Crying Wolf in the Lab

Arya Gaduh, Peter McGee, Alexander Ugarov*

September 15, 2023

Abstract

Abstract is here —-

Keywords: alarms, value of information, information economics, information design, —

1 Introduction

The 2010 gas blowout on Deep Horizon oil rig has killed 11 workers and caused one of the largest oil spills in history. The death toll was possibly aggravated by switching off a general safety alarm because its sirens interfered with workers' sleep.[1] This illustrates the trade-off between false-positive and false-negative test results with false-positive rates leading to higher false alarm costs and false-negative resulting in missed events.

Many real-life situations involve choosing binary tests to discover and prevent a negative outcome. Most binary tests transform continuous signals about the likelihood of an adverse state into simple yes/no prediction. This transformation relies on choosing a threshold for positive classification. Holding a continuous signal constant, a decrease in probability of no alarm in an adverse state (false-negative rate) corresponds to an increase in probability of alarm in a non-adverse state (false-positive rate). This trade-off motivates multiple discussions in medical diagnostics, alarm systems and extreme weather alerts. Despite ubiquity of binary alarms, there is little empirical evidence on how users evaluate alarms with different false-positive and false-negative rates.

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.

Some recent studies observe that many people put non-zero value on information about ego-relevant beliefs or future utility even if it has no apparent effect on subsequent decisions (all the citations). These preferences is not the focus of our study and hence we use relatively low stakes and ego-neutral information. As a result, our findings might not apply to settings with changing identity beliefs or to settings with delayed resolution of uncertainty and large potential payoffs.

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 (2020) similarly finds that individuals(?) do not properly account for priors and often choose tests not affecting optimal decisions even then more instrumental 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

[1] https://www.nytimes.com/2010/07/24/us/24hearings.html

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 **?** similarly employs a prediction game but varies priors on top of signal charasteristics. reducing prior uncertainty makes more signals non-instrumental in the sense that there should be no effect from a signal to optimal decisions. She find that many subjects choose non-instrumental over instrumental signals which is consistent with

Our setup differs in two important aspects from (Ambuehl and Li, 2018; **?**), 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 undervalue accurate signals, but we do not find a premium for certain signals. And similar to **?** we find that subjects do not properly account for interaction between prior probabilities and signal characteristics.

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 which differ in false-positive and false-negative rates but in 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 very few mammologists understand mamogram results and tend to overestimate 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 largely responsive to test accuracy (Liang et al., 2003; Howard and Salkeld, 2009; Neumann et al., 2012). But there are several apparent violations of rationality. 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 though the effect may come from its relatively low variation in the population.

This 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-probabiity events. Field evidence indicates that people underinsure 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 underinsurance 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.

2 Model

Environment. Consider a decision to purchase of threat assessment information. Let $\omega \in \{0,1\}$ denote the state of world, where 1 corresponds to an adverse event happening with probability π. The decision-maker has a lower utility in the adverse state, but only if she does not take the protective action. Denote the action to protect as $a \in \{0,1\}$. The protection technology is perfect: protected agents bear no losses but pay protection costs c regardless of the state ω. Decision-maker preferences are described by the utility function which depends on wealth Y, protective action a and potential damage in the adverse state $\omega(1-a)$. Utility is separable in wealth, protection costs $c > 0$ and potential loss in the adverse state $L > c$ [2]:

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

The decision-maker can purchase a binary informative signal $s \in \{0,1\}$ about the state of the world before making a decision. Let $P_{ij} \equiv P(s=i|\omega=j)$ be the probability of a signal s taking value i conditional on the state of the world being j. After receiving the signal, the decision-maker updates her belief on the likelihood of the bad state to $\mu(s)$. Unless specified otherwise, we assume that the decision-maker forms her posterior beliefs by using the Bayes rule. Hence the posterior belief equals:

$$\mu(s) = \frac{\pi P_{s1}}{\pi P_{s1} + (1-\pi)P_{s0}} \tag{2}$$

We also assume without loss of generality that a higher signal means a higher posterior probability of an adverse event $\mu(1) \geq \mu(0)$. Otherwise we can always re-label the signals.

Preferences. If there is no signal, the decision-maker protects if and only if it increases their expected utility:

$$EU_0 = \max[u(Y-c), \pi u(Y-L) + (1-\pi)u(Y)] \tag{3}$$

[2]Separability condition does not impose additional restrictions on the utility function U as long as the variation in wealth has limited range. More specifically, if $Y \in [Y_{min}, Y_{max}]$ and $c < Y_{max} - Y_{min}, L < c + (Y_{max} - Y_{min})$, then the function $u(\cdot)$ can be constructed from segments of $U(\cdot,0,0), U(\cdot,1,0), U(\cdot,0,1)$. While the resulting function $u(\cdot)$ is not necessarily monotonic, it is likely to be monotonic if protective actions and potential damages are relatively high.

The signal can increase expected utility if the decision-maker reacts differently to positive and negative signals. Under these assumptions, her expected utility with a signal is:

$$EU_s = \pi P_{11} u(Y - c) + \pi P_{01} u(Y - L) + (1 - \pi) P_{10} u(Y - c) + (1 - \pi) P_{00} u(Y) \tag{4}$$

We consider the maximum amount b which the decision-maker is willing to pay for the signal. In our framework, it is a price paid with a signal such that a decision-maker is indifferent between having a signal and paying b and not having a signal. Because the decision-maker can always ignore a useless signal, the signal's value is bounded from below by zero. Hence it equals to the maximum between zero and the solution to the following equation:

$$\begin{aligned} P(s = 1) u(Y - b - c) + \pi P_{01} u(Y - b - L) + (1 - \pi) P_{00} u(Y - b) = \\ = \max[u(Y - c), \pi u(Y - L) + (1 - \pi) u(Y)] \end{aligned} \tag{5}$$

Here we use $P(s = 1) \equiv \pi P_{11} + (1 - \pi) P_{10}$ to denote the probability of a positive signal (alert). 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 the equation (5) above has a at most one positive solution.

