Preferences for Warning Signal Quality: Experimental Evidence

Alexander Ugarov[*]
Hivereview

Arya Gaduh[†]
University of Arkansas and NBER

Peter McGee[‡]
University of Arkansas

November 20, 2025

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 the switching off the general safety alarm because the rig "did not want people woke up at 3 a.m. from false alarms" (**?**). 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 (**?**).

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). 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.

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 two 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 derived from a simple model. Subjects' WTP is weakly correlated with the value of information, resulting in the 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 two 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 structure. They find that 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 (**??**). For example, **?** 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, how the test's accuracy is presented strongly affects choices (**?**). We extend this literature using a context-neutral experiment to examine whether similar biases hold more generally and when choices are incentivized; and whether the demand elasticity for information responds symmetrically to false positive and false negative errors.

Second, we contribute to the emerging experimental literature on demand for information quality by studying a novel setting of demand for warnings. Previous studies in this literature employ prediction games, where subjects have to guess an optimal state under uncertainty (**???**). Generally, they find that while demand for information increases with signal quality, it increases more modestly than expected from a Bayesian decision maker. Two of these studies employed laboratory experiments. **?** 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. **?** shows that many subjects choose non-instrumental over instrumental signals, consistent with failures of contingent reasoning about the future value of information.

Our setup differs from those of **?** and **?**. We study preferences over warning signals where subjects face a costly protection decision with three distinct payoffs: full payoff, full payoff minus protection costs, and full payoff minus losses. Hence, risk preferences affect the value of information and can change sensitivities to false-positive and false-negative rates. We also directly elicit both willingness-to-pay and potential protection decisions for different combinations of priors and signal characteristics, allowing for a more general conclusions about subjects' preferences. Consistent with **?**, 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 (**?**), the demand for insurance goes up immediately after an insurable adverse event (e.g., **?**). One proffered 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 (**?**). This is consistent with underweighting prior probabilities relative to more recent signals. At the same time, however, **?** 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 6. Section 7 concludes.

2 Model

Environment. Let $\omega \in \{0, 1\}$ denote the state of 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 income 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.

Preferences. Without a signal, the agent protects if and only if it increases her expected utility:

$$EU_0 = \max[u(Y - c), \pi u(Y - L) + (1 - \pi)u(Y)]$$

Access to a signal can increase expected utility if the signal affects her protection decisions. Under these assumptions, her expected utility with a signal is:

$$EU_s = \pi P_{11}u(Y - c) + \pi P_{01}u(Y - L) + (1 - \pi)P_{10}u(Y - c) + (1 - \pi)P_{00}u(Y)$$

Denote the agent's willingness to pay for the signal by b, which means that she is indifferent between purchasing it at price b and not purchasing it and not learning its realization. The signal's value is equal to the maximum between zero and the solution to the following equation:

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

where $P(s = 1) \equiv \pi P_{11} + (1 - \pi)P_{10}$. The left-hand side expression 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 (1) 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 $u(Y - L)$ against $u(Y - c)$.

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 rates (FN) denoted correspondingly as P_{10} and P_{01}. This is the equation we use in our empirical work:

$$b = \max[0, \min[c, \pi L] - \pi(1 - P_{01})c - (1 - \pi)P_{10}c - \pi P_{01}L] \tag{2}$$

The WTP for signals b has equal sensitivity to expected FP and FN costs calculated as $\pi(1 - P_{01})c$ and $\pi P_{01}L$. When $b > 0$, its derivatives with respect to FP (P_{10}) and FN (P_{01}) rates are given by:

$$\frac{db}{dP_{10}} = -(1 - \pi)c \tag{3}$$

$$\frac{db}{dP_{01}} = -\pi(L - c) \tag{4}$$

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 false-positive (false-negative) rate.

Risk Aversion. 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 protection costs are low, (i.e., $c < \pi L$), then a strictly risk-averse decision-maker pays less than a risk-neutral one.*

Proof. See the Appendix. □

For low-enough protection costs, risk-averse decision-makers protect by default without using a signal. Things are more ambiguous with low risks or higher protection costs. For example, risk aversion increases the value of a perfect signal as long as a risk-averse decision-maker chooses to not protect without a 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 a signal's false-positive and false-negative rates on the 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:

$$\frac{db}{dP_{10}} = -\frac{(1-\pi)(u(Y-b)-u(Y-c-b))}{D(\pi, P_{01}, P_{10}, b)} \tag{5}$$

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

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 with FP and FN rates, i.e., $\frac{db}{dP_{10}}$ and $\frac{db}{dP_{01}} < 0$. We can also say a bit more about the sensitivity to FN rates:

Proposition 2. *A risk-averse and imprudent decision-maker has higher sensitivity to FN rates compared to a risk-neutral decision-maker.*

Proof. See the Appendix. □

Equation 5 shows that risk aversion can either increase or decrease a decision-maker's sensitivity to FP 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. We can only say that it is very likely that for low protection costs and small priors π (leading to no automatic blind protection) the ratio of sensitivities to FP rates over FN rates should be lower for risk-averse subjects.

Proposition 3. *For low protection costs c and small risks π, risk aversion lowers relative sensitivity to FP rates.*

Proof. See the Appendix. □

At the same time, equation 6 shows that the sensitivity to FN rates depends on weighing the marginal utility of consumption when experiencing losses after paying for a signal against the expected marginal utility after paying for a signal (denominator). The former brings lower-than-average payoffs that corresponds to lower marginal utility for risk-averse decision makers, but the ratio also depends on how the decision-maker perceives lotteries in payoff changes described by *prudence*. The average marginal utility is going to be higher only when the decision-maker is both risk-averse and dislikes lotteries in payoff changes (i.e., imprudent with $u'''() < 0$).

* * *

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 2).

3 Experimental Design

We conducted the experiment in the Behavioral Business Research Lab (BBRL) at the University of Arkansas between October and November 2021. A total of 105 subjects participated in an individual decision task implemented using Qualtrics. On average, including a $5 show-up fee, subjects earned $26 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. The experimental outcome was determined by 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.[1] 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 must 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 and started at the lowest probability. Subjects did not receive feedback on the decision's outcome.

[1]Incorrect quiz answers for the Informed Protection section resulted in subjects facing three additional multiple choice questions. In our opinion, 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.

Informed Protection (IP). Similar to the BP task, subjects must make a protection decision given the prior probability of drawing a black ball. Subjects learn a prior and signal's accuracy. Following **?**, we use a group of hinting gremlins to convey the signal's accuracy, where a randomly selected gremlin from a group provides a hint (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 the different gremlin types were presented to the subjects. The composition of the group of gremlins determines signal's accuracy: 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 a hint in any particular round. The prior probability of drawing a black ball and the composition of gremlins vary across the rounds.

