Willingness to Pay for Signals of Rare Events

Arya Gaduh, Peter McGee, Alexander Ugarov*

March 21, 2024

Abstract

Designing multiple kinds of signals involves trade-offs between false-positive and false-negative costs. We conduct a laboratory experiment to evaluate preferences over these trade-offs in a controlled environment. We find that the choices significantly diverge from the predictions of the model with a risk-neutral decision-maker as well as from some predictions of expected utility frameworks. Relative to a risk neutral decision-maker, willingness-to-pay overreacts to false-negative rates for low priors, but underreacts for high priors. Subjects' preferences demonstrate a reverse bias for false-positive rates. This causes overpaying for signals with positive FP rates when the prior is low, and overpaying for all priors for low-quality signals with positive FP and FN rates. We find that this pattern is not consistent with the EU framework, but most consistent with a decision-making heuristic in which subjects do not differentiate between false-positive and false-negative rates when choosing signals.

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

1 Introduction

The 2010 gas blowout on Deep 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" (Brown, 2010). The United States Preventive Services Task Force has to periodically update its cancer screening guidelines as it weighs the costs from failing to detect cancer early against the potential harms from overdiagnosis or overtreatment due to false positive results. These are real-world examples of the trade-off from the two types of errors inherent to all probabilistic warning systems, namely false-positive and false-negative rates.

Most real-world warning systems — e.g., medical diagnostics, security alarms, or 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 trigger in an adverse state (false-negative rate) but increases the probability of incorrect trigger in a safe state (false-positive rate).

In order to understand preferences over these trade-offs, we study the demand for information in the framework with a potential protection action. The subject, first, receives a signal about the probability of an adverse event. Then she decides to protect or not. This environment describes several practically important scenarios including extreme weather alerts, medical testing and safety alarms.

We find that the value of information in our setup weakly correlates with the willingness-to-pay. First, subjects on average underreact to quality of the signal, resulting in overpaying for low-quality signal and underpaying for high-quaity signals. Second, subjects tend to overreact to false-negative rates when the prior probability is low and overreact to false-positive positive rates when priors are high. We show that this pattern is most consistent with failure to estimate the effect of frequencies of false-positive and false-negative outcomes on the costs of using the signal. (Xu, 2022) similarly finds that individuals do not properly account for priors and often choose tests not affecting optimal decisions even when more useful tests are available.

Our work is one of a few experimental studies measuring demand for information used for decision-making (instrumental information). Previous experimental studies studies the demand for signals in the prediction game in which subjects have to choose an optimal state under uncertainty. The field experiment conducted by (Hoffman, 2016) finds that the demand for information increases with initial uncertainty, but decreases with the signal's accuracy. However, the decrease in accuracy is more modest than expected for a Bayesian decision-maker resulting in subjects underpaying for high-quality signals. The laboratory experiment of Ambuehl and Li (2018) finds that subjects tend to underreact to the accuracy of the binary signal about state of the world, but put a premium on completely certain signals. The paper of Xu (2022) also employs a prediction game setup to measure information preferences but varies priors on

top of signal characteristics. She finds that many subjects choose non-instrumental over instrumental signals which is most consistent with failures of contingent reasoning on future value of information.

Our setup differs in two important aspects from (Ambuehl and Li, 2018; Xu, 2022), because we study alerts and not prediction tasks. The subject faces a costly protection decision and not a prediction decision, resulting in three distinct payoffs: full payoff, full payoff minus protection costs and full payoff minus losses. It means that risk preferences affect the value of information and can change sensitivities to false-positive and false-negative rates. Our findings however are similar to prediction game findings. Consistent with Ambuehl and Li (2018) we also find that subjects overvalue inaccurate signals, but we do not find a premium for certain signals. And similar to Xu (2022) we find that subjects commit reasoning errors leading to lower correlation between preferences and signal's usefulness in terms of cost reduction.

Due to its applicability for studying preferences over expectations, there is a larger stream of literature on the demand for non-instrumental information. Eliaz and Schotter (2010) find that subjects are willing to pay for signals even when these signals are excessive for making optimal choices. Their design involves subjects choosing between two boxes with one box containing a prize of \$20. Most subjects pay just to know the probability of finding \$20 in box A even if this box is more likely to contain a prize in all the possible states. This finding is inconsistent with expected utility maximization but indicates instead having preferences for certainty before making choices. Similar to this paper, Masatlioglu et al. (2017) also study preferences over information structures differing in false-positive and false-negative rates but their setup allows for a larger role of expectations. They find that for a positive potential outcome, most subjects prefer facing high false-negative rates rather than high false-positive rates. In other words, they tolerate uncertainty after negative signals better than uncertainty after positive signals. These preferences are salient: subjects require an average payment of 18-35 cents to switch to their least preferred information structure.

There is some mixed evidence that people update beliefs differently when these beliefs are ego-relevant or concern future gains and losses. Eil and Rao (2011) find asymmetry in updating ego-relevant beliefs such as beauty and IQ. Subjects update more after receiving positive signals and do not update enough after negative signals. Additionally, subjects with high posterior ego-relevant beliefs are willing to pay to receive a more precise signals, but require a compensation for learning when their beliefs are low. In contrast, Coutts (2019) does not find any updating asymmetry with respect to either ego-relevant beliefs or beliefs about future payoffs.

Our paper is the first to measure value of information in the experimental setting of diagnostic tests or alarms. Previous work studies the use of alarms in context of medical testing, medical monitoring, safety alarms and extreme weather. Early literature on decision-making of medical professionals finds that doctors suffer from multiple biases when ordering testing, including inaccurate posterior probability estimation due to availability heuristics, hindsight bias and regret (Bornstein and Emler, 2001). Gigerenzer et al. (2007) find that most mam-

mologists tend to overestimate the probability of cancer based on a positive result. Providing practitioners with natural frequencies instead of probabilities tends to reduce this bias.

Patients' willingness-to-pay for medical tests is large and sensitive to test accuracy (Liang et al., 2003; Howard and Salkeld, 2009; Neumann et al., 2012). But test preferences also exhibit several abnormalities. First, users are willing to pay for tests having little or zero diagnostic value (Schwartz et al., 2004; Neumann et al., 2012). For example, Schwartz et al. (2004) find that 73% of Americans in their survey prefer a free full-body CT scan versus one thousand USD cash. However, medical professional do not recommend full-body CT scans for healthy people due to extreme likelihood of false-positive findings. Second, the framing of test accuracy seems to matter a lot. Howard and Salkeld (2009) conduct a discrete-choice experiment to measure willingness-to-pay for the colorectal cancer screening. Their subjects agree to get 23 unnecessary colonoscopies in order to find one additional true cancer, but only 10.4 for reducing the number of cancers missed by one even though these descriptions are equivalent. Surprisingly, the perceived risk of cancer (prior) did not significantly affect the WTP in their study.

