can resolve the Allais Paradox. Assume without loss of generality that $U(0) = 0$, then we have:

$$A \succ B \Rightarrow U(3000) > w(0.8)U(4000)$$

$$C \prec D \Rightarrow w(0.25)U(3000) < w(0.2)U(4000)$$

Suppose we use the functional form described above, $w(p) = \frac{p^\gamma}{[p^\gamma + (1-p)^\gamma]^{\frac{1}{\gamma}}}$ with $\gamma = 0.61$, then we get the following values (rounded to two decimals):

$$w(0.8) = 0.61$$

$$w(0.25) = 0.29$$

$$w(0.2) = 0.26$$

Thus, we find that:

$$0.61 = w(0.8) < \frac{U(3000)}{U(4000)} < \frac{w(0.2)}{w(0.25)} = 0.90$$

which is consistent with rank-dependent expected utility theory.

Chapter 3

Solving for True Beliefs

In this chapter we present our main contributions: Assuming a quadratic scoring rule and rank-dependent expected utility theory, we characterize the ordering of the optimal report given by the agent and a general solution for finding an agent's true beliefs. Extending the solution to any number of outcomes enables correcting an agent's report in a wider array of settings than the solutions presented by [15, 2], which consider situations with only two possible outcomes.

3.1 Preliminaries

We strive to find as general a solution as possible and thus, make few assumptions about the functional form of the utility and probability weighting functions. We have a set-up as described in the previous chapter. There is a future event with n mutually exclusive and exhaustive outcomes. The agent believes in the probability distribution $\vec{p} = (p_1, p_2, \dots, p_n)$ over the outcomes with all $p_i > 0$. As described before, she is incentivized with the general quadratic scoring rule defined in section 2.1 with $b > 0$ and reports $\vec{r} = (r_1, r_2, \dots, r_n)$ such that \vec{r} maximizes her rank-dependent expected utility. Note that $\sum_{i=1}^n p_i = \sum_{i=1}^n r_i = 1$.

We assume that her utility function $U(x)$ is concave, strictly increasing, and bounded over

$[b(a-1), b(a+1)]$, the interval in which all possible payments fall into. Additionally, $U(x)$'s first derivative is defined and bounded over the same interval. For notational convenience, we let x denote the payment of the scoring rule and omit the endowed wealth of the agent. We also assume that the probability weighting function $w(\cdot)$ is strictly increasing. Thus, we have that there exists a well-defined inverse of $w(\cdot)$ over $[0, 1]$.

We define an ordering function σ as follows: $\sigma(i) = j$ if p_j is the i th largest element in \vec{p} . Ties are broken so that the element with the smaller subscript is ranked higher. Intuitively, $\sigma(1)$ is the index of the largest element, $\sigma(2)$ is the index of the 2nd largest element, and so on so that $\sigma(n)$ is the index of the smallest element.

Assuming that the optimal report has the same ordering as \vec{p} , a claim we will prove in the next section, the maximization problem the agent faces is:

$$\max_{\vec{r}} RDEU(\vec{r}) = \sum_{i=1}^n \pi_i U(s(\vec{r}, \sigma(i))) \quad s.t. \quad \sum_{i=1}^n r_i = 1$$

3.2 Ordering Property of the Optimal Report

The ordering of preferences over outcomes is needed for calculating an agent's RDEU. Thus, we note that if r_i is the k th largest element of \vec{r} , then outcome i is the k th most preferred outcome because the payment in outcome i increases with r_i . Thus, if r_i is the largest element of \vec{r} , then outcome i is the most preferable outcome. In this section we prove the order of the optimal report which enables us to apply rank-dependent expected utility theory to the quadratic scoring rule in the general case.

We need an intermediary result about the concavity of the objective function:

Lemma 1. *$RDEU(\vec{r})$ is strictly concave over the convex feasible set $\sum_{i=1}^n r_i = 1$.*

Proof. It is sufficient to show that $RDEU(\vec{r})$ is the strictly positive weighted sum of strictly concave functions [6]. We have that each π_i is strictly positive, because all $p_i > 0$ and $w(\cdot)$

is strictly increasing. Thus, all that remains to be shown is that for each i and all \vec{r} in the feasible set, $U(s(\vec{r}, i))$ is strictly concave.

Let A be a convex set. Let $f : \mathbb{R} \rightarrow \mathbb{R}$ be a concave and strictly increasing function. Let $g : A \rightarrow \mathbb{R}$ be a strictly concave function. By definition of strict concavity, $g(\lambda x + (1 - \lambda)y) > \lambda g(x) + (1 - \lambda)g(y)$, for all $\lambda \in (0, 1)$ and $x, y \in A$, $x \neq y$. Because f is strictly increasing, $f(g(\lambda x + (1 - \lambda)y)) > f(\lambda g(x) + (1 - \lambda)g(y))$. By f 's concavity we have $f(\lambda g(x) + (1 - \lambda)g(y)) \geq \lambda f(g(x)) + (1 - \lambda)f(g(y))$. Combining these, we have

$$f(g(\lambda x + (1 - \lambda)y)) > \lambda f(g(x)) + (1 - \lambda)f(g(y))$$

showing that the composition $f \circ g$ is strictly concave. We have equality between the two sides only if $\lambda \in \{0, 1\}$.

$U(\cdot)$ is concave and strictly increasing. Thus, all that remains to be shown is that $s(\vec{r}, i)$ is strictly concave for each i . Let \vec{x}, \vec{y} be distinct reports in the feasible set and $\lambda \in (0, 1)$.

$$\begin{aligned} \vec{x} \neq \vec{y} &\Leftrightarrow \sum_{k=1}^n (x_k - y_k)^2 > 0 \\ &\Leftrightarrow \sum_{k=1}^n x_k^2 - 2x_k y_k + y_k^2 > 0 \\ &\Leftrightarrow \sum_{k=1}^n \lambda(1 - \lambda)x_k^2 - 2\lambda(1 - \lambda)x_k y_k + \lambda(1 - \lambda)y_k^2 > 0 \\ &\Leftrightarrow \sum_{k=1}^n (\lambda - \lambda^2)x_k^2 - 2\lambda(1 - \lambda)x_k y_k + (1 - \lambda)(1 - (1 - \lambda))y_k^2 > 0 \\ &\Leftrightarrow \sum_{k=1}^n \lambda x_k^2 + (1 - \lambda)y_k^2 > \sum_{k=1}^n \lambda^2 x_k^2 + 2\lambda(1 - \lambda)x_k y_k + (1 - \lambda)^2 y_k^2 \\ &\Leftrightarrow -\sum_{k=1}^n \lambda x_k^2 + (1 - \lambda)y_k^2 < -\sum_{k=1}^n (\lambda x_k + (1 - \lambda)y_k)^2 \end{aligned}$$

Recall that we require $b > 0$:

$$\begin{aligned}
& \Leftrightarrow b \left(a + 2(\lambda x_i + (1 - \lambda)y_i) - \sum_{k=1}^n \lambda x_k^2 + (1 - \lambda)y_k^2 \right) \\
& < b \left(a + 2(\lambda x_i + (1 - \lambda)y_i) - \sum_{k=1}^n (\lambda x_k + (1 - \lambda)y_k)^2 \right) \\
& \Leftrightarrow \lambda s(\vec{x}, i) + (1 - \lambda)s(\vec{y}, i) < s(\lambda \vec{x} + (1 - \lambda)\vec{y}, i) \\
& \Leftrightarrow s(\vec{r}, i) \text{ is strictly concave}
\end{aligned}$$

Thus, $RDEU(\vec{r})$ is strictly concave over the convex feasible set $\sum_{i=1}^n r_i = 1$ as desired. \square

Now we can prove an ordering property about the report that the agent will select. Recall the definition of the ordering function σ as follows: $\sigma(i) = j$ if p_j is the i th largest element in \vec{p} . Ties are broken so that the element with the smaller subscript is ranked higher. Thus, $\sigma(1)$ is the index of the largest element, $\sigma(2)$ is the index of the second largest element, and so on. For example, say $\vec{p} = (0.4, 0.2, 0.4)$ then $\sigma(1) = 1$, $\sigma(2) = 3$, and $\sigma(3) = 2$. Since σ is a bijection from $\{1, 2, \dots, n\}$ to $\{1, 2, \dots, n\}$, we have well-defined inverse. If $\sigma^{-1}(i) = j$, then p_i is the j th largest element in \vec{p} . Essentially, $\sigma^{-1}(i)$ gives the ranking of p_i .

Theorem 2. *There exists a unique optimal $\vec{r}^* = (r_1^*, r_2^*, \dots, r_n^*)$, such that*

$$r_{\sigma(1)}^* \geq r_{\sigma(2)}^* \geq \dots \geq r_{\sigma(n)}^*$$

Proof. The feasible set for \vec{r} is convex and the objective function, rank-dependent expected utility, is bounded over the feasible set. Thus, we know that an optimal \vec{r}^* that maximizes RDEU exists. Because the objective function is strictly concave, any extrema is a unique global maximum [10]. Let \vec{r}^* be this optimal solution.

For convenience we define another ordering function δ that is defined in a similar way as σ except that $\delta(i) = j$ if r_j^* is the i th largest element in \vec{r}^* . Ties are broken in a modified way. If $r_i^* = r_j^*$, but $\sigma^{-1}(i) < \sigma^{-1}(j)$ then i is ranked higher. Equivalently, $\delta^{-1}(i) < \delta^{-1}(j)$.

Intuitively, δ breaks ties by whichever index is ranked first in the sequence defined by σ . It is sufficient to show that $\sigma(i) = \delta(i)$, for $i = 1, 2, \dots, n$.

For the sake of contradiction, assume that $\sigma(i) \neq \delta(i)$ for some i . Let x be the smallest integer such that $\sigma(x) \neq \delta(x)$. Note then $\sigma(x) = \delta(y)$ for some $y > x$. Similarly, $\delta(y-1) = \sigma(z)$ for some $z > x$. The following illustrative diagram helps understand the relationship:

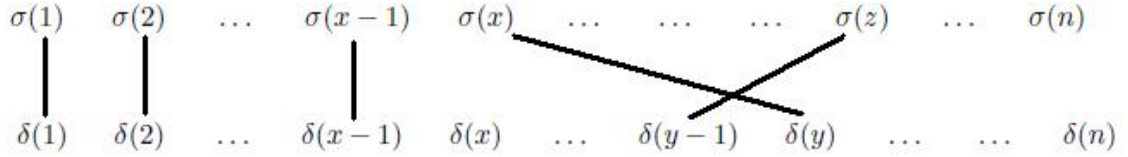


Figure 3.1: The solid lines represent equality.