Figure 1: Signals Presentation

Belief Elicitation (BE). As in the IP task, subjects know the prior probability of drawing a black ball and the composition of the group of gremlins providing hints. 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 **?** and **?** 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.[2] Motivated by **?**, who find that providing a detailed explanation of payoffs can lower truthful reporting, we instead explain that reporting one's true belief μ maximizes their payoffs, and give an example of payoff calculation under different reporting strategies.

[2]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 (**?**) 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.

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 composition of the group of gremlins giving hints. Unlike the IP task, subjects do not automatically receive a hint, instead they provide their WTP for a hint 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 of a hint will be 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 the WTPE task, subjects answered a few demographic questions: gender, age, and number of statistics courses they have taken. The payment task and the payment round were then randomly chosen to calculate the subject's payoff.

For tasks other than BP, subjects go through two different priors and three types of signals. The order is such that subjects go consecutively over all three signal types starting from the honest one for each prior. The order of priors and signals stays constant for each subject across tasks, but can vary between subjects. Table 1 summarizes our treatments.

Table 1: List of Treatments

	Gremlins composition				
Prop. of black balls (p)	Honest	Black-eyed	White-eyed	FP rate	FN rate
0.1, 0.2, 0.3, 0.5	2	0	0	0	0
0.1, 0.2, 0.3, 0.5	3	1	0	0.33	0
0.1, 0.2, 0.3, 0.5	3	0	1	0	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	0.2	0
0.1, 0.2, 0.3, 0.5	5	0	1	0	0.2
0.1, 0.2, 0.3, 0.5	5	1	1	0.2	0.2

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. We briefly discuss patterns of subject decisions below. They suggest that subjects understand these tasks reasonably well.

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 (72/105), 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.[3]

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 (24) 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.[4]

4.2 Informed Protection

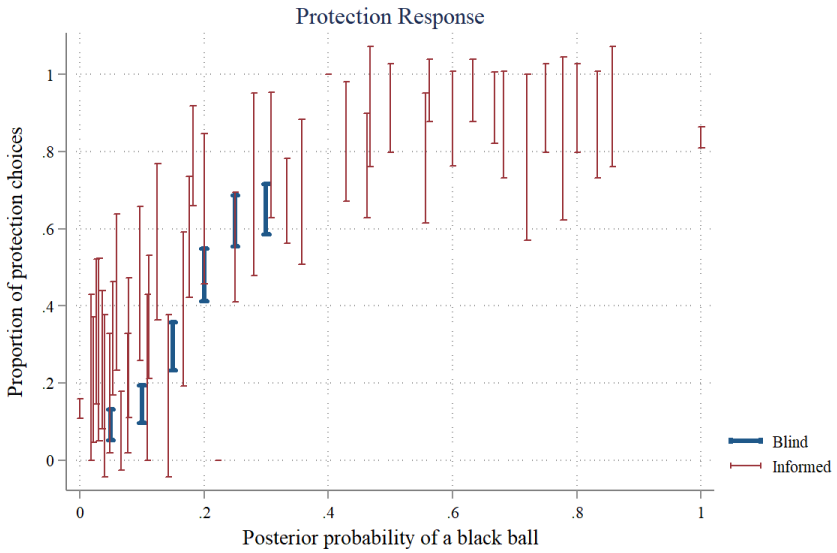
Recall that, in the IP task, subjects not only know the prior but also receive a hint about the ball color. Figure 2 shows that protection actions are increasing in the posterior probability of an adverse event, though roughly 28% of subjects break monotonicity in their protection responses with respect 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.[5]

[3]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. **?** 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 where options are presented in increasing payoff order, which they are not here. Among our switchers, however, 83% (24/39) skip only a single increment of the presented probability scale, suggesting an inattention error.

[4]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 $\theta = -0.54$.

[5]We use a linear probability model to estimate this relationship, and while the coefficient on the total number of protection choices is significant at the 1% level, the R^2 only increases from 0.295 to 0.3.

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's characteristics and the hint provided. Column 4 shows the posterior probability of a black ball averaged across all the treatments within a group, column 5 — the subjects' share of empirical protection responses, while column 6 presents the theoretical optimum for a risk-neutral decision maker. Finally, column 7 presents the p-value for a test of equality between empirical and theoretical protection responses.

Three observations emerge from the table. First, regardless of the signal's FP and FN rates, black hints 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 hints (rows 1–4), while significantly underprotecting when facing black hints without false positives (rows 5–6). Subjects overprotect for black hints with false positives, though the difference is not statistically significant.

Third, we find deviations that cannot be explained by the 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 hint is white), the protection rate increases by 6 percentage points (pp).[6] Similarly, comparing rows 3 and 4, we see that introducing false negatives to a signal that also generates false positives raises the protection rate increases to 56 percent — even though the average posterior probability given the signal's characteristics is merely 13 percent. As a benchmark, with no signal in the BP task, only 13 (32) percent of subjects choose to protect when the probability is 10 (15) percent.

[6]The difference is significant at 5%

Table 2: Average Protection by Signal Type

Row	Signal Characteristics		Hint	Posterior	Share Protect	Share Optimal	p
	False Positive	False Negative					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1)	No	No	White	0.000	0.049	0.000	0.000
(2)	No	Yes	White	0.112	0.262	0.041	0.000
(3)	Yes	No	White	0.000	0.255	0.000	0.000
(4)	Yes	Yes	White	0.117	0.454	0.096	0.000
(5)	No	No	Black	1.000	0.824	1.000	0.000
(6)	No	Yes	Black	1.000	0.855	1.000	0.000
(7)	Yes	No	Black	0.520	0.810	0.869	0.043
(8)	Yes	Yes	Black	0.517	0.875	0.900	0.367

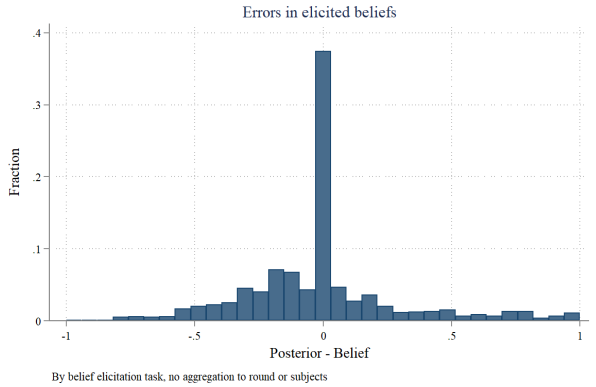
Notes: The p-value in column 7 is for the test of equality between the theoretical prediction (column 6) and the observed share of protection (column 5).