Our work also relates to the vast literature on demand for insurance and protection. Similar to our findings, several studies observe that the demand for insurance goes up after the recent experience with low probability events. Field evidence indicates that people under-insure with respect to rare natural disasters (Friedl et al., 2014). Laury et al. (2009) find no under-insurance for low-probability events in the laboratory setting. One offered explanation (Volkman-Wise, 2015) 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. It is consistent with underweighting prior probabilities relative to more recent signals.

The bias we are finding is similar to the base-rate and signal neglect phenomenons. Psychology researchers Hammerton (1973) and Kahneman and Tversky (1973) first observed that subjects underweighted prior probabilities (base rates) when calculating posteriors. This phenomenon had received the name of *base-rate neglect*. Multiple studies in economics then confirmed (Grether, 1992; Holt and Smith, 2009) this phenomenon in incentivized laboratory experiments. Most of these studies find that subjects also underweight signals on top of priors. We observe both phenomenons in responses to our belief elicitation task, but the calculation of signals' values differs substantially from the calculation of posterior probabilities. While the calculation of posterior probabilities would require using a Bayes formula, signal's value depends only on products of prior probabilities. However, we observe that subjects underestimate the effect of priors compared to theoretical predictions for an expected-utility decision-maker.

As we use use a strategic approach to elicit both beliefs and hypothetical choices, our subjects has to participate in contingent reasoning. Aina et al. (2023) recently find that contingent reasoning increases bias in belief elicitation. Belief biases and protection decisions can be reduces by eliciting responses after presenting a signal. Decisions to acquire information however fundamentally rely on thinking through contingencies.

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 only realized when $\omega(1-a)=1$.

The agent's preferences are described by a utility function which depends on income Y, protective action a, and the protective outcome $\omega(1-a)$. Taking the protective action costs $c>0$. Utility is separable in wealth, protection cost $c>0$ 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 considers a purchase of a testing instrument (hereafter, a tester) that produces 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 receiving the signal, 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 tester, the agent protects if and only if it increases her expected utility:

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

The tester can increase expected utility if its signal informs her posterior. 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 as b the agent's willingness to pay for the tester, to wit, she is indifferent between purchasing it at price b and not having its signal. Its 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 tester because the payoff distribution with the signal first-order stochastically dominates the distribution without the signal. However, determining the value of an imperfect tester 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 tester'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 tester, 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:

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

When $b > 0$, its with respect to FP (P_{10}) and FN (P_{01}) rates is given by:

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

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

The tester'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 Effects. In an expected utility framework, risk aversion can both increase and decrease an agent's valuation of the tester. More specifically, risk aversion decreases her WTP when protection costs are low:

Proposition 1. *If protection costs are low $c < \pi L$, then a strictly risk-averse decision-maker pays less than a risk-neutral one.*

Proof. See the Appendix. □

Things are more ambiguous when risks are low or protection costs are high. For example, risk aversion increases the value of a perfect tester as long as a risk-averse decision-maker still

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 tester is isomorphic to the insurance problem with deductible c.

Next, we study the effect of a tester'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 and false-negative rates:

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

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

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

Proof. See the Appendix. □

However, risk aversion can both increase and decrease subject's sensitivity to FP rates depending on the utility function's curvature and the signal's characteristics. Intuitively, an expected marginal utility of a strongly risk-averse subject with a bad tester 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 tester 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.

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

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. *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. Moreover, we show above that the relative weight of false-negative costs can be either below or above one depending only on risk preferences.

3 Experimental Design

We conduct 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, an outcome which was determined by a series of 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 answers 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 is randomly selected as the payment round. The instructions can be found in the appendix.

Blind Protection (BP). Subjects must decide whether to protect against an adverse event: a random draw of a black ball. Subjects know the prior probability that a black ball is drawn. Protection costs $5. 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 realization of the decision.

Informed Protection (IP). Similar to the BP task, subjects must make a protection decision given the prior probability of drawing a black ball. Subjects receive a prior and a signal produced by a tester with varying degrees of inaccuracy. Following Coutts (2019), we use a group of hinting gremlins to convey tester accuracy: a gremlin, randomly drawn from a group, gives out the signal. 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 tester 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 the hint. We vary the proportion of prior probability of drawing a black ball and the composition of gremlins across rounds.

[1] Incorrect answers in quiz for the Informed Protection section results in subjects facing additional questions. In our opinion, clear understanding of the Informed Protection task is essential for subsequent tasks, hence the added requirement. These questions consist of XXX; complete details 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 composition of the group of gremlin 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 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.[2] Motivated by Danz et al. (2020), who find that providing a detailed explanation of payoffs can lower trustful reporting, we instead explain that reporting true belief μ maximizes their payoffs, and give an example 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 composition of the group of gremlin providing 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 exceeded 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.[3] The payment task and the payment round were then randomly chosen to calculate the subject's payoff.

[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 (Karni, 2009). The only requirements a subject must satisfy are probabilistic sophistication and dominance: they rank lotteries based on their probabilities only and prefer higher probabilities of higher payoffs.

[3]These were the questions on subjects' gender, age, and experience of taking statistics classes.

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 signals starting from the honest one for each prior. The order of priors and signals stays constant for each subjects 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 reveals 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 generally understand these tasks reasonably well.

4.1 Blind Protection

Figure 2 plots the likelihood of choosing to protect against the posterior probability of a drawing a black ball for the BP task, where the posterior is equivalent to the prior, and in the IP task. On aggregate in the BP task, subjects' likelihood of protecting increases in the probability of a negative 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. Among these switchers, however, 83% (24/39) skip only a single increment of the presented probability scale, suggesting an inattention error.[4]

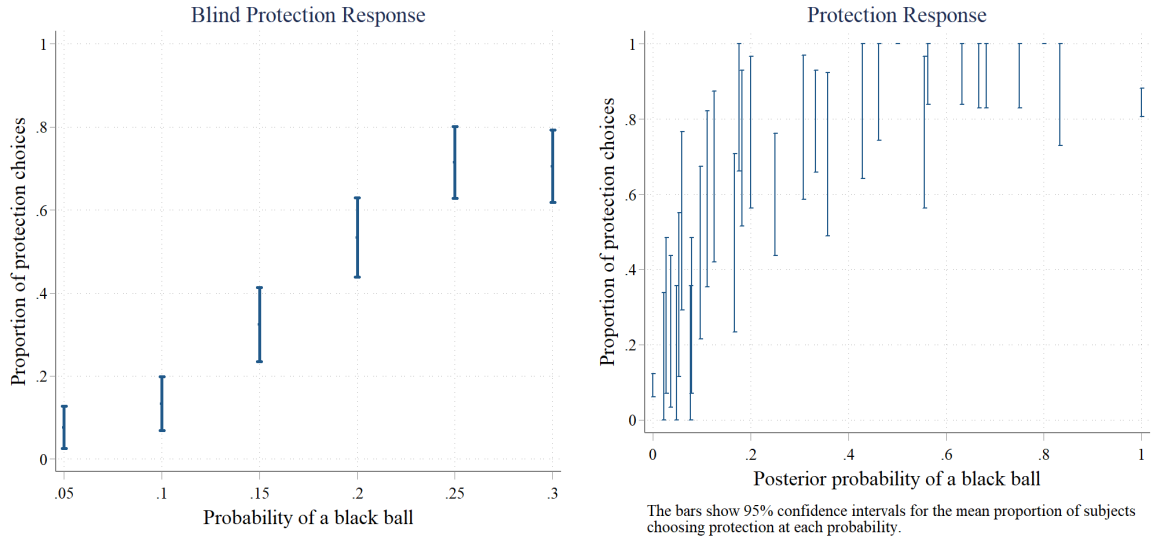
[4]For comparison, Holt and Laury (2002) for a similar instrument find that 28 of 212 subjects (13%) switched back to a low-risk option with an increasing likelihood of high payoffs in a risky option at least once when decisions were presented in increasing order, which they are not here.

