

# Sample(s) Structure

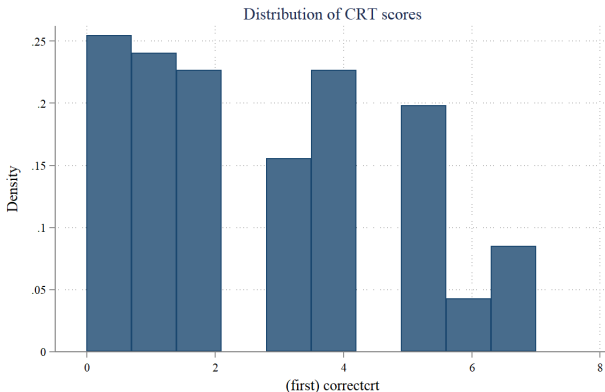
	All		$p \in \{0.1, 0.3\}$		$p \in \{0.2, 0.5\}$	
	N	%	N	%	N	%
All waves						
Male	96	47	49	46	47	47
Age>23yrs old	16	8	8	7	8	8
Students	174	84	90	84	84	85
Had statistics classes	128	62	71	66	57	58
First waves						
Male	43	21	22	21	21	21
Age>23yrs old	14	7	6	6	8	8
Students	88	43	46	43	42	42
Had statistics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

# Treatments

Prop. of black balls ( $p$ )	Gremlins composition			FP rate	FN rate
	Honest	Black-eyed	White-eyed		
0.1, 0.2, 0.3, 0.5	2	0	0	0	0
0.1, 0.2, 0.3, 0.5	1	1	0	0.5	0
0.1, 0.2, 0.3, 0.5	1	0	1	0	0.5
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
New treatments					
0.1, 0.2, 0.3, 0.5	1	1	0	0.5	0
0.1, 0.2, 0.3, 0.5	1	0	1	0	0.5
0.1, 0.2, 0.3, 0.5	5	2	0	0.29	0
0.1, 0.2, 0.3, 0.5	5	0	2	0	0.29
0.1, 0.2, 0.3, 0.5	5	1	1	0.14	0.14

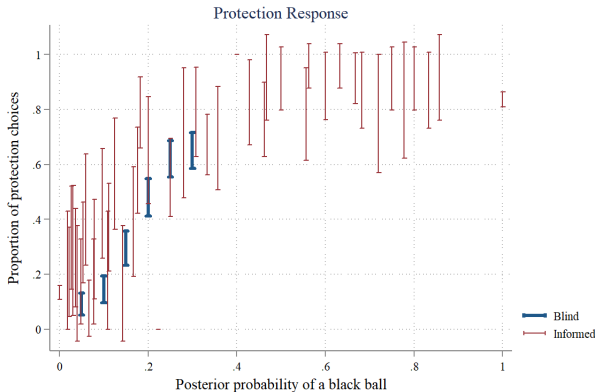
# CRT scores: just for the reference

- Extended CRT scores are lower than I expected (2 out of 7 median), slightly higher for college graduates (4 median). But this beats some previous studies (Toplak et al, 2014) finding about 1.5 items answered correctly on average.



# Blind and Informed Protection

- Tighter confidence intervals for blind protection (BP) as expected
- More points in IP, narrower confidence intervals for existing points, still roughly correlates with BP



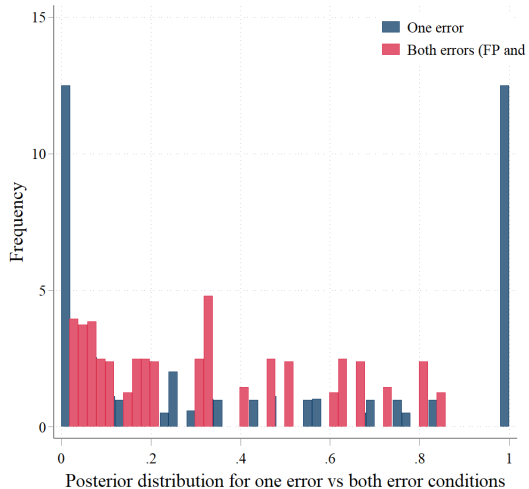
## Comment: intermediate priors with both conditions

*"For instance, on page 16, Result 1 shows that both-error conditions have systematically lowest WTP. This pattern might be suspicious since single-error conditions often produce extreme posteriors (0 or 1) while both-error conditions tend to produce intermediate posteriors. The complexity level is different. Likelihood insensitivity, rather than belief updating, might also explain the valuation. In addition, almost all the both-error conditions generate very low WTPs, thus the apparent overvaluation for them might "simply be due to reversion to the mean.""*

Response: We added new treatments with both error (technically one extra combination of gremlins but for different priors). The distribution of WTP for both errors doesn't concentrate near zero.