By definition of σ we have $p_{\sigma(x)} \geq p_{\sigma(z)}$, implying that $p_{\delta(y)} \geq p_{\delta(y-1)}$. Similarly, by definition of δ we have that $r_{\delta(y-1)}^* \geq r_{\delta(y)}^*$. However, consider if $r_{\delta(y-1)}^* = r_{\delta(y)}^*$, by the tie-breaking rule, $\sigma^{-1}(\delta(y-1)) < \sigma^{-1}(\delta(y)) \Rightarrow \sigma^{-1}(\sigma(z)) < \sigma^{-1}(\sigma(x)) \Rightarrow z < x$. This is a contradiction, thus, $r_{\delta(y)}^* \neq r_{\delta(y-1)}^*$.

We will construct a new report \vec{t} such that $RDEU(\vec{t}) \geq RDEU(\vec{r}^*)$, leading to a contradiction. Define \vec{t} as follows:

$$t_{\delta(i)} = \begin{cases} r_{\delta(y)}^* & \text{if } i = y-1 \\ r_{\delta(y-1)}^* & \text{if } i = y \\ r_{\delta(i)}^* & \text{otherwise} \end{cases}$$

The RDEU of \vec{r}^* is as follows:

$$\begin{aligned} RDEU(\vec{r}^*) &= \sum_{i=1}^{y-2} \pi_i U(s(\vec{r}^*, \delta(i))) \\ &\quad + \pi_{y-1} U(s(\vec{r}^*, \delta(y-1))) \end{aligned}$$

$$\begin{aligned}
& + \pi_y U(s(\vec{r}^*, \delta(y))) \\
& + \sum_{i=y+1}^n \pi_i U(s(\vec{r}^*, \delta(i)))
\end{aligned}$$

with $\pi_i = w\left(\sum_{j=1}^i p_{\delta(j)}\right) - w\left(\sum_{j=1}^{i-1} p_{\delta(j)}\right)$. Now we consider \vec{t} :

$$\begin{aligned}
RDEU(\vec{t}) &= \sum_{i=1}^{y-2} \theta_i U(s(\vec{t}, \delta(i))) \\
& + \theta_{y-1} U(s(\vec{t}, \delta(y))) \\
& + \theta_y U(s(\vec{t}, \delta(y-1))) \\
& + \sum_{i=y+1}^n \theta_i U(s(\vec{t}, \delta(i)))
\end{aligned}$$

where we use the definition of decision weights from RDEU to define $\vec{\theta}$. Because none of the first $y-2$ terms have changed, we have that for $i \leq y-2$, $\theta_i = \pi_i$. Note then that:

$$\theta_{y-1} = w\left(p_{\delta(y)} + \sum_{j=1}^{y-2} p_{\delta(j)}\right) - w\left(\sum_{j=1}^{y-2} p_{\delta(j)}\right)$$

and

$$\theta_y = w\left(p_{\delta(y-1)} + p_{\delta(y)} + \sum_{j=1}^{y-2} p_{\delta(j)}\right) - w\left(p_{\delta(y)} + \sum_{j=1}^{y-2} p_{\delta(j)}\right)$$

Thus, we can see that:

$$\begin{aligned}
\theta_{y-1} + \theta_y &= \left(w\left(p_{\delta(y)} + \sum_{j=1}^{y-2} p_{\delta(j)}\right) - w\left(\sum_{j=1}^{y-2} p_{\delta(j)}\right) \right) \\
& + \left(w\left(p_{\delta(y-1)} + p_{\delta(y)} + \sum_{j=1}^{y-2} p_{\delta(j)}\right) - w\left(p_{\delta(y)} + \sum_{j=1}^{y-2} p_{\delta(j)}\right) \right) \\
&= w\left(p_{\delta(y-1)} + p_{\delta(y)} + \sum_{j=1}^{y-2} p_{\delta(j)}\right) - w\left(\sum_{j=1}^{y-2} p_{\delta(j)}\right)
\end{aligned}$$

$$\begin{aligned}
&= \left(w \left(p_{\delta(y-1)} + \sum_{j=1}^{y-2} p_{\delta(j)} \right) - w \left(\sum_{j=1}^{y-2} p_{\delta(j)} \right) \right) \\
&\quad + \left(w \left(p_{\delta(y-1)} + p_{\delta(y)} + \sum_{j=1}^{y-2} p_{\delta(j)} \right) - w \left(p_{\delta(y-1)} + \sum_{j=1}^{y-2} p_{\delta(j)} \right) \right) \\
&= \pi_{y-1} + \pi_y
\end{aligned}$$

Therefore, for $i \geq y+1$, $\theta_i = \pi_i$. The first and last terms of $RDEU(\vec{t})$ and $RDEU(\vec{r})$ are the same since the decision-weights and utilities are equal. Additionally, $U(s(\vec{t}, \delta(y))) = U(s(\vec{r}^*, \delta(y-1)))$ and $U(s(\vec{t}, \delta(y-1))) = U(s(\vec{r}^*, \delta(y)))$ given that \vec{t} and \vec{r}^* simply swap those terms and keep the other elements constant. Therefore,

$$\begin{aligned}
RDEU(\vec{t}) - RDEU(\vec{r}^*) &= \theta_{y-1}U(s(\vec{t}, \delta(y))) + \theta_y U(s(\vec{t}, \delta(y-1))) \\
&\quad - \pi_{y-1}U(s(\vec{r}^*, \delta(y-1))) - \pi_y U(s(\vec{r}^*, \delta(y))) \\
&= (\theta_{y-1} - \pi_{y-1})U(s(\vec{r}^*, \delta(y-1))) + (\theta_y - \pi_y)U(s(\vec{r}^*, \delta(y)))
\end{aligned}$$

Since $\pi_y + \pi_{y-1} = \theta_y + \theta_{y-1}$, we have that $(\theta_y - \pi_y) = -(\theta_{y-1} - \pi_{y-1})$. This gives us:

$$RDEU(\vec{t}) - RDEU(\vec{r}^*) = (\theta_{y-1} - \pi_{y-1})(U(s(\vec{r}^*, \delta(y-1))) - U(s(\vec{r}^*, \delta(y))))$$

Since $r_{\delta(y-1)}^* > r_{\delta(y)}^*$, we have that $U(s(\vec{r}^*, \delta(y-1))) > U(s(\vec{r}^*, \delta(y)))$ because the utility function and quadratic scoring rule are strictly increasing. Using $p_{\delta(y)} \geq p_{\delta(y-1)}$, we see:

$$\begin{aligned}
&\theta_{y-1} - \pi_{y-1} \\
&= \left(w \left(p_{\delta(y)} + \sum_{j=1}^{y-2} p_{\delta(j)} \right) - w \left(\sum_{j=1}^{y-2} p_{\delta(j)} \right) \right) \\
&\quad - \left(w \left(p_{\delta(y-1)} + \sum_{j=1}^{y-2} p_{\delta(j)} \right) - w \left(\sum_{j=1}^{y-2} p_{\delta(j)} \right) \right)
\end{aligned}$$

$$\begin{aligned}
&= w \left(p_{\delta(y)} + \sum_{j=1}^{y-2} p_{\delta(j)} \right) - w \left(p_{\delta(y-1)} + \sum_{j=1}^{y-2} p_{\delta(j)} \right) \\
&\geq 0
\end{aligned}$$

Thus, $RDEU(\vec{t}) \geq RDEU(\vec{r}^*)$. This contradicts the fact that \vec{r}^* is the unique optimal solution. Thus, $\sigma(i) = \delta(i)$ for $i = 1, 2, \dots, n$. \square

For most agents it seems natural and intuitive that the ranks for the report and probability beliefs for outcome i be the same. Since the proper scoring rule $s(\vec{r}, i)$ increases payment with the probability assigned to each event, the event with the highest payoff is $\sigma(1)$, then $\sigma(2)$ and so on. In fact, in our experiment 26 out of 31 subjects gave reports with this ordering property in the situations with known objective probabilities. The 5 remaining subjects displayed convex utilities, so our earlier lemma about the concavity of the objective function no longer held. With the ordering property we can more specifically define the agent's maximization problem:

$$\max_{\vec{r}} RDEU(\vec{r}) = \sum_{i=1}^n \pi_i U(s(\vec{r}, \sigma(i))) \quad s.t. \quad \sum_{i=1}^n r_i = 1$$

3.3 Rank-Dependent Expected Utility Maximization

Since we have that $r_{\sigma(1)}^* \geq r_{\sigma(2)}^* \geq \dots \geq r_{\sigma(n)}^*$ and the quadratic scoring rule is strictly increasing, outcome $\sigma(1)$ is the most preferred outcome. Then $\sigma(2)$ is the second most preferred outcome and so on. We can use first-order conditions to find the agent's optimal report because the objective function is strictly concave over a convex feasible set:

$$\max_{\vec{r}} RDEU(\vec{r}) = \sum_{i=1}^n \pi_i U(s(\vec{r}, \sigma(i))) \quad s.t. \quad \sum_{i=1}^n r_i = 1$$

Using a Lagrangian we have:

$$\max_{\vec{r}} RDEU(\vec{r}) = \sum_{i=1}^n \pi_i U(s(\vec{r}, \sigma(i))) + \lambda \left(\sum_{i=1}^n r_i - 1 \right)$$

We substitute the Quadratic Scoring Rule for $s(\vec{r}, i)$:

$$s(\vec{r}, i) = b \left(a + 2r_i - \sum_{k=1}^n r_k^2 \right)$$

and take the first-order conditions:

$$\begin{aligned} [r_{\sigma(k)}] : & \sum_{i \neq k} \pi_i U'(s(\vec{r}^*, \sigma(i))) (-2br_{\sigma(k)}^*) + \pi_k U'(s(\vec{r}^*, \sigma(k))) (2b - 2br_{\sigma(k)}^*) + \lambda = 0 \\ [\lambda] : & \sum_{i=1}^n r_i^* = 1 \end{aligned}$$

The first FOC simplifies into:

$$r_{\sigma(k)}^* \left(\sum_{i=1}^n -2b\pi_i U'(s(\vec{r}^*, \sigma(i))) \right) + 2\pi_k b U'(s(\vec{r}^*, \sigma(k))) + \lambda = 0$$

Now we sum over all $k \in \{1, 2, \dots, n\}$ and get:

$$\begin{aligned} 0 &= \sum_{k=1}^n \left(r_{\sigma(k)} \sum_{i=1}^n -2b\pi_i U'(s(\vec{r}^*, \sigma(i))) \right) + \sum_{k=1}^n 2b\pi_k U'(s(\vec{r}^*, \sigma(k))) + n\lambda \\ &= \left(\sum_{i=1}^n -2b\pi_i U'(s(\vec{r}^*, \sigma(i))) \right) \left(\sum_{k=1}^n r_{\sigma(k)} \right) + \sum_{k=1}^n 2b\pi_k U'(s(\vec{r}^*, \sigma(k))) + n\lambda \\ &= \left(\sum_{i=1}^n -2b\pi_i U'(s(\vec{r}^*, \sigma(i))) \right) (1) + \sum_{k=1}^n 2b\pi_k U'(s(\vec{r}^*, \sigma(k))) + n\lambda \\ &= 0 + n\lambda \\ &= \lambda \end{aligned}$$

Thus, we have that $\lambda = 0$.