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 start protecting at lower priors, indicating strict risk aversion.[5] A smaller group of subjects makes choices consistent with risk loving by protecting at a probability of 0.3 or never protecting.

4.2 Informed Protection

Recall that, in the IP task, subjects receive a signal about the color of the ball in addition to the prior. Figure 2 shows that protection actions are increasing in the posteriors, though roughly 28% of subjects break monotonicity in their protection responses with respect to posterior probabilities — approximately the percentage of non-monotonic responses in the BP task.[6] At the individual level, we also find that the total number of times subjects protect in the BP task significantly correlates with their likelihood to protect in the IP task conditional on posteriors, but this explains only a very small part (<1%) of variation in the IP decisions.[7]

Figure 2: Average Protection Response



(a) The bars show 95% confidence intervals for the mean proportion of subjects choosing protection at each prior probability.

(b) 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 signal type and tester characteristics. The first three columns summarize the tester accuracy information by the signal produced. Column 4 shows the posterior probability of a black ball averaged across all the treatments

[5] As a reference, switching at the probability 0.1 corresponds to a CRRA risk aversion $\theta = 2$, while switching at 0.2 corresponds to $\theta = 0.57$.

[6] That is, subjects do not protect for some treatments with posterior probability P while protecting for a posterior probability $P' < P$.

[7] 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.

within a group. Column 5 shows the subjects' share of empirical protection responses, next to the theoretical optimum for risk-neutral subjects in Column 6. Column 7 presents the p-value for a test of equality between empirical and theoretical protection responses.

We make three notable observations. First, regardless of the tester'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 tend to overprotect when facing white signals (rows 1–4). Subjects underprotect when facing black signals, except if the signals were produced by a tester with positive FP rates (rows 5–8).

Third, we find that some deviations 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 tester's FP rate does not change the posterior (because the signal is white), the protection rate increases by 6 percentage points (pp). Similarly, row 4 shows that when both FP and FN are positive, the protection rate increases to 56 percent — even though the average posterior probability given the tester 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.

Table 2: Average Protection by Signal Type

Row	**Tester Characteristics**		**Signal**	**Posterior**	**Share Protect**	**Share Optimal**	**P-val** (H_0 : (5) = (6))
	False Positive	**False Negative**					
	(1)	(2)	(3)	(4)	(5)	(6)	(7)
(1)	No	No	White	0.000	0.067	0.000	0.000
(2)	No	Yes	White	0.100	0.333	0.000	0.000
(3)	Yes	No	White	0.000	0.130	0.000	0.000
(4)	Yes	Yes	White	0.131	0.564	0.121	0.000
(5)	No	No	Black	1.000	0.846	1.000	0.000
(6)	No	Yes	Black	1.000	0.841	1.000	0.000
(7)	Yes	No	Black	0.550	0.833	0.870	0.355
(8)	Yes	Yes	Black	0.483	0.886	0.871	0.685

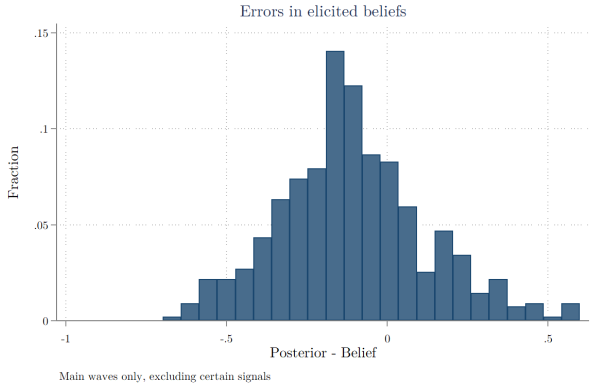
Notes:

4.3 Belief Elicitation

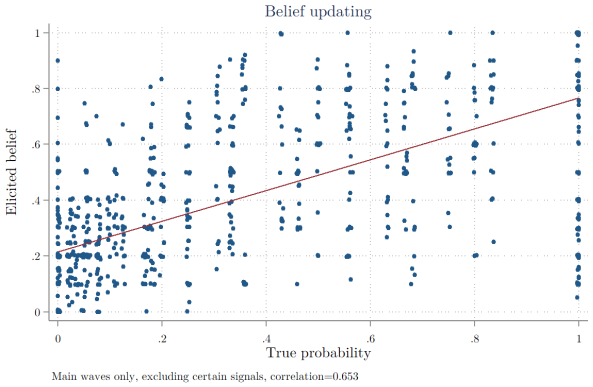
Subject decisions in the IP task capture the use of signals in protection decisions, but decisions reflect but risk preferences and (potentially erroneous) beliefs. The BP task can be used to construct a measure of the former; the BE task to measure the latter.