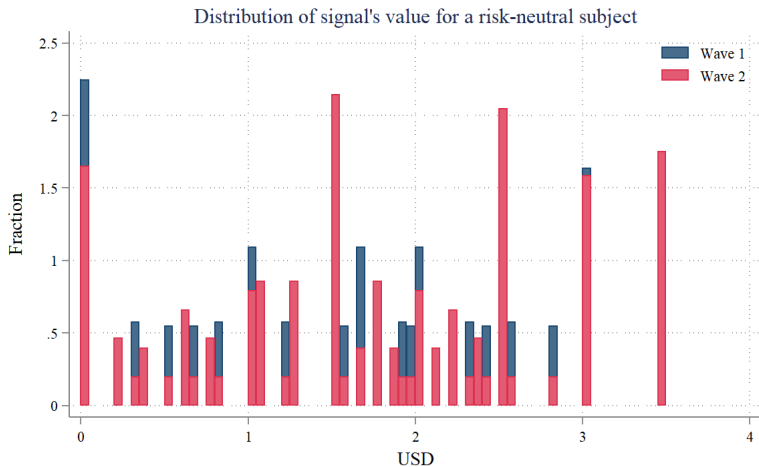
# Distribution of posteriors (both errors vs one error)

- The majority of uncertain cases has both errors and they are not concentrated near zeros/edges



# Distribution of theoretical values

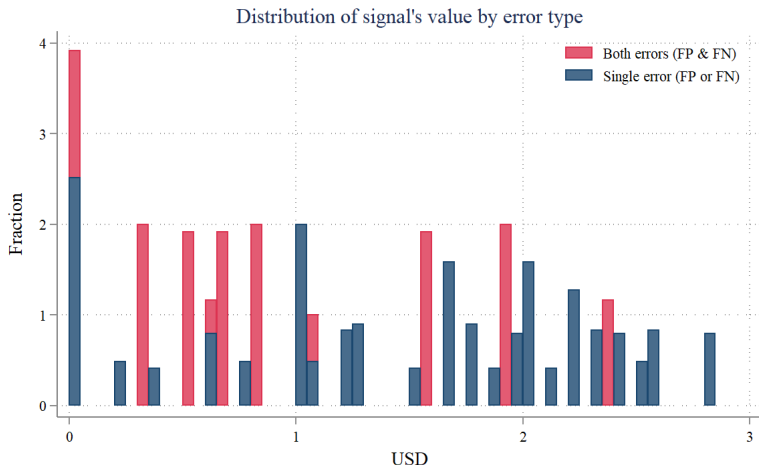
- New wave significantly beefs up treatments with intermediate value



Value = expected change in costs from BP to IP

# Distribution of theoretical values

- Both error conditions often result in significant WTP



Value = expected change in costs from BP to IP



# Comments: pooling in summary tables

*1) More importantly, their approach contradicts their own theory since they average responses across all subjects and conditions, but their theory predicts that different types of people (risk-averse versus risk-neutral) should show different patterns of FP/FN sensitivity. ""*

*2) I hope the authors can revise Tables 2 and 3 accordingly since the pooling of priors and FN and FP structures may be uninformative. Given that, it will be more straightforward to check how the elicited posteriors, protection actions and WTPs change for different priors and error types.*

Response: Split it by prior too? Could be two tables. The danger is that readers would take a more detailed table too seriously, it's original goal was just to give very rough first impression.

# Protection Summary

- Similar: overprotection for white hints, underprotection for black hints with no FP; new - slight underprotection for  $FP > 0, FP = 0$

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

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

# Belief Errors Summary

- Overestimation for white hints, black hints with FP, underestimation if there are FN or no error. Very similar.

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

- Overpaying: low priors - if there are FP, high priors - if there are FN; overpaying if both errors.

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

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

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

# Just IP responses regression for review

- Previous insights hold: FP/FN rates affect protection controlling on posteriors and beliefs

**Table:** Informed Protection Response

	(1)	(2)	(3)	(4)
FP rate x (S=White)	0.895*** (10.011)	0.943*** (10.145)	0.525*** (5.518)	0.571*** (5.850)
FN rate x (S=White)	0.537*** (3.709)	0.532*** (3.631)	0.307** (2.139)	0.299** (2.048)
p>0.2	0.039** (2.408)	0.041** (1.965)	0.024 (1.558)	0.029 (1.457)
S=Black	0.531*** (5.161)	0.542*** (4.758)	0.383*** (3.653)	0.374*** (3.268)
FP rate x (S=Black)	-0.032 (-0.158)	0.025 (0.123)	-0.065 (-0.330)	-0.000 (-0.001)
FN rate x (S=Black)	0.103 (1.398)	0.069 (0.860)	-0.005 (-0.055)	-0.021 (-0.229)
FP rate x (p>0.2)		-0.081 (-1.066)		-0.085 (-1.081)
FN rate x (p>0.2)		0.088 (0.891)		0.048 (0.521)
N	2424	2424	2424	2424
Pseudo R-squared	0.505	0.505	0.538	0.539
Log-likelihood	-830.188	-829.168	-773.731	-773.018
Subject FE	Yes	Yes	Yes	Yes
Flexible controls for:				
Posterior	Yes	Yes	Yes	Yes
Beliefs	No	No	Yes	Yes

