Preferences for Warning Signal Quality: Experimental Evidence

Alexander Ugarov[*]
Hivereview

Arya Gaduh[†]
University of Arkansas and NBER

Peter McGee[‡]
University of Arkansas

February 28, 2026

Abstract

We use a laboratory experiment to study preferences over false-positive and false-negative rates of warning signals for an adverse event with a known prior. We find that subjects decrease their demand with signal quality, but less than predicted by our theory. They disproportionately reduce their demand for signals with high false-negative rates for rare events, while the opposite holds true for frequent events. We show that neither risk preference nor Bayesian updating skills can fully explain our results. Our results are most consistent with a decision-making heuristic in which subjects do not distinguish between false-positive and false-negative errors.

JEL Classification: C91, D81, D84, D91
Keywords: alarms, value of information, information economics, information design, medical tests

[*] Email: augarov@hivereview.org.
[†] Email: agaduh@walton.uark.edu.
[‡] Email: pmcgee@walton.uark.edu.

1 Introduction

The trade-offs between false-positive and false-negative errors of warning systems often have life-and-death consequences. The 2010 gas blowout on the Deepwater Horizon oil rig killed 11 workers and caused one of the largest oil spills in history. The death toll was possibly aggravated by switching off the general safety alarm because the rig "did not want people woke up at 3 a.m. from false alarms" (Brown, 2010). In medicine, different expert groups often disagree with the cancer screening guidelines issued by the U.S. Preventive Services Task Force, in large part over their perceived trade-offs between the costs from missed early detection against the potential harms from overdiagnosis or overtreatment due to false positive results (Rabin, 2024).

Most real-world warning systems — medical diagnostics, security alarms, extreme weather alerts — transform continuous signals about the likelihood of an adverse state into a yes/no binary signal. This transformation requires choosing a threshold for a positive classification. A lower threshold lowers the probability of failing to warn of an adverse state (false-negative rate) but increases the probability of warning in a safe state (false-positive rate), reflecting the standard trade-off characterized in ROC analysis (Fawcett, 2006). While the optimal threshold depends on user preference over the costs of these probabilistic errors, currently there is no guidance on what this threshold might be beyond assuming that decision-makers weigh false-positive and false-negative costs equally.[1]

To address this gap, we conduct a laboratory experiment to measure the demand for warning signals with varying quality. In the experiment, subjects receive information about the prior probability of an adverse event and are asked to take a protective action after receiving a signal with known false-positive and negative rates. We then elicit our main outcome, i.e., their willingness-to-pay (WTP) for each signal. To account for subject heterogeneity, we use separate experimental tasks to measure their risk preference and Bayesian updating skill.

We compare the behavior of our subjects to that of a risk-neutral, utility-maximizing decision maker that we derive from a simple model. Subjects' WTP is weakly correlated with the value of information, resulting in overpaying for low-quality signals and underpaying for high-quality signals. Importantly, we find asymmetric under-responsiveness by prior: with a low (high) prior, their WTP does not fully adjust for the increase in the false-positive (false-negative) costs. We provide evidence that this pattern is most consistent with a failure to estimate the effect of the frequencies of false-positive and false-negative outcomes on the potential costs of using the signal.

We contribute to the literature in three ways. First, we provide novel evidence on the demand for warning systems using an incentivized experiment. Existing studies of warning systems, which mostly focus on medical diagnostic tests, use unincentivized surveys to measure WTP and do not explore preferences over the tests' information structures. They find that

[1] We define false positive and false negative *costs* formally in Section 2, but it is important to keep in mind how they differ from false positive and false negative *rates*. False positive (negative) costs are the expected monetary costs to the decision maker as a function of the false positive (negative) rates and prior probabilities.

preferences over diagnostic tests correlate with their accuracy, but respondents exhibit two significant biases. First, they are willing to pay for tests with little or no diagnostic value (Schwartz et al., 2004; Neumann et al., 2012). For example, Schwartz et al. (2004) find that 73% of Americans prefer a free full-body CT scan versus $1,000 in cash even though full-body scans for healthy people are not recommended by physicians. Second, the way the information about the test's accuracy is presented strongly affects choices (Howard and Salkeld, 2009). We extend this literature using a context-neutral experiment to examine whether similar biases hold more generally and when choices are incentivized, as well as whether the demand elasticity for information responds symmetrically to false positive and false negative errors.

Second, we contribute to the emerging experimental literature on the demand for information by studying signals that affect protection decisions. Previous studies in this literature employ prediction games, in which subjects must guess an optimal state under uncertainty (Hoffman, 2016; Ambuehl and Li, 2018; Xu, 2022; Montanari and Nunnari, 2023). Generally, they find that while the demand for information increases with signal quality, it increases more modestly than expected from a Bayesian decision maker. Two of these studies employ laboratory experiments. Ambuehl and Li (2018) find that subjects underreact to the accuracy of a binary signal about the state of the world, but put a premium on completely certain signals. Xu (2022) shows that many subjects choose non-instrumental signals over instrumental signals, consistent with failures in contingent reasoning about the future value of information and with partial failures to distinguish FP and FN rates. Both studies employ a prediction game.

We extend this literature by moving beyond a prediction game and instead incorporating protection decisions with three distinct payoffs (i.e., full payoff, full payoff minus protection costs, and full payoff minus losses). With asymmetric payoffs, risk preferences affect the value of information and can change sensitivities to false-positive and false-negative rates. Asymmetric payoffs also creates asymmetry in how subjects value different types of errors which is not observed in (symmetric) prediction games. We also directly elicit both willingness-to-pay and potential protection decisions for different combinations of priors and signal characteristics, allowing for more general conclusions about subjects' preferences. Consistent with Ambuehl and Li (2018), our subjects overvalue inaccurate signals, but we do not find a premium for signals with high certainty.

Additionally, the subject's choices after receiving a signal in our experiment are equivalent to insurance decisions with full coverage. Hence our results also apply to insurance problems when subjects receive additional signals of their risks (such as flood zone designations). While on average people under-insure with respect to rare natural disasters (Friedl, Lima de Miranda and Schmidt, 2014), the demand for insurance goes up immediately after an insurable adverse event (e.g., Kousky, 2011). One suggested explanation is that subjects overweight recent evidence leading to under-insurance when there were no negative events in the recent past and to overinsurance after the fact (Volkman-Wise, 2015). This is consistent with underweighting prior probabilities relative to more recent signals. At the same time, however, Laury, McInnes and Swarthout (2009) find no under-insurance for low-probability events in the laboratory setting. We similarly find that, on average, subjects do not under-insure after receiving a signal even though we see potential over-protection for negative signals.

The paper proceeds as follows. The next section sets up a simple model and outlines our hypotheses. Section 3 describes the experimental design. Given the novelty of some of the experimental tasks, we present our results in three consecutive sections. First, we describe a theory-free exposition of subjects' choices in all treatments in Section 4. Then, Section 5 describes the results for main empirical tests of this paper with regards to the willingness-to-pay for signals. Finally, we explore potential explanations for the observed pattern of underreaction to false-positive rates for low initial probabilities in Section 5.3. Section 6 concludes.

2 Model

Environment. Let $\omega \in \{0,1\}$ denote the state of the world, where 1 corresponds to an adverse event that happens with probability π and induces a loss, L. An agent can take protective action $a \in \{0,1\}$ to avoid losing L under the adverse state. The loss is realized only when $\omega(1-a)=1$.

The agent's preferences are described by a utility function that depends on wealth Y, protective action a, and the protective outcome $\omega(1-a)$. Taking the protective action costs $c>0$ as given, utility is separable in wealth, protection cost, and the potential loss $L>c$ in the adverse state if not protected:

$$U = U(Y, a, \omega(1-a)) = u(Y - ac - \omega(1-a)L)$$

The agent can purchase information in the form of a binary signal $s \in \{0,1\}$ about the state of the world. Let $P_{ij} \equiv P(s=i|\omega=j)$ be the probability that signal s takes the value i conditional on the state of the world being j. After learning the signal's value, the agent updates her belief on the likelihood of the adverse event to $\mu(s)$. We assume that she is Bayesian and her posterior belief equals to:

$$\mu(s) = \frac{\pi P_{s1}}{\pi P_{s1} + (1-\pi)P_{s0}}$$

where a larger $\mu(s)$ implies a higher posterior probability of the adverse event.

Without loss of generality, we assume information structure in which a positive signal $s=1$ is more likely in the adverse state of the world $P_{11} > P_{10}$.

Strategy. Without a signal, the agent can choose only between always protecting and never protecting resulting in the following expected utility:

$$U_0 = \max[u(Y-c), \pi u(Y-L) + (1-\pi)u(Y)]$$

With a signal, the subject's strategy $\sigma : \{0,1\} \to \{0,1\}$ describes protection action for each signal's value. We can represent it as a tuple of numbers (a_0, a_1). Let $b \geq 0$ be the price of a signal paid beforehand from initial wealth and use $U(\sigma, b)$ to describe expected utility with a signal conditional on strategy and information price. Note that if the strategy is not responsive to the signal, payoffs cannot improve upon payoffs available without a signal because for any $b > 0$

$$U(\sigma = (0,0), b) = \pi u(Y - L - b) + (1 - \pi)u(Y - b) < \pi u(Y - L) + (1 - \pi)u(Y)$$

i.e., choosing not to protect irrespective of the signal makes one worse off than just not purchasing the signal, and

$$U(\sigma = (1,1), b) = u(Y - c - b) < u(Y - c)$$

i.e., choosing to always protect irrespective of the signal makes one worse off than just not purchasing a signal. If a subject follows the signal her expected utility is:

$$U(\sigma = (0,1), b) = \pi P_{11} u(Y - c - b) + \pi P_{01} u(Y - L - b) + (1 - \pi)P_{10} u(Y - c - b) + (1 - \pi)P_{00} u(Y - b)$$

Given our assumption that $P_{11} > P_{10}$, the expected utility of following the signal is higher than the expected utility of following the "do the opposite" strategy:

$$U[\sigma = (1,0), b] = \pi P_{01} u(Y - c - b) + \pi P_{11} u(Y - L - b) + (1 - \pi)P_{00} u(Y - c - b) + (1 - \pi)P_{10} u(Y - b)$$

Hence, value of the signal, i.e., the maximum willingness to pay for it, is determined by comparing the expected utility of following the signal (the LHS of eqn. 1) with the expected utility of not using a signal (the RHS of eqn. 1):

$$\max_{\sigma} U(\sigma, b) = U_0 = \max[u(Y - c), \pi u(Y - L) + (1 - \pi)u(Y)] \tag{1}$$

As noted above, the signal's value cannot be positive if it is not followed, hence it should satisfy:

$$\begin{aligned} P(s = 1)u(Y - b - c) + \pi P_{01} u(Y - b - L) + (1 - \pi)P_{00} u(Y - b) = \\ = \max[u(Y - c), \pi u(Y - L) + (1 - \pi)u(Y)] \end{aligned} \tag{2}$$

where $P(s = 1) \equiv \pi P_{11} + (1 - \pi)P_{10}$. The expression on the left-hand side of this equation is a strictly decreasing function of b. Additionally, for $b \to \infty$ the left-hand side is smaller than the right-hand side. It implies that equation (2) has at most one positive solution.

Obviously, $b > 0$ for a perfectly accurate signal because the payoff distribution with the signal first-order stochastically dominates the distribution without it. However, determining the value of an imperfect signal is non-trivial, as it requires more restrictions on preferences to allow weighing outcomes with losses vs outcomes with protection costs.

Risk-neutral agent. If the agent is risk-neutral, the expression above collapses to:

$$b + P(s = 1)c + \pi P_{01}L = \min[c, \pi L]$$

The signal's value is just:

$$b^* = \max[0, \min[c, \pi L] - P(s = 1)c - \pi P_{01}L]$$

We can express the WTP for the signal, b, as a function of priors, false-positive (FP), and false-negative (FN) *rates* denoted correspondingly as P_{10} and P_{01}. The following equation serves as our benchmark in the empirical work:

$$b = \max[0, \min[c, \pi L] - \pi(1 - P_{01})c - (1 - \pi)P_{10}c - \pi P_{01}L] \tag{3}$$

The signal's value is decreasing in both FP and FN rates. The effect is proportional to the non-adverse (adverse) state probability for the FP (FN) rate. When WTP is positive ($b > 0$), its derivatives with respect to FP (P_{10}) and FN (P_{01}) rates are given by:

$$\frac{db}{dP_{10}} = -(1 - \pi)c \tag{4}$$

$$\frac{db}{dP_{01}} = -\pi(L - c) \tag{5}$$

Note also that WTP is always equally affected by FP and FN *costs* which are defined as $\pi(1 - P_{01})c$ and $\pi P_{01}L$, respectively.

Risk Preferences. In an expected utility framework, risk aversion can either increase or decrease an agent's valuation of the signal. More specifically, risk aversion decreases her WTP when protection costs are low:

Proposition 1. *If the probability of the adverse state is high (i.e., $\pi > c/L$), then a strictly risk-averse decision-maker pays less than a risk-neutral one.*

Proof. See the Appendix. □

Intuitively, risk-averse decision-makers protect by default without using a signal when protection costs are sufficiently low. Things are less clear with low risks or higher protection costs. For example, if a risk-averse decision-maker chooses to not protect without a signal, risk aversion increases the value of a perfect signal. This follows from the standard argument that demand for insurance increases with risk aversion, and the fact that the protection problem with a perfect signal is isomorphic to the insurance problem with deductible c.

Next, we examine the effect of the false-positive and false-negative rates of a signal on WTP, b. Assuming a differentiable utility function $u(.)$, we use implicit differentiation to derive sensitivities of b to false-positive (FP) and false-negative (FN) rates:

$$\frac{db}{dP_{10}} = -\frac{(1-\pi)(u(Y-b)-u(Y-c-b))}{D(\pi, P_{01}, P_{10}, b)} \tag{6}$$

$$\frac{db}{dP_{01}} = -\frac{\pi(u(Y-c-b)-u(Y-L-b))}{D(\pi, P_{01}, P_{10}, b)} \tag{7}$$

with the denominator equal to the expected marginal utility:

$$D(\pi, P_{01}, P_{10}, b) \equiv P(S=1)u'(Y-c-b) + \pi P_{01}u'(Y-L-b) +$$

$$+(1-\pi)P_{00}u'(Y-b) = E[MU] > 0$$

The signal's value decreases in FP and FN rates, i.e., $\frac{db}{dP_{10}}$ and $\frac{db}{dP_{01}} < 0$. Equations 6-s 7 indicate that risk aversion can either increase or decrease the decision-maker's sensitivity to FP and FN rates depending on the utility function's curvature and the signal's characteristics. Intuitively, the expected marginal utility of a strongly risk-averse subject with an imperfect signal can be lower than the average slope of the utility function between $(Y-c-b)$ and $(Y-b)$ which reduces sensitivity to FP rates. It can also be higher if either the signal is perfect or the curvature is small. Similarly, sensitivity to FN rates also depends on the form of the utility function.

However, we can say more about the ratio of sensitivities to FP and FN rates. For information structures in which both risk-averse and risk-neutral subjects purchase a signal, the ratio of sensitivities to FP rates over FN rates should be lower for strictly risk-averse subjects.

Proposition 2. *Conditional on both risk-neutral and risk-averse subjects purchasing a signal, strictly risk averse subjects have lower relative sensitivity to FP rates as compared to risk-neutral ones.*

Proof. See the Appendix. □

* * *

The model offers two testable hypotheses on the WTP that can be brought to the experiment. *First,* as a natural starting point, we can test whether subjects' WTPs are equal to the values predicted for risk-neutral expected-utility maximizers given in equation 2. *Second,* the model of a risk-neutral agent suggests that subjects' WTP should have equal sensitivity to *costs* from false-positive and false-negative signals (equation 3).

3 Experimental Design

We conducted the experiment using Qualtrics in the Behavioral Business Research Lab (BBRL) at the University of Arkansas. We conducted two waves of data collection with a total of 207 subjects. The first wave was conducted between October and November 2021 with 105 subjects, while the second wave was conducted between October and November 2025 with 102 subjects. On average, including a \$5 show-up fee, subjects earned \$25.50 for a session lasting around 45 minutes.

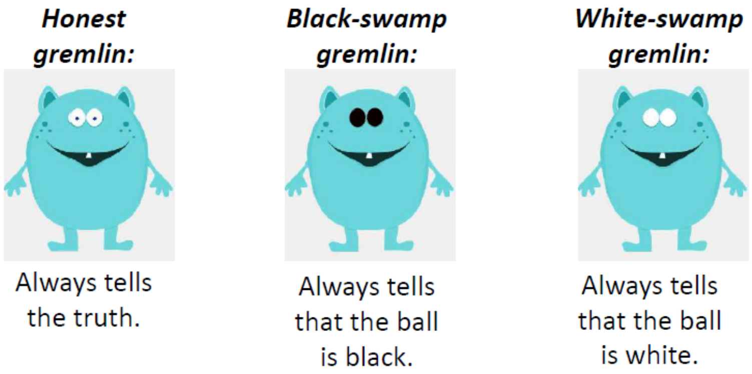
Subjects were endowed with \$25 (on top of the show-up fee) that they could potentially lose in the experiment. Their payoff was determined by their decisions in four sets of tasks played in the following order: (i) Blind Protection; (ii) Informed Protection; (iii) Belief Elicitation; and (iv) Willingness-to-Pay Elicitation. Subjects took a quiz of understanding prior to each task; the correct answer and an explanation were provided if a subject answered a question incorrectly.[2] Each task consisted of 6 rounds, resulting in 24 total rounds. At the end of the experiment, one of these 24 rounds is randomly selected as the payment round. The instructions can be found in the appendix.

Blind Protection (BP) Subjects decide whether to pay \$5 to protect against an adverse event: a random draw of a black ball. Subjects know the prior probability that a black ball is drawn. A subject who draws a black ball will lose nothing if they chose to protect and \$20 if they did not. The prior probability of drawing a black ball across the 6 rounds is denoted as $p \in \{0.05, 0.10, ..., 0.3\}$. The order was common for all the subjects, increasing monotonically from the lowest probability. Subjects did not receive feedback on the decision's outcome.

Informed Protection (IP) As in the BP task, subjects must make a protection decision before learning if they drew a black ball. Prior to the protection choice, subjects learn the prior probability of drawing a black ball, receive a signal about the ball, and learn the signal's information structure. Following Coutts (2019), we represent this information structure using a group of hinting gremlins, where the hint is provided by a randomly selected gremlin (mapping to a signal realization in the model). The gremlin is one of three types: (i) honest; (ii) "black-swamp" who always says that the ball is black; and (iii) "white-swamp" who always says that the ball is white. Figure 1 illustrates how these gremlin types were presented to the subjects. The group composition determines the information structure: a higher share of black(white)-swamp gremlins produces a signal with higher FP (FN) rate. Subjects know the group composition, but do not know which gremlin provides the signal in any particular round. Subjects are asked to make two protection decisions, to wit, for black and white signals.