Figure 3: Errors in Bayesian Updating

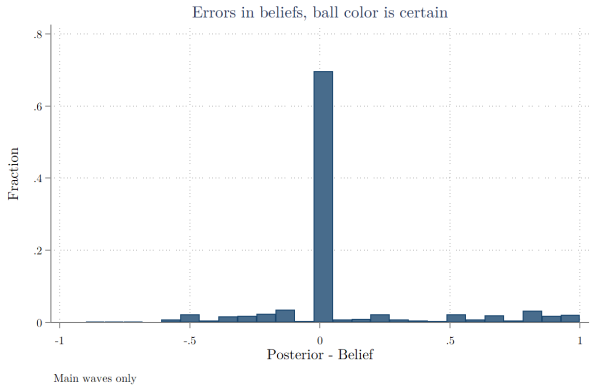
(a) Error Distribution



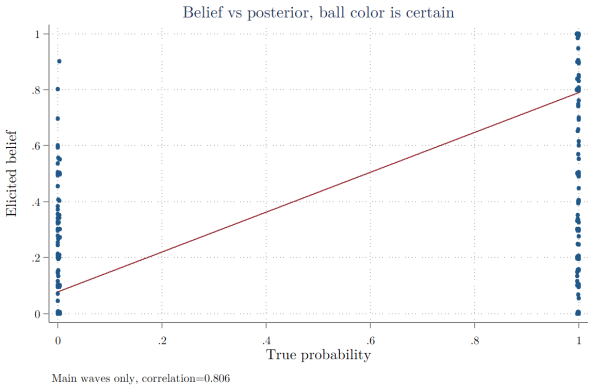
(b) Error v. Posterior



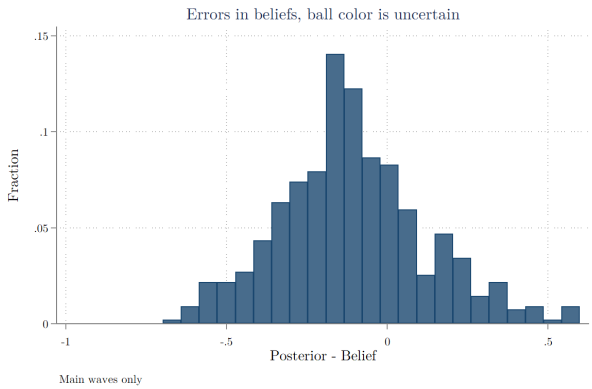
(c) Error Distribution, Certain Posterior



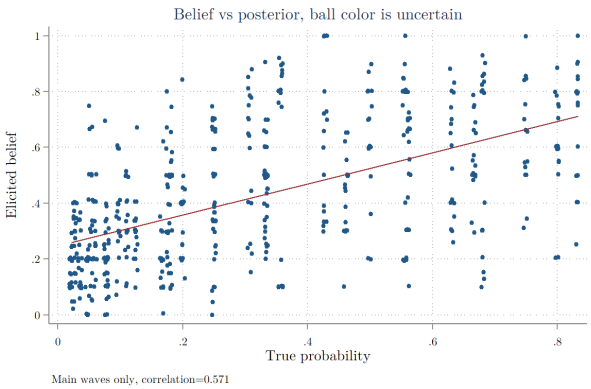
(d) Error v. Posterior, Certain Posterior



(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 uses all observations and suggests that, while errors occur, beliefs are still sensible. The distribution

of updating errors is centered at 0, with roughly one-half (51%) concentrated within +/- 0.1 interval around zero. Overall, the correlation between the elicited beliefs and the true posteriors was 0.653.

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 gremlins comprising honest and white-swamp gremlins 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., with uncertain posteriors). The median error in Panel C is -0.12, with 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.[8]

Table 3: Average Updating Error by Signal Type

Row	**Tester Characteristics**		**Signal**	**Posterior**	**Updating Error***	**P-val** (H_0 : $Error = 0$)
	False Positive	**False Negative**				
	(1)	(2)	(3)	(4)	(5)	
(1)	No	No	White	0.000	0.050	0.000
(2)	No	Yes	White	0.100	0.122	0.000
(3)	Yes	No	White	0.000	0.122	0.000
(4)	Yes	Yes	White	0.131	0.218	0.000
(5)	No	No	Black	1.000	-0.163	0.000
(6)	No	Yes	Black	1.000	-0.279	0.000
(7)	Yes	No	Black	0.550	0.039	0.130
(8)	Yes	Yes	Black	0.483	0.048	0.021

Notes: *Updating error = $Belief - Posterior$.

Table 3 summarizes how updating errors vary with tester characteristics. We find that subjects overestimate the probability of a black ball when given a white signal. This upward bias for a white signal increases in the FP/FN rate of the tester. To illustrate, consider the change between rows 1 and 3, where introducing a FP rate would not change the posterior

[8]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 9). 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

because the signal 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 v. 2).

The updating bias for black signals, however, varies by the information structure. Subjects slightly underestimate the probability with a perfectly accurate tester, 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' belief. With non-zero FP rate, subjects again overestimate the probability of a black ball. The difference in the updating errors for black signals coming from FP-only (row 7) v. FP-FN testers (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?

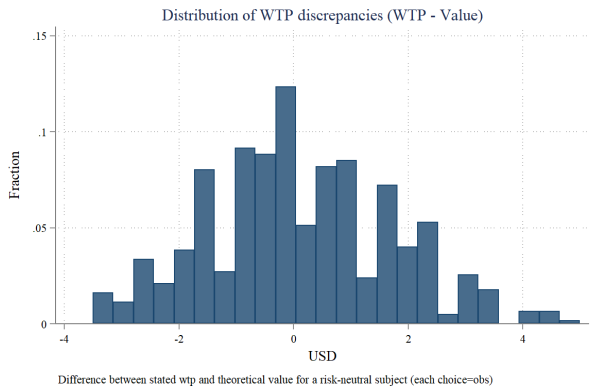
Hypothesis 1. *Subjects' WTPs for signals are equal to their value for risk-neutral agents.*

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

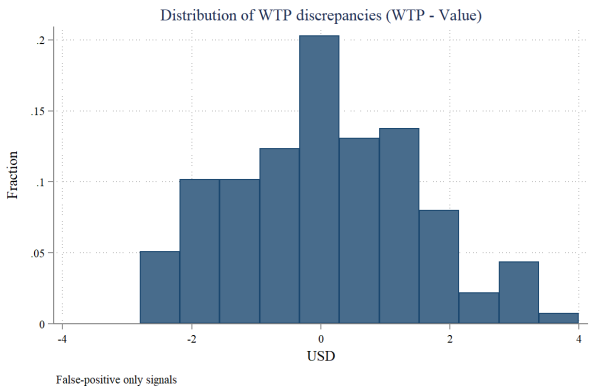
Overall, the theoretical value of a tester for a utility maximizing risk-neutral subject (hereafter, the risk-neutral WTP) in equation 2 is a useful benchmark of our subjects' WTP. Figure 4 plots the distribution of the differences between subjects' WTP and this value. The WTP is centered around the risk-neutral WTP, 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 that shift the average upward.