3.4 Finding the System of Linear Equations

From this we notice that the first-order conditions have become a set of linear equations with the variables $\pi_1, \pi_2, \dots, \pi_n$. They take the form of, for $k \in \{1, 2, \dots, n\}$:

$$\begin{aligned} 0 &= \pi_k(2b(1 - r_{\sigma(k)}^*)U'(s(\vec{r}^*, \sigma(k)))) + \sum_{i \neq k} \pi_i(-2br_{\sigma(k)}^*U'(s(\vec{r}^*, \sigma(i)))) \\ &= \pi_k((1 - r_{\sigma(k)}^*)U'(s(\vec{r}^*, \sigma(k)))) + \sum_{i \neq k} \pi_i(-r_{\sigma(k)}^*U'(s(\vec{r}^*, \sigma(i)))) \end{aligned}$$

Put into matrix form we get the system of linear equations is:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n,1} & a_{n,2} & \cdots & a_{n,n} \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{bmatrix}$$

where

$$a_{i,j} = \begin{cases} (1 - r_{\sigma(i)}^*)U'(s(\vec{r}^*, \sigma(j))) & \text{if } i = j \\ -r_{\sigma(i)}^*U'(s(\vec{r}^*, \sigma(j))) & \text{otherwise} \end{cases}$$

However, we can see that this system is solved by the trivial solution $\pi_1 = \pi_2 = \dots = \pi_n = 0$. This is because the sum of the first $n - 1$ rows is identical to the n th row since $r_{\sigma(n)} = 1 - \sum_{i=1}^{n-1} r_{\sigma(i)}$. Thus, we remove the n th row and add in the following constraint: $\sum_{i=1}^n \pi_i = 1$. This is shown in the section 2.2. The final system is:

$$\begin{bmatrix} a_{1,1} & a_{1,2} & \cdots & a_{1,n} \\ a_{2,1} & a_{2,2} & \cdots & a_{2,n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n-1,1} & a_{n-1,2} & \cdots & a_{n-1,n} \\ 1 & 1 & \cdots & 1 \end{bmatrix} \begin{bmatrix} \pi_1 \\ \pi_2 \\ \vdots \\ \pi_{n-1} \\ \pi_n \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ \vdots \\ 0 \\ 1 \end{bmatrix}$$

3.5 Solving the System of Linear Equations

In this section we present the general solution for $\vec{\pi}$.

Claim 3. For $i = 1, 2, \dots, n$, we have:

$$\pi_i = \frac{r_{\sigma(i)}^* \prod_{j \neq i} U'(s(\vec{r}^*, \sigma(j)))}{\sum_{k=1}^n r_{\sigma(k)}^* \prod_{j \neq k} U'(s(\vec{r}^*, \sigma(j)))}$$

Proof. We simply substitute this solution in and show that it solves the system of linear equations. We have that for $i = 1, 2, \dots, n-1$:

$$\begin{aligned} 0 &= \pi_i(1 - r_{\sigma(i)}^*)(U'(s(\vec{r}^*, \sigma(i)))) + \sum_{j \neq i} -\pi_j r_{\sigma(i)}^* U'(s(\vec{r}^*, \sigma(j))) \\ &= \pi_i U'(s(\vec{r}^*, \sigma(i))) + \sum_{j=1}^n -\pi_j r_{\sigma(i)}^* U'(s(\vec{r}^*, \sigma(j))) \\ &= \frac{r_{\sigma(i)}^* \prod_{k=1}^n U'(s(\vec{r}^*, \sigma(k)))}{\sum_{k=1}^n r_{\sigma(k)}^* \prod_{j \neq k} U'(s(\vec{r}^*, \sigma(j)))} - \sum_{j=1}^n \frac{r_{\sigma(i)}^* r_{\sigma(j)}^* \prod_{k=1}^n U'(s(\vec{r}^*, \sigma(k)))}{\sum_{k=1}^n r_{\sigma(k)}^* \prod_{j \neq k} U'(s(\vec{r}^*, \sigma(j)))} \\ &= \left(\frac{r_{\sigma(i)}^* \prod_{k=1}^n U'(s(\vec{r}^*, \sigma(k)))}{\sum_{k=1}^n r_{\sigma(k)}^* \prod_{j \neq k} U'(s(\vec{r}^*, \sigma(j)))} \right) \left(1 - \sum_{j=1}^n r_{\sigma(j)}^* \right) \\ &= \left(\frac{r_{\sigma(i)}^* \prod_{k=1}^n U'(s(\vec{r}^*, \sigma(k)))}{\sum_{k=1}^n r_{\sigma(k)}^* \prod_{j \neq k} U'(s(\vec{r}^*, \sigma(j)))} \right) (0) \\ &= 0 \end{aligned}$$

We can also see that $\sum_{i=1}^n \pi_i = 1$ since the denominator of each π_i is the sum of all the numerators, satisfying the last constraint. Thus, this solution solves the system of linear equations derived earlier. \square

A more convenient formulation is:

$$\pi_i = \left(\frac{\prod_{j=1}^n U'(s(\vec{r}^*, \sigma(j)))}{\sum_{k=1}^n r_{\sigma(k)}^* \prod_{j \neq k} U'(s(\vec{r}^*, \sigma(j)))} \right) \frac{r_{\sigma(i)}^*}{U'(s(\vec{r}^*, \sigma(i)))}$$

3.6 Finding True Beliefs from Decision Weights

Given $\vec{\pi}$ in terms of known values we can now solve for \vec{p} . We can do so inductively using the inverse of the probability weighting function.

Claim 4. For $k \in \{1, 2, \dots, n\}$, $p_{\sigma(k)} = w^{-1} \left(\sum_{i=1}^k \pi_i \right) - w^{-1} \left(\sum_{i=1}^{k-1} \pi_i \right)$

Proof. We proceed with a proof by induction. Let $f(k) = \sum_{i=1}^k p_{\sigma(i)}$. Then we have that $p_{\sigma(k)} = f(k) - f(k-1)$. Thus, it is sufficient to show that $f(k) = w^{-1}(\sum_{i=1}^k \pi_i)$. We consider the base case $f(1) = p_{\sigma(1)}$. By definition $\pi_1 = w(p_{\sigma(1)})$, so taking the inverse $f(1) = p_{\sigma(1)} = w^{-1}(\pi_1)$ as desired.

By the inductive hypothesis, assume $f(k-1) = w^{-1}(\sum_{i=1}^{k-1} \pi_i)$. Now we show that $f(k) = w^{-1}(\sum_{j=1}^k \pi_j)$ to complete the proof:

$$\begin{aligned} \pi_k &= w \left(\sum_{i=1}^k p_{\sigma(i)} \right) - w \left(\sum_{i=1}^{k-1} p_{\sigma(i)} \right) \Rightarrow \pi_k = w(f(k)) - w(f(k-1)) \\ &\Rightarrow w^{-1}(\pi_k + w(f(k-1))) = f(k) \\ &\Rightarrow w^{-1} \left(\pi_k + w \left(w^{-1} \left(\sum_{i=1}^{k-1} \pi_i \right) \right) \right) = f(k) \\ &\Rightarrow w^{-1} \left(\pi_k + \sum_{i=1}^{k-1} \pi_i \right) = f(k) \\ &\Rightarrow w^{-1} \left(\sum_{i=1}^k \pi_i \right) = f(k) \end{aligned}$$

□

3.7 The General Solution

We simply combine are results from the previous two sections to get a general solution. Let

$$c = \frac{\prod_{j=1}^n U'(s(\vec{r}^*, \sigma(j)))}{\sum_{k=1}^n r_{\sigma(k)}^* \prod_{j \neq k} U'(s(\vec{r}^*, \sigma(j)))}$$

For $i = 1, 2, \dots, n$:

$$p_{\sigma(i)} = w^{-1} \left(c \sum_{k=1}^i \frac{r_{\sigma(k)}^*}{U'(s(\vec{r}^*, \sigma(k)))} \right) - w^{-1} \left(c \sum_{k=1}^{i-1} \frac{r_{\sigma(k)}^*}{U'(s(\vec{r}^*, \sigma(k)))} \right)$$

The final step to find $\vec{p} = (p_1, p_2, \dots, p_n)$ is to use σ^{-1} to assign the probabilities to their respective outcomes. Thus, we have found a solution for finding the agent's true beliefs for any number of outcomes.

Chapter 4

Experimental Design

This chapter describes the experiment we conducted on Amazon Mechanical Turk to test our theoretical results. We describe the various sections of the study and the reasons for design choices. Section 4.1 gives a description of both the theory and implementation in eliciting the utility and probability weighting functions. Section 4.2 outlines the types of questions we used to elicit probability reports. And section 4.3 is a discussion of the pros and cons of implementing the experiment on Amazon Mechanical Turk and how we adjusted the design accordingly. The instructions for the study can be found in the appendix A.

There were 5 sections to our study. Section 1 was used to elicit each subject's utility and probability weighting functions. Sections 2-5 were prediction tasks with two predictions in each section. Thus, we solicited eight predictions from each subject.

All subjects were paid a base rate of \$0.50 for completing the study. Subjects in the treatment group were incentivized in the prediction tasks with a quadratic scoring rule. They were told one of the eight predictions they made would be chosen at random to determine their bonus. Thus, they were told to try their best on all of the questions. The bonus paid was between \$0.00 and \$1.00. Those in the control group performed the same tasks as those in the treatment group, but the prediction tasks were framed such that the

bonuses were hypothetical.

4.1 Finding the Utility and Probability Weighting Functions

This section will describe the theory and implementation of section 1 in the study, where we elicited the probability weighting and utility functions.

4.1.1 Theoretical Method

We used the Trade-Off method described in [5, 23] to solicit the utility and probability weighting functions of an agent. This method was selected because it is robust against probability weighting, making it compatible with rank-dependent expected utility. Additionally, it does not assume any functional form of either function. Instead, we are able to calculate each function's value at specific points. This provides flexibility because we can simply linearly interpolate between points, fit a parametric-model by setting parameters, or both.