4.3 Belief Elicitation

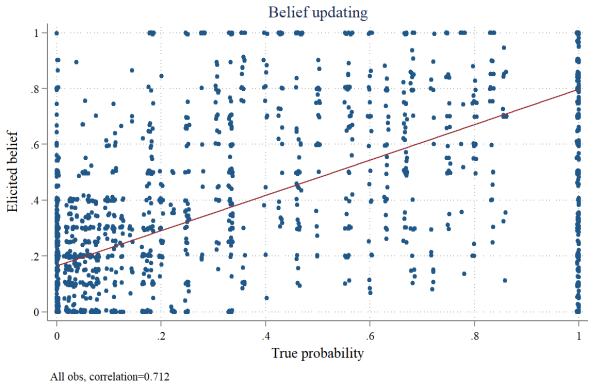
Subject decisions in the IP task capture the use of signals in protection decisions, but decisions reflect both risk preferences and (potentially erroneous) beliefs. The BP task can be used to construct a measure of the former; the BE task to measure the latter.

Figure 3: Errors in Bayesian Updating

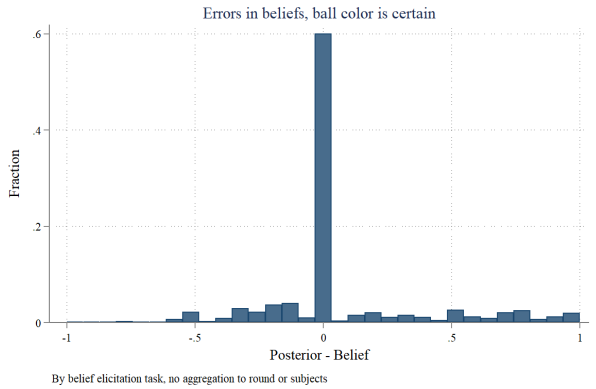
(a) Error Distribution



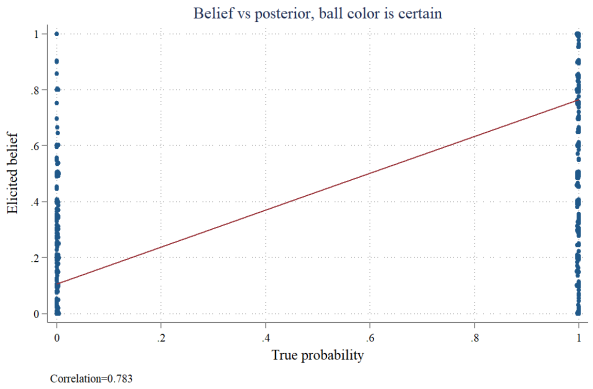
(b) Error v. Posterior



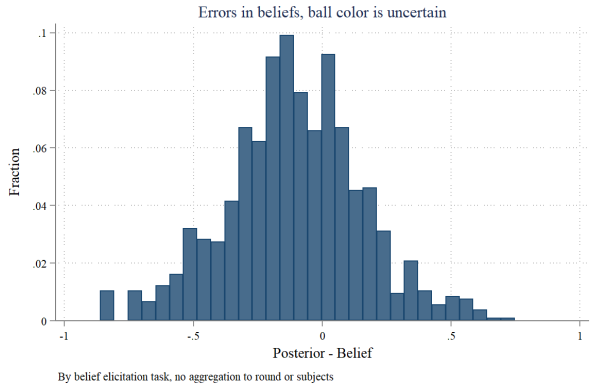
(c) Error Distribution, Certain Posterior



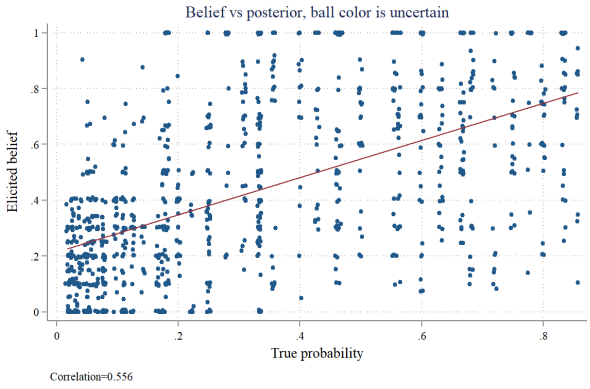
(d) Error v. Posterior, Certain Posterior



(e) Error Distribution, Uncertain Posterior



(f) Error v. Posterior, Uncertain Posterior



We define updating errors as the difference between the subjects' elicited belief and the actual 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 with a fitted line. Panel A indicates that beliefs are still sensible despite errors. The distribution of updating errors is centered

at 0, with roughly one-half (51%) of errors concentrated within +/- 0.1 interval around zero. Overall, the correlation between the elicited beliefs and the true posteriors was 0.748 (see Panel B for the scatter plot).

For some combinations of priors and signals, updating should be trivial and posteriors are completely certain. Panel B plots such cases, which account for 56% 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 honest and white-swamp gremlins only announcing that the ball is black — or vice versa). Reassuringly, 69% of reported beliefs are correct. About half of the errors involve reporting a probability of one when it should have been zero.

Meanwhile, Panel C plots the remaining observations, i.e., those with uncertain posteriors. The median error in Panel C is -0.12, with 90% of errors lying between -0.48 and 0.3, 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 0.571.[7]

Table 3: Average Updating Error by Signal Type

Row	**Signal Characteristics**		**Hint**	**Posterior**	**Updating Error***	p
	False Positive	**False Negative**				
	(1)	(2)	(3)	(4)	(5)	(6)
(1)	No	No	White	0.000	0.047	0.000
(2)	No	Yes	White	0.112	0.061	0.000
(3)	Yes	No	White	0.000	0.189	0.000
(4)	Yes	Yes	White	0.117	0.201	0.000
(5)	No	No	Black	1.000	-0.145	0.000
(6)	No	Yes	Black	1.000	-0.362	0.000
(7)	Yes	No	Black	0.520	0.139	0.000
(8)	Yes	Yes	Black	0.517	0.036	0.043

Notes: The updating error is defined as Belief − Posterior. 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 signal's characteristics. We find that subjects overestimate the probability of a black ball when receiving a white hint, which is consistent with overprotection noted for the IP task. This upward bias for a white hint increases in both the FP and FN rates of the signal. To illustrate, consider the change between rows 1 and 3, where introducing a FP rate would not change the posterior because the signal

[7]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 (**???**) — is known as *representativeness bias* or *base-rate neglect*. Using the regression approach of **?**, 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 A2). These values are within the range found by the meta-analysis of **?** 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

realization is white. Yet, subjects update their posterior 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 signal realizations (hints), however, varies by information structure. Subjects slightly underestimate the probability with a perfectly accurate signal, but introducing FN rates exacerbates subjects' underestimation. Rows 5 and 6 suggest that the introduction of a FN rate without changing the posterior further reduces subjects' beliefs. With non-zero FP rate, subjects again overestimate the probability of a black ball. The difference in the updating errors for black hints coming from FP-only (row 7) v. FP-FN signals (row 8) is negligible. The magnitude of subjects' adjustments to their beliefs was smaller than the actual change to the posteriors due to the FP rates.