Obviously, perfectly accurate signals always have positive value $b > 0$ because the payoff distribution with the signal first-order stochastically dominates the distribution without the signal. If the decision-maker protects without a signal, a perfect signal reduces the protection costs and if she takes chances, then it reduces losses in the adverse outcome from L to $c < L$. However, it is harder to determine the value of the imperfect signal without imposing more restrictions on preferences as it requires weighing $u(Y - L)$ against $u(Y - c)$.

Risk-neutral agent. If the decision-maker 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:

$$b = \max[0, \min[c, \pi L] - P(s = 1)c - \pi P_{01} L] \tag{6}$$

We can express WTP b as a function of priors, false-positive and false-negative rates. This is the equation we use in our empirical work:

$$b = \max[0, \min[c, \pi L] - \pi(1 - P_{01})c - (1 - \pi) P_{10} c - \pi P_{01} L] \tag{7}$$

The sensitivity of (positive) value b with respect to false-positive P_{10} and false-negative P_{01}

rates is given by:

$$\frac{db}{dP_{10}} = -(1-\pi)c \tag{8}$$

$$\frac{db}{dP_{01}} = -\pi(L-c) \tag{9}$$

Both false-positive and false-negative rates decrease the (positive) signal's value. The effect is proportional to the adverse state probability for the false-negative rate and to the non-adverse state probability for the false-positive rates.

Risk Aversion Effects. In a more general expected utility framework, risk aversion can both increase and decrease the signal's value. More specifically, risk aversion decreases the value when the 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. 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 $b > b_0$. □

It is harder to make definite statements for lower risks or higher protection costs. For example, risk aversion increases value of a perfect signal as long as risk-averse decision-maker still chooses to not protect without a signal. This follows from the standard argument of increasing demand for insurance 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 study the effect of false-positive and false-negative rates on the signal's value b. Assuming a differentiable utility function $u()$ we use implicit differentiation to derive sensitivities of WTP 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:

$$D(\pi, P_{01}, P_{10}, b) \equiv P(S=1)u'(Y-c-b) + \pi P_{01}u'(Y-L-b)+$$
$$+(1-\pi)P_{00}u'(Y-b) = E[MU] > 0$$

It is clear that the signal's value decreases with false-positive and false-negative rates $\frac{db}{dP_{10}}, \frac{db}{dP_{01}} < 0$. We can also say a bit more about the sensitivity to false-negative rates:

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

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 decision-maker 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 a decision-maker 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 $(Y-b)$ reducing sensitivity to false-positive rates. It can also be higher if either the signal is good or the curvature is small.

If both protection costs c and risks π are low enough (leading to small WTP b) then we can use Taylor expansion to provide an approximate argument on the effect of risk aversion on sensitivities to false-positive and false-egative rates. For this case, risk aversion decreases sensitivity to false-positive rates and increases the ratio of false-negative to false-positive rates. Risk-loving individuals respond in the opposite way.

Start by rewriting the equilibrium condition for willingness-to-pay as the expected sum of utility differences:

$$P(0,0)(u(Y-b)-u(Y))+p(0,1)(u(Y-b-L)-u(Y-L))+P(1,0)(u(Y-c-b)-u(Y))+ \\ +P(1,1)(u(Y-b-c)-u(Y-L))=0 \tag{10}$$

Here I use $P(x,y)$ as 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:

$$P(0,0)[u'(Y)(-b)+o(b)]+p(0,1)[u'(Y-L)(-b)+o(b)]+P(1,0)[u'(Y)(-c-b)+o(c+b)]+ \\ +P(1,1)[u(Y)-u'(Y)(b+c)+o(b+c)-u(Y-L)]=0 \tag{11}$$

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

$$P(0,0)u'(Y)b+P(0,1)(u'(Y)+[u'(Y-L)-u'(Y)])b+P(1,0)u'(Y)(c+b)+ \\ +P(1,1)(-u'(Y)(b+c)-(u(Y-L)-u(Y))=0 \tag{12}$$

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:

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

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.

Hypotheses. Our theoretical considerations motivates two testables hypotheses for the experiment:

1. Risk-neutral model gives an accurate description of preferences over different alarm signal structures.
2. Subjects preferences put equal relative weights on the costs of false-positive and false-negative events as consistent with the risk-neutral model.

3 Experimental Design

Subjects received a USD 5 show-up fee and were endowed with USD 25 that they might lose in the experiment. Subjects must then make a series of decisions in four sets of tasks: (i) Blind Protection; (ii) Informed Protection; (iii) Belief Elicitation; and (iv) Willingness to Pay Elicitation. To verify their comprehension, subjects took a quiz before each task. For every wrong response, the correct answer and explanation are given. Additionally, subjects receive extra questions if they give wrong answers in a 5-question quiz given before the Informed Protection task. We do this because we consider Informed Protection as a first challenging task in the sequence which understanding is essential for the rest of the tasks. Each set of tasks has 6 rounds, for a total of 24 rounds. One of these rounds is selected at random as the payment round. A copy of the instruction is included in Appendix XX.

Blind Protection (BP). In each BP round, subjects must decide whether to insure (or "protect") against an adverse event (i.e., drawing a black ball from a box). Subjects were informed of the prior probability of drawing a black ball before making their decision. The cost to protect is USD 5. If a black ball is drawn, an unprotected subject will lose USD 20. Subjects then played six rounds, where the probability of drawing a black ball was varied between XX and XX percent in each round. During the BP task, subjects did not receive any feedback on how that round would have been realized were it chosen as the payment round.

Informed Protection (IP). For the IP task, subjects make a protection decision as in BP. However, before each decision, subjects are given a signal that was generated with varying degrees of inaccuracy. Following Coutts (2019), we present the signal-generation process using groups of "gremlins" that represent three types of signals: accurate (an honest gremlin), false positive (a black-swamp gremlin that always announces that the ball is black), and false negative (a white-swamp gremlin that always announces that the ball is white). Figure XX illustrates how the different gremlin types were presented to the subjects. Subjects knew the composition of the group from which the hint came from, but did not know which gremlin provided the hint. We vary the proportion of black balls in the box (prior probability of a black ball) and the composition of gremlins (signal quality) between rounds.