The first step is to solicit the utility function. We select the following four parameters R, r, x_0 , and p such that $R \succ r \succ x_0$ and $p \in [0, 1]$. We ask the agent for a value of x_j that makes him indifferent between the following lotteries: $[p, R; 1 - p, x_{j-1}]$ and $[p, r; 1 - p, x_j]$ for j equal to 1, 2, and so on. By rank-dependent expected utility theory, we have the following:

$$\begin{aligned}
 & [p, R; 1 - p, x_{j-1}] \sim [p, r; 1 - p, x_j] \\
 & \Leftrightarrow \pi_1 U(R) + \pi_2 U(x_{j-1}) = \pi_1 U(r) + \pi_2 U(x_j) \\
 & \Leftrightarrow w(p)U(R) + (1 - w(p))U(x_{j-1}) = w(p)U(r) + (1 - w(p))U(x_j) \\
 & \Leftrightarrow w(p)(U(R) - U(r)) = (1 - w(p))(U(x_j) - U(x_{j-1})) \\
 & \Leftrightarrow \frac{w(p)}{1 - w(p)}(U(R) - U(r)) = U(x_j) - U(x_{j-1})
 \end{aligned}$$

$$\Rightarrow \forall 0 < i, j \leq t, U(x_j) - U(x_{j-1}) = U(x_i) - U(x_{i-1})$$

Note that in the binary case $\pi_1 = w(p_1) - w(0) = w(p)$ and $\pi_2 = w(p_1 + p_2) - w(p_1) = w(1) - w(p) = 1 - w(p)$. We repeat this process until we get to a value t such that $x_{t+1} \succ r \succ x_t$. At this point, the rankings over outcomes in the second lottery have changed so that the analysis no longer holds. Given this result, we can simply define $U(x_0) = 0$ and $U(x_1) = 1$ and thus, for all $0 \leq j \leq t$, $U(x_j) = j$. As a result, we have a utility function defined over the interval $[x_0, x_t]$. Of course, the utility function can be scaled as long as $U(x_1) - U(x_0) = U(x_{j+1}) - U(x_j)$, a fact we make use of in section 5.1.

Now that we have the utility function, we elicit the probability weighting function. For a given value p that we want to find the value of $w(p)$, we do the following: if p is small, we ask for z_r such that they are indifferent between $[p, y_i; 1 - p, y_j]$ and $[p, y_k; 1 - p, z_r]$ where $y_k \geq y_i \geq y_j$. If p is large, we ask for z_s such that they are indifferent between $[p, y_m; 1 - p, y_n]$ and $[p, z_s; 1 - p, y_q]$ where $y_m \geq y_n \geq y_q$. Then we have:

$$\begin{aligned} [p, y_i; 1 - p, y_j] &\sim [p, y_k; 1 - p, z_r] \\ \Leftrightarrow w(p)U(y_i) + (1 - w(p))U(y_j) &= w(p)U(y_k) + (1 - w(p))U(z_r) \\ \Leftrightarrow w(p) &= \frac{U(y_j) - U(z_r)}{U(y_j) - U(z_r) + U(y_k) - U(y_i)} \end{aligned}$$

Similarly,

$$\begin{aligned} [p, y_m; 1 - p, y_n] &\sim [p, z_s; 1 - p, y_q] \\ \Leftrightarrow w(p) &= \frac{U(y_n) - U(y_q)}{U(y_n) - U(y_q) + U(z_s) - U(y_m)} \end{aligned}$$

We need that the y 's, z_r , and z_s are within $[x_0, x_t]$ in order to ensure that the utility is known. Thus, $z_r = x_0$ causes the first equation to reach the upper bound. Similarly, $z_s = x_t$ causes the the second equation to have a lower bound. Steps to avoid this potential problem

described in section 4.1.2. The two different formulas for small and large p are to minimize the chances of exceeding the bounds.

We have explained how to calculate the value of both the utility function and the probability weighting function of an agent under rank-dependent expected utility theory. In the next section we will discuss practical considerations when implementing the method described above.

4.1.2 Practical Implementation

In the experiment, all possible bonuses fell into the range \$0.00 to \$1.00. Thus, we sought to ensure that each subject's utility function is defined over that interval $[0, 1]$. Thus, we set $x_0 = 0$, guaranteeing the lower bound. In order to reduce the risk that $x_t < 1$, we chose r to be substantially greater than 1. We did not have any subjects that had $x_t < 1$. We measured the probability weighting function for $p \in \{0.1, 0.25, 0.5, 0.75, 0.9\}$. Values of $p < 0.5$ were considered small and values of $p \geq 0.5$ were considered large.

The following is a summary of the parameters chosen for the experiment:

$$R = 1.8$$

$$r = 1.55$$

$$x_0 = 0$$

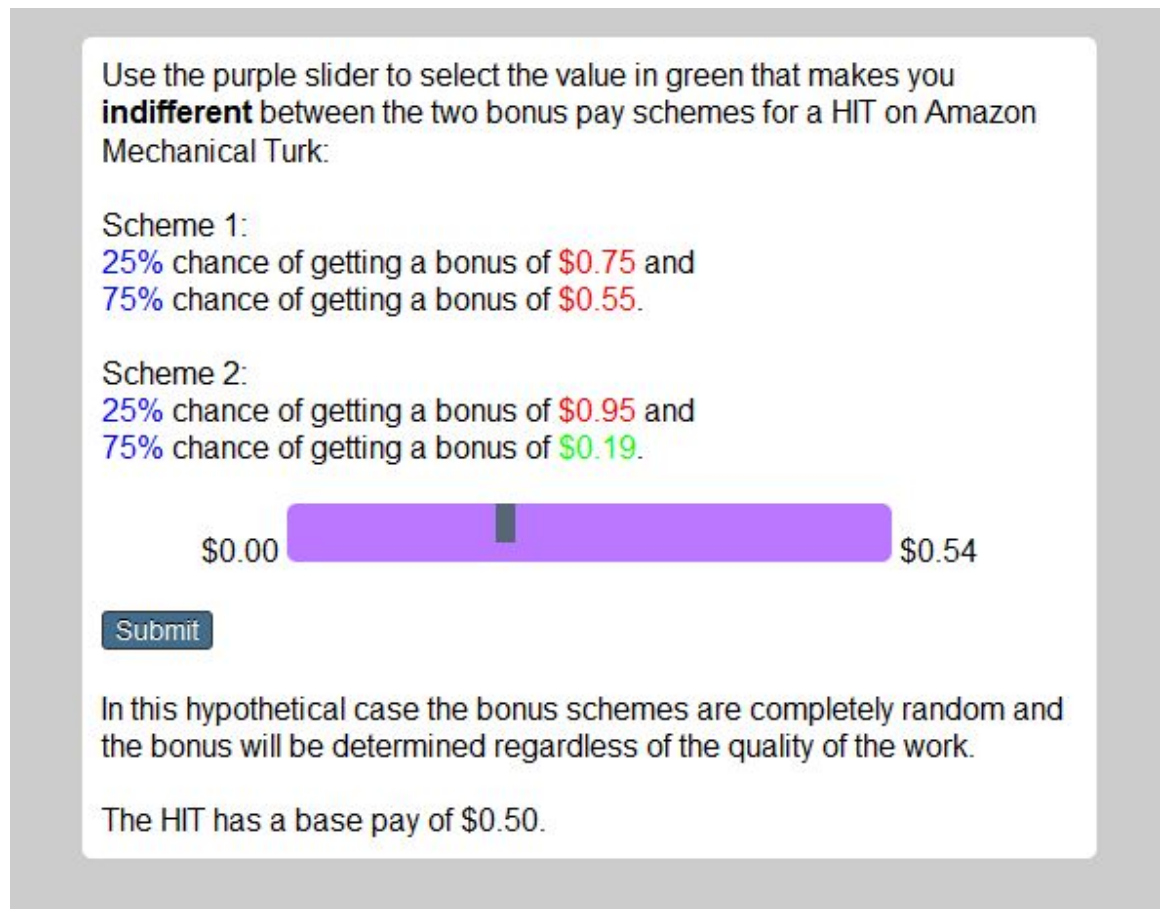
$$p = 0.6$$

As mentioned in section 4.1.1, there is a risk of subjects choosing values for z_r or z_s that hit the boundary conditions. Thus, to minimize this risk we constructed lotteries that provided a large interval of valid choices. As a result, only 2 subjects specified values at or beyond the valid bounds.

There is no need, in practice, for the utility function elicitation to precede the probability

weighting function elicitation. In order to reduce the repetitiveness of the task, we alternated questions asking for the utility function and the probability weighting function.

Below is a screenshot of a question from the utility and probability weighting functions elicitation process:



Use the purple slider to select the value in green that makes you **indifferent** between the two bonus pay schemes for a HIT on Amazon Mechanical Turk:

Scheme 1:
25% chance of getting a bonus of \$0.75 and
75% chance of getting a bonus of \$0.55.

Scheme 2:
25% chance of getting a bonus of \$0.95 and
75% chance of getting a bonus of \$0.19.

\$0.00 \$0.54

In this hypothetical case the bonus schemes are completely random and the bonus will be determined regardless of the quality of the work.

The HIT has a base pay of \$0.50.

Figure 4.1: A screenshot of a question in the first section of the study. Subjects could move the slider to change the value in green.

They are specifying their utility in the context of getting a bonus for a HIT on top of a base pay of \$0.50. This was to ensure that the utility function derived would match up with their utility function in the prediction task.

4.2 Eliciting Probability Reports

In the final four sections of the study we solicited probability predictions from subjects. We followed a 2×2 design by having sections defined by subjective vs. objective probabilities and two vs. three outcomes. Sections 2 and 3 were subjective probability contexts, while sections 4 and 5 were objective probability contexts. Similarly, sections 2 and 4 had two outcomes, while sections 3 and 5 had three outcomes.


In the subjective probability sections, we wanted to simulate a situation where people have ambiguous signals about a future outcome, but cannot calculate a precise probability. This could be similar, for example, determining the probabilities that a particular candidate will win an election. We leveraged the fact that people have difficulty performing bayesian updating without an external aid. The basic setup was that we had k coins and told the subject the probability that each coin lands on heads. We select one of the coins at random, with each coin equally likely to be selected. We then flip it 5 times and tell the subject the number of heads and tails that landed. Then, we ask the subject to give his probability estimates as to which coin was selected. Below is a subjective prediction task with three outcomes. The two outcome situation was the same except it used a slider as in section 1 and two coins.

Coin 1 lands on heads 70% of the time.
Coin 2 lands on heads 50% of the time.
Coin 3 lands on heads 30% of the time.