## Comment: coefficients interpretation

*The authors write: "subjects tend to overvalue false-negative costs for low probability events and overvalue false-positive costs for high probability events." Where do we see that in Table 5? The coefficients of FP costs, FN costs on column 4 and 5 are all positive. Should it be "subjects tend to overvalue more false-positive costs (coeff: 0.800 vs 0.204) for low probability events and overvalue more false-negative (coeff: 0.407 vs 0.150) costs for high probability events."? When comparing coefficients, the authors should also report results in statistical tests*

The coefficients had changed, still underreacting to FP for low priors, no real difference for high priors on average. Should be reframed + tests when needed.

# Main WTP regression

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

	All			Prior	
				{ .1, .2 }	{ .3, .5 }
	(1)	(2)	(3)	(4)	(5)
FP costs	0.421 (0.081)***	0.487 (0.126)***	0.643 (0.158)***	0.577 (0.180)***	0.303 (0.308)
FN costs	0.287 (0.046)***	0.327 (0.084)***	0.357 (0.088)***	0.016 (0.216)	0.367 (0.085)***
Risk-averse $\times$ FP costs		-0.329 (0.225)	-0.415 (0.257)	-0.243 (0.285)	-0.576 (0.427)
Risk-averse $\times$ FN costs		-0.355 (0.124)***	-0.361 (0.135)***	-0.352 (0.292)	-0.288 (0.128)**
Risk-loving $\times$ FP costs		0.048 (0.179)	0.018 (0.213)	0.008 (0.290)	0.318 (0.408)
Risk-loving $\times$ FN costs		0.080 (0.107)	0.110 (0.117)	0.361 (0.341)	0.119 (0.118)
Obs	1230	1230	1230	615	615
Subject FE	Yes	Yes	Yes	Yes	Yes
Inaccurate Belief Interactions	No	No	Yes	Yes	Yes
Prior Probability FE	No	No	No	Yes	Yes

# Accounting for WTP bounds (Tobit), WTP as dependent variable, in theory sensitivity=-1

	All			Prior	
	(1)	(2)	(3)	{.1,.2}	{.3,.5}
model					
FP costs	-0.804 (0.121)***	-0.653 (0.012)***	-0.262 (0.016)***	-0.389 (0.015)***	0.101 (0.019)***
FN costs	-0.321 (0.061)***	-0.329 (0.007)***	-0.211 (0.009)***	-0.791 (0.017)***	0.386 (0.006)***
Risk-averse × FP costs		-0.346 (0.013)***	-0.360 (0.023)***	-0.069 (0.021)***	-0.796 (0.027)***
Risk-averse × FN costs		-0.342 (0.008)***	-0.276 (0.013)***	-0.307 (0.028)***	-0.441 (0.009)***
Risk-loving × FP costs		0.114 (0.012)***	0.058 (0.017)***	0.046 (0.015)***	0.251 (0.025)***
Risk-loving × FN costs		0.102 (0.008)***	0.143 (0.010)***	0.463 (0.020)***	0.141 (0.008)***
Constant	2.233 (0.154)***	-7.971 (0.008)***	-13.035 (0.003)***	-5.754 (0.004)***	-9.871 (0.004)***
sigma					
Constant	1.990 (0.077)***	1.302 (0.001)***	1.270 (0.001)***	0.994 (0.001)***	0.835 (0.001)***
*.subject_id	No	Yes	Yes	Yes	Yes
Prob(FP=FN)	0.000	0.000	0.000	0.000	0.000
Obs	1230	1230	1230	615	615
Risk-Averse Subjects:					
False Positive		-1.999	-1.622	-1.458	-1.694
se		(0.024)	(0.037)	(0.034)	(0.044)
p-value		[0.000]	[0.000]	[0.000]	[0.000]
False Negative		-1.671	-1.487	-2.098	-1.055
se		(0.014)	(0.020)	(0.042)	(0.014)
p-value		[0.000]	[0.000]	[0.000]	[0.000]
Risk-Loving Subjects:					
False Positive		-1.538	-1.203	-1.343	-0.647
se		(0.022)	(0.031)	(0.028)	(0.042)
p-value		[0.000]	[0.000]	[0.000]	[0.000]
False Negative		-1.227	-1.068	-1.328	-0.473
se		(0.013)	(0.018)	(0.036)	(0.013)
p-value		[0.000]	[0.000]	[0.000]	[0.000]