[2] Incorrect quiz answers for the Informed Protection section resulted in subjects facing three additional multiple choice questions. We believe that a clear understanding of the Informed Protection task is essential for subsequent tasks, hence the additional questions. Complete details of the comprehension questions are in the appendix.

Figure 1: Signals Presentation



Belief Elicitation (BE). As in the IP task, subjects know the prior probability of drawing a black ball and the information structure (represented by the composition of the group of gremlins). Instead of making a protection decision, however, subjects are asked to estimate the probability that: (i) the ball is black when the gremlin says that it is white; (ii) the ball is black when the gremlin says that it is black.

To elicit incentive-compatible responses, we follow the stochastic version of the Becker-DeGroot-Marshak mechanism developed by Grether (1992) and Holt and Smith (2009) but stated equivalently in terms of losses rather than gains. Subjects submit their beliefs about the probability of the adverse event $\mu \in [0, 1]$. If μ is above some uniform random number $r \in [0, 1]$, they lose \$20 only if this event happens (i.e., a black ball is drawn). If $r > \mu$, then they draw an independent lottery that will lose \$20 with probability r and 0 otherwise.[3] Danz, Vesterlund and Wilson (2020) find that providing a detailed explanation of payoffs in belief elicitation lowers truthful reporting. Accordingly, our instructions begin by pointing out that reporting one's true belief μ maximizes payoffs, followed by an explanation of payoff calculation under different reporting strategies.

Willingness-to-Pay Elicitation (WTPE). The WTPE task measures a subject's willingness to pay (WTP) for a signal. As before, subjects know the prior probability of drawing a black ball and the information structure of a potential signal. Unlike the IP task, subjects do not automatically receive a signal, instead they provide their WTP for a signal by choosing a value $\in (\$0, \$5)$ in \$0.50 increments. The elicitation is incentive compatible: if a WTPE round is selected as the payment round, a random price for the signal is drawn. If that price exceeds the subject's WTP, they will play a BP round, otherwise the subject pays their WTP and plays an IP round after receiving a signal with the given information structure.

[3] The benefit of this mechanism versus other probability elicitation mechanisms (e.g., quadratic scoring) is that reporting truthfully is a dominant strategy regardless of risk preferences (Karni, 2009) as long as a subject's preferences adhere to probabilistic sophistication and dominance i.e., they rank lotteries based on their probabilities only and prefer higher probabilities of higher payoffs.

Post-Experiment Questionnaire and Payment. After the WTPE task, subjects in both waves answered a few demographic questions: gender, age, education level, and on taking any statistics or probability classes. Subjects in the second wave also answered questions about their GPA, and the seven-item extended Cognitive Reflection Test (CRT) to measure their cognitive skills (Toplak, West and Stanovich, 2014). The payment task and the payment round were then randomly chosen to calculate the subject's payoff.

Treatments and Treatment Assignment. Table 1 summarizes our treatments for tasks other than BP. The treatments combine four priors with 12 different information structures for a total of 48 treatments. These treatments allow for a distribution of posteriors (for black and white signals) that spans a wide range of interior values and allows for a wide range of theoretical WTP values (Appendix Figure A1).

Subjects were assigned to treatments in the following manner. Each subject was randomly assigned six treatments that combined a pair of priors of either (0.1, 0.3) or (0.2, 0.5) with three information structures. The presentation of the treatments were organized by prior, and the order of the priors — i.e., whether the first was larger than the second — was randomly assigned. For each prior, the first treatment was always the one with a fully truthful information structure. The order of priors and signals stays constant for each subject across tasks, but can vary between subjects.

Table 1: List of Treatments

	Number of Gremlins				
Prior (p)	Total	Black-Swamp	White-Swamp	FP rate	FN rate
	(1)	(2)	(3)	(4)	(5)
0.1, 0.2, 0.3, 0.5	2	0	0	0.00	0.00
0.1, 0.2, 0.3, 0.5	2	1	0	0.50	0.00
0.1, 0.2, 0.3, 0.5	2	0	1	0.00	0.50
0.1, 0.2, 0.3, 0.5	3	1	0	0.33	0.00
0.1, 0.2, 0.3, 0.5	3	0	1	0.00	0.33
0.1, 0.2, 0.3, 0.5	3	1	1	0.33	0.33
0.1, 0.2, 0.3, 0.5	5	1	0	0.20	0.00
0.1, 0.2, 0.3, 0.5	5	0	1	0.00	0.20
0.1, 0.2, 0.3, 0.5	5	1	1	0.20	0.20
0.1, 0.2, 0.3, 0.5	7	2	0	0.29	0.00
0.1, 0.2, 0.3, 0.5	7	0	2	0.00	0.29
0.1, 0.2, 0.3, 0.5	7	1	1	0.14	0.14

4 Subject Decisions By Task

Decisions in the Blind Protection (BP), Informed Protection (IP), and Belief Elicitation (BE) tasks measure determinants of WTP in our model. Protection choices in the BP task reveal subjects' risk preferences with known probabilities. Choices in the IP task demonstrate how subjects use signals given their characteristics. Finally, the BE task provides insight into subjects' beliefs for given signals.

Given the complexities of some of these tasks, this section illustrates the extent of subject comprehension using simple descriptive analyses. We find that subjects understand these tasks reasonably well. This section also provides a qualitative overview of how subjects tend to deviate from the theoretical predictions for each task. We provide a more formal regression analysis of how signal characteristics affect subjects' WTP later in Section 5.

4.1 Blind Protection

Figure 2 plots the likelihood of protecting against the posterior probability of drawing a black ball for the BP task, where the posterior is equivalent to the prior (the thick line), and in the IP task (the thin line). On aggregate in the BP task, subjects' likelihood of protecting increases in the probability of an adverse outcome: only 13% subjects protect when the probability of a black ball is 10% in contrast to 70% protecting when the probability is 30%.

At the individual level, BP responses indicate significant heterogeneity in risk preferences. For approximately 70% of subjects (143/206), protection action increases monotonically in probability. The remaining 30% make at least one switch from protecting to not protecting and back, which is inconsistent with EU maximization.[4]

Risk-neutral agents who maximize their expected utility should start protecting when the prior exceeds 0.25, i.e., at the ratio of the protection cost to the potential loss ($5/$20). Many of our subjects (61, or 29%) start protecting at lower priors (0.05-0.15), indicating strict risk aversion. A smaller group of subjects makes choices consistent with risk loving by protecting at a probability of 0.3 or never.[5]

4.2 Informed Protection

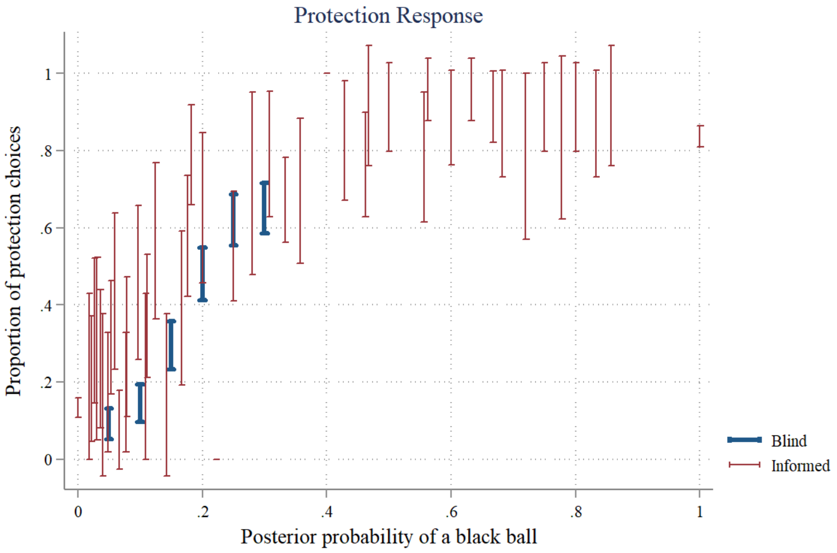
Recall that, in the IP task, subjects not only know the prior but also receive a signal about the ball color. Figure 2 shows that protection actions are increasing in the posterior probability of an adverse event, though roughly 28% of subjects break monotonicity in their protection responses

[4] That is, subjects do not protect for some treatments with posterior probability P while protecting for a posterior probability $P' < P$. Inconsistency on risk preference measures is well known. Filippin and Crosetto (2016) found that 17.1% of more than 6,300 subjects in 54 published papers made inconsistent switches on the Holt and Laury (2002) paired-lottery measure. Among our switchers, however, 76% (48/63) skip only a single increment of the presented probability scale, suggesting an inattention error.

[5] As a reference using a CRRA utility function, switching at the probability 0.1 corresponds to a coefficient of relative risk aversion $\theta = 2$, switching at 0.2 corresponds to $\theta = 0.57$, and switching at 0.3 corresponds to $\theta = -0.54$.

with respect to posterior probabilities. This is approximately the percentage of non-monotonic responses in the BP task. Breaking monotonicity here is not particularly surprising as subjects are not directly given their posterior probabilities and may estimate them incorrectly. At the individual level, we also find that the total number of times subjects protect in the BP task significantly correlates with their likelihood of protection in the IP task conditional on posteriors, but this explains only a very small part (<1%) of variation in the IP decisions.[6]

Figure 2: Average Protection Response



The bars show 95% confidence intervals for the mean proportion of subjects choosing protection at each posterior probability.

Table 2 presents the average protection decisions by prior and signal type. The first three columns indicate the signal and its information structure. Column 4 shows the posterior probability of a black ball averaged across all the treatments within a group. We calculate it from priors and FP and FN rates using the Bayes formula for each treatment and each signal value and then average across the subjects. Column 5 display the proportion of subjects who protect for corresponding treatment, and column 6 presents the weighted across treatments average of optimal responses for a risk-neutral decision maker. Finally, column 7 presents the p-value for a test of equality between empirical and theoretical protection responses.[7]

Three observations emerge from the table. First, regardless of the signal's FP and FN rates, black signals substantially increase the likelihood of protection. Second, subjects' protection decisions deviate significantly from what is optimal for risk-neutral subjects in most treatments, as evidenced by column 7. Subjects significantly overprotect when facing white signals (rows

[6] We use a linear probability model to estimate this relationship with controls for posterior probability, reported belief, and hint, and while the coefficient on the total number of protection choices is significant at the 1% level, the R^2 only increases from 0.41 to 0.42.

[7] In Appendix Table A2, we disaggregate the results by prior. The disaggregated results are largely similar to those in Table 2.

1–4), while significantly underprotecting when facing black signals without false positives (rows 5–6).

Third, we find deviations that cannot be explained by expected utility maximization for any degree of risk aversion. For example, consider rows 1 and 3: even though an increase in the FP rate does not change the posterior (because the signal is white), the protection rate increases by 23 percentage points. Similarly, comparing rows 3 and 4, we see that introducing FN rates to an information structure with an FP rate raises the protection rate increases to 41 percent — even though the average posterior probability given the signal's characteristics is merely 8 percent. As a benchmark, with no signal in the BP task, only 15(30) percent of subjects choose to protect when the probability is 10 (15) percent.

Table 2: Average Protection by Signal Type

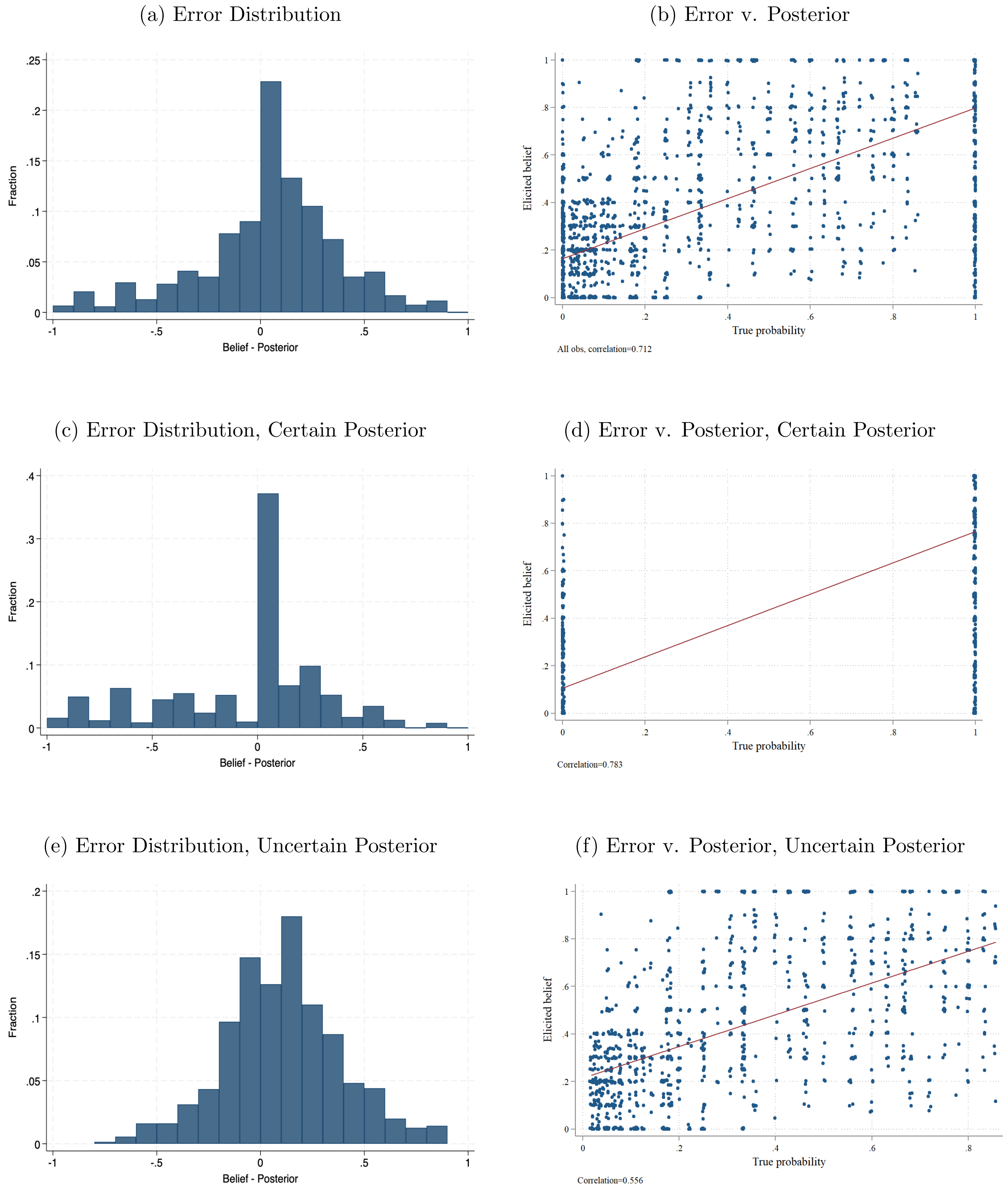
		Signal Characteristics					
Row	**Signal**	**False Positive**	**False Negative**	**Posterior**	**Share Protect**	**Share Optimal**	p
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1)	White	No	No	0.000	0.049	0.000	0.013
(2)	White	No	Yes	0.119	0.239	0.063	0.036
(3)	White	Yes	No	0.000	0.280	0.000	0.000
(4)	White	Yes	Yes	0.110	0.413	0.083	0.001
(5)	Black	No	No	1.000	0.825	1.000	0.001
(6)	Black	No	Yes	1.000	0.858	1.000	0.000
(7)	Black	Yes	No	0.512	0.809	0.875	0.427
(8)	Black	Yes	Yes	0.538	0.883	0.917	0.678

Notes: The p-value in column 7 is for the test of equality between the theoretical prediction (column 6) and the observed share of protection (column 5).

4.3 Belief Elicitation

Subject decisions in the IP task capture the use of signals in protection decisions, but decisions reflect both risk preferences and (potentially erroneous) beliefs. The BP task can be used to construct a measure of the former; the BE task measures the latter.

Figure 3: Errors in Bayesian Updating



(a) Error Distribution

(b) Error v. Posterior

(c) Error Distribution, Certain Posterior

(d) Error v. Posterior, Certain Posterior

(e) Error Distribution, Uncertain Posterior

(f) Error v. Posterior, Uncertain Posterior

Notes: The updating error is defined Belief - Posterior. Each observation is weighted by the inverse of its treatment frequency.

We define updating errors as the difference between the subjects' elicited belief and the Bayesian posterior probability of drawing a black ball for a given signal. The left-hand column of Figure 3 shows the distribution of the updating errors, while its right-hand column presents a scatter plot of the elicited beliefs against the true posterior[8] with a fitted line. We re-weight the observations by the inverse its treatment frequency to ensure balance by treatment.[9]

Panel A indicates that beliefs are still sensible despite errors. The distribution of updating errors is centered at 0, with roughly one-half (53%) of errors concentrated within +/- 0.2 interval around zero. Overall, the correlation between the elicited beliefs and the true posteriors was 0.71 (Panel (b)).

For some combinations of priors and signals, updating should be trivial and posteriors are completely certain. Panel (c) plots such cases, which account for 57% of the sample and include: (i) treatments with all-honest gremlins; and (ii) treatments with obviously irrelevant dishonest gremlins (e.g., a group comprising of only honest and white-swamp gremlins announcing that the ball is black — or vice versa). Reassuringly, 78.3% of reported beliefs in those cases are correct. About half of the errors involve reporting a probability of one when it should have been zero.

Meanwhile, Panel (e) plots the remaining observations, i.e., those with uncertain posteriors. The median error in Panel (e) is 0.11, with 90% of errors lying between -0.36 and 0.57, suggesting that, on average, subjects overestimate the likelihood of adverse events for uncertain posteriors. The correlation between beliefs and posteriors in this sub-sample falls to 0.56.[10]

[8] Posteriors are calculated from priors and information structures using the Bayes formula.

[9] Since our treatment assignment was not balanced across treatments, we may give too much weight to information structures with certain characteristics absent this reweighting. For example, we would have put too much weight on treatments with a truthful information structure since there will always be a total of two treatments with (one for each prior) in every session.