With each coin having a one-third chance of being selected, I chose one coin at random and discarded the other 2 coins. I flipped the selected coin 5 times. Out of 5 flips, there were 3 heads and 2 tails.

Estimate how likely the selected coin is coin 1, coin 2, or coin 3. Your bonus will depend on how accurate you are.

Coin 1



Coin 2

Payment if it is coin 1: \$0.88

Payment if it is coin 2: \$0.53

Payment if it is coin 3: \$0.40

Your estimate of the probability that it is coin 1: 61%

Your estimate of the probability that it is coin 2: 26%

Your estimate of the probability that it is coin 3: 13%

Valid entry

Submit

Moving the slider towards Coin 1's label increases your estimate for coin 1. Similarly, for coin 2 and coin 3.

Only the green area is valid.

Figure 4.2: A screenshot of a question in the third section of the study. Subjects could move the two-dimensional slider to specify their report.

In order to give subjects some intuition on how to update their beliefs based on the coin flips, we explained some basic probability theory. You can find our explanation in appendix A. This was to increase the likelihood that they understood the basic ideas of updating beliefs without going into the mathematics.


In the objective probability sections, we wanted to see, without the potential noise of improper updating, the predictions that the agent would give. Thus, we essentially gave subjects the probability of an outcome occurring and had them specify the payments they would like to receive in each outcome. We omitted the probabilities that they were implicitly specifying from the screen, because we thought this would lead to subjects choosing the ‘objectively’ correct answer rather than optimizing their utility.

The basic setup is that we have a jar with 100 marbles. There are k different colors and we tell the subject how many of each color are in the jar. We will draw one of the marbles at random, with each marble equally likely to be drawn. The subject must choose the amount of payment he wants in the event each color is drawn. He is implicitly giving a probability report by specifying his optimal payments. As the subject moved the slider, we determined the subject’s implicit report and then converted it to the payments according to the quadratic scoring rule. Below is a screenshot from section 4. Section 5 had three colors and used the two-dimensional slider like in section 3.

I have a jar with 100 marbles. There are 60 red marbles and 40 blue marbles.

I will draw one marble from the jar with each of the 100 marbles equally likely to be chosen. You will receive a bonus based on which color I draw.

Choose the bonus that you would like to receive for each color with the slider below.

Red  Blue

Payment if it is red: \$0.86
Payment if it is blue: \$0.62

Moving the slider towards Red's label increases your payoff for red. Similarly, for blue.

Figure 4.3: A screenshot of a question in the fourth section of the study. Subjects could move the slider to specify the payments they wanted to receive.

4.3 Amazon Mechanical Turk

Amazon Mechanical Turk (AMT) [1] is an online marketplace that allows requesters to post Human Intelligence Tasks (HITs). Requesters specify a base fee for completing the task. Workers can view and search a list of HITs and find one to accept. Once they complete a HIT, workers submit their work. Once the requester has approved the work, the workers are paid. At this time, a requester can choose to pay a bonus to the worker. We used this bonus option to pay the payment from the quadratic scoring rule. A more thorough

description of AMT can be found in [12].

We chose AMT to conduct our study because of the low monetary costs and the speed of data collection. One-hundred subjects (65 in the treatment group and 35 in the control group) completed the study in three days at a cost of \$105.46 (including fees to Amazon). These characteristics enabled us to put our study through many pilot studies to improve subject understanding of the instructions and quality of the data.

In addition to these benefits, however, there are drawbacks to conducting behavioral economics studies on AMT. Since workers are on AMT to earn money, they try to complete as many HITs as possible in a given amount of time to maximize their earning rate. Thus, there is a potential problem that many workers will rush through the study to complete it as quickly as possible. Second, because the study needed to be hosted on a website and the subjects were in different locations, it was not feasible to answer any questions the subject may have while taking the study.

We adjusted to these concerns in a number of ways. To discourage workers that would rush through the task, we explicitly stated in the task preview that the study would take 10 to 15 minutes. In addition, we stated that they would have to spend at least 30 seconds reading each instructions page and 10 seconds on each question. We enforced this on the website and would not allow subjects to click to the next section or question until the minimum amount of time had been passed.

We sought to make the instructions as short as possible, so that subjects would spend time to understand the task and not skip the instructions. In pilot studies, we found that giving the mathematical description of the quadratic scoring rule was not helpful. Testers told us that it was intimidating and unhelpful in understanding how the payment worked. As result, we allowed subjects to voluntarily read the mathematical description of the quadratic scoring rule, but hid it by default. We also, automatically calculated and displayed the payments the subjects would receive for each outcome as they moved the slider around.

Subjects were tested for understanding of the task with some testing questions. These questions had objectively correct answers and can be seen in the appendix A. Workers were permitted to finish the study even if they failed the test questions, but their data was omitted in the analysis. Only subjects that passed all the test questions were used for data analysis.

Chapter 5

Empirical Results

In this section we discuss the empirical findings from the experiment. Thirty-one subjects in the treatment group and 17 subjects in the control group passed all of the test questions. The analyses done refer to the 31 subjects in the treatment group. We make use of the control group when assessing the impact of incentives on performance in section 5.3.

5.1 Utility Function

We solicited the utility function as described in section 4.1.1. Given a sequence x_0, x_1, \dots, x_t where $U(x_i) = i$, we linearly interpolated between the points. We found that every subject had a strictly increasing utility function and $x_t > 1$. However, we focus on the utility function over the interval $[0, 1]$. In order to make a comparison between individuals easier, we normalized the utilities such that $U(0) = 0$ and $U(1) = 1$. Over the interval of \$0.00 to \$1.00, we find that the mean utility of all subjects is remarkably close to $U(x) = x$.

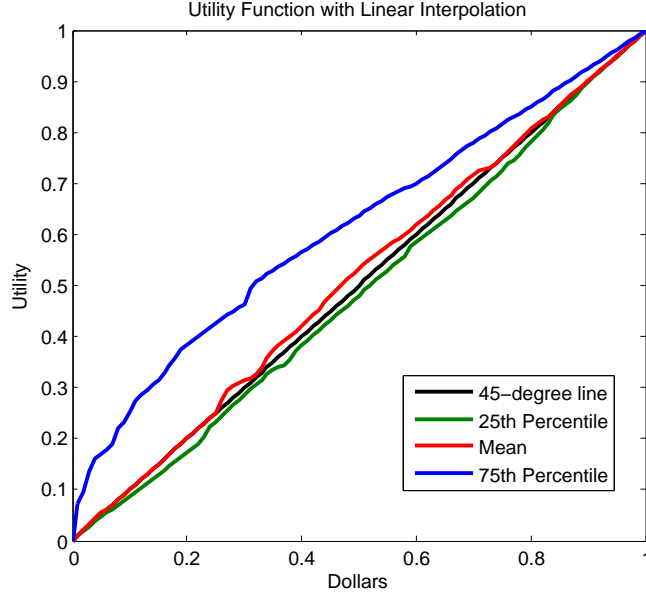


Figure 5.1: The 25th, 50th, and 75th percentiles of the utility functions for subjects in the treatment group.

For the linearly interpolated utility function, the derivative at point x was calculated by using the following approximation:

$$U'(x) = \frac{U(x + \$0.01) - U(x - \$0.01)}{\$0.02}$$

In the special case for $x = 0$, we took the one-sided approximation. We were still able to use the above equation for $x = 1$ because we had the utility functions defined over $[0, x_t]$ where $x_t > 1$.

We also fitted each subject's utility function to the model $U(x) = x^\theta$. We took the linearly interpolated utility function and found the values of $U(x)$ for $x = x_0, x_1, \dots, x_t$. Then we fit the value of θ that minimized the least squared error for these pairs of $(x_i, U(x_i))$.

The sample mean was 0.928 with standard error of 0.058. The distribution is not symmetric and instead, we find that the distribution is skewed towards smaller θ . The following figure gives a histogram of our results:

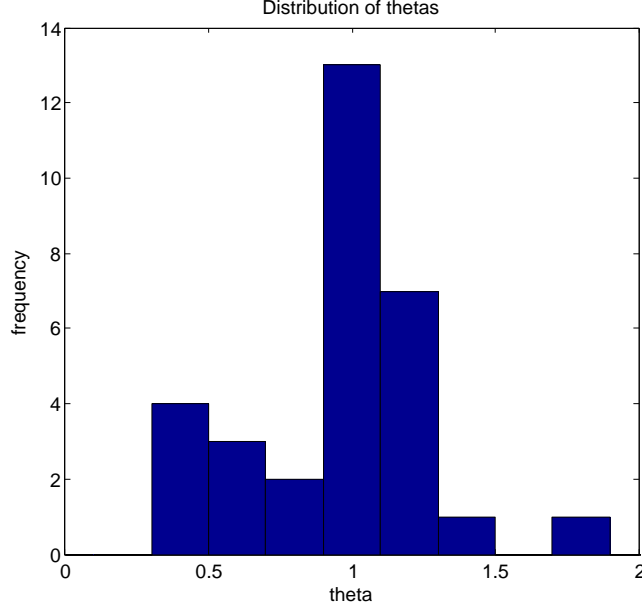


Figure 5.2: A histogram of the distribution of θ for subjects in the treatment group.

5.2 Probability Weighting Function

Using the methodology discussed in section 4.1.1 we were able to find the value of $w(p)$ for $p = 0.1, 0.25, 0.5, 0.75, 0.9$. We attempted to linearly interpolate between these points as done in [15], but found that there were violations in strict monotonicity. In order to find w^{-1} , strict monotonicity is required. Thus, we assume the popular functional form used in [22]:

$$w(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{\frac{1}{\gamma}}}$$

With this functional form, we performed a least squares fitting for each subject using each of the utility functions described in the last section. There was little difference between the sample mean for γ , although there were differences at the individual-level.

Utility Function	sample mean for γ	standard error
$U(x) = x$	0.618	0.046
Linear Interpolation	0.591	0.043
$U(x) = x^\theta$	0.605	0.032

Our findings agree strongly with [22], which estimated $\gamma = 0.61$ as a population mean in the context of gains.

5.3 Assessing Performance

We adopt the Kullback-Leibler divergence between a report and the corresponding bayesian beliefs as a measure of the quality of the report. The K-L divergence between two probability mass functions $p(x)$ and $q(x)$ is defined as:

$$D(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

with the convention that $0 \log \frac{0}{0} = 0$, $0 \log \frac{0}{q(x)} = 0$, and $p(x) \log \frac{p(x)}{0} = \infty$. We let the report being evaluated be p and the bayesian belief be q . This assignment was done to avoid the infinite case since we could guarantee that all the bayesian beliefs were non-zero, but some subjects reported zero for some outcomes. Note that the K-L divergence is always non-negative and zero only if $p = q$. When comparing two reports, the one with a lower K-L divergence is considered the more accurate of the two.