Belief Elicitation (BE). We use the BE task to elicit subjects' beliefs about the likelihood of an adverse event and an adverse signal conditional on prior and signal characteristics in an incentive-compatible way. Similar to the IP task, subjects were informed of the prior probability of a black ball and the composition of the group of gremlins that would provide an additional signal. However, instead of asking subjects to make a protection decision, we asked them to estimate the probability of two events, to wit: (i) the ball is black ball when a randomly drawn gremlin says that it is white; (ii) the ball is black when a randomly drawn gremlin says that it is black.

We follow the stochastic version of the Becker-DeGroot-Marshak mechanism developed by Grether (1992) and Holt and Smith (2009) to elicit incentive-compatible responses: the subject submits their belief of the probability of the event $\mu \in [0, 1]$. If this belief is above some uniform random number $r \in [0, 1]$, they receive the payoff x only if the stated event happens. Otherwise their payoff is determined by an independent lottery which pays x with probability r and 0 otherwise.[3] We also provide our subjects with the heuristics that under this mechanism, truthful reporting of beliefs is the dominant strategy.

Willingness to Pay Elicitation (WTPE). The WTPE task measures subjects' willingness to pay (WTP) for signals. Subjects know the prior probability of a black ball and the group

[3]The benefit of this mechanism versus other probability elicitation mechanism (for example, quadratic scoring) is that reporting truthfully is a dominant strategy regardless of risk preferences (Karni, 2009). The only requirements a subject needs to satisfy are probabilistic sophistication and dominance: they rank lotteries based on their probabilities only and prefer higher probabilities of higher payoffs.

composition of the gremlins that will determine signal quality. We then ask subjects for their WTP to receive a hint from a randomly drawn gremlin. Subjects can choose a value from USD 0 to 5 with USD 50-cent increments. Their decisions are 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 subjects' WTP, they will play a BP round. Otherwise, the subject would pay for the hint and play an IP round. After completing the WTPE task, subjects were asked a few demographic questions. The session concluded with the random selection and realization of the payment round, after which subjects were paid and dismissed.

The first three tasks were designed to provide measures of the different components of WTP described in Section XX and use them to examine the extent to which they explain subjects' WTP measured in the WTPE task. We use the BP task both to measure subjects' responses to the prior and their risk aversion. Next, we use the IP task to examine how signals affect protection decisions. Finally, we use the BE task as a measure of subjects' ability to estimate the probability of a signal for a given quality and to perform Bayesian updating. To construct these measures, we presented our subjects with 6 different priors for the BP task, and 3 priors and 2 gremlin groupings for the IP, BE, and WTPE tasks. Table reftab:treatments XX shows the values of the different priors in our treatments, as well as the gremlin groupings (along with the associated false positive and false positive rates) that we used for the different tasks.

We conducted this experiment in the Behavioral Business Research Lab (BBRL) at the University of Arkansas between October and November 2021. The experiment was implemented using Qualtrics. There were a total of 105 subjects. 84 percent of the subjects were university students and 41 percent were male. About 60 percent of the subjects had taken at least one statistics course. On average, including the show-up fee, subjects received around USD 26 for a session lasting around 45 minutes.

I think we need to have panels for parameter values in BP to help understand BP in Fig.1

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 Subjects Decisions

4.1 Blind Protection

Figure 1 plots the probability of protection decision against posterior probability of a black ball for the BP task, where the posterior is equivalent to the prior, and the IP task. On aggregate, subjects protect more with a higher 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 terms of risk aversion. For approximately 70% of subjects (X/Y), the probability of choosing protection 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]

Risk-neutral subjects maximize their expected utility by protecting whenever the prior probability exceeds 0.25, which is the ratio of the protection cost ($5) to the potential loss ($20). In contrast, many of our subjects start protecting for lower probabilities of 0.1 or 0.2 indicating strict risk aversion. As a point of reference, switching at the probability 0.1 corresponds to CRRA risk aversion of $\theta = 2$, while switching at 0.2 corresponds to $\theta = 0.573$ (see **??** in Appendix). A smaller group of subjects makes choices consistent with risk loving by never protecting or protecting for the probability of 0.3.

4.2 Informed Protection

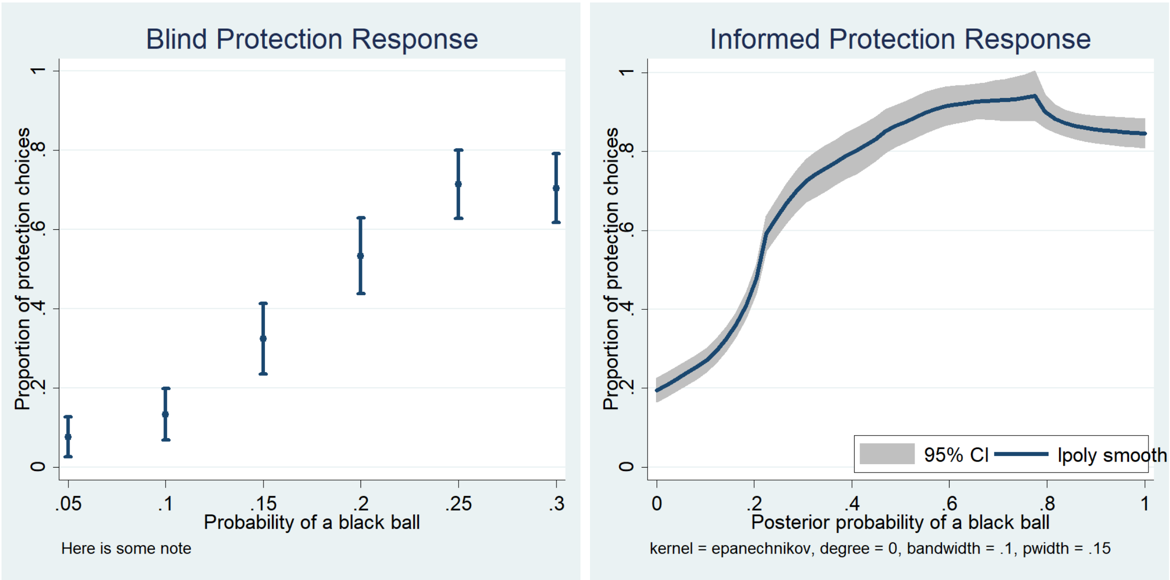
Subjects receive more information in the IP task than in the BP task, though using that information requires that subjects engage in Bayesian updating. Figure 1 shows that, consistent with their behavior in the BP task, the share of subjects protecting in the IP task is increasing in the posterior probabilities. Roughly 28% of subjects break monotonicity in their protection responses with respect to posterior probabilities[5] which roughly equals the percentage of non-monotonic responses for the BP task. At the individual level, we also observe that the total number of times subjects choose protection 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[6]