Figure 4: Discrepancy (Observed WTP - Signal value) by Signal Type

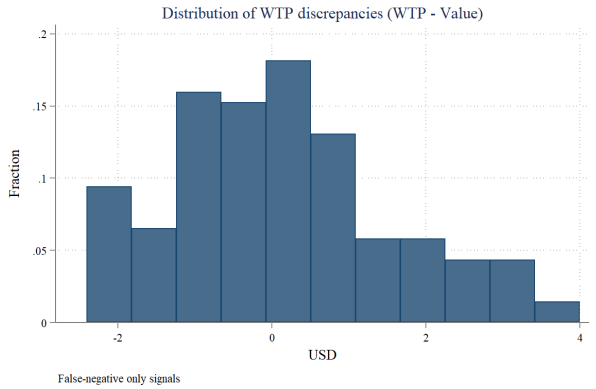
(a) All signals



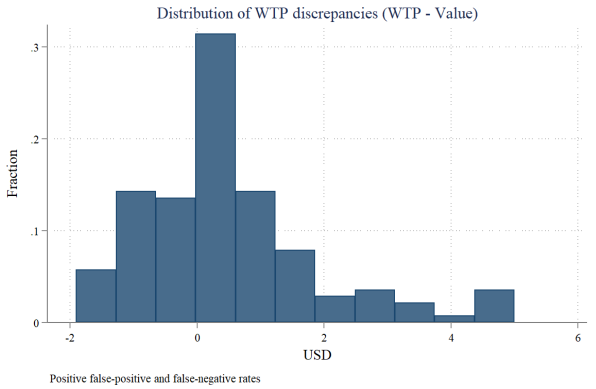
(b) FP only



(c) FN only



(d) Both FP and FN



Our non-parametric analysis in Table 4 also finds no differences on average between the observed WTP and the risk-neutral WTP for 3 out of 4 tester categories: honest (i.e., perfectly accurate), FP-only, and FN-only. With both FP and FN rates, however, subjects' WTPs are significantly higher than the risk-neutral WTP. Subjects overvaluations were similar for both low and high priors. Note, that these tester characteristics induce overprotection in the IP task. Subjects tend to overpay for testers with positive FP rates when the prior is low ($\in \{0.1, 0.2\}$), and for testers with positive FN rates when the prior is high ($\in \{0.3, 0.5\}$).

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

Priors	**Honest**	**FN only**	**FP only**	**FP and FN**
All priors	-0.106	0.143	0.081	0.492***
Low priors	-0.135	-0.209	0.465**	0.437**
High priors (>0.2)	-0.077	0.496*	-0.303	0.547**

*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.*

Next, we examine how the WTP responds to tester 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. All specifications include subject fixed effects, with standard errors clustered at the subject level. If subjects are risk-neutral expected-utility-maximizers, we expect $\beta_1 = 0$ and $\beta_2 = 0$. The result, reported in column 1 of Table 5, shows positive and statistically significant coefficients for both FP and FN costs. In other words, subjects deviate by overpaying for inaccurate testers.

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 indicating higher sensitivity to FP costs, but we cannot reject the hypothesis that the two coefficients are equal. However later we note that 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 model's predictions. Positive and significant coefficients on FP and FN costs indicate that subjects reduce their WTP with growing FP and FN rates by less than a risk-neutral decision-maker would, i.e., subjects' WTP did not adjust enough to decreasing tester quality.

As our benchmark model assumes both perfect updating and risk neutrality, assumptions which open two channels through which deviations could occur. First, Proposition 2 suggests that risk preferences can influence the sensitivity of WTP to these tester characteristics. Second, systematic biases during updating can also lead to deviations.

We find that risk preferences matter for sensitivity to tester quality. 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.[9] 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 WTPs of both risk-neutral and risk-loving subjects increase with FP/FN costs — suggesting that they did not downward-adjust

[9]We classify subjects based on the total number of protection choices made in the BP task with 2 or 3 choices corresponding to risk-neutrality (protecting starting from 0.2 or 0.25), but exclude subjects making more than one choice at odds with a consistent risk preference.

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

	All			Prior	
				$\{.1, .2\}$	$\{.3, .5\}$
	(1)	(2)	(3)	(4)	(5)
FP costs	0.231	0.316	0.615	0.800	0.204
	(0.126)*	(0.195)	(0.252)**	(0.239)***	(0.488)
FN costs	0.319	0.425	0.426	0.150	0.407
	(0.070)***	(0.118)***	(0.127)***	(0.279)	(0.106)***
Risk-averse × FP costs		-0.297	-0.429	-0.491	-0.707
		(0.291)	(0.352)	(0.385)	(0.679)
Risk-averse × FN costs		-0.410	-0.367	-0.321	-0.264
		(0.174)**	(0.175)**	(0.343)	(0.151)*
Risk-loving × FP costs		0.053	-0.084	-0.431	0.253
		(0.339)	(0.413)	(0.449)	(0.718)
Risk-loving × FN costs		-0.016	0.027	0.083	0.035
		(0.166)	(0.192)	(0.387)	(0.142)
Constant	-0.182	-0.191	-0.443	-0.219	-0.332
	(0.083)**	(0.080)**	(0.135)***	(0.179)	(0.191)*
R^2	0.480	0.492	0.504	0.739	0.753
Obs	624	624	624	312	312
Risk-Averse Subjects:					
False Positive		0.019	0.186	0.309	-0.503
se		(0.216)	(0.246)	(0.302)	(0.472)
p-value		[0.928]	[0.451]	[0.309]	[0.289]
False Negative		0.014	0.060	-0.171	0.143
se		(0.127)	(0.120)	(0.200)	(0.108)
p-value		[0.910]	[0.621]	[0.393]	[0.189]
Risk-Loving Subjects:					
False Positive		0.369	0.531	0.369	0.457
se		(0.277)	(0.328)	(0.381)	(0.526)
p-value		[0.186]	[0.109]	[0.334]	[0.388]
False Negative		0.409	0.453	0.232	0.442
se		(0.117)	(0.143)	(0.267)	(0.094)
p-value		[0.001]	[0.002]	[0.387]	[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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

their WTP enough to account for lower quality testers. In contrast, the WTPs of risk-averse subjects show hardly any sensitivity to FP and FN costs.

Accounting both for belief accuracy and risk preferences does little to explain the pattern of underreacting to FP and FN rates. We use data from the BE task to construct a measure of subjects' belief accuracy.[10] 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 did not downward-adjust their WTP enough to increasing FP/FN costs.[11]

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 prior by below/above 0.25 (= protection cost/potential loss). We incorporate prior-probability fixed effects to the aforementioned flexible specification.

Column 4 of Table 5 presents the results for low-prior WTPE tasks. With low priors, deviations from the risk-neutral WTP increase with FP costs: subjects overvalue testers that would induce them to overprotect. This overvaluation is similar for different risk preference profiles. It should be noted that while coefficients' magnitudes are relatively large, none of the coefficients or predicted sensitivities here (bottom panel) is significant due to relatively small sample size, so these results should be interpreted with caution.

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. These subjects did not downward-adjust their WTP enough to account for increasing FN rates and overvalue testers that would induce them to underprotect.

To sum up, most subjects underreact to false-positive costs with low priors and underreact to false-negative costs for high priors. In practice, and given low priors implied by many alert systems, it means that users would tend to overpay for alert signals with high false-positive costs, while excessively discounting signals with significant false-negative rates. For example,

[10] 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.