5 WTP and Signal Characteristics

5.1 Are Subjects Risk Neutral, Expected Utility Maximizers?

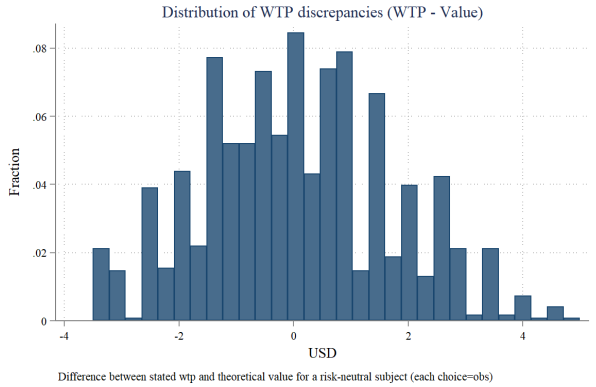
Hypothesis 1. *Subjects' WTPs for signals are equal to their value for risk-neutral agents.*

Result 1. *On average, there are no significant discrepancies between WTP and predicted value for risk-neutral agents. When splitting by a signal type, a difference emerges only for signals with both false-positive and false-negative rates.*

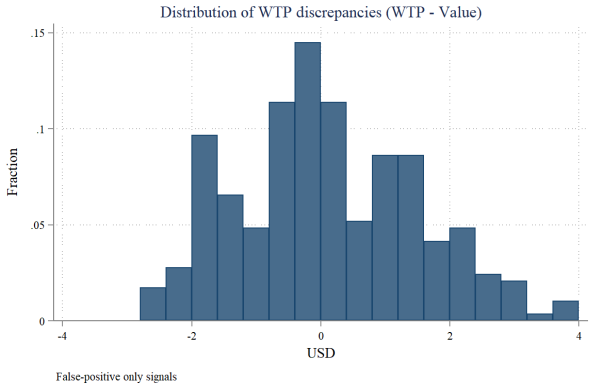
Overall, the theoretical signal value for a utility maximizing risk-neutral subject (hereafter, the risk-neutral WTP) in equation 2 is a useful benchmark of our subjects' WTP. Figure 4 plots the distribution of the differences between subjects' WTP and this value. The distribution is centered around 0, indicating that average choices do not fall far from the choices of a risk-neutral utility maximizer. However, there is substantial variation: only 25% of reported WTP are within \$0.50 of the risk-neutral signal value, and subjects overvalue signals by at least \$1.5 in 22% of cases and undervalue by at least \$1.5 in 19% 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: Discrepancy (Observed WTP - Signal value) by Signal Type

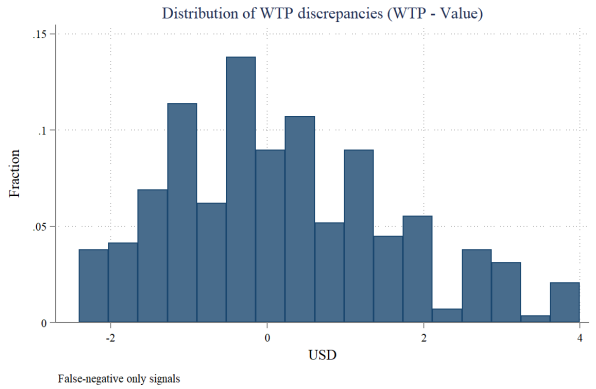
(a) All signals



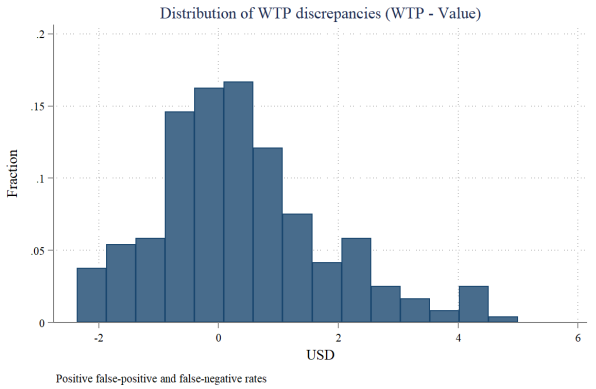
(b) FP only



(c) FN only



(d) Both FP and FN



Our comparisons in Table **??** also find no differences on average between the observed WTP and the risk-neutral WTP for 3 out of 4 signal types: honest (i.e., perfectly accurate), FP-only, and FN-only. For signals having both FP and FN rates, however, subjects' reported WTP is significantly higher than the risk-neutral WTP. Subjects' overvaluations were similar for both low and high priors. Note, that these signals also induce overprotection in the IP task. Additionally, subjects tend to overpay for signals with positive FP rates when the prior is low (0.1 or 0.2), and for signals with positive FN rates when the prior is high (0.3 or 0.5).

Table 4: Average WTP discrepancy (WTP-Value) by Signal Type

Priors	**Honest**	**FN only**	**FP only**	**FP and FN**
All priors	-0.261**	0.183*	0.099	0.434***
Low priors	-0.039	-0.033	0.593***	0.451***
High priors (>0.2)	-0.483***	0.399**	-0.394***	0.417**

*The number of stars represents statistical significance (0.05, 0.01, 0.001)

Hypothesis 2. *Subjects' preferences demonstrate equal sensitivity to costs generated by false-positive and false-negative events.*

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

To examine how the WTP responds to signal quality, we estimate the relationship between WTP biases and signal characteristics with the following regression:

$$\Delta b_{is} = \beta_0 + \beta_1 FP + \beta_2 FN + \varepsilon_{is}$$

where $\Delta b_{is} = (b_{is} - b_s^*)$ is the difference between the WTP of individual i for signal s and b_s^* is the risk-neutral WTP; FP (FN) is the false positive (false negative) cost. Note here that false positive or false negative *costs* are functions of both the rates and the costs of the consequences of those rates. For example, a high false-negative rate imposes fewer costs when priors are low because the adverse outcome is already unlikely, whereas high false-negative rates carry substantial costs when the prior probability of the adverse outcome is high. All specifications include subject fixed effects, with standard errors clustered at the subject level.