[10] The overall pattern of belief updating is consistent with the existing literature which shows that despite updating in the correct direction, people tend to underreact both to the priors and to the signals. The effect of underweighting priors — first noted in the psychology literature (Phillips and Edwards, 1966; Tversky and Kahneman, 1971; Kahneman and Tversky, 1972) — is known as *representativeness bias* or *base-rate neglect.* Using the regression approach of Grether (1980), we find both base-rate neglect and signal underweighting. Our estimates of these parameters are significantly below one with $\hat{\alpha} = 0.43$ $\hat{\beta} = 0.25$ (see Column 1 in Appendix Table A3). These values are within the range found by the meta-analysis of Benjamin (2019) which calculates the average $\hat{\alpha}$ estimate to be around 0.22 (0.4 for incentivized studies only) and the average $\hat{\beta}$ to be 0.6 (0.43 for incentivized) for studies (like ours) that presented their signals simultaneously. Such experiments are known as *bookbag-and-poker-chip* experiments

Table 3: Average Updating Error by Signal Type

Row	**Signal**	**Signal Characteristics**		**Posterior**	**Updating Error**	p
		False Positive	**False Negative**			
	(1)	(2)	(3)	(4)	(5)	(6)
(1)	White	No	No	0.000	0.048	0.023
(2)	White	No	Yes	0.119	0.051	0.008
(3)	White	Yes	No	0.000	0.210	0.000
(4)	White	Yes	Yes	0.110	0.200	0.000
(5)	Black	No	No	1.000	-0.145	0.000
(6)	Black	No	Yes	1.000	-0.384	0.000
(7)	Black	Yes	No	0.512	0.161	0.007
(8)	Black	Yes	Yes	0.538	0.022	0.692

Notes: The updating error is defined as Belief - Posterior, where Posterior is the Bayesian probability estimate for the treatment based on its information structure. The p-value in column 6 is for the test of the null hypothesis that the updating error in column 5 is equal to 0.

Table 3 summarizes how updating errors vary with the information structure.[11] We find that subjects overestimate the probability of a black ball when receiving a white signal, which is consistent with the overprotection in the IP task. This upward bias for a white signal increases in both the FP and FN rates of its information structure. To illustrate, consider the change between rows 1 and 3, where introducing an FP rate would not change the posterior (theoretical probability) because the signal is white. Yet, subjects update their beliefs upward, magnifying their updating error; we find a similar effect for the introduction of the FN rate (row 1 vs. 2).

The updating bias for black signals, however, varies by the information structure. Subjects slightly underestimate the probability with a perfectly accurate signal. However, a comparison of rows 5 and 6 suggests introducing FN rates to the information structure exacerbates subjects' underestimation. Black signals from an information structure with a non-zero FP rate leads to an overestimation of the probability of a black ball. Interestingly, adding a FN rate to the information structure attenuates this overestimation.

Subjects' choices in BE task are overall highly consistent with their choices in the IP task. We use the Mann-Whitney U-statistic to examine the number of cases when a subject inconsistently protects for a belief μ_1 but does not protect for a belief $\mu_2 \geq \mu_1$. The median U-statistic is 0.93, indicating an almost perfect rank correlation with just 1 or 2 discordant choices per subject; 35% of subjects report beliefs which are fully consistent with their protection decisions. At the same time, there is a minority of subjects with significant discrepancies: about 25% of subjects have U-statistic below 0.71 indicating multiple discrepancies.

* * *

[11] Appendix Table A4 disaggregates the results further by prior

In general, subjects choices vis-a-vis risks in BP are sensible, as is their use of signals. While subject beliefs about posterior probabilities are correlated with true posteriors, subjects err on average in ways that are related to the information structure of the signal. All of this suggests that subjects understand the experimental environment well enough to allow us to draw reasonable inferences from our WTP elicitation.

5 WTP and the Information Structure

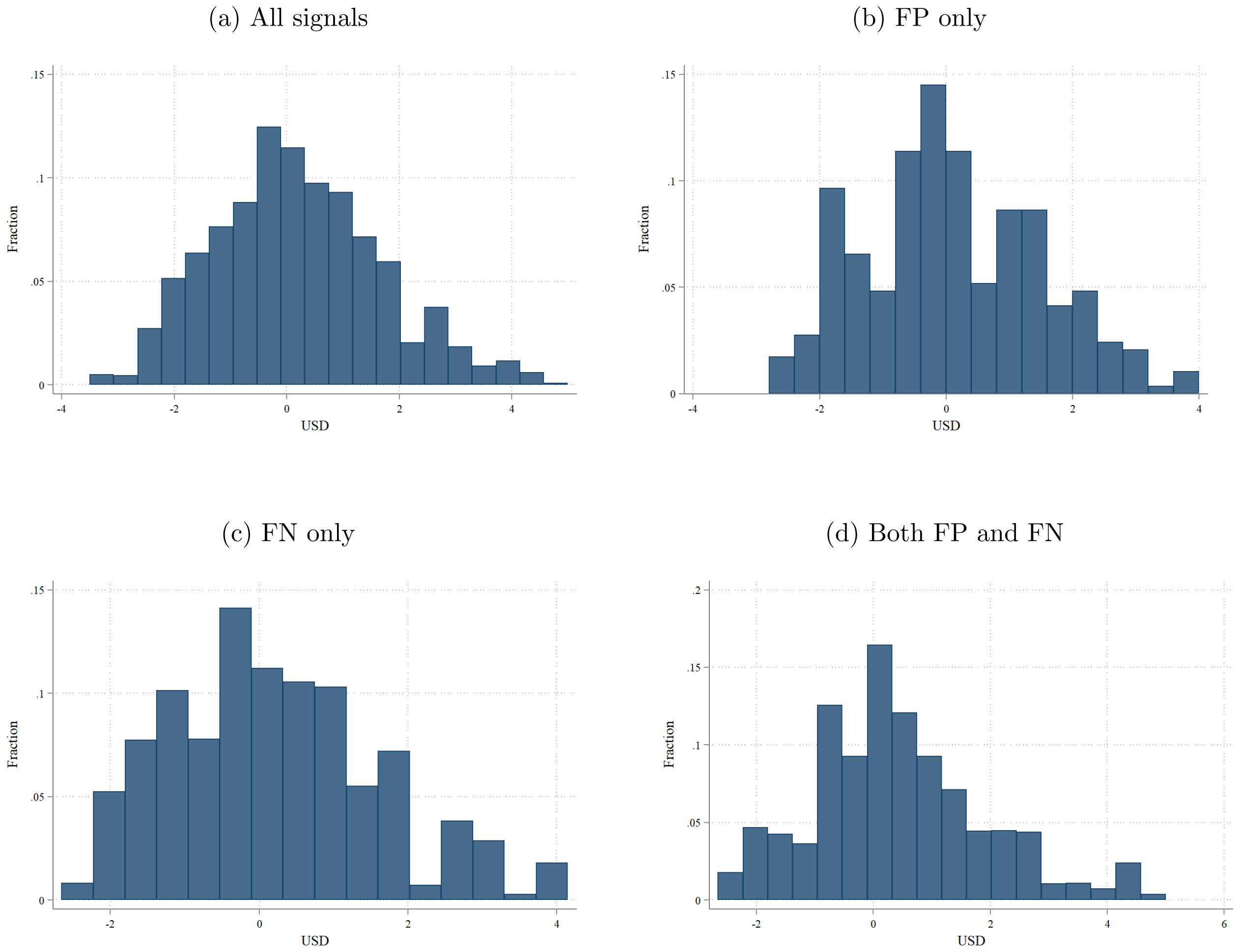
5.1 Are Subjects Risk Neutral, Expected Utility Maximizers?

Hypothesis 1. *Subjects' WTPs for signals are equal to their value for risk-neutral agents.*

Result 1. *On average, there are no significant discrepancies between the empirical WTP and the risk-neutral benchmark. When split by information structure, a discrepancy emerges only for signals with both false-positive and false-negative rates. However, subjects overvalue signals with only FP rates for low priors and undervalue signals with only FN rates for high priors.*

Overall, the theoretical signal value for a utility maximizing risk-neutral subject (hereafter, the risk-neutral WTP) in equation 3 is a reasonable benchmark of our subjects' WTP. Figure 4 plots the distribution of the differences between subjects' empirical WTP and the theoretical risk-neutral WTP. As in Figure 3, we re-weight the observations using inverse of treatment frequencies. The distribution is centered around 0, indicating that average choices do not fall far from the choices of a risk-neutral utility maximizer. However, there is substantial variation: only 22% of reported WTP are within \$0.50 of the risk-neutral signal value, and subjects overvalue signals by at least \$1 in 25% of cases and undervalue by at least \$0.83 in 25% of cases. Introducing FP and FN rates does not increase the range or variation of discrepancies, but introduces a long tail of positive discrepancies shifting the average upward.

Figure 4: WTP Discrepancy by Information Structure


(a) All signals

(b) FP only

(c) FN only

(d) Both FP and FN

Notes: WTP Discrepancy is defined as Empirical WTP - Risk-Neutral WTP. Each observation is weighted by the inverse of its treatment frequency.

Our comparisons in Table 4 also do not find differences between the empirical WTP and the risk-neutral WTP for 3 out of 4 signal types: honest, FN-only, and FP-only. For signals from information structures with both FP and FN rates, however, the empirical WTP is significantly higher than the risk-neutral WTP. Subjects' overvaluations were similar for both low and high priors, but they are not statistically significant. Additionally, subjects tend to overvalue signals with positive FP-only rates in low-prior environments ($p \in \{0.1, 0.2\}$) and undervalue signals in high-prior environments ($p \in \{0.3, 0.5\}$).

Willingness-to-pay for signals increases with priors starting from \$0.86 on average when $\pi = 0.1$ and reaching \$2.10 for the highest prior of 0.5. This stands in contrast to the theoretical value, which is largest for a prior of 0.3 (followed closely by the value for for a prior of 0.2). Hence subjects tend to overpay for signals with low priors and underpay for medium priors.

Table 4: Average WTP Discrepancy by Signal Type

Priors	Honest	FN only	FP only	FP and FN
All priors	-0.260	0.192	0.085	0.417*
	(0.377)	(0.206)	(0.203)	(0.232)
Low priors	-0.059	-0.011	0.609**	0.436
	(0.232)	(0.232)	(0.232)	(0.232)
High priors (>0.2)	-0.461	0.395	-0.440*	0.398
	(0.232)	(0.232)	(0.232)	(0.232)

Notes: The coefficient for each cell comes from a regression of WTP discrepancy on a constant by group. Standard errors in parentheses (clustered by subject and treatment). */**/*** indicates statistical significance at 10/5/1 percent.

5.2 The Responsiveness of WTP to Error Costs

Hypothesis 2. *Subjects' preferences demonstrate equal sensitivity to the expected costs from false-positive and false-negative outcomes.*

Result 2. *On average, we cannot reject the hypothesis of equal sensitivity. However, we observe significant heterogeneity with respect to priors: subjects tend to overvalue false-positive costs for low prior probability events and overvalue false-negative costs for high prior probability events.*

Ultimately, we want to learn how people respond to the expected *costs* from signals generated by different information structures and whether our simple theoretical model adequately captures this behavior. We therefore first examine whether and how subjects deviate from the benchmark WTP of our risk-neutral, Bayesian-updating, utility-maximizing agent. We then test whether these deviations can be explained by risk preference or updating accuracy. Finally, we describe our findings on the heterogeneous sensitivity of the WTP by prior — which motivates our discussion on the potential mechanisms behind these deviations that we address in the next section.

We begin by estimating the relationship between the deviations of the empirical WTP from the benchmark model on FP and FN costs. The FP and FN costs studied here — defined above as $\pi(1-P_{01})c$ and $\pi P_{01}L$ — differ from their respective rates because they incorporate the cost associated with each error type. For example, a high false-negative rate imposes fewer costs when priors are low because the adverse outcome is already unlikely, whereas a high false-negative rate carries substantial costs when the prior probability of the adverse outcome is high. We estimate the following regression:

$$\Delta b_{is} = \beta_0 + \beta_1 FPC + \beta_2 FNC + \delta_i + \varepsilon_{is}$$

where $\Delta b_{is} = (b_{is} - b_s^*)$ is the difference between individual i's WTP and the risk-neutral WTP for signal s, and FPC (FNC) is the false positive (false negative) cost. All specifications include subject fixed effects δ_i, with standard errors clustered at the subject and treatment levels.

Table 5: Deviations from Signal Value (WTP - Value) and Signal Characteristics

	All			Prior	
				$\{.1, .2\}$	$\{.3, .5\}$
	(1)	(2)	(3)	(4)	(5)
FP costs	0.421	0.487	0.643	0.577	0.303
	(0.258)	(0.267)*	(0.264)**	(0.186)***	(0.246)
FN costs	0.287	0.327	0.357	0.016	0.367
	(0.117)**	(0.131)**	(0.133)**	(0.199)	(0.070)***
Risk-averse × FP costs		-0.329	-0.415	-0.243	-0.576
		(0.174)*	(0.241)*	(0.241)	(0.355)
Risk-averse × FN costs		-0.355	-0.361	-0.352	-0.288
		(0.111)***	(0.112)***	(0.204)*	(0.103)**
Risk-loving × FP costs		0.048	0.018	0.008	0.318
		(0.126)	(0.182)	(0.308)	(0.327)
Risk-loving × FN costs		0.080	0.110	0.361	0.119
		(0.073)	(0.093)	(0.339)	(0.101)
Constant	-0.308	-0.310	-0.436	-0.076	-0.517
	(0.263)	(0.263)	(0.230)*	(0.172)	(0.139)***
R^2	0.480	0.491	0.500	0.731	0.745
Prob>F	0.0559	0.0433	0.0197	0.0003	0.0000
Obs	1230	1230	1230	615	615
FP=FN	0.539	0.495	0.204	0.000	0.746
Risk-Averse Subjects:					
False Positive		(0.158)	(0.228)	(0.334)	(-0.273)
se		[0.308]	[0.280]	[0.221]	[0.324]
p-value		0.609	0.421	0.144	0.408
False Negative		(-0.028)	(-0.003)	(-0.337)	(0.079)
se		[0.138]	[0.140]	[0.213]	[0.104]
p-value		0.841	0.982	0.127	0.454
Risk-Loving Subjects:					
False Positive		(0.535)	(0.661)	(0.585)	(0.621)
se		[0.257]	[0.221]	[0.248]	[0.219]
p-value		0.043	0.004	0.027	0.009
False Negative		(0.406)	(0.467)	(0.377)	(0.487)
se		[0.117]	[0.121]	[0.306]	[0.072]
p-value		0.001	0.000	0.231	0.000
Subject FE	Yes	Yes	Yes	Yes	Yes
Inconsistent Risk Pref. Interactions	No	Yes	Yes	Yes	Yes
Inaccurate Belief Interactions	No	No	Yes	Yes	Yes
Prior Probability FE	No	No	No	Yes	Yes

Notes: Standard errors in parentheses (clustered at the subject and treatment levels). */**/*** indicates statistical significance at 10/5/1 percent. The bottom panels include tests of whether the total coefficient values (baseline + interaction) are different from zero.

Table 5 reports the results of our regression. Column 1 presents the basic results that do not account for risk preference and belief accuracy. If subjects are risk-neutral expected-utility maximizers, we expect all coefficients to be jointly and individually insignificant. Instead, column 1 shows positive coefficients for both FP and FN costs, albeit only significant for the latter. As costs increase, subjects should downward-adjust their WTP accordingly. These positive coefficients show that relative to the risk-neutral benchmark, our subjects did not react enough to rising costs (by adjusting their WTPs), resulting in a WTP discrepancy that increases with costs.

Theoretically, risk-neutral subjects value the marginal costs of false-negative and false-positive events symmetrically. Although column 1 shows that the coefficient on FN costs is slightly lower, the difference between the coefficients on FP and FN costs is not statistically significant.

Risk Preference and Belief Accuracy Since our benchmark model assumes both risk neutrality and perfect updating, the finding that our subjects do not adequately adjust their WTP to rising error costs could arise from two channels. First, Proposition 2 suggests that risk preferences can influence the sensitivity of WTP to FP and FN rates. Second, systematic biases during updating can also cause deviations.

We find that risk preferences affect the sensitivity to a signal's quality, but fall short in explaining the systematic WTP biases reported above. We use data from the BP task to categorize subjects by their risk preference. We classify all the subjects with internally consistent BP choices into three categories: risk averse, risk neutral, and risk loving.[12]

Column 2 explores the heterogeneity of subject responses to FP and FN costs by their risk preferences, with risk-neutral as the default category. The total coefficient values for risk-averse and risk-loving subjects, including their respective standard errors and p-values, are presented at the bottom panel. Among risk-neutral and risk-loving subjects, the WTP discrepancies increase with both FP and FN costs — suggesting that they do not downward-adjust their WTP enough in the face of increasing error costs due to the lower signal quality. In contrast, the WTP discrepancies of risk-averse subjects are uncorrelated with rising error costs.

Subjects' under-reaction to deteriorating signal quality remains even after we further control for their ability to Bayesian update. We use data from the BE task to construct a measure of subjects' (posterior) belief accuracy.[13] Column 3 presents the most flexible specification that controls for belief accuracy and risk preference by including triple interactions of belief accuracy, risk preference, and signal characteristics. The baseline group is the group of risk-

[12] We classify most subjects to risk-averse, risk-loving or risk-neutral based on the total number of protection choices made in the BP task, with 2 and 3 choices corresponding to risk-neutrality (protecting starting from 0.2 or 0.25). Subjects that make more than one inconsistent choice in BP are included as their own category.

[13] We calculate a belief error as the absolute value of the difference between the subject's belief and the true posterior probability and then average these errors across all the decisions with identical priors, false positive, and false negative rates. A subject's posterior belief for a decision is defined as accurate if its error is less than the median error across all the subjects making the same decision.

neutral subjects with relatively accurate beliefs. We find that subjects with more accurate beliefs are not more likely to adjust optimally to rising FP and FN costs on average.[14]

Heterogeneity by Prior We motivate our experiment with a real-world problem of designing warning systems — often for events with low probabilities. With a low prior, the default action of risk-neutral subject would be not to protect, and vice versa with a high prior. The signal would help risk-neutral subjects decide whether to maintain or switch away from the default action. We split the priors into two groups using the threshold of 0.25 (= protection cost/potential loss), and we incorporate prior-probability fixed effects to the aforementioned flexible specification.

Column 4 of Table 5 presents the results for low-prior WTPE tasks. For a low prior, i.e, $p \in \{0.1, 0.2\}$, the WTP discrepancies of risk-neutral and risk-loving subjects from the risk-neutral benchmark increase with FP costs. These subjects do not optimally reduce their WTP enough as signal quality deteriorates, leading to the overvaluation of FP signals. Column 5 presents the results for high-prior WTPE tasks. With high priors, the WTP discrepancies of risk-neutral and risk-loving subjects once again increase with FN costs, leading to the overvaluation of FN signals.