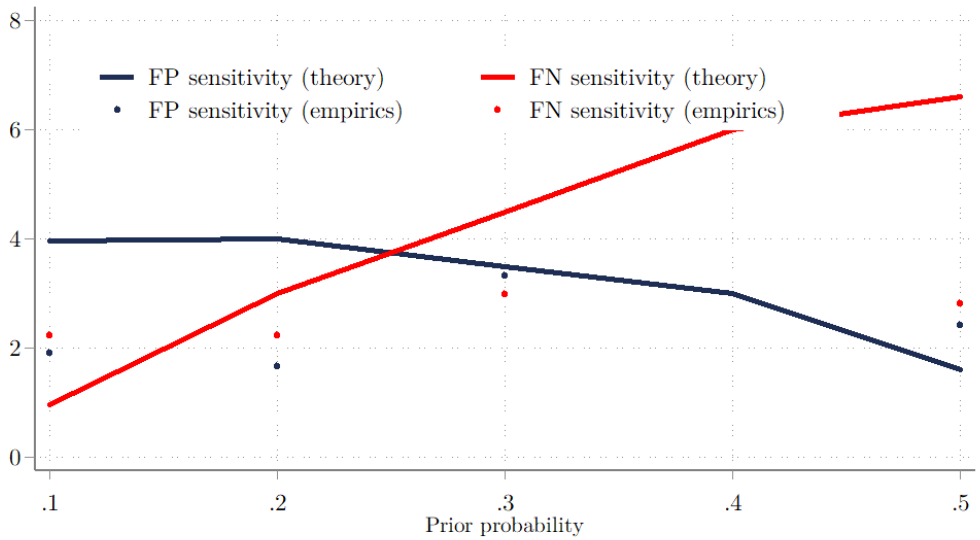
[11] Aside from these theoretically motivated individual differences, we investigate several other characteristics. Heterogeneity is not driven by demographic characteristics (e.g., age, gender) or prior statistical coursework. These results are in Appendix A Table A.

they would prefer a smoke alarm which never misses fires even if it involves higher expected costs of false alarms. Risk preferences seem to affect this pattern with risk-averse subjects moving closer to a risk-neutral benchmark, but most interaction coefficients are not statistically significant despite high 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 suggest that WTP should respond more to FP rates (relative to FN rates) for low priors and vice versa for high priors(XXX Alex, this is correct, right? XXX). Intuitively, for a given FN rate, false-negative events are much less likely with low priors and hence impose less costs on the agent. As priors increase, FN rates become more salient while FP rates become less salient. 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. XXX Hence the puzzle can be reframed as the uniformity of WTP response to FP and FN rates for different priors. XXX I THINK THE FLOW TO THIS LAST SENTENCE ISN'T QUITE THERE. SOMETHING IS MISSING XXX

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. Two sensitivities are also surprisingly close to each other.[12]

[12]For none of the priors can we reject the hypothesis that two sensitivities are equal to each other.

We consider four candidate explanations. First, risk preference. Second, anchoring bias. Third, bias from valuation of non-instrumental information (Eliaz and Schotter, 2010; Masatlioglu et al., 2017). Finally, subjects may fail to distinguish how FP and FN error rates should affect how they calculate their posteriors differently.

Our evidence suggests that risk preference cannot explain this behavior. We test the risk preference hypothesis using subjects' BP choices. Columns 4–5 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 Table **??** 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. (XXX ALEX: Which table is this? XXXX)

The evidence also does not support the anchoring hypothesis, to wit, that subjects anchored on previous priors. Each subject goes through two sets of treatments with two different priors and a fixed order of priors, so anchoring could be possible. 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 increase subjects' confusion. Most importantly, the uniformity in coefficient ratio is present even if we limit our attention only to the first priors in each sequence.[13]

There is evidence in the literature of people valuing "non-instrumental information" that does not affect their decisions. (XXX ALEX: MAYBE ILLUSTRATE OF WHAT THAT MEANS XXX For example, Eliaz and Schotter (2010) finds that XXXX while Masatlioglu et al. (2017)....) Most information in our experiment is instrumental by design (it helps to choose actions) and indeed enters into subjects' decisions as evidenced by choices in the IP task. Nonetheless, we find many subjects choosing positive WTP for signals that cannot affect their IP decisions (159 out of 624 total choices). It is therefore plausible that the reported WTPs includes some non-instrumental components.

However, we think that preferences for non-instrumental information cannot provide a full explanation of our results. First, the sensitivity of WTP to FP rates is much lower in the experiment compared to the theory when priors are low. Suppose that the information value $b = b(\pi, P_{01}, P_{10}) + n(\pi, P_{01}, P_{10})$ with $n(.)$ describing the non-instrumental component. If the discrepancy in sensitivities to FP rates (XXX between what and what? XXX) comes from the non-instrumental component n, it needs to be increasing in FP rates. In other words, subjects would have been putting higher non-instrumental values on worse testers — which seems implausible. Second, the closeness of coefficients for FP and FN rates seems also apriori

[13]Depending on session, the first 3 WTP treatments use either 0.1 or 0.2 as the prior, so there is no anchoring on the previous prior or something special about a particular prior.

implausible based only on the non-instrumental information value story.

Instead, we argue that subjects' observed behaviors arise from confusing FN and FP rates. We use as evidence subjects' own proffered explanation. At the end of the experiment, we asked subjects to explain to us how they made their protection choices. Out of 105 subjects in the main waves of the experiment, 39 refer to the *percentage* of dishonest gremlins as their rationale for choosing protection.[14] For example:

- "*I took into consideration how many honest there were and looked at the chances of picking a ball.*"
- "*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. The closeness of the coefficient estimates for FP and FN rates in Table **??** are certainly 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. Note also that 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.

In order to test this hypothesis, we use choices from the BE and IP tasks where subjects also 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 (positive) signals. This happens because some subjects 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.

In Table 6 we estimate a linear regression of updating error (actual posterior - reported belief) on FP and FN rates by signal color. We use fixed effects to control for individual updating biases. Consistent with our conjecture, we observe that the FP rate has a significant positive effect on the error when the signal is white (negative), and that FN rate has a significant negative effect when the signal is black (positive).

In Table 7, we regress IP decisions on FP and FN rates and flexible controls of both posteriors and reported beliefs:[15]

[14] IS IT POSSIBLE TO PUT A TABLE IN THE APPENDIX WITH THESE STATEMENTS, WITH THOSE IN YOUR 39 HIGHLIGHTED?

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

Table 6: Updating Errors in BE Task

	All	Signal Received	
		White	Black
	(1)	(2)	(3)
FP rate	.6***	.292***	.908***
	(0.1)	(0.1)	(0.1)
FN rate	.0108	.273***	-.251***
	(0.1)	(0.1)	(0.1)
Constant	-.0784***	.314***	-.47***
	(0.0)	(0.0)	(0.0)
Subject FE	Yes	Yes	Yes
Observations	1248	624	624
Adjusted R^2	0.15	0.41	0.52
Subject FE	Yes	Yes	Yes

Notes: Standard errors in parentheses. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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

where 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/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.