[4]For comparison, this reference on the Holt and Laury (2002) instrument suggests the XX% (YY%) of subjects switch at least (at most) once.

[5]They do not protect for some treatments with posterior probability P while protecting for a posterior probabilty $P' < P$.

[6]We use a LPM to estimate this relationship, and while the coefficient on the total number of protection choices is significant at 99%, R^2 increases from 0.295 to 0.3.

Figure 1: Average Protection Response



Eventually we need to add notes to the figures

In Table 2 we break out average protection decisions by signal characteristics. The first three columns summarize the information available to the subject, i.e., the signal as well as whether the signal might be either a false positive or false negative. Column 4 shows the posterior probability of a black ball averaged across all the treatments within a group, Column 5 the share protection among actual IP responses, Column 6 the share of protection under the RN optimum, and Column 7 the p-value of a t-test that actual and optimal choices use the same probability of protection.

First, we note that regardless of FP and FN rates, a hint that the ball is black substantially increases the share of protection decisions. **Second, subjects' protection decisions in the majority of treatments significantly deviate from what is optimal for risk-neutral subjects.** In general, subjects tend to overprotect when facing white signals (rows 1–4) and underprotect when facing black signals (rows 5–8). The exceptions are treatments with black signals and positive FP rates in which we cannot reject the hypothesis that the protection responses matches the response of a risk-neutral subject.

In light of BP decisions it is not surprising that subjects do not behave as risk-neutral agents, but some biases cannot be explained by the expected utility maximization for any degree of risk aversion. For example, consider the change in the protection rates between rows 1 and 3: the signal is white, so an increase in the signal's FP rate does not change the posterior, but 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 (maximum) posterior probability for the signal characteristics is just 13 percent. As a benchmark, with no signal in the BP task, only 13 (32) percent of subjects chose to protect when the probability is 10 (15) percent.

Table 2: Average Protection by Signal Type

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

Notes:

4.3 Belief Elicitation

While the IP task gives us a sense for how subjects utilize signals in making protection decisions, we observe only whether or not they choose to protect, which conflates preferences with potential errors in updating posteriors. The BP task gives provides insight into subjects' risk preferences, while the BE task allows us to better understand to what extent updating errors influence decisions.

We define updating errors as the difference between the posterior and subjects' elicited belief on the posterior probability of a black ball for a given signal. We plot the distribution of the updating errors in the left hand column of Figure 2, while the right hand column provides a scatter plot of the elicited beliefs against the true posterior with a fitted line. Panel A of Figure 2 uses all elicited beliefs 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.

Using all the observations, however, obscures an important distinction: in many cases the ball color is completely certain based on priors and signals and so the updating should be trivial. Panel B of Figure 2 includes only those 44% beliefs elicited for an uncertain posterior. The median error is now -0.12, with with 90% of errors between -0.48 and 0.3, suggesting that subjects tend to overestimate the likelihood of adverse events for uncertain posteriors, which is consistent with what we see in Figure 1. The correlation between beliefs and posteriors in this subset of observations is only 0.571. Panel C of Figure 2 plots the distribution of updating errors with certain posteriors, which includes: (i) treatments with all-honest gremlins; and (ii) treatments with obviously irrelevant dishonest gremlins (e.g., a group with honest and white-eyed gremlins with a hint that the ball is black — or vice versa). Reassuringly, 69% of reported beliefs are correct, but subjects still err in about 30% of cases. About half of these errors involve reporting a probability of between one and zero, with the other half reporting a probability of one when it should have been zero.

Overall, the pattern of belief updating is consistent with previous literature which finds that while humans usually update beliefs in a correct direction, they tend to underreact both to priors and 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), and is known under the names of representativeness bias or base rate neglect. Subjects sensitivity both to priors and signals is easy to measure through estimating the following equation ((first introduced by Grether (1980)) which links the posterior probabilities $\mu(B|S)$ of the state B conditional on signal S with the prior log-odds $\log\left(\frac{P(S|B)}{P(S|W)}\right)$ of the signal and signal log-odds $\log\left(\frac{P(B)}{P(W)}\right)$:

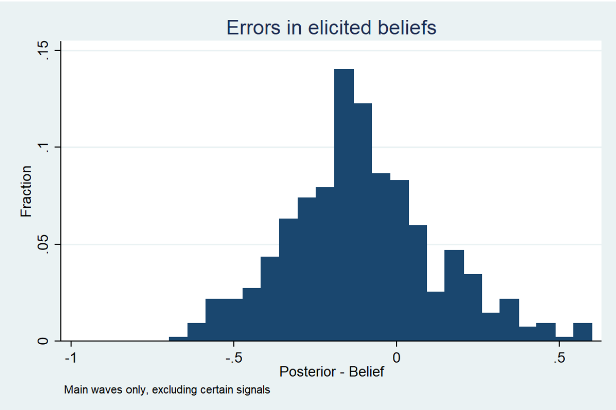
$$\log\left(\frac{\mu(B|S)}{1-\mu(B|S)}\right) = \alpha \log\left(\frac{P(S|B)}{P(S|W)}\right) + \beta \log\left(\frac{P(B)}{P(W)}\right) \tag{13}$$