* * *

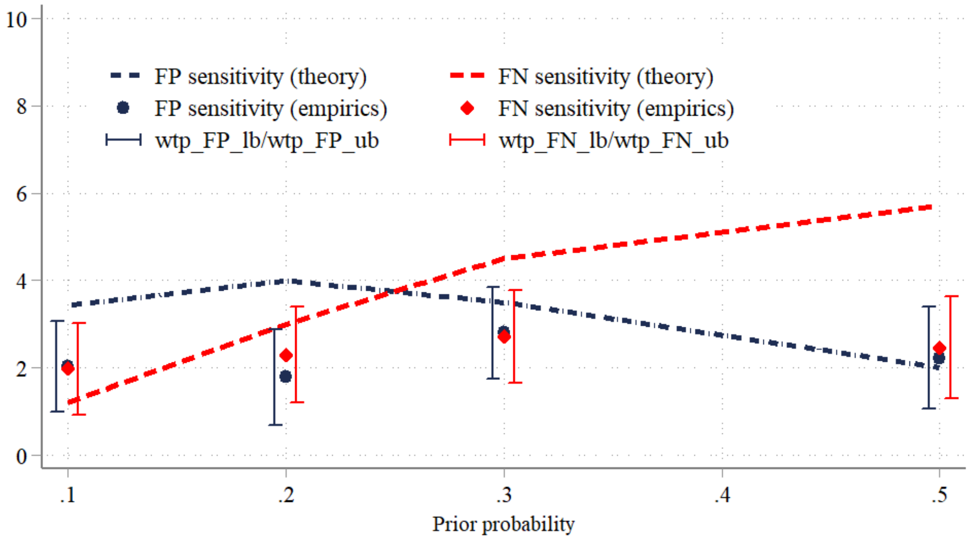
To address the potential bias from WTP censoring both at zero and at the upper bound of \$5, we also estimate the same relationship using a tobit regression (see Appendix Table A7). To maintain comparability, we use the difference between the elicited WTP and theoretical value as the dependent variable and account for varying upper and lower bounds. We find that our essential results do not change: subjects underreact to FP and FN costs for all the priors on average, and underreact to FP costs for low priors and to FN costs for high priors.

5.3 Understanding Source(s) of Empirical Deviations

The under-reactions of subjects' WTP to FP (FN) costs for low (high) priors are inconsistent with our risk-neutral model. Equations 4 and 5 in Section 2 suggest that WTP should respond more to FP rates (relative to FN rates) for low priors and vice versa for high priors. That is, for a given FN rate, false negative outcomes are much less likely with low priors and hence impose lower costs on the agent. As priors increase, FN rates become more salient while FP rates become less salient. Instead, our subjects react very similarly to FP and FN rates for both low and high priors.

[14] Aside from these theoretically motivated individual differences, we investigate several other characteristics. Heterogeneity is not driven by demographic characteristics (e.g., age, gender) or prior statistical coursework. The complete set of results are in Appendix Table A5.

Figure 5: Theoretical and Estimated WTP Sensitivities to Error Rates by Prior



OLS estimates of sensitivity to FP and FN rates by prior probability of a black ball.

To illustrate, we obtained the sensitivities of WTP on FP and FN rates by estimating the following regression for each prior for the elicited WTP (b^*):

$$b^* = \alpha + \beta_{FP} FP + \beta_{FN} FN + \epsilon \tag{8}$$

In Figure 5, we overlay these estimates on the line plots of the theoretical sensitivities of WTP on FP and FN rates across different priors (based off Equation 8).[15] The figure shows that the point estimates of the sensitivity coefficients of WTP on FP rates for lower priors ($p \in \{0.1, 0.2\}$) are below their theoretical benchmarks. In contrast, the point estimates of WTP sensitivity on FN rates for higher priors ($p \in \{0.3, 0.5\}$) are significantly below. Surprisingly, across all priors, the empirical estimates of the sensitivities on FP and FN rates are very close to each other.

We consider six candidate explanations for these puzzling results: (i) risk preferences; (ii)probability weighting; (iii) anchoring; (iv) valuing non-instrumental information; (v) the anticipation of bias in the protection decision; and (vi) a failure to distinguish how FP and FN error rates ought to affect calculated posteriors differently.

Risk Preferences. Our evidence suggests that risk preferences do not explain this behavior. We test the risk preference hypothesis using subjects' BP choices. Columns 4 and 5 of Table 5 already show that, even after controlling for subjects' risk preferences, the coefficients on FP and FN costs remain very different for low and high priors. We augment this analysis

[15] Given that theoretical sensitivities depend on information structures which are pooled within priors, we fit the same regression to the theoretical values and report coefficients as theoretical sensitivities. With tobit estimates, the empirical coefficients are generally higher but are also very close to each other and follow the same pattern. However, we cannot fit them easily for the theoretical values.

by explicitly testing for interactions between risk-preferences, priors, and FP and FN rates (Appendix Table A6). We find that these interactions are mostly insignificant, except for the interactions between FN rates and risk aversion in low-prior environments. The heterogeneity largely remains after controlling for risk preferences, but the interaction between high priors and FP rates becomes insignificant.

Probability Weighting. Rank-dependent probability weighting assumes that subjects evaluate lotteries using transformed outcome probabilities that account for the preference ranking of each outcome and cumulative probability distribution. It helps to explain extreme low probability outcomes receiving disproportionate weight in decision-making (Quiggin, 1982) and hence is applicable to protection decisions in which losses can be both rare and extreme.

Assuming that each lottery p has ordered outcomes $x_1 \succ x_2 \succ ... \succ x_n$ with probabilities p_i for the outcome i, rank-dependent utility re-weights cumulative probabilities of each outcome using a weighting function $\phi : [0,1] \to [0,1]$ in the following way:

$$U(p) = \phi(p_1)u(x_1) + \sum_{i=2}^{n} \left(\phi\left(\sum_{j=1}^{i} p_j \right) - \phi\left(\sum_{j=1}^{i-1} p_j \right) \right) u(x_i)$$

As follows from the equation, the weights depend on the ranking of the outcomes and the probabilities within each lottery. In the signal purchase decision, subjects compare a simple blind protection lottery with the compound lottery of first drawing a signal and then choosing a trivial protection outcome or a simple no protection lottery with a signal. We assume that the reduction of compound lotteries applies so that the compound lottery of informed protection is ranked equivalently to a reduced simple lottery with weighting applied to all the possible outcomes. This means that the utility with a signal can be written as:

$$\begin{gathered} U_S = \phi(\pi P_{01})u(Y-b-L) + [\phi((1-\pi)P_{10} + \pi P_{11}) - \phi(\pi P_{01})]u(Y-b-c) + \\ + [1 - \phi((1-\pi)P_{10} + \pi P_{11}) - \phi(\pi P_{01})]u(Y-b) \end{gathered}$$

After applying implicit differentiation by FP and FN rates we obtain that the ratio of sensitivities to FN and FP rates should satisfy:

$$\frac{db/dP_{01}}{db/dP_{10}} = \frac{\pi}{(1-\pi)} \Big(\frac{\phi'(\pi P_{00})}{\phi'(\pi + (1-\pi)P_{10})} \Big) \frac{(u(Y-c-b) - u(Y-L-b))}{(u(Y-b) - u(Y-c-b))} \tag{9}$$

Probability weighting with reasonable parameter choices can explain higher relative sensitivity to FN costs when the priors are low, though it does not explain lower relative sensitivity when the priors are high. First, note that this ratio is equivalent to the expression we derived for the EU maximizers in the general case, except the additional multiplier $\Lambda \equiv \left(\frac{\phi'(\pi P_{00})}{\phi'(\pi + (1-\pi)P_{10})} \right)$ which is equal to a ratio of derivatives of the weighting functions at two different points. We know that $\pi P_{01} < \pi < \pi + (1-\pi)P_{10}$ and that, for our experimental parameters, $\pi + (1-\pi)P_{10} \leq 0.75$

meaning that the value is still likely to fall in the mid-range of an S-shaped weighting function. Thus, for small priors π or for small FN rates P_{01}, we have $\Lambda > 1$, implying higher relative sensitivity to FN rates. For higher priors and FN rates Λ would be closer to one making little difference in relative sensitivity[16].

Anchoring. The evidence also does not support the hypothesis that anchoring on previous priors explains higher sensitivity to FN rates for low priors. The prior order is randomized with some subjects starting with lower priors and other starting with higher priors. Most subjects (159 out of 206) also change their reported WTP when facing a new prior in round 4, demonstrating responsiveness to information on priors. The average belief error in the BE task is actually *lower* for the second set of priors rather than the first (0.201 vs 0.186), which suggests that changing priors does not make subjects more confused. Most importantly, our main results on WTP sensitivity hold even when we limit the sample only to the first prior in each sequence.

Non-instrumental information. There is evidence in the literature of people valuing "non-instrumental information" that does not affect their decisions. For example, Eliaz and Schotter (2010) find that subjects are willing to pay to know the probability of their choice being correct even if this information cannot affect their choice, while Ganguly and Tasoff (2017) document that most people are willing to pay a small amount to know their pre-determined experimental payoffs at the beginning of the experiment rather than at the end. Similarly, Masatlioglu, Orhun and Raymond (2023) reveal a strong and systematic preference for positively skewed information, even when such signals are less informative, suggesting that anticipatory emotions such as hope and anxiety shape information demand independently of instrumental value.[17] Most information in our experiment is instrumental by design, i.e., it informs their choices, and indeed enters into subjects' decisions as evidenced by choices in the IP task. Nonetheless, many subjects have a positive WTP for signals that cannot affect their IP decisions (295 out of 1242 total choices). It is therefore plausible that the reported WTPs include some non-instrumental components.

Preferences for non-instrumental information cannot, however, provide a full explanation of our results, mainly because there is no time delay between receiving a signal and learning the outcome. If the WTP task round is selected as a payment round, the subject receives a signal,

[16] Gonzalez and Wu (1999) report very diverse estimates of shapes of weighting functions estimated individually for 10 subjects. We calculate the ratio using linear in odds probability weighting function for all their parameter estimates and find that the ratio is greater than one for all our treatments except the treatment with prior $\pi = 0.5$ and the FN rate of 0.5. Using their power weighting function estimates produces ratios which are always greater than one.

[17] Unlike Masatlioglu, Orhun and Raymond (2023), Oliveros, Zultan and Llorente-Saguer (2025) find that decision makers exhibit a striking and persistent preference for symmetric sources, i.e., those offering a balanced distribution of positive and negative signals, even when asymmetric sources would yield more decision-relevant information. The authors argue that symmetry exerts an intrinsic appeal, suggesting that individuals prefer information structures that "feel fair" or representative, independent of their actual informativeness.

chooses an action, and then immediately learns their payoffs. This leaves essentially no window for anticipatory feelings assumed to be the causal mechanism behind the demand for non-instrumental information in most of the literature. Additionally, the closeness of coefficients for FP and FN rates also seems a priori implausible based only on the non-instrumental information value story because, in contrast to our last explanation, no theory of preferences for non-instrumental information suggests the effects should be so similar to one another.

Anticipation of Bias in Protection Decision. Since both IP and WTP task responses exhibit certain unexpected biases, subjects might adjust their bids in the WTP task in anticipation of the projected distortions in their use of the signal in the IP task. For example, a subject expecting to always protect for a particular information structure might report a lower WTP. This should be expected if subjects make consistent choices across tasks. We find that their IP biases do affect their bids, but fail to explain the main patterns of sensitivities to FP and FN costs that we observe.

To show this, we define two variables measuring the discrepancies in expected payoffs due to elicited strategies in IP (a_{iw}, a_{ib}) and BP tasks (a_i) versus optimal decisions for a risk-neutral decision-making $(a_i^*, a_{ib}^*, a_{iw}^*)$. The first variable measures the protection-cost component and the second variable the loss component of the shortfall in the value of the signal under the subject's elicited strategies relative to the optimal benchmark[18]:

$$\Delta_i^{\text{prot}} \equiv c\Big\{(a_i - a_i^*) - \big[P(S=W)(a_{iw} - a_{iw}^*) + P(S=B)(a_{ib} - a_{ib}^*)\big]\Big\}$$

$$\Delta_i^{\text{loss}} \equiv \pi L\Big[(a_i^* - a_i) - \big(P_{11}(a_{iw}^* - a_{iw}) + P_{01}(a_{ib}^* - a_{ib})\big)\Big]$$

Here $P(S=B) = \pi P_{11} + (1-\pi)P_{10}$ is the probability of a black signal, and $P(S=W) = \pi P_{01} + (1-\pi)P_{00}$ is the opposite. The sum of these two variables equals a decrease in the signal's value for a risk-neutral subject if using actual subject i strategies. Defining the expected benefit in this way allows for an asymmetry in responding to large losses versus relatively small protection costs.

Appendix Table A10 reports the estimation results of the deviation of the empirical WTP from the signal value (similar to Table 5) with the two new measures of bias in informed protection decisions. The coefficients on the added covariates are statistically significant but relatively small in magnitude. At the same time, both coefficients on FP costs for low priors and FN costs for high priors remain highly statistically significant. Thus, we conclude that subjects' biases in how they will utilize signals when facing an IP task do not explain the observed pattern of sensitivities to FP and FN rates.

[18] Given that there is no elicited blind protection choice for prior of $\pi = 0.5$, we assume full protection. For context, 65% of subjects protect for the highest prior probability of 0.3 in our BP elicitation.

Failure to distinguish between FP and FN Instead, we argue that the observed sensitivity patterns emerges because many subjects neglect the difference between FN and FP rates, treating these two characteristcs as if they were the same thing. Subjects' own proffered explanations motivate our consideration of this possible mechanism. At the end of the experiment, we asked subjects to explain to us how they made their protection choices.[19] Out of 206 subjects, 70 refer to the *percentage* or presence of dishonest gremlins as their rationale for choosing protection. For example:

- "*my strategy for this task was to only buy protection if there was a white or black gremlin and not if there was a truth gremlin*"
- "*If there were only honest gremlins then I never protected but even if there was one white-swamp gremlin or one black-swamp gremlin then I payed* (sic) *for protection.*"

Among the other 136 subjects, some may use this heuristic without describing it and many subjects make decisions solely based on priors.[20] The similarity of the coefficient estimates for FP and FN rates for each prior as reported in Figure 5 is consistent with these statements. If subjects neglect the difference between FP and FN risks when choosing their WTP, it would explain both the coefficients' similarity and the lack of variation with respect to priors in Figure 5. Indeed, if subjects treat FP and FN rates the same and consider only the total proportion of false signals, they would assign equal weights to each of them, and the best-fit line of the signal value on the sum of FP and FN rates should be relatively flat because priors affect FP and FN costs in opposite ways.[21]

We can test this hypothesis using choices from the BE and IP tasks when signals come from information structures that are obviously irrelevant in the wake of the signal. If subjects systematically neglect the difference between FP and FN rates, we expect subjects to (unexplainably) react to irrelevant information structures that would not have affected posteriors, namely that subjects would still react to FP (FN) rates when given a white (black) signal. This could happen if subjects react to FP rates as if they are FN rates, and vice versa. If present, this pattern cannot be explained by any of the aforementioned mechanisms.

Table 6 presents the results from linear regressions of updating error (i.e., belief - posterior) on FP and FN rates by signal color with individual fixed effects to control for updating biases. Consistent with our conjecture, we observe that the FP rate has a significant positive effect on

[19] The text of all the responses are in the appendix.

[20] Several studies find that decision makers use heuristics in choice environments like this. For example, Montanari and Nunnari (2023) investigate selective exposure to information in an environment where two biased information sources differ in direction and reliability and find deviations from Bayesian rationality including reliance on simple heuristics such as "listen to the more reliable source." Similarly, Charness, Oprea and Yuksel (2021) study how individuals choose between two information sources with opposing biases when forming beliefs about a binary state and find that subjects rely on intuitive, but flawed, heuristics about which sources are more "useful."

[21] The equality of coefficients on FP and FN rates is a necessary prediction of this explanation, but it can also emerge by chance with (some) heterogeneous risk preferences and probability weighting.

Table 6: Updating Errors in BE Task

	All	Signal Received	
		White	Black
	(1)	(2)	(3)
FP rate	0.797***	0.532***	1.063***
	(0.069)	(0.051)	(0.127)
FN rate	-0.243***	0.042	-0.528***
	(0.060)	(0.057)	(0.101)
Observations	2460	1230	1230
Adjusted R^2	0.203	0.407	0.543
Subject FE	Yes	Yes	Yes

Notes: Standard errors in parentheses (clustered at the subject and treatment level). */**/*** indicates statistical significance at 10/5/1 percent.

the error when the signal is *white*, and that the FN rate has a significant negative effect when the signal is *black*.

To further explore this hypothesis, in Table 7, we regress IP decisions on FP and FN rates and flexible controls of both posteriors and reported beliefs:[22]

$$Prob(s_{ij} = 1) = \alpha_i + \beta_1 P_{10} + \beta_2 P_{01} + Z(P_{ij}) + Z(\mu_{ij}) + \epsilon_{ij}$$

Here s_{ij} is the protection decision of subject i in treatment j, α_i is subject FE, P_{10}, P_{01} are FP and FN rates, and $Z(P_{ij})$ and $Z(\mu_{ij})$ are the splines of FP or FN rates and reported beliefs μ_{ij} to control for these variables in a flexible way. Each spline is a function $Z(x)$ which is just linear $x + C$ within one interval, and constant everywhere else. The splines are constructed so that their linear intervals cover the whole domain of probabilities and beliefs $[0, 1]$.[23] Columns 1 and 2 include only the flexible controls of the true posteriors. Columns 3 and 4 add further flexible controls to account for subjects' (often incorrect) beliefs, inferred from their BE responses.

Columns 1 and 2 show that, even after we condition on posterior and subject FEs that account for risk preferences, IP choices still changed with FP and FN rates. Notably, we find that FP rates increase the tendency to overprotect for a white signal. The effect remains after allowing for heterogeneity of sensitivities to FP and FN rates with respect to priors (Column 2). Columns 3 and 4 show that adding flexible controls for subjects' beliefs reduces the coefficient magnitudes, but they remain statistically significant. This suggests that beliefs offer only a partial explanation to these protection anomalies.

Overall, we observe a striking uniformity in sensitivity of WTP to both FP and FN rates.

[22] Given that the true functional form is unknown, we use a linear probability model to get unbiased coefficient estimates.

[23] We use Stata's `mkspline` command to create 5 splines $z_1(x), z_2(x), ..z_5(x)$ of initial variable x over the range $[0, 1]$ such that $z_k(x) = \min[0, x - x_{k-1}, x_k - x_{k-1}]$ with x_k being equally spaced knot values. Splines account for potential nonlinear effects of posteriors and beliefs on protection decision with limited effect on degrees of freedom.