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 signal, FP and FN rates increase the tendency to overprotect while the FP rate had an opposite effect with comparable magnitude but without statistical significance for a black signal. Hence the first prediction of a conjecture of indiscriminate FP/FN rate use holds: FP rates increase protection when the signal 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 signals 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 signals (Columns 3 and 4), but the coefficients still remains significant. This indicates that while beliefs partially contribute to these protection anomalies,

[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.

they cannot explain them completely.

Table 7: Informed Protection Response

	(1)	(2)	(3)	(4)
FP rate x (S=White)	.461***	.494**	.282**	.286
	(3.3)	(2.4)	(2.0)	(1.4)
FN rate x (S=White)	.544***	.474**	.195	.125
	(2.9)	(2.1)	(1.0)	(0.5)
S=Black	.42***	.429***	.316**	.336**
	(2.7)	(2.7)	(2.0)	(2.1)
FP rate x (S=Black)	-.256	-.225	-.379	-.389
	(-0.5)	(-0.4)	(-0.8)	(-0.7)
FN rate x (S=Black)	.0494	-.027	-.00394	-.0879
	(0.5)	(-0.2)	(-0.0)	(-0.6)
p=0.2	.113***	.101***	.09***	.0723**
	(4.2)	(2.8)	(3.6)	(2.1)
FP rate x (p=0.2)		-.0363		.00218
		(-0.2)		(0.0)
FN rate x (p=0.2)		.122		.127
		(0.9)		(0.9)
N	1224	1224	1224	1224
Pseudo R-squared	.551	.552	.578	.578
Log-likelihood	-379	-378	-356	-356
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. t-statistics in parentheses. Standard errors are clustered at the subject level. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

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 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.

7 Conclusion

References

Aina, Chiara, Andrea Amelio, and Katharina Brütt (2023) "Contingent Belief Updating," *ECONtribute Discussion Papers Series*, https://ideas.repec.org//p/ajk/ajkdps/263.html, Number: 263 Publisher: University of Bonn and University of Cologne, Germany.

Ambuehl, Sandro and Shengwu Li (2018) "Belief updating and the demand for information," *Games and Economic Behavior*, 109, 21–39, 10.1016/j.geb.2017.11.009.

Benjamin, Daniel J. (2019) "Chapter 2 - Errors in probabilistic reasoning and judgment biases," in Bernheim, B. Douglas, Stefano DellaVigna, and David Laibson eds. *Handbook of Behavioral Economics: Applications and Foundations 1*, 2 of Handbook of Behavioral Economics - Foundations and Applications 2, 69–186: North-Holland, 10.1016/bs.hesbe.2018.11.002.

Bornstein, B. H. and A. C. Emler (2001) "Rationality in medical decision making: a review of the literature on doctors' decision-making biases," *Journal of Evaluation in Clinical Practice*, 7 (2), 97–107, 10.1046/j.1365-2753.2001.00284.x, Number: 2.

Brown, Robbie (2010) "Oil Rig's Siren Was Kept Silent, Technician Says," *New York Times*, 1, https://www.nytimes.com/2010/07/24/us/24hearings.html.

Coutts, Alexander (2019) "Good news and bad news are still news: experimental evidence on belief updating," *Experimental Economics*, 22 (2), 369–395, 10.1007/s10683-018-9572-5, Number: 2.

Danz, David, Lise Vesterlund, and Alistair J. Wilson (2020) "Belief Elicitation: Limiting Truth Telling with Information on Incentives," June, 10.3386/w27327.

Eil, David and Justin M. Rao (2011) "The Good News-Bad News Effect: Asymmetric Processing of Objective Information about Yourself," *American Economic Journal: Microeconomics*, 3 (2), 114–138, 10.1257/mic.3.2.114, Number: 2.

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, 10.1016/j.geb.2010.01.006, Number: 2.

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, 10.1007/s11166-014-9189-9.

Gigerenzer, Gerd, Wolfgang Gaissmaier, Elke Kurz-Milcke, Lisa M. Schwartz, and Steven Woloshin (2007) "Helping Doctors and Patients Make Sense of Health Statistics," *Psychological Science in the Public Interest: A Journal of the American Psychological Society*, 8 (2), 53–96, 10.1111/j.1539-6053.2008.00033.x, Number: 2.

Grether, David M. (1980) "Bayes Rule as a Descriptive Model: The Representativeness Heuristic," *The Quarterly Journal of Economics*, 95 (3), 537–557, 10.2307/1885092, Publisher: Oxford University Press.

——— (1992) "Testing bayes rule and the representativeness heuristic: Some experimental evidence," *Journal of Economic Behavior & Organization*, 17 (1), 31–57, 10.1016/0167-2681(92)90078-P, Number: 1.

Hammerton, M. (1973) "A case of radical probability estimation," *Journal of Experimental Psychology*, 101 (2), 252–254, 10.1037/h0035224, Number: 2 Place: US Publisher: American Psychological Association.

Hoffman, Mitchell (2016) "How is Information Valued? Evidence from Framed Field Experiments," *The Economic Journal*, 126 (595), 1884–1911, 10.1111/ecoj.12401, Number: 595.

Holt, Charles A. and Angela M. Smith (2009) "An update on Bayesian updating," *Journal of Economic Behavior & Organization*, 69 (2), 125–134, 10.1016/j.jebo.2007.08.013, Number: 2.

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, 10.1111/j.1524-4733.2008.00417.x, Number: 2.

Kahneman, Daniel and Amos Tversky (1972) "Subjective probability: A judgment of representativeness," *Cognitive Psychology*, 3 (3), 430–454, 10.1016/0010-0285(72)90016-3.

——— (1973) "On the psychology of prediction," *Psychological Review*, 80 (4), 237–251, 10.1037/h0034747, Number: 4 Place: US Publisher: American Psychological Association.

Karni, Edi (2009) "A Mechanism for Eliciting Probabilities," *Econometrica*, 77 (2), 603–606, `https://ideas.repec.org/a/ecm/emetrp/v77y2009i2p603-606.html`, Number: 2 Publisher: Econometric Society.

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, 10.1007/s11166-009-9072-2, Number: 1.

Liang, Wenchi, William F. Lawrence, Caroline B. Burnett, Yi-Ting Hwang, Matthew Freedman, Bruce J. Trock, Jeanne S. Mandelblatt, and Marc E. Lippman (2003) "Acceptability of diagnostic tests for breast cancer," *Breast Cancer Research and Treatment*, 79 (2), 199–206, 10.1023/a:1023914612152, Number: 2.

Masatlioglu, Yusufcan, A. Yesim Orhun, and Collin Raymond (2017) "Intrinsic Information Preferences and Skewness," September, 10.2139/ssrn.3232350, Issue: 3232350.

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, 10.1002/hec.1704, Number: 3.