Table 5 reports the results of our regression. If subjects are risk-neutral expected-utility-maximizers, we expect all coefficients to be jointly and individually insignificant. Instead, column 1 shows positive and statistically significant coefficients for both FP and FN costs with highly significant model's F-test. In other words, subjects deviate by overpaying for inaccurate signals.

The risk-neutral model predicts that subjects should value the marginal costs of false-negative and false-positive events symmetrically. Table 5 shows that the coefficient on FN costs is slightly larger, but we cannot reject the hypothesis that the two coefficients are equal. However, as we will show, this equivalency breaks down when considering specific priors.

5.2 Risk Preference and Belief Accuracy

Our baseline estimation in column 1 indicates significant deviations from the behavior of a risk-neutral agent as predicted by the model. The positive and significant coefficients suggest that these deviations are increasing in FP and FN costs. In other words, as signal quality deteriorates (with increasing FP or FN rates), our subjects do not reduce their valuation of the signals as quickly as a risk-neutral agent would.

Because our benchmark model assumes both perfect updating and risk neutrality, the deviations could occur through 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 signal's quality, but fall short in explaining WTP systematic biases reported above. We use data from the BP task to categorize

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
	(0.081)***	(0.126)***	(0.158)***	(0.180)***	(0.308)
FN costs	0.287	0.327	0.357	0.016	0.367
	(0.046)***	(0.084)***	(0.088)***	(0.216)	(0.085)***
Risk-averse × FP costs		-0.329	-0.415	-0.243	-0.576
		(0.225)	(0.257)	(0.285)	(0.427)
Risk-averse × FN costs		-0.355	-0.361	-0.352	-0.288
		(0.124)***	(0.135)***	(0.292)	(0.128)**
Risk-loving × FP costs		0.048	0.018	0.008	0.318
		(0.179)	(0.213)	(0.290)	(0.408)
Risk-loving × FN costs		0.080	0.110	0.361	0.119
		(0.107)	(0.117)	(0.341)	(0.118)
Constant	-0.308	-0.310	-0.436	-0.076	-0.517
	(0.059)***	(0.057)***	(0.093)***	(0.144)	(0.142)***
R^2	0.480	0.491	0.500	0.731	0.745
Prob>F	0.0000	0.0001	0.0001	0.0000	0.0000
Obs	1230	1230	1230	615	615
Risk-Averse Subjects:					
False Positive		0.158	0.228	0.334	-0.273
se		(0.186)	(0.203)	(0.220)	(0.295)
p-value		[0.395]	[0.263]	[0.131]	[0.356]
False Negative		-0.028	-0.003	-0.337	0.079
se		(0.091)	(0.102)	(0.196)	(0.096)
p-value		[0.760]	[0.976]	[0.087]	[0.411]
Risk-Loving Subjects:					
False Positive		0.535	0.661	0.585	0.621
se		(0.127)	(0.143)	(0.227)	(0.268)
p-value		[0.000]	[0.000]	[0.011]	[0.021]
False Negative		0.406	0.467	0.377	0.487
se		(0.066)	(0.077)	(0.264)	(0.082)
p-value		[0.000]	[0.000]	[0.155]	[0.000]
Subject FE	Yes	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 level). In the bottom panels, we also test whether the total coefficient value (baseline+interaction) are different from zero.

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.[8]

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 WTP discrepancies of both risk-neutral and risk-loving subjects increase with FP (statistically insignificant) and FN costs — suggesting that they do not downward-adjust their WTP enough to account for lower quality signals. In contrast, the WTP discrepancies of risk-averse subjects show little sensitivity to FP and FN 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.[9] 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-neutral subjects with relatively accurate beliefs. We find a lower sensitivity to FP costs for risk-neutral subjects with accurate beliefs and very little change to the corresponding sensitivity to FN costs. This indicates that even relatively accurate Bayesians do not downward-adjust their WTP enough to increasing FP or FN costs.[10]

5.3 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 keep the default action or to switch. We split the priors in 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 (10% and 20%). With low priors, deviations from the risk-neutral WTP increase with FP costs: subjects overvalue signals that would induce them to overprotect. This overvaluation is similar for different risk preference profiles. Column 5 presents the results for high-prior WTPE tasks. With high priors, the deviations of risk-neutral and risk-loving subjects from the risk-neutral WTP increase with FN costs. By not reducing their WTP enough to account for the increasing FN rates, subjects

[8]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.

[9]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.

[10]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 A Table A4.

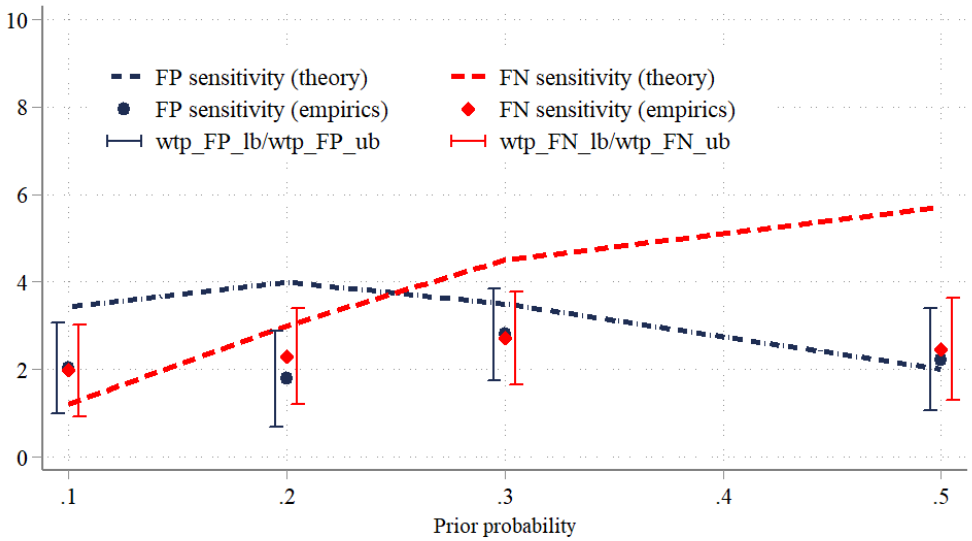
overvalue signals that would induce them to underprotect.