In the prediction task with coins, the bayesian beliefs were just the posterior probabilities using the given information and updating with Bayes' Rule. In the marble case, the bayesian

beliefs are simply the proportion of marbles of each color.

Using this measure of performance, we evaluated all the reports given by the treatment and control groups. We tested for a difference in the median performance with the Wilcoxon Rank-sum test for two unpaired samples. There was no statistically significant difference in any individual prediction task nor in the aggregate.

Prediction Task	treatment median	control median	p-value	Significant at 5%
2 coins	0.072	0.030	0.222	No
3 coins	0.145	0.123	0.722	No
2 colors	0.017	0.042	0.083	No
3 colors	0.056	0.079	0.878	No
Total	0.053	0.053	0.786	No

Next we adjusted the reports given by each subject in the treatment group using the theoretical results in chapter 3. We did 6 different adjustment processes, one for each combination of the 3 utility functions and 2 probability weighting functions. The three utility functions were $U(x) = x$, linearly interpolated, and $U(x) = x^\theta$. The probability weighting functions were $w(p) = p$ and $w(p) = \frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{\frac{1}{\gamma}}}$. However, the adjusted reports were significantly worse than the unadjusted ones in the non-linear probability weighting case. There was no significant change if the linear probability weighting was used. Below we summarize our findings, using the Wilcoxon signed-rank test. All of the prediction tasks were aggregated.

$U(x)$	$w(p)$	Adjusted median	Original median	p-value	Sig. at 5%
$U(x) = x$	linear	0.053	0.053	1	No
	$\frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{\frac{1}{\gamma}}}$	0.289	0.053	< 0.0001	Yes
Linear	linear	0.077	0.053	0.076	No
Interpolation	$\frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{\frac{1}{\gamma}}}$	0.303	0.053	< 0.0001	Yes
$U(x) = x^\theta$	linear	0.063	0.053	0.748	No
	$\frac{p^\gamma}{(p^\gamma + (1-p)^\gamma)^{\frac{1}{\gamma}}}$	0.287	0.053	< 0.0001	Yes

5.4 Inferring the Probability Weighting Function from Predictions

We did not offer any incentives for subjects to answer truthfully in section 1 of the experiment, which was used to infer the utility and probability weighting functions. Given that the probability weighting function that we inferred had violations of monotonicity, it is possible that subjects did not answer truthfully. Thus, in this section, we use the reports from the two objective probability tasks to infer the probability weighting function.

In order to infer the probability weighting function from predictions, we need to assume that $U'(x) = 1$ for all $x \in [0, 1]$, implying $U(x) = x$ as the utility function for all subjects. With this assumption the result from section 3.5 simplifies significantly:

$$\begin{aligned}
\pi_i &= \left(\frac{\prod_{j=1}^n U'(s(\vec{r}^*, \sigma(j)))}{\sum_{k=1}^n r_{\sigma(k)}^* \prod_{j \neq k} U'(s(\vec{r}^*, \sigma(j)))} \right) \frac{r_{\sigma(i)}^*}{U'(s(\vec{r}^*, \sigma(i)))} \\
&= \left(\frac{1}{\sum_{k=1}^n r_{\sigma(k)}^*} \right) \frac{r_{\sigma(i)}^*}{1} \\
&= r_{\sigma(i)}^*
\end{aligned}$$

Since we know the objective probabilities, we can use \vec{p} to find $w(p)$ at multiple points by using the following result.

Claim 5. For $k = 1, 2, \dots, n$:

$$\sum_{i=1}^k \pi_i = w\left(\sum_{i=1}^k p_{\sigma(i)}\right)$$

Proof. We can show this by induction. The base case of $k = 1$, holds because by definition $\pi_1 = w(p_{\sigma(1)}) - w(0) = w(p_{\sigma(1)})$. Assume that the property holds for $k - 1$. We will now consider property for k :

$$\begin{aligned} \sum_{i=1}^k \pi_i &= \pi_k + \sum_{i=1}^{k-1} \pi_i \\ &= \pi_k + w\left(\sum_{i=1}^{k-1} p_{\sigma(i)}\right) \\ &= \left(w\left(\sum_{i=1}^k p_{\sigma(i)}\right) - w\left(\sum_{i=1}^{k-1} p_{\sigma(i)}\right)\right) + w\left(\sum_{i=1}^{k-1} p_{\sigma(i)}\right) \\ &= w\left(\sum_{i=1}^k p_{\sigma(i)}\right) \end{aligned}$$

as desired. □

Combining these two results enables us to use \vec{r}^* and \vec{p} , which are known to us and the subjects in the objective prediction tasks, to find values of $w(\cdot)$ at multiple points. We then fit these points to the functional form described in section 5.2. The values of γ in this case were close to 1, implying a minimal amount of probability weighting. The mean of these γ 's was 1.053 with a standard error of 0.03. We found again that adjusting for this weighting function of the predictions in the subjective case did yield any significant results. This is not surprising because for $\gamma \approx 1$, the probability weighting function is almost linear. Thus, there is little adjustment.

$U(x)$	Adjusted median	Original median	p-value	Sig. at 5%
$U(x) = x$	0.059	0.053	0.921	No
Linear Interpolation	0.076	0.053	0.066	No
$U(x) = x^\theta$	0.060	0.053	0.579	No

5.5 Distribution of Prediction Errors

We found that reports given by subjects were surprisingly close to the Bayesian beliefs, especially in aggregate. For every report \vec{r} and the bayesian belief \vec{b} , we found the mean of $r_1 - b_1$ for all subjects in all prediction tasks. We did not compare the other elements of the reports, in order to maintain independence between each observation. We found that the mean difference was 0.004 with a standard error of 0.011. Below is a distribution of the error of the reports to the bayesian beliefs.

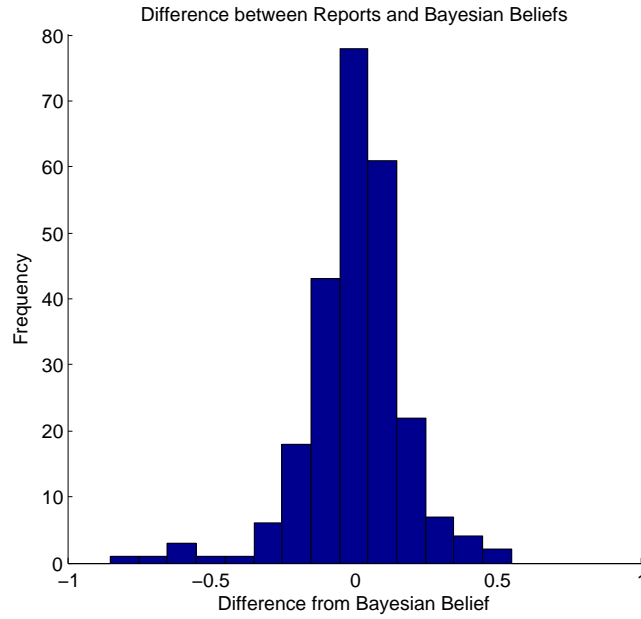


Figure 5.3: Histogram of $r_1 - b_1$ for all subjects and all predictions.

This plot suggests that reports are centered around the bayesian belief with a random error term with mean zero. Assuming normality, the 95% confidence interval for the mean of $r_1 - b_1$ is $0.004 \pm 1.96(0.011) = [-0.214, 0.222]$.

This accuracy did not depend on the bayesian value. Additionally, there was no significant difference between each of the prediction tasks in terms of performance. When we plotted r_1 against b_1 we found that line of best fit was $\hat{r}_1 = 0.887b_1 + 0.065$ with standard errors of 0.027 and 0.0464, respectively. Below is a scatter plot of this data:

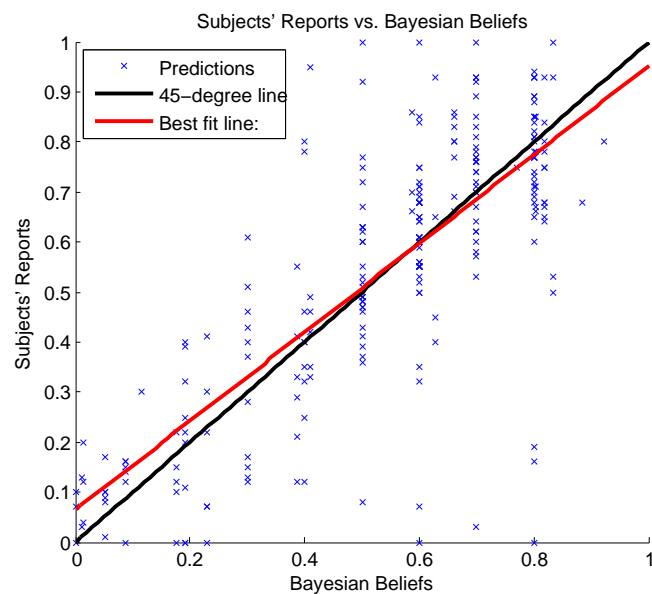


Figure 5.4: Scatter plot of r_1 vs b_1 for all subjects and all predictions.

Chapter 6

Discussion

The empirical results are surprising for a number of reasons. We expected that the treatment group would perform better than the control group because of the presence of incentives. However, we found no such effect; both groups performed equally well. This finding is echoed in [22]; they find that monetary incentives in choices between prospects had little effect on the results.

The second surprising finding is that the adjusted reports performed significantly worse than the unadjusted reports when probability weighting was used. Even adjusting using the inferred probability weighting function from the predictions in the objective probability setting did not yield significant improvements. Predictions were centered around the bayesian beliefs with no apparent bias of overestimation or underestimation. According to our results, a valid method of finding good probability predictions is to simply ask many agents and average their reports.

6.1 Potential Explanations

We believe that there are three potential explanations for the results. We will consider each in this section and discuss their strengths and weaknesses.

The first potential explanation is that stakes in our experiment were too small. This explanation is appealing because it explains both surprising results well. Subjects did not pay attention to the potential difference in payments, but rather reported their best prediction. The control and treatment group exerted similar effort because they both perceived little monetary incentive. The maximum bonus was only \$1.00 and we picked only one out of the eight predictions for evaluation. Thus, the stakes on any one prediction task were extremely small. For the incentives to have a meaningful impact, the agent must be sensitive to differences of a few cents. When designing this experiment, we had hypothesized that workers on AMT would have a utility functions at this level of sensitivity given the low wages found on the site. The one drawback to this explanation is that it cannot explain why predictions in the objective task performed well given that the predictions were not explicitly stated.