Coefficients α and β should equal one if subjects perfectly follow Bayesian updating. Our estimates of these parameters are significantly below one with $\hat{\alpha} = 0.43$ $\hat{\beta} = 0.25$ (see Column 1 in 10). This is consistent with the meta-analysis in 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 presenting signals simultaneously (consistent with this study)[7]. Hence our subjects also demonstrate both the base-rate neglect and the signals underweighting. These effects lower the correlation between posteriors and reported beliefs and reduce sensitivity of beliefs to signal characteristics.

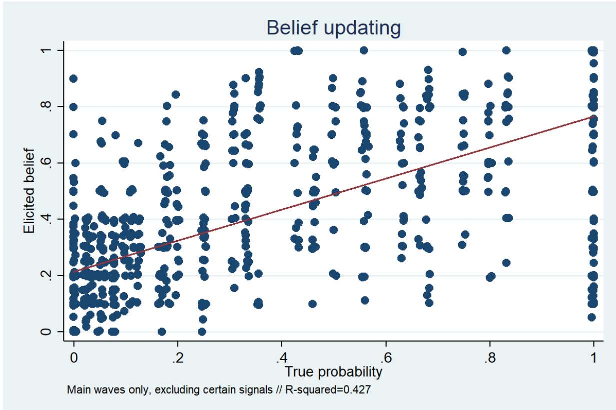
[7]The common name for this kinds of experiments is *bookbag-and-poker-chip experiments*

Figure 2: Errors in Bayesian Updating

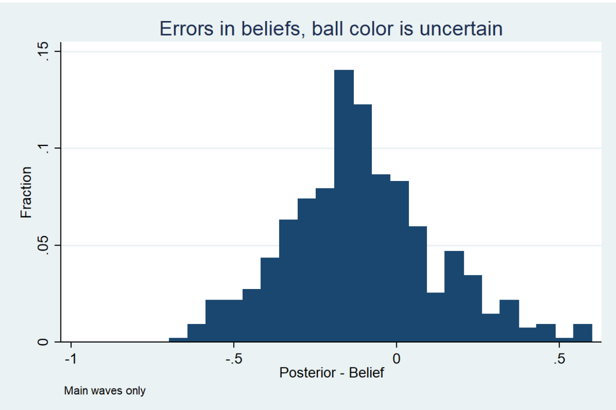
(a) Error Distribution



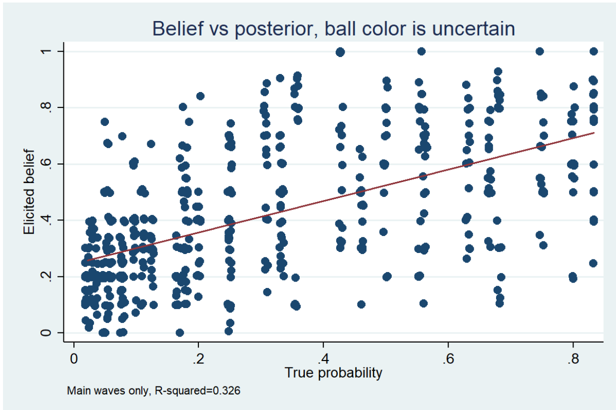
(b) Error v. Posterior



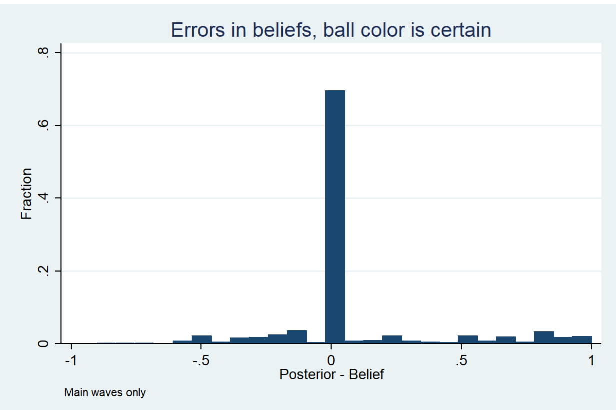
(c) Error Distribution, Uncertain Color



(d) Error v. Posterior, Uncertain Color



(e) Error Distribution, Certain Color



(f) Error v. Posterior, Certain Color

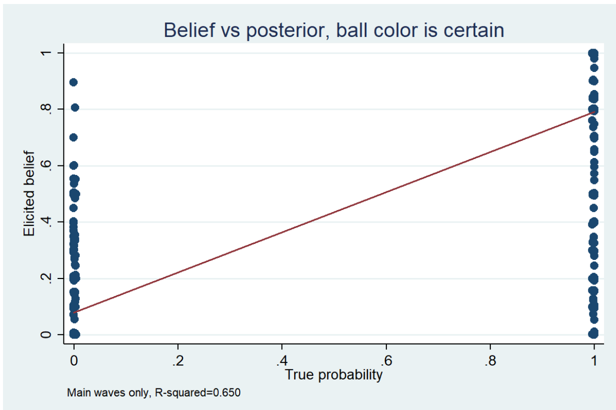


Table 3 summarizes how the updating errors vary with signal characteristics. We find that subjects overestimate the probability of a black ball with a white signal. Introducing FP rates to the signal exacerbated their upward bias. To illustrate, consider the change between rows 1 and 3, where introducing a FP rate would not change the posterior because the signal is white. Yet, subjects update their posterior upward, magnifying their updating error. The FN rates also have a similar effect of exacerbating this upward bias for a white signal.