In summary, with low priors, subjects do not reduce their WTP enough to account for the deteriorating signal quality captured by the false-positive costs. In contrast, with high priors, they underreact to false-negative costs. In practice, most warning systems are designed for low probability events. In such cases, users would tend to overpay for signals with high false-positive costs and underpay for signals with high false-negative rates. For example, they would prefer a smoke alarm that never misses fire incidents, even when its expected cost of false alarms is high. Risk preferences affect this pattern with risk-averse subjects moving closer to a risk-neutral benchmark, but most interaction coefficients are not statistically significant despite large magnitudes.

6 Discussion

Subjects' underreactions to false-positive (false-negative) costs for low (high) priors present a puzzle. These behaviors are inconsistent with our risk-neutral model: Equations 3 and 4 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. As stated earlier, for a given FN rate, false-negative events 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. The divergence between our subjects' WTP and the risk-neutral WTP explains changing signs on FP and FN costs in the previous regressions of WTP differences.

Figure 5: Theoretical and Empirical WTP Sensitivities to FP and FN rates



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

Figure 5 illustrates this puzzling behavior. This figure plots estimates from the regression of reported WTP — instead of its deviation from the risk-neutral WTP — on FP and FN rates.

We find that the sensitivities of subjects' WTP to both FP and FN rates increase with priors and that the change occurs relatively smoothly, and the two sensitivities are also surprisingly close to each other.[11]

We consider four candidate explanations for these puzzling results: risk preferences, anchoring, valuing non-instrumental information, and finally a failure to distinguish how FP and FN error rates ought to affect calculated posteriors differently.

Risk Preferences. Our evidence suggests that risk preference cannot 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 in Appendix Table A5 by explicitly testing for interactions between risk-preferences, priors, and FP and FN rates. We find that these interactions are mostly insignificant, with the exception of interactions between FN rates and risk aversion for some specifications. The heterogeneity largely remains after controlling for risk preferences, but the interaction between high priors and FP rates becomes insignificant.

Anchoring. The evidence also does not support the hypothesis that subjects anchored on previous priors. Each subject goes through two sets of treatments with two different priors in a fixed order, so anchoring could occur. We find, however, that most subjects (92 out of 104) change their decisions when going from one prior to another, and the average belief error in the BE task is actually *lower* for the second set of priors rather than the first, which suggests that changing priors does not make subjects more confused. Most importantly, we do not see statistically significant differences between FP and FN coefficient estimates even if we limit our attention only to the first priors in each sequence.[12]

Non-instrumental information. There is evidence in the literature of people valuing "non-instrumental information" that does not affect their decisions. For example, **?** find that subjects are willing to pay to know the probability of their choice being correct even if this information cannot affect their choice. Similarly, **?** 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. 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 (159 out of 624 total choices). It is therefore plausible that the reported WTPs include some non-instrumental components.

[11] We cannot reject the hypothesis that two sensitivities are equal to each other for any of the priors.

[12] The first 3 WTP treatments use either 0.1 or 0.2 as the prior (depending on the treatment), so there is no anchoring on the previous prior or something special about a particular prior.

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, chooses an action, and then immediately learns their payoffs. Hence, there is practically no window for subjects to experience any anticipatory feelings that the literature assumes to be the standard causal mechanism behind the demand for non-instrumental information. 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 next explanation, no theory of preferences for non-instrumental information suggests the effects should be so similar to one another.

Failure to distinguish between FP and FN. Instead, we argue that subjects' observed behaviors arise from treating FN and FP as 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. Out of 105 subjects, 39 refer to the *percentage* 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 for protection.*"

Among the other 66 subjects, some may use this heuristic without describing it.[13] 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 their lack of variation with respect to priors. 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 signal's value with the respect to the sum of FP and FN rates should be relatively flat because priors affect FP and FN costs in opposite ways.[14]

We can test this hypothesis using choices from the BE and IP tasks where subjects face imperfect signals. If subjects systematically neglect the difference between FP and FN rates, we expect to find the pattern of unexplained reaction to FP and FN rates in cases when they do not affect the posterior. Namely, subjects would show sensitivity to FP rates when the signal is white and sensitivity to FN rates with black signals. This happens because some subjects

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

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

react to FP rates as if they are FN rates, and vice-versa. If present, this pattern cannot be explained by any distribution of risk preferences or by anchoring on previous priors.

Table 6 presents the results from linear regressions of updating error, i.e., actual posterior - reported belief, 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–the fraction of the group of gremlin that are black swamp gremlins who always say black–has a significant positive effect on the error when the signal is *white*, and that FN rate–the fraction of the group of gremlin that are white swamp gremlins who always say white–has a significant negative effect when the signal is *black*.

Table 6: Updating Errors in BE Task

	All	Signal Received	
		White	Black
	(1)	(2)	(3)
FP rate	0.797***	0.532***	1.063***
	(0.040)	(0.042)	(0.073)
FN rate	-0.243***	0.042	-0.528***
	(0.038)	(0.045)	(0.061)
Constant	-0.058***	0.328***	-0.444***
	(0.005)	(0.006)	(0.009)
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 level).

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:[15]

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

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 $[0, 1]$.[16] 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.

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

[16] We use Stata 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.

Columns 1 and 2 show that, even conditional on posterior and subject FEs that account for risk preferences, IP responses are still affected by FP and FN rates. For a white hint, FP and FN rates increase the tendency to overprotect while the FP rate has an opposite effect with comparable magnitude but without statistical significance for a black hint. Hence the first prediction of a conjecture of indiscriminate FP and FN rate use holds: FP rates increase protection when the signal realization is white conditional on the posterior. The effect holds if allowing for heterogeneity of sensitivities to FP and FN rates with respect to priors (Column 2), though the effect of the FN rate for black hints is small in magnitude and not statistically significant at conventional levels. Adding flexible controls for subjects' beliefs reduces the coefficient magnitude on FP rate for white hints (Columns 3 and 4), but the coefficients still remain significant. This indicates that while beliefs partially contribute to these protection anomalies, they cannot explain them completely.

Table 7: Informed Protection Response

	(1)	(2)	(3)	(4)
FP rate x (S=White)	0.895***	0.943***	0.525***	0.571***
	(10.011)	(10.145)	(5.518)	(5.850)
FN rate x (S=White)	0.537***	0.532***	0.307**	0.299**
	(3.709)	(3.631)	(2.139)	(2.048)
p>0.2	0.039**	0.041**	0.024	0.029
	(2.408)	(1.965)	(1.558)	(1.457)
S=Black	0.531***	0.542***	0.383***	0.374***
	(5.161)	(4.758)	(3.653)	(3.268)
FP rate x (S=Black)	-0.032	0.025	-0.065	-0.000
	(-0.158)	(0.123)	(-0.330)	(-0.001)
FN rate x (S=Black)	0.103	0.069	-0.005	-0.021
	(1.398)	(0.860)	(-0.055)	(-0.229)
FP rate x (p>0.2)		-0.081		-0.085
		(-1.066)		(-1.081)
FN rate x (p>0.2)		0.088		0.048
		(0.891)		(0.521)
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).