Table 7: Informed Protection Response

	(1)	(2)	(3)	(4)
FP rate x (S=White)	0.895***	0.943***	0.525***	0.571***
	(0.089)	(0.093)	(0.095)	(0.098)
FN rate x (S=White)	0.537***	0.532***	0.307**	0.299**
	(0.145)	(0.146)	(0.144)	(0.146)
FP rate x (S=Black)	-0.032	0.025	-0.065	-0.000
	(0.201)	(0.204)	(0.196)	(0.202)
FN rate x (S=Black)	0.103	0.069	-0.005	-0.021
	(0.074)	(0.080)	(0.087)	(0.093)
S=Black	0.531***	0.542***	0.383***	0.374***
	(0.103)	(0.114)	(0.105)	(0.115)
p>0.2	0.039**	0.041**	0.024	0.029
	(0.016)	(0.021)	(0.015)	(0.020)
FP rate x (p>0.2)		-0.081		-0.085
		(0.076)		(0.079)
FN rate x (p>0.2)		0.088		0.048
		(0.099)		(0.093)
N	2424	2424	2424	2424
Pseudo R-squared	0.505	0.505	0.538	0.539
Log-likelihood	-830.188	-829.168	-773.731	-773.018
Subject FE	Yes	Yes	Yes	Yes
Flexible controls for:				
Posterior	Yes	Yes	Yes	Yes
Beliefs	No	No	Yes	Yes

Notes: Coefficients are average marginal effects. Standard errors in parentheses (clustered at the subject level). */**/*** indicates statistical significance at 10/5/1 percent.

This pattern is consistent with subjects neglecting the difference between FP and FN signals, a behavior that is supported by subjects' explanations of their decision making and the anomalous sensitivities to FP and FN rates in other treatments in which they do not affect posterior probabilities.[24]

6 Conclusion

We study how information structures affect valuation of a warning signal. While the risk-neutral benchmark model does a good job of describing the average elicited WTP, it masks an important underlying heterogeneity. Relative to the risk-neutral WTP, individual valuations of warning signals inadequately adjust for declining signal quality as false-positive (false-negative) costs increases for low (high) prior probability events. These deviations can be partly explained by risk preferences and the inability of subjects to discriminate between false-negative and false-positive errors. Probability weighting might also play a role in the case of low priors.

Our findings have direct applications for the design of warning signals. Bias in preferences for signals arising from risk aversion means that prioritizing false-negative outcomes while designing warning signals can be welfare enhancing, even when it increases the expected costs of using the signal. At the same time, finding ways to better present the signal's information structure to help users discriminate between false-positive and false-negative payoffs should improve individual decision making. Studies on Bayesian updating, for instance, show that medical professionals make better decisions if information on medical tests is presented in the form of expected frequencies rather than as a tuple of prior conditional probabilities (Hoffrage et al., 2015; McDowell and Jacobs, 2017).

Understanding the source of this asymmetry can also help reduce externalities associated with signal (mis)use. We find that people overvalue signals with excessive false alarms for a typical case with low-probability events. In some cases, individuals may not incur the full cost of false positive signals. For example, the cost of medical overtreatment from tests with high false positives may be an externality absorbed by the healthcare system. Similarly, first responders may be required by law to respond to automatic fire alarms installed in commercial buildings resulting in excess costs borne by taxpayers when false positive rates are high.

Future research should also explore the role of warning signals as an experience good, to wit, how prolonged exposure to a signal with a given information structure can influence individual biases. For example, buying a smoke alarm that never misses may seem like a good idea at first, but may seem less so after a year experiencing repeated false alarms. Given our finding

[24] This pattern is consistent with greater bias in belief elicitation when subjects have to engage in contingent reasoning versus smaller belief biases when eliciting responses after presenting a signal results, as Aina, Amelio and Brütt (2023) found. This result implies that decisions to acquire information, such as decisions made in our experiment where subjects determining their WTP have to reason through contingencies, might suffer from persistent inherent biases. Indeed, we find that subjects commit reasoning errors which reduces the correlation between their WTP for a signal and its usefulness for decreasing expected potential costs.

of insufficient attention to false-positive costs for rare events, ex-ante preferred warning signals could later be considered overly sensitive and ignored, leading to *alarm fatigue* and increasing the risk of potentially preventable disasters.

References

Aina, Chiara, Andrea Amelio, and Katharina Brütt. 2023. "Contingent Belief Updating." *ECONtribute Discussion Papers Series.*

Ambuehl, Sandro, and Shengwu Li. 2018. "Belief updating and the demand for information." *Games and Economic Behavior*, 109: 21–39.

Benjamin, Daniel J. 2019. "Chapter 2 - Errors in probabilistic reasoning and judgment biases." In *Handbook of Behavioral Economics: Applications and Foundations 1.* Vol. 2 of *Handbook of Behavioral Economics - Foundations and Applications 2*, , ed. B. Douglas Bernheim, Stefano DellaVigna and David Laibson, 69–186. North-Holland.

Brown, Robbie. 2010. "Oil Rig's Siren Was Kept Silent, Technician Says." *The New York Times.*

Charness, Gary, Ryan Oprea, and Sevgi Yuksel. 2021. "How Do People Choose Between Biased Information Sources? Evidence from a Laboratory Experiment." *Journal of the European Economic Association*, 19(3): 1656–1691.

Coutts, Alexander. 2019. "Good news and bad news are still news: experimental evidence on belief updating." *Experimental Economics*, 22(2): 369–395.

Danz, David, Lise Vesterlund, and Alistair Wilson. 2020. "Belief Elicitation: Limiting Truth Telling with Information on Incentives." National Bureau of Economic Research w27327, Cambridge, MA.

Eliaz, Kfir, and Andrew Schotter. 2010. "Paying for confidence: An experimental study of the demand for non-instrumental information." *Games and Economic Behavior*, 70(2): 304–324.

Fawcett, Tom. 2006. "An introduction to ROC analysis." *Pattern Recognition Letters*, 27(8): 861–874.

Filippin, Antonio, and Paolo Crosetto. 2016. "A reconsideration of gender differences in risk attitudes." *Management Science*, 62(11): 3138–3160.

Friedl, Andreas, Katharina Lima de Miranda, and Ulrich Schmidt. 2014. "Insurance demand and social comparison: An experimental analysis." *Journal of Risk and Uncertainty*, 48(2): 97–109.

Ganguly, Ananda, and Joshua Tasoff. 2017. "Fantasy and Dread: The Demand for Information and the Consumption Utility of the Future." *Management Science*, 63(12): 4037–4060.

Gonzalez, Richard, and George Wu. 1999. "On the Shape of the Probability Weighting Function." *Cognitive Psychology*, 38(1): 129–166.

Grether, David M. 1980. "Bayes Rule as a Descriptive Model: The Representativeness Heuristic." *The Quarterly Journal of Economics*, 95(3): 537–557.

Grether, David M. 1992. "Testing bayes rule and the representativeness heuristic: Some experimental evidence." *Journal of Economic Behavior & Organization*, 17(1): 31–57.

Hoffman, Mitchell. 2016. "How is Information Valued? Evidence from Framed Field Experiments." *The Economic Journal*, 126(595): 1884–1911.

Hoffrage, Ulrich, Stefan Krauss, Laura Martignon, and Gerd Gigerenzer. 2015. "Natural frequencies improve Bayesian reasoning in simple and complex inference tasks." *Frontiers in Psychology*, 6.

Holt, Charles A., and Angela M. Smith. 2009. "An update on Bayesian updating." *Journal of Economic Behavior & Organization*, 69(2): 125–134.

Holt, Charles A, and Susan K Laury. 2002. "Risk Aversion and Incentive Effects." *American Economic Review*, 92(5): 1644–1655.

Howard, Kirsten, and Glenn Salkeld. 2009. "Does Attribute Framing in Discrete Choice Experiments Influence Willingness to Pay? Results from a Discrete Choice Experiment in Screening for Colorectal Cancer." *Value in Health*, 12(2): 354–363.

Kahneman, Daniel, and Amos Tversky. 1972. "Subjective probability: A judgment of representativeness." *Cognitive Psychology*, 3(3): 430–454.

Karni, Edi. 2009. "A Mechanism for Eliciting Probabilities." *Econometrica*, 77(2): 603–606.

Kousky, Carolyn. 2011. "Understanding the Demand for Flood Insurance." *Natural Hazards Review*, 12(2): 96–110.

Laury, Susan K., Melayne Morgan McInnes, and J. Todd Swarthout. 2009. "Insurance decisions for low-probability losses." *Journal of Risk and Uncertainty*, 39(1): 17–44.

Masatlioglu, Yusufcan, Yeşim Orhun, and Collin Raymond. 2023. "Intrinsic Information Preferences and Skewness." *American Economic Review*, 113(10): 2615–2644.

McDowell, Michelle, and Perke Jacobs. 2017. "Meta-analysis of the effect of natural frequencies on Bayesian reasoning." *Psychological Bulletin*, 143(12): 1273–1312.

Montanari, Giovanni, and Salvatore Nunnari. 2023. "Audi Alteram Partem: An Experiment on Selective Exposure to Information." CESifo Working Paper Series 10699.

Neumann, Peter J., Joshua T. Cohen, James K. Hammitt, Thomas W. Concannon, Hannah R. Auerbach, Chihui Fang, and David M. Kent. 2012. "Willingness-to-pay for predictive tests with no immediate treatment implications: a survey of US residents." *Health Economics*, 21(3): 238–251.

Oliveros, Santiago, Ro'i Zultan, and Aniol Llorente-Saguer. 2025. "Beyond Value: On the Role of Symmetry in Demand for Information." *Management Science.*

Phillips, Lawrence D., and Ward Edwards. 1966. "Conservatism in a Simple Probability Inference Task." *Journal of Experimental Psychology*, 72(3): 346.

Quiggin, John. 1982. "A theory of anticipated utility." *Journal of Economic Behavior & Organization*, 3(4): 323–343.

Rabin, Roni Caryn. 2024. "In Reversal, Expert Panel Recommends Breast Cancer Screening at 40." *New York Times*, 16.

Schwartz, Lisa M., Steven Woloshin, Floyd J. Fowler, and H. Gilbert Welch. 2004. "Enthusiasm for cancer screening in the United States." *JAMA*, 291(1): 71–78.

Toplak, Maggie E., Richard F. West, and Keith E. Stanovich. 2014. "Assessing miserly information processing: An expansion of the Cognitive Reflection Test." *Thinking & Reasoning*, 20(2): 147–168.

Tversky, Amos, and Daniel Kahneman. 1971. "Belief in the law of small numbers." *Psychological Bulletin*, 76: 105–110.

Volkman-Wise, Jacqueline. 2015. "Representativeness and managing catastrophe risk." *Journal of Risk and Uncertainty*, 51(3): 267–290.

Xu, Yan. 2022. "Revealed Preferences Over Experts and Quacks and Failures of Contingent Reasoning."

Appendix A Tables and Figures

Table A1: Demographic Characteristics of Subjects

	Priors					
	All		$\{0.1, 0.3\}$		$\{0.2, 0.5\}$	
	N	%	N	%	N	%
	A. All waves					
Male	96	47	49	46	47	47
Age>23yrs old	16	8	8	7	8	8
Students	174	84	90	84	84	85
Had statistics classes	128	62	71	66	57	58
	B. First Wave					
Male	43	21	22	21	21	21
Age>23yrs old	14	7	6	6	8	8
Students	88	43	46	43	42	42
Had statistics classes	63	31	37	35	26	26
	C. Second Wave					
Male	53	26	27	25	26	26
Age>23yrs old	2	1	2	2	0	0
Students	86	42	44	41	42	42
Had statistics classes	65	32	34	32	31	31

Table A2: Average Protection by Signal Type and Prior

Row	**Prior**	**Signal Characteristics**		**Signal**	**Posterior**	**Share Protect**	**Share Optimal**	p
		False Positive	**False Negative**					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)	(8)
(1)	Low	No	No	White	0.000	0.034	0.000	0.111
(2)	Low	No	Yes	White	0.055	0.184	0.000	0.001
(3)	Low	Yes	No	White	0.000	0.290	0.000	0.003
(4)	Low	Yes	Yes	White	0.051	0.334	0.000	0.002
(5)	Low	No	No	Black	1.000	0.806	1.000	0.074
(6)	Low	No	Yes	Black	1.000	0.852	1.000	0.003
(7)	Low	Yes	No	Black	0.357	0.778	0.750	0.863
(8)	Low	Yes	Yes	Black	0.387	0.873	0.833	0.812
(9)	High	No	No	White	0.000	0.064	0.000	0.070
(10)	High	No	Yes	White	0.183	0.294	0.125	0.312
(11)	High	Yes	No	White	0.000	0.269	0.000	0.000
(12)	High	Yes	Yes	White	0.169	0.492	0.167	0.079
(13)	High	No	No	Black	1.000	0.844	1.000	0.022
(14)	High	No	Yes	Black	1.000	0.865	1.000	0.001
(15)	High	Yes	No	Black	0.667	0.840	1.000	0.007
(16)	High	Yes	Yes	Black	0.689	0.894	1.000	0.007

Notes: Column (8) reports the p-value for the test of equality between the theoretical prediction (Share Optimal) and the observed share of protection (Share Protect). Priors are grouped as low ($p \in \{0.1, 0.2\}$) v. high ($p \in \{0.3, 0.5\}$).

Table A3: Error Decomposition

	(1) OLS	(2) FE	(3) OLS	(4) FE	(5) OLS	(6) FE
Prior	0.265***	0.170**	0.270***	0.154*	0.281***	0.051
	(6.5)	(2.9)	(5.3)	(2.0)	(3.7)	(0.6)
Signal	0.449***	0.450***	0.307***	0.308***	0.456***	0.456***
	(9.5)	(9.0)	(4.2)	(4.0)	(4.6)	(4.4)
Good quiz × Prior			-0.011	0.033		
			(-0.1)	(0.4)		
Good quiz × Signal			0.289***	0.288**		
			(3.0)	(2.8)		
Stat. class × Prior					-0.023	0.180*
					(-0.3)	(2.1)
Stat. class × Signal					-0.011	-0.010
					(-0.1)	(-0.1)
Observations	473	473	473	473	473	473
Adjusted R^2	0.33	0.30	0.35	0.33	0.33	0.30

Notes: Decomposition works only for imperfect signals, hence the table excludes the responses to certain signals. Standard errors in parentheses (clustered at the subject and treatment level). */**/*** indicates statistical significance at 10/5/1 percent.

Table A4: Average Updating Error by Signal Type and Prior

			Signal Characteristics				
Row	**Prior**	**Signal**	**False Positive**	**False Negative**	**Posterior**	**Updating Error**	p
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1)	Low	White	No	No	0.000	0.038	0.262
(2)	Low	White	No	Yes	0.055	0.084	0.000
(3)	Low	White	Yes	No	0.000	0.199	0.001
(4)	Low	White	Yes	Yes	0.051	0.230	0.000
(5)	Low	Black	No	No	1.000	-0.148	0.083
(6)	Low	Black	No	Yes	1.000	-0.417	0.000
(7)	Low	Black	Yes	No	0.357	0.316	0.001
(8)	Low	Black	Yes	Yes	0.387	0.169	0.017
(9)	High	White	No	No	0.000	0.058	0.177
(10)	High	White	No	Yes	0.183	0.019	0.518
(11)	High	White	Yes	No	0.000	0.220	0.000
(12)	High	White	Yes	Yes	0.169	0.170	0.000
(13)	High	Black	No	No	1.000	-0.142	0.003
(14)	High	Black	No	Yes	1.000	-0.351	0.000
(15)	High	Black	Yes	No	0.667	0.007	0.860
(16)	High	Black	Yes	Yes	0.689	-0.126	0.024

Notes: The updating error is defined as Belief - Posterior, where Posterior is the Bayesian probability estimate for the treatment based on its information structure. The p-value in column 6 is for the test of the null hypothesis that the updating error in column 5 is equal to 0.

Table A5: Deviations from Signal Value (WTP - Value) and Demographic Determinants

	(1)	(2)	(3)	(4)	(5)	(6)
FP costs	.25	.246**	.423*	.435***	.348	.336***
	(0.2)	(0.1)	(0.2)	(0.1)	(0.2)	(0.1)
FN costs	.267*	.271***	.342**	.321***	.272**	.293***
	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
Male	-.398**	-.26				
	(0.2)	(0.2)				
Male × FP costs	.196	.192				
	(0.1)	(0.1)				
Male × FN costs	.0728	.089**				
	(0.1)	(0.0)				
Stat. class			.218	.306		
			(0.2)	(0.3)		
Stat. class × FP costs			-.122	-.152		
			(0.2)	(0.1)		
Stat. class × FN costs			-.0727	-.0204		
			(0.1)	(0.1)		
>23 yrs					-.377	-.875***
					(0.3)	(0.3)
>23 yrs × FP costs					-.117	.015
					(0.2)	(0.2)
>23 yrs × FN costs					.393**	.257*
					(0.2)	(0.2)
Constant	-.0972	.546***	-.418	.222	-.252	.491***
	(0.3)	(0.2)	(0.3)	(0.2)	(0.3)	(0.2)
Prior dummies	No	Yes	No	Yes	No	Yes
Observations	1230	1230	1230	1230	1230	1230
Adjusted R^2	0.05	0.19	0.04	0.20	0.04	0.19

Notes: Standard errors in parentheses (clustered at subject and treatment levels). */**/*** indicates statistical significance at 10/5/1 percent.

Table A6: WTP Deviations from Signal Value: Risk Aversion and Sensitivity to Error Costs

	(1)	(2)	(3)	(4) FE	(5) FE
p>0.2	-.513	-.516	-.363	-.516	-.363
	(0.5)	(0.5)	(0.5)	(0.5)	(0.5)
FN costs	-.181	-.249	-.134	-.249	-.134
	(0.2)	(0.3)	(0.2)	(0.3)	(0.2)
p>0.2 × FN costs	.615**	.743**	.63**	.743**	.63**
	(0.3)	(0.3)	(0.3)	(0.3)	(0.3)
FP costs	.462**	.564**	.659***	.564**	.659***
	(0.2)	(0.2)	(0.2)	(0.2)	(0.2)
p>0.2 × FP costs	-.501	-.504	-.591	-.504	-.591
	(0.4)	(0.4)	(0.4)	(0.4)	(0.4)
Risk-loving × p>0.2 × FN costs		-.0371	-.183	-.0371	-.183
		(0.1)	(0.2)	(0.1)	(0.2)
Risk-averse × p>0.2 × FN costs		-.236***	.241***	-.236***	.241***
		(0.1)	(0.0)	(0.1)	(0.0)
Risk-loving × p>0.2 × FP costs		-.186	.167***	-.186	.167***
		(0.2)	(0.0)	(0.2)	(0.0)
Risk-averse × p>0.2 × FP costs		-.0898	.16	-.0898	.16
		(0.2)	(.)	(0.2)	(.)
Full risk pref interactions	No	No	Yes	No	Yes
Observations	1230	1230	1230	1230	1230
Adjusted R^2	0.08	0.08	0.08	0.08	0.08

Notes: Risk preferences interactions include inconsistent preferences (not shown), interacted with priors, FP, and FN costs. Standard errors in parentheses (clustered at subject and treatment levels). */**/*** indicates statistical significance at 10/5/1 percent.