The updating bias for black signals, however, varies by information structure. Subjects slightly underestimate the probability even when the signal is honest, but introducing FN rates lead subjects to underestimate it further. To illustrate, the introduction of a FN rate given a black signal does not change the posterior rows 5 and 6, but subjects decrease their beliefs. When there is a risk of a false-positive (i.e., FP>0), subjects again overestimate the probability of a black ball with little difference in errors between treatments with FP events only and with both FP and FN events. It seems that, because the false-positive rate negatively affects the posterior, subjects fail to adjust their beliefs enough in response to FP rates.

Table 3: Average Updating Error by Signal Type

Row	**Signal Characteristics**			**Posterior**	**Updating Error***	**P-val** (H_0 : $Error = 0$)
	False Positive	**False Negative**	**Signal**			
	(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 = $Posterior - Belief$.

Willingness-to-pay for Signals Elicitation. We first examine basic correlations between the WTP and signal characteristics to show that subjects understand the WTPE task reasonably well. One of the model's basic prediction is that signal value decreases with false positive and false negative rates. If subjects understands the basic premise of the WTPE, then we expect a negative correlation between the WTP and the signal's false positive and false negative rate.

The theoretical value of a signal for a utility maximizing risk-neutral subject (hereafter, risk-neutral signal value) derived in equation 2 provides a useful benchmark of our subjects' WTP. Figure 3 plots the distribution of the differences between subjects' WTP and the signal value. We find that subjects' WTP is centered around the risk-neutral signal value, which provides a reassurance that on average, subjects understand the task. However, we find a substantial variation: only 25% of actual WTP are within \$0.50 of the risk-neutral signal value, and subjects overvalue by at least \$1.5 in 22% of cases and undervalue it by at least \$1.5 in 19% of cases.

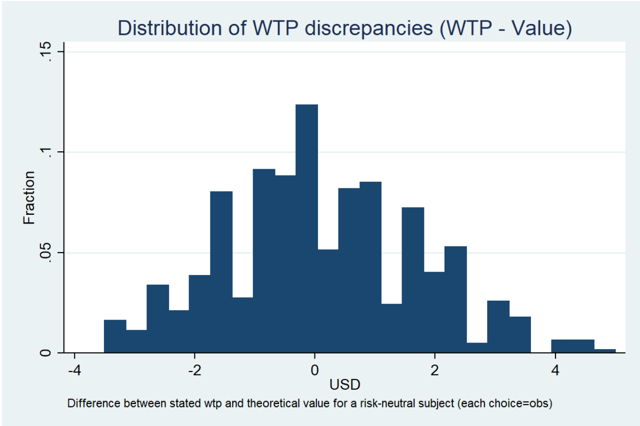


Figure 3: Discrepancy (Observed WTP - Signal value)

Our non-parametric analysis in Table 4 also finds no differences between willingness-to-pay and theoretical value for 3 out of 4 large categories of signals: Honest signals, Signals with only false-positive events and Signals with only false-negative events. When there are both false positives and false negatives, however, subjects significantly overvalue signals relative to the theoretical benchmark.Can we plot the distributions of deviations for these 4 cases?. These signals also tend to cause overprotection in the Informed Protection task.

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

False-positive	False-negative	Mean WTP discrepancy	P(= 0)
No	No	-0.106	0.433
No	Yes	0.143	0.250
Yes	No	0.081	0.502
Yes	Yes	0.492	0.000

5 WTP and Signal Characteristics

Hypothesis 1. *Subjects' preferences demonstrate equal sensitivity to costs generated by false-positive and false-negative events.*

Result 1. *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 understand how subjects' willingness-to-pay deviates from these risk-neutral signal value, we use a regression analysis of discrepancies between reported WTP and theoretical values to understand, first, the validity of a risk-neutral model and, second, relative weights put on false-negative and false-positive costs. Our main regression has the following form:

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

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 signal value; 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-maximizing subjects, we expect $\beta_1 = 0$ and $\beta_2 = 0$.

Table 5 shows that on average subjects underreact to both FP and FN cost when choosing their willingness-to-pay as compared to a risk-neutral decision maker. If subjects were to behave consistently with our (oversimplified) model, the coefficients on FP and FN costs should equal zero. And yet, Column 1 reports positive statistically significant coefficients for both FP and FN costs indicating that subjects tend to overpay for inaccurate signals. At the same time, a negative and statistically significant constant suggests underpayment for honest signals.

With regards to Hypothesis 2, we find that the coefficients on FN costs is slightly larger indicating higher sensitivity to FP costs, but we cannot reject the hypothesis that the two coefficients are equal. It means that on average and given our priors (!), the risk-neutral model still provides good guidance with choosing optimal tradeoff between false-positive and false-negative costs. However later we note that this equivalency breaks down when considering specific priors.

Our baseline estimation in Column 1 indicates significant deviations from the model's predictions. Since the benchmark signal value is based on the optimal choice of a risk-neutral Bayesian updater, deviations can arise from at least two sources. First, Proposition 2 suggests that individual risk preference can influence the sensitivity of WTP to these signal characteristics. Second, systematic belief biases can also lead to deviations.

We find that risk preferences have limited explanatory power for sensitivity to signal characteristics. We use data from the BP game to categorize subjects by their risk preference. We classify all the subjects with internally consistent BP choices into three risk-preference 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 preference, with risk-neutral as the default category. The point estimates of sensitivities among risk-neutral subjects are about 1/3 smaller than for the baseline results in Column 1 but enough to become statistically significant. This indeed indicates convergence towards prediction of a risk-neutral model, but with still sizeable remaining discrepancies. Given the magnitudes of coefficients on the interaction of risk preferences with FP and FN costs, only risk-loving subjects have plausibly different sensitivities though the interaction terms are not statistically significant.

Accounting both for belief accuracy and for risk preferences also does little to explain the

[8]Subjects who switched from no protection to protection at exactly the cost-loss ratio $\pi = 0.2$ are considered risk-neutral, while switching at lower (higher) levels indicates risk aversion (risk-loving). In addition, we created a dummy variable for subjects whose BP choices are inconsistent.