Overall, we observe a striking uniformity in sensitivity of WTP to both false-positive and false-negative rates that cannot be explained by risk preferences or anchoring. This pattern is, however, consistent with subjects neglecting the difference between false-positive and false-

negative signals, a behavior that is supported by subjects' explanations of their decision making and the odd sensitivities to false-positive and false-negative rates in other treatments in which they do not affect posterior probabilities.[17]

7 Conclusion

We study how error characteristics of warning signals affect user's valuations for warning signals. While the risk-neutral benchmark model does a good job of describing average elicited willingness-to-pay, it masks an important underlying heterogeneity. Relative to the risk-neutral WTP, individual valuations of warning signals inadequately adjust for worsening signal quality from false-positive (false-negative) costs for low (high) prior probability events. These deviations cannot be explained by either risk preferences or belief-updating accuracy. Instead, they seem to be consistent with a decision heuristic where subjects do not distinguish between false-negative and false-positive errors.

Understanding the source of this asymmetry can improve social efficiency. As we find here, 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 from 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. Being armed with a better understanding of the deviations we document is the first step in either designing better signals or implementing interventions that may help users understand the costs of various types of signals.[18]

[17] 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 **?** 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.

[18] For example, evidence to confirm that this bias comes from confusing different types of probabilistic errors can motivate interventions to improve how information about these errors is presented. Studies on Bayesian updating, for instance, show that medical professionals make better decisions if the information on medical tests is presented in the form of expected frequencies rather than a tuple of prior conditional probabilities (**??**).

A Tables

Table A1: Demographic Characteristics of Subjects

	All		$p \in \{0.1, 0.3\}$		$p \in \{0.2, 0.5\}$	
	N	%	N	%	N	%
Male	43	41	22	41	21	41
Age>23yrs old	14	13	6	11	8	16
Students	88	84	46	85	42	82
Had statistics classes	63	60	37	69	26	51
Total	105	100	54	100	51	100

Table A2: Error Decomposition

	(1) OLS	(2) FE	(3) OLS	(4) FE	(5) OLS	(6) FE
Prior	0.246	0.202	0.175	0.191	0.140	0.040
	(0.044)	(0.051)	(0.056)	(0.076)	(0.062)	(0.063)
Signal	0.430	0.430	0.327	0.327	0.539	0.539
	(0.069)	(0.069)	(0.103)	(0.103)	(0.101)	(0.101)
Good quiz × Prior			0.143	0.021		
			(0.086)	(0.102)		
Good quiz × Signal			0.193	0.193		
			(0.138)	(0.138)		
Stat. class × Prior					0.162	0.264
					(0.085)	(0.094)
Stat. class × Signal					-0.166	-0.166
					(0.134)	(0.135)
Observations	280	280	280	280	280	280
Adjusted R^2	0.31	0.31	0.33	0.32	0.32	0.32

Notes: Decomposition works only for imperfect signals, hence the table excludes the responses to certain signals. Standard errors in parentheses (clustered at the subject level).

Table A3: Informed Protection Response: Logit with Flexible Control for Posteriors

	(1)	(2)	(3)	(4)
FP rate	0.365	0.472	0.593	0.573
	(0.111)	(0.140)	(0.147)	(0.157)
FN rate	0.168	0.611	0.150	0.565
	(0.093)	(0.219)	(0.098)	(0.223)
p>0.2	0.026	0.066	0.047	0.055
	(0.017)	(0.024)	(0.026)	(0.028)
S=Black	0.004	0.426	-0.023	0.473
	(0.063)	(0.170)	(0.067)	(0.195)
FP rate x (S=Black)		-0.655		-0.690
		(0.472)		(0.470)
FN rate x (S=Black)		-0.561		-0.608
		(0.267)		(0.278)
FP rate x (p>0.2)			-0.293	-0.160
			(0.129)	(0.136)
FN rate x (p>0.2)			0.084	0.264
			(0.155)	(0.163)
Observations	1248	1224	1224	1224
Adjusted R^2				

Notes: Specifications include flexible controls of posterior probability. The table reports average marginal effects, includes subject FE, errors are clustered by subject. Standard errors in parentheses.

Table A4: WTP minus Value of Information: Demographic Determinants

	(1)	(2)	(3)	(4)	(5)	(6)
FP costs	0.283	0.352	0.117	0.215	0.248	0.291
	(0.176)	(0.180)	(0.200)	(0.207)	(0.141)	(0.142)
FN costs	0.322	0.247	0.395	0.303	0.303	0.249
	(0.096)	(0.090)	(0.111)	(0.104)	(0.076)	(0.068)
Male	-0.193	-0.157				
	(0.315)	(0.404)				
Male × FP costs	-0.153	-0.193				
	(0.244)	(0.248)				
Male × FN costs	0.079	0.114				
	(0.144)	(0.127)				
Stat. class			-0.240	-0.142		
			(0.324)	(0.435)		
Stat. class × FP costs			0.198	0.124		
			(0.259)	(0.266)		
Stat. class × FN costs			-0.083	-0.023		
			(0.146)	(0.133)		
>23 yrs					-0.366	-0.647
					(0.400)	(0.372)
>23 yrs × FP costs					-0.068	0.024
					(0.303)	(0.345)
>23 yrs × FN costs					0.350	0.277
					(0.212)	(0.209)
Constant	-0.126	0.391	-0.058	0.419	-0.157	0.397
	(0.205)	(0.266)	(0.257)	(0.361)	(0.171)	(0.221)
Prior dummies	No	Yes	No	Yes	No	Yes
Observations	624	624	624	624	624	624
Adjusted R^2	0.05	0.21	0.05	0.21	0.05	0.21

Notes: Standard errors in parentheses (clustered by subject).