Table A7: Deviations from Signal Value and Signal Characteristics (Tobit)

	All			Prior	
				$\{.1, .2\}$	$\{.3, .5\}$
	(1)	(2)	(3)	(4)	(5)
FP costs	0.130	0.315	0.547	0.446	0.204
	(0.122)	(0.187)*	(0.221)**	(0.236)*	(0.406)
FN costs	0.235	0.246	0.295	-0.137	0.309
	(0.065)***	(0.109)**	(0.112)***	(0.282)	(0.106)***
Risk-averse × FP costs		-0.379	-0.462	-0.136	-0.760
		(0.326)	(0.353)	(0.351)	(0.560)
Risk-averse × FN costs		-0.425	-0.416	-0.304	-0.375
		(0.173)**	(0.177)**	(0.373)	(0.160)**
Risk-loving × FP costs		0.087	0.022	0.068	0.330
		(0.262)	(0.297)	(0.362)	(0.561)
Risk-loving × FN costs		0.121	0.154	0.475	0.155
		(0.142)	(0.148)	(0.421)	(0.151)
Constant	-0.291	-10.296	-14.116	-8.177	-11.738
	(0.158)*	(0.417)***	(1.625)***	(0.481)***	(1.872)***
/					
var(e.wtp_diff)	4.473	2.172	2.118	1.009	1.209
	(0.340)***	(0.180)***	(0.173)***	(0.120)***	(0.134)***
*.subject_id	No	Yes	Yes	Yes	Yes
Prob(FP=FN)	0.308	0.648	0.119	0.001	0.750
Obs	1230	1230	1230	615	615
Risk-Averse Subjects:					
False Positive		-0.064	0.085	0.310	-0.556
se		(0.268)	(0.276)	(0.262)	(0.385)
p-value		[0.811]	[0.759]	[0.237]	[0.149]
False Negative		-0.180	-0.121	-0.441	-0.066
se		(0.135)	(0.137)	(0.247)	(0.120)
p-value		[0.184]	[0.377]	[0.074]	[0.582]
Risk-Loving Subjects:					
False Positive		0.402	0.569	0.514	0.534
se		(0.186)	(0.199)	(0.276)	(0.388)
p-value		[0.031]	[0.004]	[0.063]	[0.170]
False Negative		0.367	0.449	0.338	0.463
se		(0.092)	(0.096)	(0.313)	(0.108)
p-value		[0.000]	[0.000]	[0.281]	[0.000]
Subject FE	Yes	Yes	Yes	Yes	Yes
Inconsistent Risk Pref. Interactions	No	Yes	Yes	Yes	Yes
Inaccurate Belief Interactions	No	No	Yes	Yes	Yes
Prior Probability FE	No	No	No	Yes	Yes

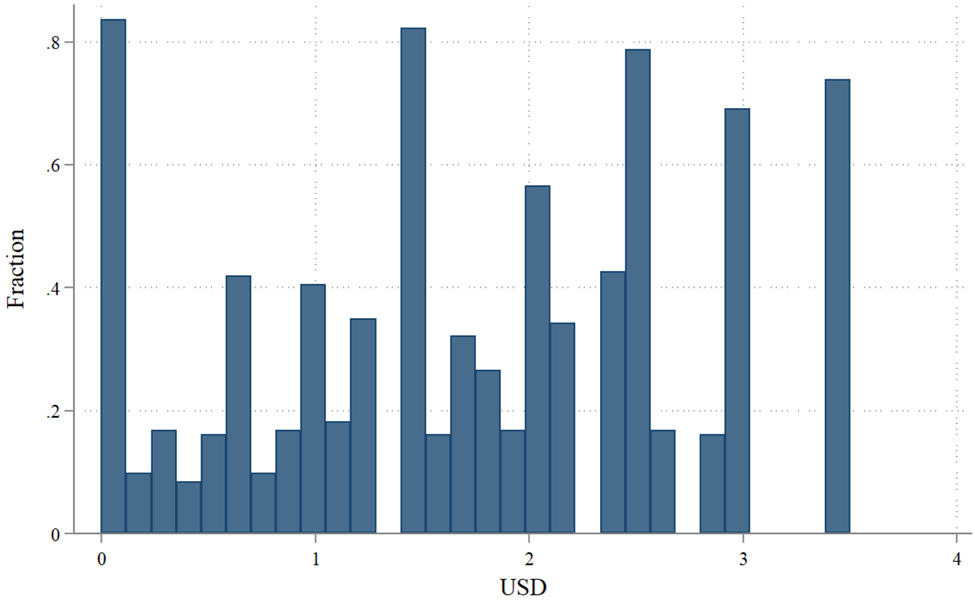
Notes: Standard errors in parentheses (clustered at the subject and treatment levels). */**/*** indicates statistical significance at 10/5/1 percent. The bottom panels include tests of whether the total coefficient values (baseline + interaction) are different from zero.

Table A8: WTP and Signal Characteristics, IP bias

	All			Prior	
				{.1, .2}	{.3, .5}
	(1)	(2)	(3)	(4)	(5)
FP costs	0.461	0.503	0.654	0.588	0.298
	(0.234)*	(0.270)*	(0.286)**	(0.177)***	(0.243)
FN costs	0.323	0.361	0.389	0.062	0.348
	(0.109)***	(0.135)**	(0.137)***	(0.185)	(0.073)***
Extra empirical IP protection	-0.127	-0.121	-0.120	-0.133	-0.198
	(0.057)**	(0.056)**	(0.055)**	(0.044)***	(0.061)***
Extra empirical IP loss	-0.018	-0.023	-0.018	-0.053	-0.075
	(0.033)	(0.032)	(0.031)	(0.065)	(0.040)*
Risk-averse × FP costs		-0.275	-0.347	-0.140	-0.640
		(0.182)	(0.256)	(0.245)	(0.380)
Risk-averse × FN costs		-0.297	-0.289	-0.302	-0.304
		(0.124)**	(0.123)**	(0.212)	(0.120)**
Risk-loving × FP costs		0.066	0.009	0.038	0.276
		(0.128)	(0.186)	(0.298)	(0.339)
Risk-loving × FN costs		0.045	0.063	0.317	0.087
		(0.076)	(0.099)	(0.324)	(0.109)
Constant	-0.332	-0.328	-0.440	-0.160	-0.249
	(0.242)	(0.242)	(0.228)*	(0.187)	(0.163)
R^2	0.500	0.507	0.516	0.738	0.753
Prob>F	0.0045	0.0082	0.0045	0.0001	0.0000
Obs	1230	1230	1230	615	615
FP=FN	0.482	0.548	0.266	0.000	0.792
Risk-Averse Subjects:					
False Positive		(0.228)	(0.307)	(0.447)	(-0.343)
se		[0.278]	[0.272]	[0.221]	[0.335]
p-value		0.417	0.265	0.055	0.318
False Negative		(0.064)	(0.100)	(-0.240)	(0.044)
se		[0.134]	[0.129]	[0.220]	[0.125]
p-value		0.635	0.440	0.286	0.729
Risk-Loving Subjects:					
False Positive		(0.569)	(0.663)	(0.625)	(0.573)
se		[0.227]	[0.204]	[0.231]	[0.232]
p-value		0.016	0.002	0.013	0.021
False Negative		(0.406)	(0.452)	(0.379)	(0.435)
se		[0.110]	[0.118]	[0.290]	[0.080]
p-value		0.001	0.000	0.204	0.000
Subject FE	Yes	Yes	Yes	Yes	Yes
Inconsistent Risk Pref. Interactions	No	Yes	Yes	Yes	Yes
Inaccurate Belief Interactions	No	No	Yes	Yes	Yes
Prior Probability FE	No	No	No	Yes	Yes

Notes: Standard errors in parentheses (clustered at the subject and treatment levels). */**/*** indicates statistical significance at 10/5/1 percent. The bottom panels include tests of whether the total coefficient values (baseline + interaction) are different from zero.

Figure A1: Distribution of theoretical values for experimental treatments



Value = expected change in costs from BP to IP

Appendix B Proofs

B.1 Proposition 1

Proof. If protection costs are low enough $c < \pi L$ then a risk-neutral decision-maker should always protect without a signal:

$$U = \max[\pi(Y - L) + (1 - \pi)Y, Y - c] = Y - c$$

It means that a strictly risk-averse decision-maker with a utility function $u()$ should also protect:

$$\pi u(Y - L) + (1 - \pi)u(Y) < u(\pi(Y - L) + (1 - \pi)Y) \leq u(Y - c)$$

Then denote stochastic payoff with a signal as X so that expected utility with a signal is $Eu(X - b)$ where b is the willingness-to-pay solving:

$$Eu(X - b) = u(Y - c)$$

Let b_0 be the willingness-to-pay for a risk-neutral decision-maker. By (strict) Jensen's inequality:

$$Eu(X - b_0) < u(EX - b_0) = u(Y - c) = Eu(X - b)$$

Because $\dot{u()}$ is an increasing function we obtain $b < b_0$. □

B.2 Proposition 2

Proof. Use the mean value theorem to rewrite the sensitivity as:

$$\frac{db}{dP_{01}} = -\frac{\pi u'(\zeta)(L - c)}{E[MU]}, \zeta \in (Y - L - b, Y - c - b)$$

Now let X denote a (random) payoff of the agent with a signal. A risk-averse decision-maker puts a positive value on the signal only if its expected payoff is higher than the certain payoff with full protection: $EX > Y - c - b$. If an agent is imprudent ($u''' < 0$) then $u'(\cdot)$ is a strictly concave function and hence $E[MU] \equiv E[u'(X)] < u'(EX)$ by Jensen inequality. Next, u' being a strictly decreasing function due to strict risk aversion and $EX > Y - c - b$: $u'(\zeta) > u'(Y - c - b) > u'(EX)$. Hence $\frac{u'(\zeta)}{E[MU]} > 1$ and $\frac{db}{dP_{01}} < -\pi(L - c)$. □

However, risk aversion can both increase and decrease subject's sensitivity to false-positive rates depending on the utility function curvature and signal's characteristics. Intuitively, an expected marginal utility of a strongly risk-averse subject with an imperfect signal can be lower than the average slope of the utility function between $(Y - c - b)$ and $(Y - b)$ which reduces sensitivity to false-positive rates. It can also be higher if either the signal is good or the curvature is small.

B.3 Proposition 3

Proof. Sensitivities to FP and FN rates are given as:

$$\frac{db}{dP_{10}} = -\frac{(1-\pi)(u(Y-b)-u(Y-c-b))}{D(\pi, P_{01}, P_{10}, b)} \tag{10}$$

$$\frac{db}{dP_{01}} = -\frac{\pi(u(Y-c-b)-u(Y-L-b))}{D(\pi, P_{01}, P_{10}, b)} \tag{11}$$

Hence the ratio of sensitivities is:

$$\frac{db/dP_{01}}{db/dP_{10}} = \frac{\pi}{(1-\pi)} \frac{(u(Y-c-b)-u(Y-L-b))}{(u(Y-b)-u(Y-c-b))} \tag{12}$$

Based on the mean value theorem, there exist such $\zeta_1 \in (Y-L-b, Y-c-b)$, $\zeta_2 \in (Y-c-b, Y-b)$ that $u(Y-c-b)-u(Y-L-b) = u'(\zeta_1)(L-c)$ and $u(Y-b)-u(Y-c-b) = u'(\zeta_2)c$. Hence we can write:

$$\frac{db/dP_{01}}{db/dP_{10}} = \frac{\pi}{(1-\pi)} \frac{u'(\zeta_1)(L-c)}{u'(\zeta_2)c} = \frac{\pi(L-c)}{(1-\pi)L} \frac{u'(\zeta_1)}{u'(\zeta_2)} \tag{13}$$

Because $\zeta_1 < Y-c-b < \zeta_2$ and $u'' < 0$, $u'(\zeta_2) < u'(\zeta_1)$ and hence:

$$\frac{db/dP_{01}}{db/dP_{10}} = \frac{\pi}{(1-\pi)} \frac{u'(\zeta_1)(L-c)}{u'(\zeta_2)c} > \frac{\pi(L-c)}{(1-\pi)L} \tag{14}$$

Because the last expression on the right describes a ratio of sensitivities for a risk-neutral agent, we obtain that strictly risk averse subjects exhibit higher relative sensitivity to FN rates.

□

Appendix C Subjects' Explanations

The list of responses to the question *"Please explain the strategy you used for Task 2 (Informed Protection)? This is the task in which you see a hint and when decide to protect or not."*:

1. if the hint was favorable not protection and vice versa
2. I always bought protection unless I was certain that I didn't need it (i.e. both gremlins were honest or it wasn't possible to get the black/white gremlin)
3. I trusted honest golems fully, and did not put much stock in the swamp golems.
4. my strategy was to just look at what the odds were
5. I looked at the percentages of white and black balls and made my guess off of that. Also, there was no big harm in buying protection, and there was a lot of harm if you did not buy protection and got a black ball.
6. I trusted my instinct.
7. If the entire panel of gremlins was honest and they told me that the selection was white, I did not buy protection, since I could be certain that I would not lose money. In any other scenario, I bought protection. In my case, better to guarantee a $25 return every time than risk $20 for a $5 reward.
8. if it is an honest one, i don't need to buy informed protection cuz i can't trust its hint.
9. I think the gremlins were confusing, but if you see how many gremlins were. Then from that how many of each type where and what they say, after that you based that to the actual percentage of balls you get close to the answer.
10. I am a little bit more risky so I chose to not get protection if any of the monsters said it was white because I felt the probability of one of the honest ones getting picked was higher and if they said it was black I bough protection.
11. i used probablity and if the odds were more in favor i would mae a decision based on that and the ball probabilty color
12. If the hint was from one of the honest gremlins then I didn't choose to protect because they could only tell the truth. If there were any just black or just white gremlins then I decided to protect because the information they give isn't helpful
13. See the quantity of hints and the percentage of drawing the colors of the balls.
14. I would calculate the probability that the gremlins were right. So in task two, I already did task 3. Like if there were two black/white gremlins, I would add the probability that they were right to the certainly that the honest gremlin was right.
15. I would see what the probability that they are telling the truth is and then see if they were saying black. if no one was the black swamp monster then I knew it was black and therefore it would be 100%
16. I looked at the box of balls and the box of gremlins. If the gremlins were honest or white, I would not use protection for a white ball. If the ball was black I would sometimes take my chances depending on the amount of white and black balls. If they were honest or black, I would use protection for a black ball. If the ball was white, I would not use protection since there were mostly honest gremlins.
17. I weighed the cost of loosing money and percentage difference with that chances of getting a white ball.
18. I weighed my odds. I knew they were in my favor.
19. When I paid attention to the composition of the box and saw the gremlins, that helped to inform my decision on whether to buy protection. For example, if I saw the box had equal numbers of both black and white ball and two honest gremlins were there, I did not buy protection. When I saw a box with

a larger amount of black than white balls and had a white-swamp gremlin with four honest gremlins, I opted to buy protection.

20. I would the probability of one of the balls being picked. If the chances were not likely than I would not protect it.
21. I looked at what percentage of gremlins were honest and used that info in my decisions.
22. Instinct and possibility of either white or black being picked
23. I took protection when there was a higher chance of drawing out black balls.
24. If all glimpses are honest, then choose not to protect on each color. If most are honest and one is black, then choose not to protect white color. If the one is white, then choose not to protect black because we know white one always say white, so black color should be the truth.
25. I based my decision on the probability of the honest gremlin being chosen.
26. I would base my answers off of how many honest goblins there were.
27. I chose the best odds
28. If it was more than approximately a 70% chance of drawing a black ball, I decided to protect. The cost to protect outweighed the potential loss of not protecting.
29. If the gremlin was honest then I did not buy protection because they were accurate in telling me the color of the ball.
30. If there were only honest gremlins then I never protected but even if there was one white-swamp gremlin or one black-swamp gremlin then I payed for protection.
31. If the gremlins were honest, I didn't buy protection. If there were swap gremlins, I calculated the chance of getting a hint from a swap gremlin and considered that along with the chance of getting a black ball. If the total chance of getting a black ball was more than 15% I get protection.
32. I determined what the probability was that the gremlin would tell the truth. The more honest gremlins in the lineup, the less likely I was to buy protection. However, I'm risk-averse, so I was more likely to buy protection than not because the risk was too high and the cost of protection was low.
33. I just used probability in my head
34. **I took into consideration how many honest there were and looked at the chances of picking a ball**
35. I was able to calculate the odds from the hints. It was not a measurement requiring me to calculate the chance of balls, but of variance between the hints. This made it easier to calculate the probability of what the chances the gremlins would give regardless of the actual odds (14/6 white-black balls)
36. I just took into note the goblins that were listed, and then the probability of which the information could be truthful or not.
37. I just relied on the number of honest gremlins to inform my decisions
38. If there were a white swamped gremlin, I would buy the protection if it said white ball. If it said black on a white swamped I would always not buy the protection. This is vice versa if there was a black swamped gremlin.
39. I used the strategy of using the "honest gremlin" to my advantage to know when I could get away with not paying for protection
40. I relied on understanding which type of gremlin was presented and then based my decision on their bias/lack of bias. Honest gremlin were a simple binary decision (white -¿ no protection, black -¿ protection). The white gremlin would default to no protection unless the probability of black was greater than 25%. The black gremlin defaulted to protection.