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	0.204	0.720	0.604	0.004
	(0.126)*	(0.339)	(0.303)**	(0.276)**	(0.571)
FN costs	0.319	0.232	0.192	-0.516	0.275
	(0.070)***	(0.286)	(0.264)	(0.486)	(0.257)
Risk-averse × FP costs		-0.020	-0.427	-0.054	-0.398
		(0.378)	(0.365)	(0.355)	(0.653)
Risk-averse × FN costs		0.061	0.233	0.432	0.147
		(0.299)	(0.319)	(0.585)	(0.319)
Risk-loving × FP costs		0.165	-0.474	0.027	-0.238
		(0.438)	(0.426)	(0.412)	(0.751)
Risk-loving × FN costs		0.177	0.357	0.970	0.040
		(0.309)	(0.295)	(0.558)*	(0.271)
Constant	-0.182	-0.184	-0.024	-0.068	0.114
	(0.083)**	(0.084)**	(0.104)	(0.111)	(0.130)
R^2	0.480	0.482	0.504	0.738	0.750
Obs	624	624	624	312	312
Risk-Averse Subjects:					
False Positive		0.184	0.293	0.551	-0.394
se		(0.168)	(0.204)	(0.223)	(0.316)
p-value		[0.274]	[0.154]	[0.015]	[0.216]
False Negative		0.293	0.425	-0.083	0.422
se		(0.089)	(0.178)	(0.326)	(0.189)
p-value		[0.001]	[0.019]	[0.799]	[0.028]
Risk-Loving Subjects:					
False Positive		0.369	0.246	0.631	-0.234
se		(0.277)	(0.299)	(0.305)	(0.488)
p-value		[0.186]	[0.414]	[0.041]	[0.633]
False Negative		0.409	0.549	0.454	0.315
se		(0.117)	(0.132)	(0.275)	(0.086)
p-value		[0.001]	[0.000]	[0.102]	[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: */**/*** denotes 10/5/1 percent significance levels.

pattern of underreacting to FP and FN rates. To study the role of subjects' ability to Bayesian update, we use data from the BE task to categorize the WTP responses by belief accuracy. 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 accurate if the error is less than the median error across all the subjects going through the same decision. For columnn 3, we present the most flexible specification to control for accuracy and risk preference by including triple interactions of belief accuracy, risk preference, and the signal characteristics. The baseline group is the group of risk-neutral subjects with relatively accurate beliefs, but for that group deviations from the risk-neutral model become even larger in terms of (lack of) sensitivity to FP costs.

Finally, we investigate the role of the prior probability. We motivate our experiment with real world problems of designing signals for low-probability disasters. 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. Since the cost to protect is equal to ($0.2\times$ the loss from the adverse effect), we split the sample using 0.2 as the cutoff. Here, we use a regression specification that includes prior probability fixed effects.

Column 4 presents the results for low-prior WTPE tasks. With a low prior, subjects overvalue false positive signals relative to the risk-neutral baseline. This overvaluation is similar for different risk preference profiles. In other words, with low priors, subjects overvalue signals that would induce them to overprotect. Both risk-neutral and risk-averse subjects do not (over-)value false-negative signals, while risk-loving subjects value such signals more positively — with a difference that is statistically significant at 0.1 — than risk-neutral subjects. The total coefficient of false-negative cost for risk-loving subjects is large, but is barely significant at 0.1 level.

Column 6 presents the results for high-prior WTPE tasks. With a high prior, risk-neutral subjects report that aligns with the signal value. All the types do not overvalue false positive signals, but both risk-averse and risk-loving subjects overvalue false negative signals. Our sample size cannot detect statistically significant differences in the extent of the overvaluation of false negative signals between both types and the risk-neutral subjects. These results imply a slight tendency, particularly about non-risk-neutral subjects, to overvalue signals that would induce them to underprotect.

There is also a possibility that the sensitivity is affected by subjects' demographic characteristics or their statistical education. We find little evidence (Table A) that gender or previous statistical education have significant effects on sensitivities for false-positive and false-negative rates (even though the lack of significance can be due to power issues). We find weak evidence that older (23+) subjects have a lower WTP for signals though our age variation is pretty limited.

To sum up, the pattern we discover indicates that subjects overreact to false-negative rates and underreact to false-positive costs with low priors, but the pattern reverses for high priors. In practice, it would mean that users would tend to overpay for alert signals with high false-positive costs, while excessively discounting systems with significant false negative rates. For example, they would prefer a smoke alarm which never misses fires even if it involves higher expected costs of false alarms. Risk preferences have weak explanatory power for this pattern as controlling for risk preferences reduces only the extra sensitivity of false-positive costs with priors but does not explain away the interaction between false-negative costs and priors. While this finding is incidental and not the part of the original research plan, its significant practical implications prompt us to explore it further.

Our analytical work also predicts that risk-averse individuals should have even higher sensitivity to false-negative rates, hence risk aversion on its own cannot explain this pattern. Similarly, risk-loving subjects should always put a higher relative weight on false-positive costs compared to a risk-neutral decision-maker, but this should hold for all the priors as long as WTP b and protection costs c are low enough relative to the potential loss L. Hence neither risk aversion nor risk-loving are consistent with the pattern. Not sure about heterogeneity of risk preferences if this heterogeneity does not much BP

XXX ALL: What would be nice if we can derive how higher prior would affect risk-averse/loving subjects. Also let's discuss what this really means and how to interpret. Does this mean that people overvalue signals that would change their default action? It would be nice, XXX

Table 6: WTP for Information, by prior (tobit)