Another possible explanation is that subjects did not fully understand the quadratic scoring rule. Thus, they ignored the payments and simply reported their true beliefs. However, the fact that people performed well even in the objective prediction task, where their implicit probabilities were not displayed undermines this hypothesis. Additionally, we only used data from subjects that had passed all of the comprehension tests. However, given the limited amount of training and expertise, it is possible that subjects naively believed reporting the truth was their best strategy.

The final explanation, and the one we believe is most likely, is that people are expected-value maximizers for gains under \$1.00. This conclusion is supported by the fact that in aggregate, subjects reported the true bayesian probabilities. Especially considering the objective prediction task, where they simply chose the payments that maximized their utility lends support for this hypothesis. However, this explanation cannot account for the lack of difference between the control and treatment groups. If people were maximizing expected value of payment, then there would be no reason for subjects in the control group to exert effort on the prediction tasks. Instead, subjects in the control group should complete the

study as fast as possible in order to get their fixed payment and do another task on AMT.

6.2 Future Work

Despite the ineffectiveness of our correction method in our experiment, we believe that additional work is required before dismissing it given that rank-dependent expected utility theory is a better model of decision-making than expected-value maximization in many contexts. A number of variations on our study could test the hypotheses made in the previous section.

Increasing the amount of payment could test the expected-value hypothesis by moving decision-making to a domain where people are no longer maximizing expected value. If the adjusted scores perform better in the situation with larger payments, then it would provide support to the notion that different decision-theories apply at different stakes.

If increasing payments causes a performance gap between the treatment and control groups, then it would provide support to the low stakes hypothesis. The reward for effort is high enough that people in the treatment group would outperform the control.

Another variation would be to perform in-person experiments. This would test the lack of understanding hypothesis by letting the experimenter clarify concepts if subjects are confused. Additionally, the experimenter can more thoroughly explain concepts and test for understanding. If subjects started giving reports more in line with RDEU, then it would support the idea that the workers on AMT did not understand the quadratic scoring rule.

6.3 Conclusion

In conclusion, this thesis was motivated by the potential benefit of improving probability elicitation by applying a more general theory of decision-making to proper scoring rules.

Previous work had only found a solution for two outcomes. Our main theoretical contributions are proving the existence of unique, ordered optimal report for the agent under general conditions and solving for the agents' true beliefs for any number of outcomes.

Our experiment, conducted on Amazon Mechanical Turk, led to some surprising results. Most prominently, monetary incentives seemed to have little impact on performance. Additionally, our adjustment method made reports significantly worse. We hypothesize that the stakes were too small to sufficiently motivate subjects, subjects did not understand the quadratic scoring rule, or subjects are expected-value maximizers for payments up to \$1.00. Future work could test these additional hypotheses with studies with larger incentives and in-person experimentation.

Chapter 7

Acknowledgements

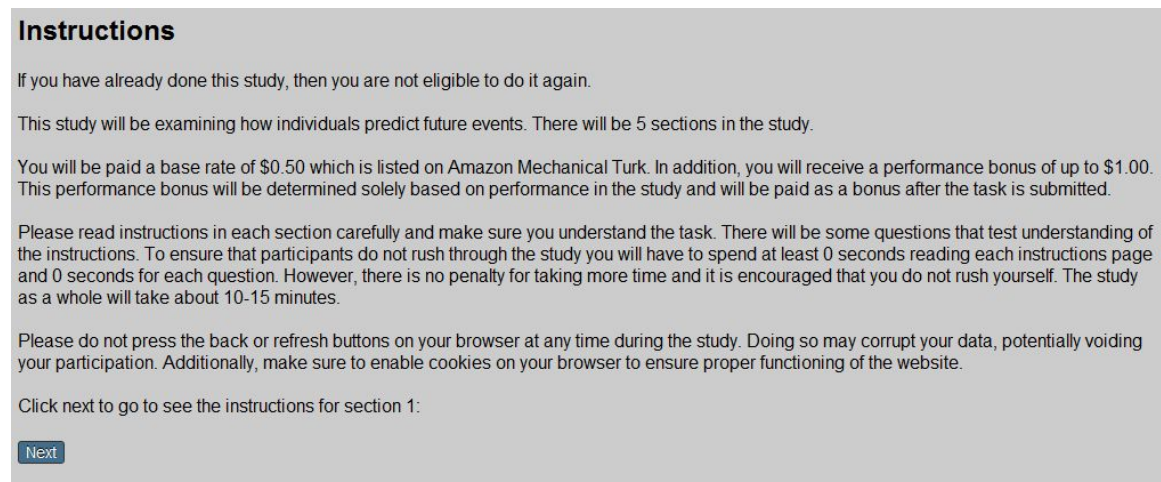
I thank my thesis advisor Yiling Chen for her consistent support throughout this project. Her help in overcoming obstacles and brainstorming new ideas was immeasurably beneficial. I thank Tom Buckley for his help in designing, implementing, and debugging the website that ran my experiment. I thank Adriana Mujal and Ashvini Thammaiah for piloting and providing feedback on my study, particularly with regards to making the instructions more understandable. I thank Haoqi Zhang for help using Amazon Mechanical Turk. I thank Ken Parreno for his advice regarding designing a behavioral experiment. I thank Harvard Computer Society for the free website domain, hosting, and support for the experiment.

Lastly, I thank my parents for their guidance and love. Their encouragement enabled me to pursue my academic interests.

Appendix A

Instructions

This chapter presents screenshots of the study. The following are the instructions for the study.



Instructions

If you have already done this study, then you are not eligible to do it again.

This study will be examining how individuals predict future events. There will be 5 sections in the study.

You will be paid a base rate of \$0.50 which is listed on Amazon Mechanical Turk. In addition, you will receive a performance bonus of up to \$1.00. This performance bonus will be determined solely based on performance in the study and will be paid as a bonus after the task is submitted.

Please read instructions in each section carefully and make sure you understand the task. There will be some questions that test understanding of the instructions. To ensure that participants do not rush through the study you will have to spend at least 30 seconds reading each instructions page and 30 seconds for each question. However, there is no penalty for taking more time and it is encouraged that you do not rush yourself. The study as a whole will take about 10-15 minutes.

Please do not press the back or refresh buttons on your browser at any time during the study. Doing so may corrupt your data, potentially voiding your participation. Additionally, make sure to enable cookies on your browser to ensure proper functioning of the website.

Click next to go to see the instructions for section 1:

[Next](#)

Figure A.1: This is the first page that subjects saw when they previewed the task on AMT.


Section 1 Instructions

All of the questions in this section have the same general format as the question below. Familiarize yourself with the wording of the question before starting section 1.

Use the purple slider to select the value in green that makes you **indifferent** between the two bonus pay schemes for a HIT on Amazon Mechanical Turk:

Scheme 1:
60% chance of getting a bonus of \$1.80 and
40% chance of getting a bonus of \$0.78.

Scheme 2:
60% chance of getting a bonus of \$1.55 and
40% chance of getting a bonus of \$1.04.

\$0.79  \$1.29

In this hypothetical case the bonus schemes are completely random and the bonus will be determined regardless of the quality of the work.

The HIT has a base pay of \$0.50.

Answer the questions in this section such that you actually have no preference between the two schemes. Different people may have different answers.

Once you have understood the instructions click next to begin section 1:

[next](#)

Figure A.2: These are the instructions for section 1.

Instructions for Performance Bonus

In sections 2 through 5 you will be given information and asked to make some predictions. One of your predictions will be selected at random and evaluated for accuracy to determine your performance bonus. Thus, you should try your best on all of the predictions because any one of them may determine your bonus.

Explanation of Probability Theory

In sections 2 and 3 you have to make a series of predictions based on coin flips. Here I will cover some basics about probability theory that will be needed to perform well in the subsequent sections.

A general version of a coin is a coin that lands on heads $X\%$ of the time. If X is 100, then the coin always lands on heads and never on tails. If X is 0, then the opposite is true.

If this coin is flipped an infinite number of times it will have $X\%$ of them as heads. However, for a small number of flips there may not be exactly $X\%$ heads. Only when X is 0 or 100 are we guaranteed that small samples have $X\%$ heads.

Let's consider an example. Coin 1 lands on heads 70% of the time and Coin 2 lands on heads 20% of the time. I pick a coin at random with each equally likely to be chosen. Out of 5 flips, I get 3 heads and 2 tails. It is possible that both coins may have produced this distribution but coin 1 is more likely to do so. In fact, there is an 86% chance that it was coin 1 and a 14% chance it was coin 2.

You will be asked to do problems like this in sections 2 and 3, but are not expected to calculate exact probabilities. However, do your best to estimate in your head what you think the probabilities are.

Click next to see the instructions for section 2:

[next](#)

Figure A.3: The instructions on how subjects will be paid as well as some basic probability theory to help with the coin prediction tasks.


Section 2 Instructions

All of the questions in this section have the same general format as the question below. Familiarize yourself with the wording of the question before starting section 2:

Coin 1 lands on heads 70% of the time.
Coin 2 lands on heads 40% of the time.

With each coin having a one-half chance of being selected, I chose one coin at random and discarded the other coin. I flipped the selected coin 5 times. Out of 5 flips, there were 1 heads and 4 tails.

Estimate how likely the selected coin is coin 1 or coin 2. Your bonus will depend on how accurate you are.

Coin 1  Coin 2

Payment if it is coin 1: \$0.75
Payment if it is coin 2: \$0.75
Your estimate of the probability that it is coin 1: 50%
Your estimate of the probability that it is coin 2: 50%

Moving the slider towards Coin 1's label increases your estimate for coin 1. Similarly, for coin 2.

When performing this task make sure to focus on the payoffs and your estimates on how likely it is to be coin 1 or coin 2. You are not expected to calculate the exact probability.

If you would like to see the exact mathematical formulas as to how you will be paid then click "show". You are not required to read the formulas and your bonuses will be calculated for you. [show](#)

Click next to begin section 2:

[next](#)

Figure A.4: These are the instructions for section 2.

Section 3 Instructions


The scenario you will be faced with in section 3 is similar to last section except now there are 3 coins. See below for an example:

All valid estimates are in the green triangle. The red area would imply that the sum of the predictions is greater than 100% which is not possible:

Coin 1 lands on heads 70% of the time.
 Coin 2 lands on heads 50% of the time.
 Coin 3 lands on heads 40% of the time.

With each coin having a one-third chance of being selected, I chose one coin at random and discarded the other 2 coins. I flipped the selected coin 5 times. Out of 5 flips, there were 2 heads and 3 tails.

Estimate how likely the selected coin is coin 1, coin 2, or coin 3.
 Your bonus will depend on how accurate you are.