Table A5: WTP minus Value of Information: Risk Aversion and Sensitivity to FP and FN Costs

	(1)	(2)	(3)	(4) FE	(5) FE
p>0.2	-0.094	-0.110	-0.041	-0.127	-0.058
	(0.189)	(0.189)	(0.286)	(0.188)	(0.283)
FN costs	-0.229	-0.442	-0.327	-0.385	-0.360
	(0.131)	(0.232)	(0.239)	(0.214)	(0.217)
p>0.2 × FN costs	0.716	0.977	0.889	0.949	0.914
	(0.104)	(0.190)	(0.191)	(0.175)	(0.177)
FP costs	0.558	0.690	0.780	0.652	0.672
	(0.118)	(0.177)	(0.188)	(0.192)	(0.191)
p>0.2 × FP costs	-0.933	-0.879	-0.899	-0.863	-0.910
	(0.182)	(0.299)	(0.337)	(0.299)	(0.322)
Risk-loving × p>0.2 × FN costs		0.037	-0.383	-0.059	-0.276
		(0.145)	(0.249)	(0.151)	(0.238)
Risk-averse × p>0.2 × FN costs		-0.245	-0.279	-0.372	-0.198
		(0.159)	(0.242)	(0.168)	(0.252)
Inconsistent × p>0.2 × FN costs		-0.074	-0.181	-0.066	-0.297
		(0.174)	(0.433)	(0.160)	(0.352)
Risk-loving × p>0.2 × FP costs		-0.287	0.097	0.179	0.259
		(0.385)	(0.473)	(0.552)	(0.477)
Risk-averse × p>0.2 × FP costs		-0.323	0.002	-0.520	0.029
		(0.370)	(0.483)	(0.453)	(0.461)
Inconsistent × p>0.2 × FP costs		0.108	-0.210	-0.480	-0.372
		(0.681)	(0.464)	(0.523)	(0.473)
Full risk pref interactions	No	No	Yes	No	Yes
Observations	624	624	624	624	624
Adjusted R^2	0.08	0.07	0.07	0.42	0.42

Notes: Standard errors in parentheses (clustered by subject).

B Proofs

B.1 Proposition 1

Proof. If protection costs are low enough $c < \pi L$ than the 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:

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

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 Jensen's inequality:

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

Because expected utility with a signal is a decreasing function of b_0 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. We can only say that it is very likely that for low protection costs and small priors π (leading to no automatic blind protection) the ratio of sensitivities to FP rates over FN rates should be lower for risk-averse subjects.

B.3 Proposition 3

Proof. The proof is approximate and relies on Taylor expansion to measure the effect of risk aversion on sensitivities to false-positive and false-negative rates. Start by rewriting the equilibrium condition for willingness-to-pay as the expected sum of utility differences:

$$P(0,0)(u(Y-b)-u(Y))+p(0,1)(u(Y-b-L)-u(Y-L))+P(1,0)(u(Y-c-b)-u(Y))+ \\ +P(1,1)(u(Y-b-c)-u(Y-L))=0 \tag{7}$$

Here, $P(x,y)$ is a shorthand for the probability of an event that the signal equals x and the state equals y. Next, we expand the utility differences of $u(Y-b)-u(Y), u(Y-c-b)-u(Y)$ as Taylor series around Y and $u(Y-L-b)-u(Y-L)$ difference around $Y-L$ to get the following equation:

$$P(0,0)[u'(Y)(-b)+o(b)]+p(0,1)[u'(Y-L)(-b)+o(b)]+P(1,0)[u'(Y)(-c-b)+o(c+b)]+ \\ +P(1,1)[u(Y)-u'(Y)(b+c)+o(b+c)-u(Y-L)]=0 \tag{8}$$

Then we drop the terms $o(b), o(b+c)$ which we expect to be small enough to neglect to obtain:

$$P(0,0)u'(Y)b+P(0,1)(u'(Y)+[u'(Y-L)-u'(Y)])b+P(1,0)u'(Y)(c+b)+ \\ +P(1,1)(-u'(Y)(b+c)-(u(Y-L)-u(Y))=0 \tag{9}$$

Now we can express the equilibrium (approximate) WTP b as:

$$b=\frac{P(1,1)\frac{(u(Y)-u(Y-L))}{u'(Y)}-P(S=1)c}{D}$$

Where the denominator $D\equiv 1-P(0,1)\left(\frac{(u'(Y)-u'(Y-L))}{u'(Y)}\right)$. Now we remember that $P(1,1)\equiv\pi P_{11}=\pi(1-P_{01}), P(S=1)=\pi(1-P_{01})+(1-\pi)P_{10}$ and take derivatives of equilibrium (approximate) WTP b with respect to false-positive and false-negative rates:

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

$$\frac{db}{dP_{10}}=-\pi\left[\frac{\frac{(u(Y)-u(Y-L))}{u'(Y)}-c}{D}-\left(\frac{P(1,1)\frac{(u(Y)-u(Y-L))}{u'(Y)}-P(s=1)c}{D^2}\right)\frac{(u'(Y)-u'(Y-L))}{u'(Y)}\right]$$

For a strictly risk-averse subject the sensitivity to false-positive rates should be lower than for a risk-neutral one because $u'(Y)-u'(Y-L)<0$ by decreasing marginal utility leading to $D>1$. The opposite is true for strictly risk-loving subjects. It is hard to say something more specific about the sensitivity to false-negative rates.

Dividing the sensitivity to FN rate to the sensitivities of FP rate, we also obtain that this ratio is greater than 1 for strictly risk-averse subjects and less than one for strictly risk-loving ones.

$$\frac{db/dP_{01}}{db/dP_{10}} = \frac{\pi}{(1-\pi)}\left[\frac{(u(Y)-u(Y-L))}{u'(Y)} - c + \frac{(P(1,1)\frac{(u(Y)-u(Y-L))}{u'(Y)} - P(s=1)c)}{D}\frac{(u'(Y)-u'(Y-L))}{u'(Y)}\right]$$

Note that the corresponding equation for the risk-neutral decision-maker puts the ratio of sensitivities to:

$$\frac{db/dP_{01}}{db/dP_{10}} = \frac{\pi}{(1-\pi)}\left[L-c\right]$$

Hence the question of comparison of two ratios is equivalent to the question of the sign of the following inequality:

$$\frac{(u(Y)-u(Y-L))}{u'(Y)} + \frac{(P(1,1)\frac{(u(Y)-u(Y-L))}{u'(Y)} + P(s=1)c)}{D}\frac{(u'(Y-L)-u'(Y))}{u'(Y)} >< L$$

However note that the first component in the left-hand sum is already greater $\frac{(u(Y)-u(Y-L))}{u'(Y)} > L$ for any strictly risk-averse decision-maker by a mean value theorem. Risk aversion also makes the second component positive as $u'(Y-L)-u'(Y) < 0$ and $P(1,1)\frac{(u(Y)-u(Y-L))}{u'(Y)} + P(s=1)c > P(1,1)L - P(s=1)c > 0$ is also positive as it equal the expected savings from using a signal. Hence the LHS is greater than the RHS L leading to the ratio of sensitivities to be greater than for a risk-neutral decision-maker. The same argument applied in reverse will show that for a strict risk-loving decision-maker the ratio of sensitivities will be lower. □

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.

D Experiment Instructions (for online appendix)