	(1)	(2)	(3)	(4)
	0.1	0.2	0.3	0.5
model				
FP rate	-2.91***	-2.08**	-4.46***	-3.25**
	(1.1)	(1.0)	(1.0)	(1.3)
FN rate	-2.48**	-2.73***	-3.7***	-3.65***
	(1.1)	(1.0)	(1.0)	(1.3)
Constant	1.79***	2.33***	2.71***	3.32***
	(0.2)	(0.2)	(0.2)	(0.3)
sigma				
Constant	1.83***	1.7***	1.72***	2.16***
	(0.1)	(0.1)	(0.1)	(0.2)
P(FP rate=FN rate)	.792	.669	.617	.832
Adjusted R^2	.	.	.	.
Observations	159	153	159	153

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

In Table 5 we estimate directly the regression of reported willingness-to-pay on false-positive and false-negative costs given to subjects directly and find that the sensitivity to both false-positive and false-negative rates, first, changes relatively smoothly with priors and, second, that the two sensitivities are surprisingly close to each other. For none of the priors can we reject the hypothesis that two sensitivities are equal to each other. This drastically contrasts with the risk-neutral model, predicting that the sensitivity to FP rates should be much higher than the sensitivity to FN rates for low priors, but lower for high priors. This happens because for a given FN rate, false-negative events are much less likely with low priors and hence impose less costs on the decision-maker. Increasing priors in theory makes FN rates more salient while the saliency of FP rates should go down.

The closeness of coefficient estimates for FP and FN rates for Table 5 suggests one possible explanation for the observed pattern of sensitivities. If subjects' neglect the difference between false-positive and false-negative risks when choosing their WTP, it would explain both coefficients' similarity and flatness of sensitivities with respect to priors. If subjects treat FP and FN rates the same and treat only about the total proportion of false signals, they would assign equal weights to each of them. And because the sensitivity of theoretical value to FP and FN rates changes in opposite ways with priors, the best fit line for the sum of FP and FN rates should be much flatter with respect to priors.

Subjects' verbal explanations of choices made in Informed Protection provide additional credence for the hypothesis that subjects bundle FP and FN rates together. For example, the explanation from one of the subjects states*"I took into consideration how many honest there were and looked at the chances of picking a ball."*, while another also states *"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."*. In total, we find that 39 subjects out of 105 subjects in the main waves refer to the percentage of dishonest gremlins as their rationale for choosing protection. Many other could rely on this heuristic without giving a complete accurate explanation.

In order to test this hypothesis, we use choices from other tasks also using imperfect signals: Belief Elicitation and Informed Protection. If subjects systematically neglect the difference between false-positive and false-negative rates, we expect to find the pattern of abnormal 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 (FN) rates as if they are FN (FP) rates.

First, we test this pattern for the BE choices by estimating the relationship between belief errors and FP and FN rates by signal type. Table 7 reports our results. We estimate a linear regression of updating error (actual posterior - reported belief) on FP and FN rates by signal color. We used fixed effects to control for individual updating biases. Consistent with our hypothesis, we observe that the FP rate has 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). False-positive rates should not affect beliefs with white signals because a white signal (negative) can never be a false positive. The significance of FN rate for black signals is similarly an anomaly inconsistent with rational updating.

Table 7: 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

In Table 8, we regress informed protection decisions on FP and FN rates and flexible controls of both posteriors and reported beliefs:[9]

$$Prob(s_{ij} = 1) = \alpha_i + \beta_1 FP + \beta_2 FN + Z(p_{ij}) + Z(\mu_{ij}) + \epsilon_{ij}$$

where s_{ij} is the protection decision of subject i in treatment j, α_i - subject FE, P_{10}, P_{01} are FP and FN false positive and false negative rates and $Z(p_{ij}), Z(\mu_{ij})$ are the splines of corresponding variables P_{ij}, μ_{ij} to control for these variables in the 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 $[0, 1]$. [10] of posteriors and reported beliefs μ_{ij} for corresponding treatments. 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) estimates of the posterior, inferred from their BE responses.

Columns 1 and 2 show that even conditional on posterior and subject FE to control 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 for a black signal, FP rate had an opposite effect with comparable magnitude but without statistical significance. Hence the first prediction

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

[10]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.

of FP/FN rate confusion hypothesis 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/FN rates with respect to priors (Column 2). The coefficient on FN rate for black signals however is small in magnitude and statistically insignificant at 10%. 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 belief partially contribute to these protection anomalies, they cannot explain them completely (possibly due to subjects re-evaluating their beliefs between tasks).

Table 8: 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. * $p < 0.10$, ** $p < 0.05$, *** $p < 0.01$.

To sum up, we observe a striking uniformity in sensitivity of WTP to both false-positive and false-negative rates. This pattern is consistent with subjects neglecting the difference between false-positive and false-negative signals. This hypothesis is supported by subjects' explanation and with abnormal sensitivities to false-positive and false-negative rates in other treatments in which they do not affect posterior probabilities. We will leave the discussion of practical implications of this finding for the conclusion.

We also believe that this pattern of sensitivites does not come from subjects's anchoring on priors or from preferences for non-instrumental information. Given that each subject goes through two sets of treatments with two different priors and the order of priors is fixed, anchoring remains a theoretical concern. However, we find that most subjects change their decisions when going from one prior to another (92 out of 104). The average belief error in the BE task is actually lower for the second set of priors rather than for the first showing that changing priors does not increase subjects' confusion. And most importantly, there is still uniformity in coefficient ratio even if we limit our attention only to the first priors in each sequence (0.1 or 0.2)[11]. Similarly, while there is multiple evidence on humans valuing information not affecting their decisions (non-instrumental information), the value of this information should increase in false-negative rates for low priors in order to countervail a higher sensitivity predicted in theory. Having the value of non-instrumental information to increase with errors seems a priori implausible.

6 Conclusion

[11] Depending on session, the first 3 WTP treatments use either the prior of 0.1 or 0.2 and hence there is no anchoring on the previous prior.

References

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.

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.

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.

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://www.jstor.org/stable/40263877`, Number: 2 Publisher: [Wiley, The 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.

A Tables

Table 9: 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 10: 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 11: 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 12: 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

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