Payment if it is coin 1: \$0.67
 Payment if it is coin 2: \$0.66
 Payment if it is coin 3: \$0.66

Your estimate of the probability that it is coin 1: 34%
 Your estimate of the probability that it is coin 2: 33%
 Your estimate of the probability that it is coin 3: 33%

Valid entry

Moving the slider towards Coin 1's label increases your estimate for coin 1. Similarly, for coin 2 and coin 3.

Only the green area is valid.

When performing this task make sure to focus on both the payoffs and your estimates on how likely it is to be coin 1, coin 2, or coin 3. You are not expected to calculate the exact probability.

If you would like to see the exact mathematical formulas as to how you will be paid then click "show". You are not required to read the formulas and your bonuses will be calculated for you. [show](#)

Click next to begin section 3:

[next](#)

Figure A.5: These are the instructions for section 3.

Section 4 Instructions

All of the questions in this section have the same general format as the question below. Familiarize yourself with the wording of the question before starting section 4:

I have a jar with 100 marbles. There are 95 red marbles and 5 blue marbles.

I will draw one marble from the jar with each of the 100 marbles equally likely to be chosen. You will receive a bonus based on which color I draw.

Choose the bonus that you would like to receive for each color with the slider below.

Red

Blue

Payment if it is red: \$0.75

Payment if it is blue: \$0.75

Moving the slider towards Red's label increases your payoff for red. Similarly, for blue.

When performing this task make sure to focus on the payoffs for red and blue. There are no right answers. Pick the payoffs that you find most attractive.

The payoffs are the same as those in Section 2, but this time you will not be able to see your implied probability estimate. Instead, you can only see the potential bonuses. This is not a test of your memory, simply answer by focusing on the potential bonuses.

If you would like to see the exact mathematical formulas as to how you will be paid then click "show". You are not required to read the formulas and your bonuses will be calculated for you. [show](#)

Click next to begin section 4:

next

Figure A.6: These are the instructions for section 4.

Section 5 Instructions

The scenario you will be faced with in section 5 is similar to last section except now there are 3 colors. See below for an example:


All valid estimates are in the green triangle. The red area would imply that the sum of the predictions is greater than 100% which is not possible:

I have a jar with 100 marbles. There are 45 red marbles, 35 blue marbles, and 20 green marbles.

I will draw one marble from the jar with each of the 100 marbles equally likely to be chosen. You will receive a bonus based on which color I draw.

Choose the bonus that you would like to receive for each color with the slider below.

Red Blue



Green

Payment if it is red: \$0.67
 Payment if it is blue: \$0.66
 Payment if it is green: \$0.66
 Valid entry

Moving the slider towards Red's label increases your payoff for red. Similarly, for blue and green.

Only the green area is valid.

When performing this task make sure to focus on the payoffs for each color. There are no right answers. Pick the payoffs that you find most attractive.

The payoffs are the same as those in Section 3, but this time you will not be able to see your implied probability estimate. Instead, you can only see the potential bonuses. This is not a test of your memory, simply answer by focusing on the potential bonuses.

If you would like to see the exact mathematical formulas as to how you will be paid then click "show". You are not required to read the formulas and your bonuses will be calculated for you. [show](#)

Click next to begin section 5:

[next](#)

Figure A.7: These are the instructions for section 5.

The following are test questions to screen subjects for understanding and effort.


This question is just to test your understanding of the directions.

It will not be used to calculate your bonus.

Use the purple slider to select the value in green that makes you **indifferent** between the two bonus pay schemes for a HIT on Amazon Mechanical Turk:

Scheme 1:
60% chance of getting a bonus of \$1.80 and
40% chance of getting a bonus of \$0.78.

Scheme 2:
60% chance of getting a bonus of \$1.80 and
40% chance of getting a bonus of \$1.02.

\$0.78  \$1.27

In this hypothetical case the bonus schemes are completely random and the bonus will be determined regardless of the quality of the work.

The HIT has a base pay of \$0.50.

Figure A.8: This is one of the test questions. The subject passed if they set the value to less than \$0.88.


This question is just to test your understanding of the directions.

It will not be used to calculate your bonus.

Coin 1 lands on heads 100% of the time.
Coin 2 lands on heads 0% of the time.

With each coin having a one-half chance of being selected, I chose one coin at random and discarded the other coin. I flipped the selected coin
Out of 5 flips, there were 5 heads and 0 tails.

Estimate how likely the selected coin is coin 1 or coin 2. Your bonus will depend on how accurate you are.

Coin 1  Coin 2

Payment if it is coin 1: \$0.75
Payment if it is coin 2: \$0.75
Your estimate of the probability that it is coin 1: 50%
Your estimate of the probability that it is coin 2: 50%

Moving the slider towards Coin 1's label increases your estimate for coin 1. Similarly, for coin 2.

Figure A.9: This is one of the test questions. The subject passed if they said it was coin 1 with greater than 90% probability.

This question is just to test your understanding of the directions.

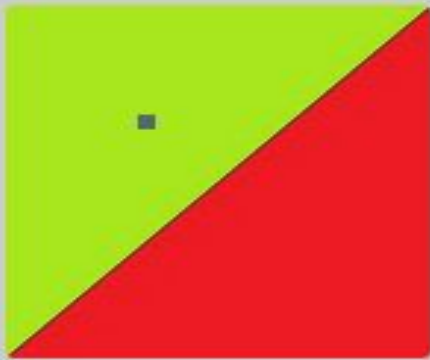
It will not be used to calculate your bonus.

Coin 1 lands on heads 100% of the time.
 Coin 2 lands on heads 50% of the time.
 Coin 3 lands on heads 0% of the time.

With each coin having a one-third chance of being selected, I chose one coin at random and discarded the other 2 coins. I flipped the selected coin 5 times. Out of 5 flips, there were 2 heads and 3 tails.

Estimate how likely the selected coin is coin 1, coin 2, or coin 3. Your bonus will depend on how accurate you are.

Coin 1



Coin 2

Coin 3

Payment if it is coin 1: \$0.67
 Payment if it is coin 2: \$0.66
 Payment if it is coin 3: \$0.66

Your estimate of the probability that it is coin 1: 34%
 Your estimate of the probability that it is coin 2: 33%
 Your estimate of the probability that it is coin 3: 33%

Valid entry

Moving the slider towards Coin 1's label increases your estimate for coin 1. Similarly, for coin 2 and coin 3.

Only the green area is valid.

Figure A.10: This is one of the test questions. The subject passed if they said it was coin 2 with greater than 90% probability.

Bibliography

- [1] Inc. Amazon.com. Amazon mechanical turk. <https://www.mturk.com/mturk/welcome>.
- [2] Steffen Andersen, John Fountain, Glenn W. Harrison, and E. Elisabet Rutström. Estimating subjective probabilities. Experimental Economics Center Working Paper Series 2010-08, Experimental Economics Center, Andrew Young School of Policy Studies, Georgia State University, June 2010.
- [3] Daniel Bernoulli. Specimen theoriae novae de mensura sortis. *Commentarii Academiae Scientiarum Imperialis Petropolitanae*, 5:175–192, 1738.
- [4] J. Eric Bickel. Some comparisons among quadratic, spherical, and logarithmic scoring rules. *Decision Analysis*, 4(2):49–65, 2007.
- [5] Han Bleichrodt and Jose Luis Pinto. A parameter-free elicitation of the probability weighting function in medical decision analysis. *Manage. Sci.*, 46:1485–1496, November 2000.
- [6] Stephen Boyd. Convex functions. <http://www.stanford.edu/class/ee364a/lectures/functions.pdf>, March 2011.
- [7] Glenn W. Brier. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3, 1950.

- [8] Gary J. Echternacht. The use of confidence testing in objective tests. *Review of Educational Research*, 42(2):217–236, Spring 1972.
- [9] Tilmann Gneiting and Adrian E. Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, March 2007.
- [10] Arne Hallam. Convexity and optimization. http://www2.econ.iastate.edu/classes/econ500/hallam/documents/Convex_Opt_000.pdf, March 2011.
- [11] Carl-Axel S. Stal Von Holstein. Probabilistic forecasting: An experiment related to the stock market. *Organizational Behavior and Human Performance*, 8(1):139 – 158, 1972.
- [12] Eric Hsin-Chun Huang. Automatic task design on amazon mechanical turk, 2010.
- [13] Daniel Kahneman and Amos Tversky. Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2):263–91, March 1979.
- [14] Nicolas S. Lambert, David M. Pennock, and Yoav Shoham. Eliciting properties of probability distributions. In *Proceedings of the 9th ACM conference on Electronic commerce*, EC '08, pages 129–138, New York, NY, USA, 2008. ACM.
- [15] Theo Offerman, Joep Sonnemans, Gijs Van De Kuilen, and Peter P. Wakker. A truth serum for non-bayesians: Correcting proper scoring rules for risk attitudes. *Review of Economic Studies*, 76(4):1461–1489, October 2009.
- [16] John Quiggin. A theory of anticipated utility. *Journal of Economic Behavior and Organization*, 3(4):323 – 343, 1982.
- [17] Allan H. Murphy Robert L. Winkler. Good probability assessors. *Journal of Applied Meteorology*, 7(5):751–758, Oct 1968.
- [18] Reinhard Selten. Axiomatic characterization of the quadratic scoring rule. *Experimental Economics*, 1(1):43–61, June 1998.

- [19] D. J. Spiegelhalter. Probabilistic prediction in patient management and clinical trials. *Statistics in Medicine*, 5(5):421–433, 1986.
- [20] P. E. Tetlock. *Expert Political Judgment*. Princeton University Press, 2005.
- [21] Jonathan Tuthill and Darren L. Frechette. Non-expected utility theories: Weighted expected, rank dependent, and cumulative prospect theory utility. 2002 Conference, April 22-23, 2002, St. Louis, Missouri 19073, NCR-134 Conference on Applied Commodity Price Analysis, Forecasting, and Market Risk Management, 2002.
- [22] A. Tversky and D. Kahneman. Advances in prospect theory: Cumulative representation of uncertainty. *Journal of Risk and Uncertainty*, 5:297–323, 1992.
- [23] P.P. Wakker and D. Deneffe. Eliciting von neumann-morgenstern utilities when probabilities are distorted or unknown. *Management Science*, 42:1131–1150, 1996.
- [24] Robert L. Winkler and Allan H. Murphy. Nonlinear utility and the probability score. *Journal of Applied Meteorology*, 9(1):143–148, 1970.
- [25] William F. Wright. Empirical comparison of subjective probability elicitation methods. *Contemporary Accounting Research*, 5(1):47–57, 1988.