41. I considered the probability of the computer selecting a white ball and a honest gremlin. If that probability was high (>70%), then I decided not to buy protection. When there were only honest and black gremlins and the hint was that the ball was white, then it was easier since that hint could only come from an honest gremlin.
42. I took into consideration which of the gremlins I got. If it were two honest ones, I would not buy protection if they said white because they were right. If they were two honest ones and a black one, and they said it was white, I would do the same thing because the black one would never say the ball is white. If any of the gremlins said the ball was black, I would buy protection because there would always be a chance that the ball was black.
43. It was really just similar to math and common sense.
44. I went with the odds. I didn't buy protection if the probability of picking a ball was really high in a situation
45. I would look at how many honest gremlins there were to see if i could trust it or not. ex: if there were only honest and white gremlins, and they said the ball was white, i would trust that.
46. If it was all honest then I 100% percent trusted it and went for no protection but if there was even a chance of dishonest gremlin then I went with protection
47. My strategy depended on the gremlins. I was willing to pay a higher price for more honest gremlins, while I was not willing to pay as much when there were not as many honest gremlins.
48. The higher the % of black balls the more likely I was to buy protection.
49. I based it off of the amount of different colored balls mainly. Because, if there was only 2 black balls and one black gremlin, then I would most likely have a white ball chose if the other two were honest.
50. I looked at the percentage and the chance of drawing which ball, and I compared it to the grimlin options/hints and made my decision based off of the numbers I was provided.
51. I am broke and I was willing to take risks to make more money.
52. I just hoped for the best and picked one
53. **If it was all honest gremlins I did not buy protection. Even if there was one un honest gremlin I was skeptical to buy protection. If there was more than one un honest gremlin I definitely bought protection.**
54. If there were more black balls I would decide to protect it because there was a higher chance it needed to be and if there were more white balls I didn't protect it because I assumed the chance of a black ball being chosen was lower.
55. My strategy in task two was primarily based on the gremlins. For example, if they were all honest then I would not buy protection if they said white but would if they said black. Furthermore, if four were honest and one was a white-choosing gremlin, then if the gremlins said the ball was black I would buy protection; Considering that the white gremlin could only say the ball would be white, then it is known that an honest gremlin said that the ball would be black and vise versa. I did not really consider the probability of the balls being chosen and rather focused on the likely hood that the hint given by the gremlins is correct.
56. I would first take into account how many white and how many black balls were in a box, and the chance of drawing each. With the gremlins then telling a hint I would not buy protection if the gremlin said white and the percent of drawing white was more than 75%. I used this kind of method for the whole task.
57. my strategy for this task was to only buy protection if there was a white or black gremlin and not if there was a truth gremlin

58. the percentages of black and white balls and which gremlins I would get to give a hint.
59. I took my chances that the gremlins telling the truth would be selected
60. If the goblins were all honest I would buy protection if they say the black was the ball chosen and not if the ball was white. If 1 of the goblins was saying the ball was black or white exclusively I would buy protection if they say it was black and not if the ball was white. If 2 of the goblins was saying the ball was black or white exclusively I would buy protection no matter what they said
61. How likely it was that it would be white
62. I mainly looked at the probability percentage of the computer choosing a white ball. If it was greater than or equal to 70%, then I would not choose protection.
63. It was pretty simple, actually. I basically based my decision off of the amount of honest gremlins there were. If there were 4/5 honest, then there was an 80% chance the hint was correct. On a situation with 50% white and 50% black, this strategy proved to be helpful.
64. I based my decision off of the makeup of the gremlins if they were all honest and said the ball was white I would not buy protection and if they said it was black I would buy protection. If there was a 1/3 chance of an honest gremlin being picked for the hint I would just buy protection because I did not like the odds of the hint being true. If the chance of an honest gremlin being picked was 2/3 I would look at the probability of a white/black ball being chosen and then make my decision to protect or not off of that.
65. I based my chances solely on the honest Gremlins.
66. I mostly would buy protection if there was an over 50 percent chance to get a black ball.
67. I thought of how many un honest gremlins there were and tried to guess the percent of accuracy I would be given based on the colors.
68. If it was mostly Honest Grimlins I took the hint
69. I looked at the different types of gremlins in each group to make my decision. If it was all of the honest gremlins, I would go from there, but even if it were 2 honest and 1 black or white swamp gremlin that would inform my decision better than if it was an equal mix of all three types
70. I looked at the % of white vs black balls then looked at how many honest grimlins there were. If there were a majority of white balls and honest grimlins I would do no protection for a white ball but buy protection for black.
71. Always went with the honest ones. When there was one white or one black, I would know it was an honest one when they said the opposite of the color. For example, two honest and one white, when it said the ball was black, I knew it would be black because the white can't say that.
72. I compared the number of balls to the gremlins hints and if the chances were higher than 50% ish I wouldn't get protection
73. I would always take the hints from honest and be skeptical of non-honest
74. I looked at the gremlins and then looked at their hint. depending on what gremlins I had, i looked at the combination of balls to see if I should risk it or not. If I had a lot of white balls and quite a few honest gremlins, I did not buy the protection plan
75. I decided weather or not to buy protection based on the gremlins
76. I am basically gambling so I would not pay attention to the Gremlins and look at the percentages
77. Sorry. My strategy was same through-out, except the very first question of task1. Risk-averse, not worried about losing $5. Also, not trusting even honest gremlins or perhaps myself if I had mis-read.
78. Just went with my gut guess. I didn't really use a strategy for any of them tbh

79. there was no need to protect if the hint were made by all honest gremlins. also no need to protect if i had a combination of honest and black gremlin and the prediciton said it's white cos a black gremlin will never give a white answer

80. I had two honest gremlins, so the hint was 100% accurate.

81. I measured my decision based off of the type of gremlin giving the hint. If I felt that the gremlin or group was highly trustworthy, I would follow the advice.

82. If it was highly likely that the gremlin was going to be correct, I chose no protection. I aimed for the highest payout each round based on the amount of black to white balls there were.

83. If there were all honest ones I would not buy protection if it was white. I bought protection on all the others so that I would not lose more money.

84. I just created a pattern in my head and looked at the percentage of the likeliness of a black ball being drawn or not.

85. I based it off the amount of honest gremlins presented

86. If the color said was the opposite of black or white eyed gremlin then I knew it was true because the rest were honest gremlins

87. Based off how many white ball there was

88. I decided what to do based on both probability of selecting a ball of off composition of colors, and the used the gremlins to add an extra level of certainty.

89. Simply used the projection of likelihood for how much risk I was willing to take.

90. If i was feeling lucky or not

91. Based off of the number of gremlins would help me determine to use protection or not

92. I used the gremlins as my strategy, i took more risks if it was the honest gremlins

93. I payed attention to the honest gremlins and I used my answers based off how many there were.

94. I would observe which of the gremlins informing me were honest and make my decision there.

95. I just tried not to risk it. I prefer getting a little bit less than the total amount than actually reducing $20

96. I figured out what the gremlins were saying and used that to calculate the probability

97. I just guessed.

98. I thought about which option would make me the most amount of money based on protection or not.

99. I just decided which ones wanted protection and not.

100. Basically if the white balls had a higher rate than the black balls I wouldn't buy protection

101. I looked mostly to whether or not I had an honest gremlin in my group. If I had gremlins which could be dishonest, I then evaluated my chances based on the percentage of black vs. white balls in the box.

102. If I knew the ball would be white then I would not protect, everything else I protected

103. I was a little more clueless about it, I tried to make sense of the question first and then see the number of balls that were black and if they were less, then I would not buy protection.

104. If the goblins were guaranteed to be honest, I followed their hint. If there was a white goblin at all, I ignored the hint completely. If there was only a black goblin, I wouldn't buy protection if the hint was white since that couldn't be correct.

105. I took a chance each time

106. If there was all honest gremlins, I would not protect unless they said the ball is black. If there were white or black gremlins in the mix, I generally chose to protect unless the chances the ball was white were greater than 90%.

107. If there was ever a question I always bought protection because I would rather have 25 dollars than 10.
108. It really depended on how many balls were in the box
109. anything under 80%. I know what my time is worth and don't want to waste it for $5 unless my return was greater then the odds I would be given.
110. I would try to think back to what the amount were if i choose protected v.s unprotected.
111. I would calculate the possibility in my head each task.
112. I decided whether to trust my informed protected based on how many gremlins were present divided by how many gremlins were honest/would say if the ball was white or black. I was more likely to trust the honest gremlins or buy protection for the dishonest gremlins.
113. If I new the odds of the hint gremlin is honest then I would protect baes on the hint they made
114. if it was mainly honest ones and the probability was an 80/20 split i just took my chances
115. Seeing the probabilities and protentional outcomes were what played a role in decision making.
116. if it is white never by protection, if it is black always buy protection. There is a higher chance that 1 a ball is white, and 2 that the gremlin is honest so it makes sense
117. examine the beasts, then make a decision
118. if all the gremlins were honest and said white i would not protect. If all the gremlins were honest and said black i would protect. If the gremlins were honest, honest, black and the hint was white, i knew it was one of the honest gremlins that was giving me the correct answer of white, so i didnt protect.
119. Odds based on the original sample times the odds of the gremlins giving bad info
120. I just thought there would no way the black ball would be chosen if there was only 1 or 2 of them. I was just trusting my guts.
121. mainly it was for the 70/30 i reled on the hits if it was black to think what are the changes it is realy black and if do then buy the protection. often it was more favored of the white being true to dont need the protection
122. If all the Grimmlies where honest I did not use protection, if they weren't I used protection.
123. I first looked at the chance and if it was 90% white 10% black, I probably didn't buy protection or 50 cents at most. Then, I looked at the gremlin and if they were only honest ones I was willing to pay up to 2.*Iftherewerewhite/blackgremlinsandtherewasahighchanceofblackballsIfiguredIwouldpayupto*1 just incase.
124. GAMBLING
125. So the strategy that I used are to look at the box and see how many honest gremlin there were and decide if I wanted to protect or not.
126. I didn't realize it told me my odds of the balls at the top of the screen so i went based off of the gremlins i had again. I got a lot of honest gremlins so anytime i had a white or black one because i never had both in on grouping it was normally 2 honest or like 4 honest and 1 black or white gremlin so then i would go off of the answer that was opposite of what the white or black gremlin was cause how could it tell me if the answer was a color it couldn't see.
127. I just guessed based on probability.
128. every time it was all honest gremlins I never used protection, but after that I based my answers off the percentage of the white ball being drawn.
129. not needed unless over 50%
130. always buy protection for guranteed money

131. if they were honest gremlins i said what they said, if there was a smaller amount of gremlins i would pay for protection but iuf there was a larger amount of gremlins i would put both as not paying for protectin

132. I would measure the chances that the hint would be useful in addition to the actual percentage that the ball would be a certain color

133. If I had majority of honest gremlins, I would most likely follow the hint. The more swamp gremlins I had the less I would follow the hint.

134. High risk, high reward

135. If the odds were favorable to me without the hint i wouldnt take it. If i got bad gremlin luck i would not take hint. If the honest gremlin said white and the white gremlin was there I would take the hint

136. If they were honest gremblins then I trusted that whatever they said was correct, but if there was one of each I added protection because it then wasn't as sure, especially if it was 90% to 10%.

137. If the gremlins were honest, then I would take the hint. If they were not, then I did not want to take that chance.

138. If they are all truth then you don't have to worry. If they are all truth and then 1 black but it says white, you dont have to worry. After that i did mental math to see what was the likelihood one of the liers get drawn and then what percentage of the remaining data would be effected.

139. Well for a white ball there's really no penalty to protect it so you would be wasting money to protect it. For a black ball it depends, because if there's a black eyed gremlin paired with a honest gremlin then take you chances with the protection, but if there's a white eyed one paired with a honest gremlin then its false because that would never happen.

140. I just used math and assigned a value to how much the hints provided help and went from there

141. Counting the ratio of white to black dots

142. I always took the hint from honest gremlins or white only gremlins

143. I would always choose protection because the risk was too high if i did not choose it. But, if i had to honest gremlins i knew what they said was correct so i would not protect.

144. I pretty much just kept a basis of what I was willing to lose. If you purchase the informed protection but at too high of a cost, you lose out on key dollars over time.

145. If it was 5 honest gremlins + 2 white gremlins and they were saying it was black, I would protect because the only option would be for the honest gremlins to be saying the truth about the ball. If it was 5 honest gremlins and 2 black gremlins and they claimed the ball was black, I would have less of a certainty getting protection, and would see the actual proportion of the balls.

146. I decided to not take the risk for most parts because 25 is a lot better than 10, even if it means you're losing $5

147. I mostly used my chances of the balls (10% black 90% white) to make my decision and would only accept the hint if it dipped below 90% and have that influence me.

148. I looked at the balls in the drawing pit and used those odds to determine my decisions.

149. All or nothing

150. I used deductive reasoning

151. Well I knew if it was all honest gremlins they were always correct, if it was all honest and like 1 or 2 black gremlins and i knew if the gremlin said white it had to be white and vice versa because that means the honest gremlin said the color as the other gremlins cannot say each others opposite colors.

152. I decided to look at the number of white vs black balls to make my main decision.

153. Kinda looking at the eye balls because if it says white then the black gremling cant say that

154. if they were both honest, i did not protect if they said white because it was white and i did protect if they said black so that i would only lose 5$opposedto$20. with the white and black eyed, i then weighed my likelihood and decided based on the percentage of pulling and the percentage of them telling the truth
155. If 2 of the three swamp gremlins were honest, and the other was white, i knew that if a random 1 said white, my chances were 66 percent, but if they said black, my chances were 100 percent
156. I took into account the probability that I got an honest hint and then also the probability of the ball being black. If there was a 90% chance of the ball being white, it was a waste of money to buy a hint regardless of the probability of it being honest or not.
157. The hints don't matter, the question was asking about % likelihood and those numbers were provided.
158. I saw the gremlins and figured out what they would tell me, the truth, only white, or only black. I decided mostly based on the percentage of the ball actually being black. Because if the ball is most likely going to be white I won't need protection.
159. I mostly paid attention to how many black and white swamp monster things there were compared to the honest ones. If there was only one possible liar in the line up then the chances of not choosing correctly felt lower.
160. Played it very risky and it paid off
161. I almost always decided to buy protection because it only cost $5.
162. I used statistics to calculate whether it was worth it or not.
163. I used the given probability to decide if a paycut was statistically worth it
164. based on the chances of drawing a black ball and the hint recieved i made my decision
165. Using both the chances of the ball being white and what monsters I had
166. If there were enough truths, i would bet more on it being truthful
167. I would assume that based off of probability i would get a truthfully dude, the odds were always 5:1 of them not lying so you play the odds
168. more white balls then black
169. This one was interesting
170. i usually chose to protect if there was 70% chance or less
171. I looked at how many honest ones and non honest ones were accounted for.
172. I used percentage of balls in given box.
173. If there were more truth grimlins i tended to trust my decision more
174. If both gremlins were truth gremlins I listened, but if one was lying I bought protection for both.
175. If there were two honest gremlins then I would protect when given a hint that it was a black ball. With the others there was a 5/7 chance that I would get an honest gremlin, I should have protected when it said "black ball" in the case with 2 white gremlins. If there was a white and black gremlin I tried to base it off of the probability of actually drawing a white ball.
176. I looked to see which gremlin was diplayed to determine my answers. If there were more honest gremlins then I would protect or not confidently. But if there were more dishonest gremlines then I would protect or not with doubt.
177. I looked at the probability of the balls colors as well as the types of things giving the hints. I would protect if there was a high chance of the ball being a certain color and if the hint could be false.
178. I was just looking at what the odds of the black and white balls where and then which kind of gremlins I had then made my decision off that.

179. i decided to use a back and forth decision by buying and not buying protection

180. I kind of just used my best judgement and guessed a little bit.

181. If the Gremlins were both good hint ones I would not protect. If there was a good hint one and a white one I would protect on the white ball and the black, since the white gremlin can't say black, it would have to be the hint one. I did the same thing but opposite for the black and hint gremlins.

182. I didn't have much of a strategy honestly I just looked at the goblins and guessed.

183. My strategy for informed protection was fairly simple. I would take the hint because in my opinion 25 dollars is better than 10, or the risk of winning 30. I am a bit of a gambling man, but when the white and black eyed gremlins could have been lying about something they said I figured better to take the confirmed money than risk it for more.

184. I did not get protection if it was likely that I would get a white ball.

185. Hints that are given from two truthful monsters are valuable so no protection is needed, because I know what they are telling be is the truth (if the ball is white). In all other cases I bought protection no matter what because I didn't want to risk losing a free $25 by simply not buying protection.

186. If the gremlins were honest, I never protected, but if there were black or white-eyed gremlins at all, I would protect.

187. I played it safe and bought protection unless I liked the odds of getting a white dot or if I had two truth gremlins.

188. i just went off of the gremlins for each round

189. Just used what are the odds and logical thinking.

190. I looked at the probabilities, then made my decision based on the gremlins presented.

191. If I saw the trust swamp monster I didn't pay for protection.

D Experiment Instructions (for Online Appendix)

Introduction

Welcome! This is a study of individual decision-making and behavior. The money you earn will be paid to you in cash at the end of this experiment.

This experiment has 4 parts. For each part, we will give you instructions just before it begins. Your choices in one part of the experiment will not affect what happens in any other part. Each part proceeds in rounds. There will be 10 rounds in total. We expect that most participants would be able to complete the experiment in about 30 minutes. The experiment will end with a short questionnaire.

At the end of the experiment, we will draw one of the rounds at random as the **Payment Round.** Each round of the experiment is equally likely to be drawn. Only the decision that you made in that Payment Round will determine your final payoff. Hence you should make every decision as if it is the one that counts, because it might be!

At the start of the experiment, you will be given $25, so with the show up fee included you will have $30 in total. The choices you make within the experiment will determine how much of this amount you may lose. It is impossible to lose more than $25, so your earnings in the worst-case scenario will be exactly your show-up fee of $5.

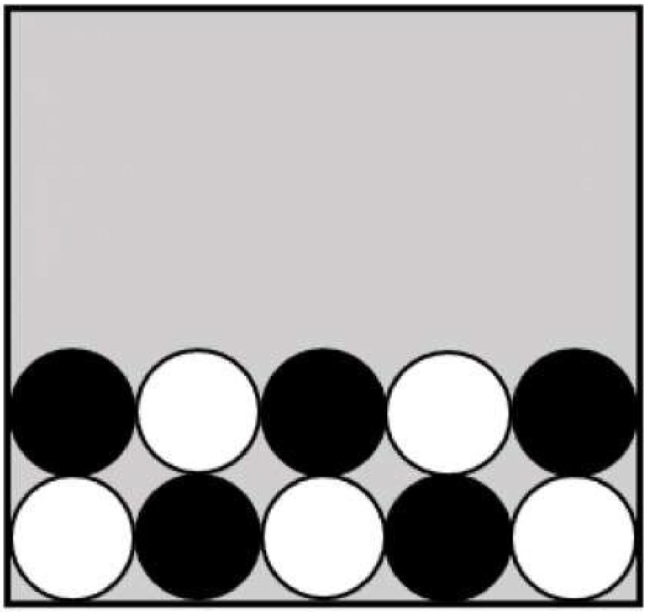
You are not allowed to talk during the study. If you have any questions, please raise your hand and we will come and answer your questions privately. Please do not use cell phones or other electronic devices until after the study is over. If we do find you using your cell phone or other electronic devices, the rules of the study require us to withhold your completion payment.