Phillips, Lawrence D. and Ward Edwards (1966) "Conservatism in a Simple Probability Inference Task," *Journal of Experimental Psychology*, 72 (3), 346, 10.1037/h0023653.

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, 10.1001/jama.291.1.71, Number: 1.

Tversky, Amos and Daniel Kahneman (1971) "Belief in the law of small numbers," *Psychological Bulletin*, 76, 105–110, 10.1037/h0031322, Place: US Publisher: American Psychological Association.

Volkman-Wise, Jacqueline (2015) "Representativeness and managing catastrophe risk," *Journal of Risk and Uncertainty*, 51 (3), 267–290, 10.1007/s11166-015-9230-7, Number: 3.

Xu, Yan (2022) "Revealed Preferences Over Experts and Quacks and Failures of Contingent Reasoning," September, 10.2139/ssrn.4560390.

A Tables

Table 8: 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 9: Error Decomposition

	(1) OLS	(2) FE	(3) OLS	(4) FE	(5) OLS	(6) FE
Prior	.246***	.202***	.175***	.191**	.14**	.0403
	(5.5)	(4.0)	(3.1)	(2.5)	(2.3)	(0.6)
Signal	.43***	.43***	.327***	.327***	.539***	.539***
	(6.3)	(6.3)	(3.2)	(3.2)	(5.3)	(5.3)
Good quiz × Prior			.143*	.0207		
			(1.7)	(0.2)		
Good quiz × Signal			.193	.193		
			(1.4)	(1.4)		
Stat. class × Prior					.162*	.264***
					(1.9)	(2.8)
Stat. class × Signal					-.166	-.166
					(-1.2)	(-1.2)
Observations	280	280	280	280	280	280
Adjusted R^2	0.31	0.31	0.33	0.32	0.32	0.32

Decomposition works only for imperfect signals

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

Table 10: Informed Protection Response: logit with flexible control for posteriors

	(1)	(2)	(3)	(4)
FP rate	.365***	.472***	.593***	.573***
	(3.3)	(3.4)	(4.0)	(3.7)
FN rate	.168*	.611***	.15	.565**
	(1.8)	(2.8)	(1.5)	(2.5)
p>0.2	.0259	.0664***	.0471*	.0547*
	(1.5)	(2.8)	(1.8)	(2.0)
S=Black	.00422	.426**	-.0229	.473**
	(0.1)	(2.5)	(-0.3)	(2.4)
FP rate x (S=Black)		-.655		-.69
		(-1.4)		(-1.5)
FN rate x (S=Black)		-.561**		-.608**
		(-2.1)		(-2.2)
FP rate x (p>0.2)			-.293**	-.16
			(-2.3)	(-1.2)
FN rate x (p>0.2)			.0843	.264
			(0.5)	(1.6)
Observations	1248	1224	1224	1224
Adjusted R^2				

t statistics in parentheses

Reporting average marginal effects, subject FE, errors are clustered by subject.

With flexible controls of posterior probability

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

Table 11: WTP minus Value of Information: demographic determinants

	(1)	(2)	(3)	(4)	(5)	(6)
FP costs	.283	.352*	.117	.215	.248*	.291**
	(0.2)	(0.2)	(0.2)	(0.2)	(0.1)	(0.1)
FN costs	.322***	.247***	.395***	.303***	.303***	.249***
	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)	(0.1)
Male	-.193	-.157				
	(0.3)	(0.4)				
Male × FP costs	-.153	-.193				
	(0.2)	(0.2)				
Male × FN costs	.0791	.114				
	(0.1)	(0.1)				
Stat. class			-.24	-.142		
			(0.3)	(0.4)		
Stat. class × FP costs			.198	.124		
			(0.3)	(0.3)		
Stat. class × FN costs			-.0834	-.0226		
			(0.1)	(0.1)		
>23 yrs					-.366	-.647*
					(0.4)	(0.4)
>23 yrs × FP costs					-.0679	.0238
					(0.3)	(0.3)
>23 yrs × FN costs					.35	.277
					(0.2)	(0.2)
Constant	-.126	.391	-.0579	.419	-.157	.397*
	(0.2)	(0.3)	(0.3)	(0.4)	(0.2)	(0.2)
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

Standard errors in parentheses

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

Table 12: WTP minus Value of Information, risk aversion and sensitivity to FP and FN costs

	(1)	(2)	(3)	(4) FE	(5) FE
p>0.2	-.0942	-.11	-.0409	-.127	-.0578
	(0.2)	(0.2)	(0.3)	(0.2)	(0.3)
FN costs	-.229*	-.442*	-.327	-.385*	-.36*
	(0.1)	(0.2)	(0.2)	(0.2)	(0.2)
p>0.2 × FN costs	.716***	.977***	.889***	.949***	.914***
	(0.1)	(0.2)	(0.2)	(0.2)	(0.2)
FP costs	.558***	.69***	.78***	.652***	.672***
	(0.1)	(0.2)	(0.2)	(0.2)	(0.2)
p>0.2 × FP costs	-.933***	-.879***	-.899***	-.863***	-.91***
	(0.2)	(0.3)	(0.3)	(0.3)	(0.3)
Risk-loving × p>0.2 × FN costs		.037	-.383	-.0593	-.276
		(0.1)	(0.2)	(0.2)	(0.2)
Risk-averse × p>0.2 × FN costs		-.245	-.279	-.372**	-.198
		(0.2)	(0.2)	(0.2)	(0.3)
Inconsistent × p>0.2 × FN costs		-.0735	-.181	-.066	-.297
		(0.2)	(0.4)	(0.2)	(0.4)
Risk-loving × p>0.2 × FP costs		-.287	.0971	.179	.259
		(0.4)	(0.5)	(0.6)	(0.5)
Risk-averse × p>0.2 × FP costs		-.323	.00169	-.52	.0291
		(0.4)	(0.5)	(0.5)	(0.5)
Inconsistent × p>0.2 × FP costs		.108	-.21	-.48	-.372
		(0.7)	(0.5)	(0.5)	(0.5)
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

Standard errors in parentheses

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

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:

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

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 payoff with full protection: $EX > Y - c - b$. If an agent is imprudent ($u''' < 0$) then $E[MU] \equiv E[u'(X)] < u'(EX)$. Next, u' being a strictly increasing function 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 a bad 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:

$$P(0,0)(u(Y-b)-u(Y))+p(0,1)(u(Y-b-L)-u(Y-L))+P(1,0)(u(Y-c-b)-u(Y))+ \\ +P(1,1)(u(Y-b-c)-u(Y-L))=0 \tag{5}$$

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:

$$P(0,0)[u'(Y)(-b)+o(b)]+p(0,1)[u'(Y-L)(-b)+o(b)]+P(1,0)[u'(Y)(-c-b)+o(c+b)]+ \\ +P(1,1)[u(Y)-u'(Y)(b+c)+o(b+c)-u(Y-L)]=0 \tag{6}$$

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

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

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. □