Often during this study, you will be shown information or asked to make decisions. After doing so, remember to click the button that says "Continue". The experiment will not proceed until you click that button.

Task 1: Protection Decision

The first part of the experiment has 6 rounds. In each round, you will make the Protection Decision as described below. Please note that after the instruction screen, there will be a short quiz to make sure you understand the experiment. Please read the instructions carefully.

At the beginning of each round, the computer will randomly draw a ball from the box, which contains white and black balls. The number of balls of each color can vary between rounds. We will not tell you which ball has been selected by the computer, but you will know the number of balls of each color as in the picture below.



In each round you must decide whether to buy ***Protection. Protection*** costs $5. If you do not buy ***Protection,*** you lose $20 of your starting money if the Ball is Black, but you do not lose anything if the Ball is White. This means that your earnings will be:

- $30-$5=$25 if you buy protection and the ball is White
- $30-$5=$25 if you buy protection and the ball is Black
- $30-$0=$30 if you do not buy protection and the ball is White
- $30-$20=$10 if you do not buy protection and the ball is Black

We would like to ask you a few questions to check your understanding of this task. Please feel free to go back to the instructions if needed.

Task 2: Informed Protection Task

The second part of the experiment has 6 rounds. Please note that after the instruction screen, there will be a short quiz to make sure you understand the experiment before you can continue to the first round. Please read the instructions carefully.

As in the first part, the computer is going to randomly select one ball from the Box with white and black balls. The computer will show you the contents of the Box but will not tell you the color of the selected ball.

Within each round, you would receive a hint about the ball's color from a gremlin. There are three types of gremlins: an honest gremlin always tells the true color of the Ball, a black-swamp gremlin always says that the Ball is black and a white-swamp gremlin always says that the Ball is white. This is how they look:

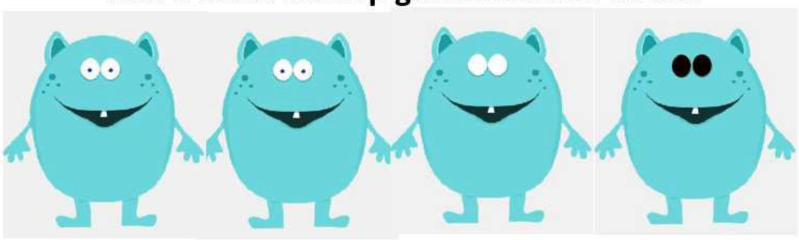
	Honest gremlin: • always tells the true color of the Ball • has regular eyes
	White-swamp gremlin: • always tells that the Ball is white (even when it is not!) • has completely white eyes
	Black-swamp gremlin: • always tells that the Ball is black (even when it is not!) • has completely black eyes

The hints of white-swamp and black-swamp gremlins do not depend on the color of the selected ball. A white-swamp gremlin always says that the Ball is white and would never say that the Ball is black; a black-swamp gremlin always says that the Ball is black and would never say that the Ball is white. Their hints can be correct only by accident.

Suppose, for example, that the Ball is black. Then an honest gremlin would say that the Ball is black. A white-swamp gremlin would say that the Ball is white. A black-swamp gremlin would say that the Ball is black.

On the other hand, if the Ball is white, then an honest gremlin would say that it is white. A white-swamp gremlin would say that it is white. A black-swamp gremlin would say that it is black. Remember that gremlins are just pre-coded computer algorithms and do not intentionally try to help or harm you.

The computer picks the hinting gremlin randomly from a group of gremlins of different types, where each individual gremlin is equally likely to be selected. You will be informed of the mixture of gremlins in this group (similar to the figure below), but you do not know which gremlin is giving the hint.



There are 2 honest gremlins, 1 white-swamp gremlin and 1 black-swamp gremlin in this round.

One of these gremlins would give you a hint, but you won't know which one. All gremlins are equally likely.

The group of gremlins from which the computer selects the hinting gremlin can change from round to round. For example, in one round, you might have two honest gremlins and one white-swamp gremlin in the group. In another round, you might have three honest gremlins and two black-swamp gremlins. You will see the group's composition before making your decisions.

There are two possible hints: either the gremlin says "The Ball is white!" or it says "The Ball is black!". We would like to know whether or not you would buy protection for each of these possible hints. That is, if the hint you receive from a gremlin randomly selected from that group says the Ball is white, would you buy protection? If the hint you receive says that the Ball is black, would you buy protection?

You will need to figure out on your own how likely it is that the hint is true given the group's composition. For example, if all the gremlins are honest, any hint from a randomly drawn gremlin is true. If all the gremlins are white-swamp or all are black-swamp, then their hints give no information. Most often though, your group will include both honest and dishonest gremlins.

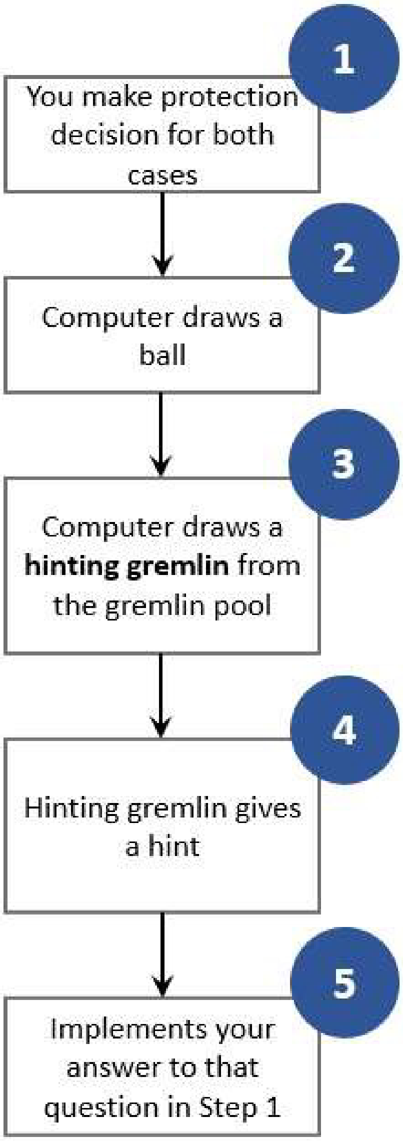
As before, protection costs $5. If you do not buy Protection, you lose $20 of your starting money if the Ball is Black, but you would not lose anything if the Ball is White. This means your earnings will be:

- $30-$5=$25 if you buy protection and the Ball is White
- $30-$5=$25 if you buy protection and the Ball is Black
- $30-$0=$30 if you do not buy protection and the Ball is White
- $30-$20=$10 if you do not buy protection and the Ball is Black

After you tell us your decision for each possible hint, the computer will draw a ball. Then it will record a hint from one randomly chosen gremlin from the group. If the gremlin says that the Ball is white, the computer will implement the choice you made for that hint. If the gremlin says that the Ball is black, the computer will implement the choice you made for that hint. The flow chart below illustrates what happens in each round. You should make your choice for each of two possible hints carefully because either one may determine your payoff if this round is chosen for payment.

Will you protect if a gremlin says the ball is **black**?

Will you protect if a gremlin says the ball is **white**?



1. You make protection decision for both cases

2. Computer draws a ball

Selects a black or white ball

3. Computer draws a **hinting gremlin** from the gremlin pool

The hinting gremlin can be:

- always honest
- always saying black
- always saying white

4. Hinting gremlin gives a hint

5. Implements your answer to that question in Step 1

Payoff when:

You decide to	Selected ball is	
	Black	White
Protect	$25	$25
Not protect	$10	$30

Task 3: Measuring Chances

In this part of the experiment, you will estimate the chance that the Ball is black based on gremlin's hints. We will first show you: 1) the box with white and black balls and 2) the group of gremlins. Imagine that the computer then randomly picks one ball from the box and one gremlin out of this group who will give you a hint. We will ask you two questions:

1. If this gremlin says that the Ball is white, what do you think are the chances that the Ball is white?
2. If this gremlin says that the Ball is black, what do you think are the chances that the Ball is black?

Your estimate each time will be a percentage between 0 and 100. To illustrate how this works, suppose that all the gremlins in the group are honest. It means that their hints are always true: if a gremlin says that the Ball is white, there is exactly 0% chance of it being black. If a gremlin says that the Ball is black, there is exactly 100% chance that the Ball is black. And the chance that the gremlin says it is Black is exactly the chance that is is Black or the proportion of black balls in the box. This case is very easy, but in most cases, the group of gremlins will include some white-swamp and/or black-swamp gremlins. You should take into account the number of white and black balls and the proportions of each type of gremlin in your group when estimating the chances.

Your payoff depends on the accuracy of your answers. All you have to understand in this task is that you make more money if your guess is closer to the actual probability of the event given your information. You make the most money if your guess is exactly equal to the actual probability of the event. For example, you want to predict the chances that the ball is black if the gremlin says that it is black. If the actual probability is 10% and you choose 20%, you payoff will be $30 with probability 90% and $10 with probability 10%. If you choose 50% instead, your payoff will be $30 with the probability of about 60%>. As you can see, you can win if your estimate is very imprecise, but chances are higher for a more accurate estimate. The next two paragraphs lay out the details of how the payoff is calculated, and you are welcome to read these details.

If any round of this task is chosen as the Payment Round, the computer would, first, draw a ball at random from the Box. Then it would record a hint from one randomly chosen gremlin from the group. Finally, it will draw one random lottery with chances between 0 and 100.

This computer will then calculate your payment based both on the hint, the actual ball color and this random lottery. This is easier to understand through an example. Suppose, that the gremlin hints that the Ball is white and you estimate that the Ball is indeed white with probability 85%. If a computer draws a lottery with chances of 85% and above, then you lose $20 if the Ball is white. If the computer draws a lottery with chances lower than 85%, then you would lose $20 with the chance specified in the lottery.

Belief Elicitation: rounds

Suppose that one of the gremlins says that the Ball is white. What do you think is the chance that the Ball is actually **white?** Please estimate to the best of your ability and make your selection on the slider below:

Impossible Completely certain
0 10 20 30 40 50 60 70 80 90 100
Chance (%) that the Ball is white

Suppose that one of the gremlins says that the Ball is black. What do you think is the chance that the Ball is actually **black?** Please estimate to the best of your ability and make your selection on the slider below:

Impossible Completely certain
0 10 20 30 40 50 60 70 80 90 100
Chance (%) that the Ball is black

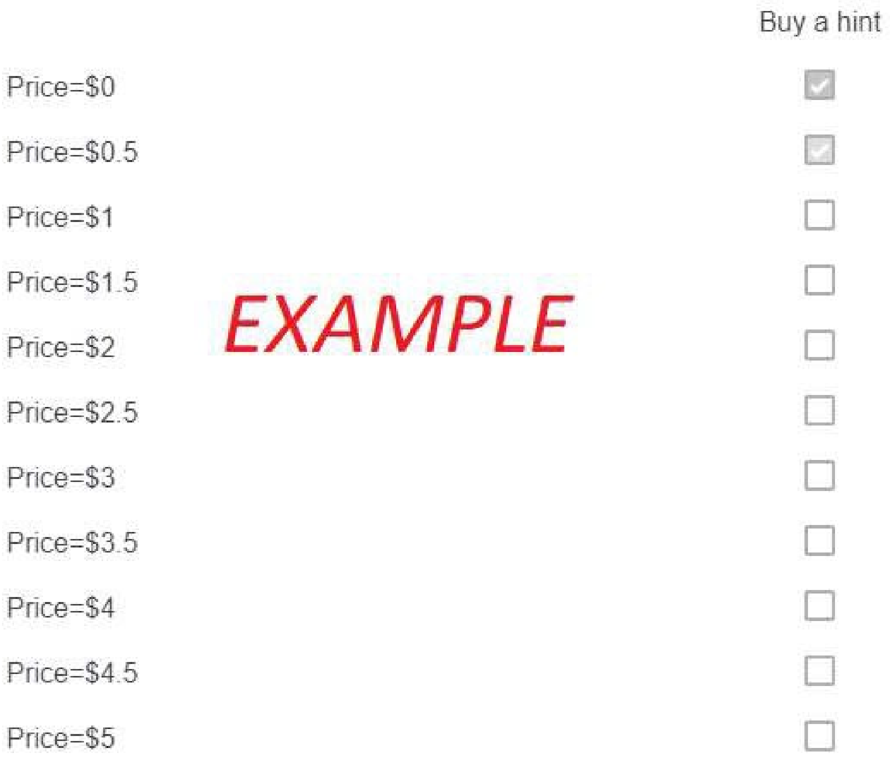
This concludes the round. You will see the outcome only if this round is selected as the Payment Round in the end of the experiment.

Task 4: Value

Were gremlins helpful for you? How much would you pay for their hints if given an opportunity?

In this task, you can buy a hint before making a protection decision. As before, the hint will come from a gremlin which is randomly selected from a group of gremlins of different types. We will show you the group composition, but not the type of the hinting gremlin.

After seeing the group of gremlins, please think about the prices you are willing to pay for the hint. You will then select all acceptable prices by filling a table such as this:



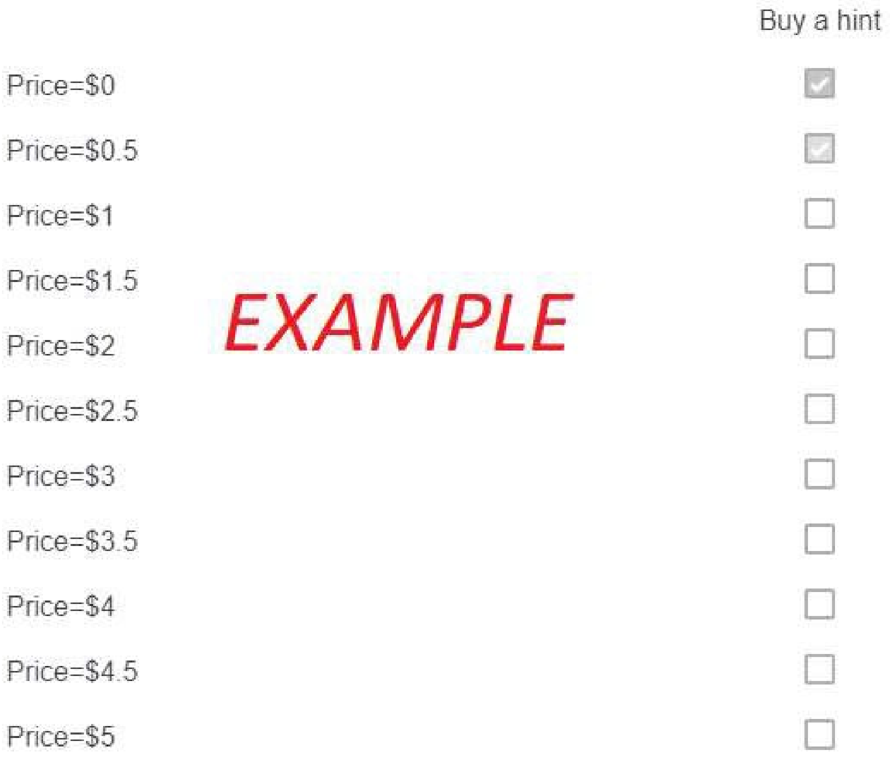
	Buy a hint
Price=$0	☑
Price=$0.5	☑
Price=$1	□
Price=$1.5	□
Price=$2	□
Price=$2.5	□
Price=$3	□
Price=$3.5	□
Price=$4	□
Price=$4.5	□
Price=$5	□

In this table, you select all the prices which you are willing to pay to receive a hint. For example, if you are willing to pay no more than $0.5, then the first and the second rows in the table should be selected as shown in the example above. If you are willing to pay no more than $3, all the rows from the first to the seventh one should be selected. For your convenience, you just need to select the maximum price you are willing to pay for the hint and the system will automatically select all prices lower than that chosen price. You can always unselect the prices by clicking on their checkboxes.

In each round, you will have a different group of gremlins. There are also six rounds in this part of the experiment. You will also have to answer a short quiz before proceeding to the rounds to make sure you understand the experiment. Please read the instructions carefully.

Payoff Calculation. If this the Payment Round, the computer will randomly select one of the prices from the Table. If you chose to buy a hint at this price, you would go through one round of the Informed Protection Task. You will make a Protection decision after receiving a hint from the gremlin. We will subtract the selected price from your payoff in that round. Note, that the price you are paying does not affect the hint's quality.

If you opted not to buy a hint at this price, you would go through one round of the Blind Protection task. In other words, you would make a Protection decision without a hint.



	Buy a hint
Price=$0	☑
Price=$0.5	☑
Price=$1	□
Price=$1.5	□
Price=$2	□
Price=$2.5	□
Price=$3	□
Price=$3.5	□
Price=$4	□
Price=$4.5	□
Price=$5	□

For example, suppose that you fill the table as shown above and this round is the Payment Round. If the computer randomly selects price $0.5 (the second line), you will pay $0.5 and go through one round of the **Informed Protection:** you will receive a hint from one of the gremlins and then choose to protect or not. Your payoff would be equal to what you would have received from the Informed Protection round minus the price of the hint. In this example, if you do not protect, then your payoff will be $30-$0.5=$29.5 if the Ball is white and ($30-$20)-$0.5=$9.5 if the Ball is black. If you decide to protect, your payoff will be ($30-$5)-$0.5=$24.5 for any color of the Selected Ball.

If, for example, the computer randomly selects $1 (line 3) instead of $0.5, you will go through one Blind Protection round and this round would determine your payoff. You will neither pay $1 nor receive a hint, because you did not want to pay this price for a hint based on your selections in the Table. The computer would calculate your payoff in the same way as in the Part 1 of the experiment (Blind Protection).

Suggestions. You should consider the composition of gremlins when selecting the prices to pay. For example, you might have only white-swamp gremlins in the group. Because white-swamp gremlins always say that the Ball is white, their hints are worthless, and most people would not pay anything for them. On another hand, a hint from a group of honest gremlins is more valuable because it tells you the Ball's color with certainty.

It is always in your best interest to select all the prices below or equal to your maximum price. Suppose, for example, that you want to pay any price up to $3 for a hint from a certain group of gremlins. If you do not select the price of $2 and this price is randomly chosen by the Computer, you would have to make the protection decision without a hint even though you prefer to pay $2 to get one. On another hand, if you select the price of $5, you might have to pay $5 which is $2 more than the maximum price you are willing to pay.