# Cognitive determinants of WTP

- No significant effects of CRT scores either on the level or on sensitivity of WTP

	(1)	(2)	(3)	(4)	(5)	(6)
GPA>3.5=1	.131 (0.4)	-.451 (0.5)				
FP costs	.367* (0.2)	.352* (0.2)	.55*** (0.1)	.515*** (0.1)	.423*** (0.1)	.435*** (0.1)
GPA>3.5=1 × FP costs	-.0409 (0.2)	-.0279 (0.2)				
FN costs	.341*** (0.1)	.442*** (0.1)	.318*** (0.1)	.417*** (0.1)	.342*** (0.1)	.321*** (0.1)
GPA>3.5=1 × FN costs	-.158 (0.1)	-.228* (0.1)				
GPA>3.5	0 (.)	0 (.)				
<4 CRT errors=1			.434 (0.5)	-.0602 (0.5)		
<4 CRT errors=1 × FP costs			-.363 (0.2)	-.326 (0.2)		
<4 CRT errors=1 × FN costs			-.134 (0.1)	-.225* (0.1)		
crt_errors			.00699 (0.1)	.0212 (0.1)		
Stat. class					.218 (0.2)	.306 (0.3)
Stat. class × FP costs					-.122 (0.2)	-.152 (0.2)
Stat. class × FN costs					-.0727 (0.1)	-.0204 (0.1)
Constant	-.345 (0.3)	.868** (0.4)	-.556 (0.7)	.47 (0.7)	-.418** (0.2)	.222 (0.2)
Prior dummies	No	Yes	No	Yes	No	Yes
Observations	492	492	606	606	1230	1230
Adjusted R <sup>2</sup>	0.02	0.18	0.04	0.20	0.04	0.20

# Demographic determinants of WTP

- Males and subjects with good quiz have slightly lower WTP. Sensitivities are higher only for subjects with the good quiz (though it is endogenous obv)

	(1)	(2)	(3)	(4)	(5)	(6)	(7)
	b	b	b	b	b	b	b
model							
Male	-.388*	-.341	-.567*	-.381*	-.401**	-.32	-.341
Stat. class	.191	.433	.171	.145	.289	.466	.456
ncorrect	.211***	.315***	.202***	.334***	.202***	.318***	.315***
correctcrt		-.0287				-.0333	.00519
gpa		-.66				-.423	-.572
FP costs			-.859***	-.431***	-.633***	-.692***	-.581***
FN costs			-.457***	-.28***	-.375***	-.279**	-.366***
Male × FP costs			.264				
Male × FN costs			.0869				
Good quiz × FP costs				-.686***			
Good quiz × FN costs				-.285***			
Stat. class × FP costs					-.17		
Stat. class × FN costs					-.0752		
GPA>3.5=1 × FP costs						.00127	
GPA>3.5=1 × FN costs						-.21	
<4 CRT errors=1 × FP costs							-.305
<4 CRT errors=1 × FN costs							-.0989
Constant	-.348	1.37	.333	-.808	.178	1	1.46
sigma							
Constant	2.01***	2.05***	1.92***	1.91***	1.92***	1.96***	1.96***
Prior dummies	Yes	Yes	Yes	Yes	Yes	Yes	Yes
Observations	1230	492	1230	1230	1230	492	492
Adjusted $R^2$							

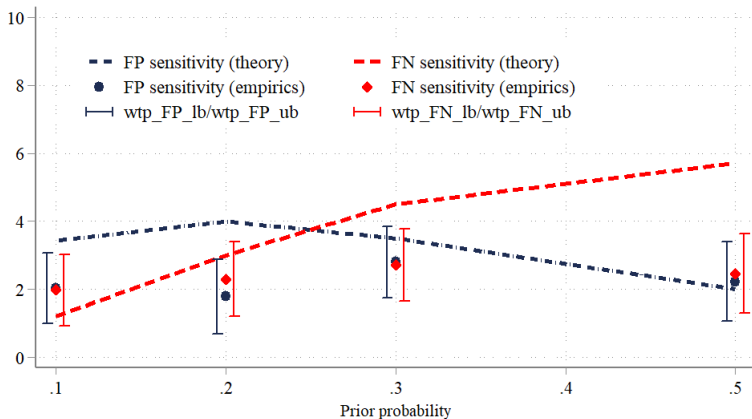
# Comment: coefficients interpretation

*"Similarly, given the importance of Figure 5, it would be nice if the authors could include confidence interval of the regression coefficients, and present in more details the regression specification."*

See the graph with confidence intervals added. And also the same graph using Tobit to estimate sensitivities. Will add regression specification either into the figure notes or into the text.

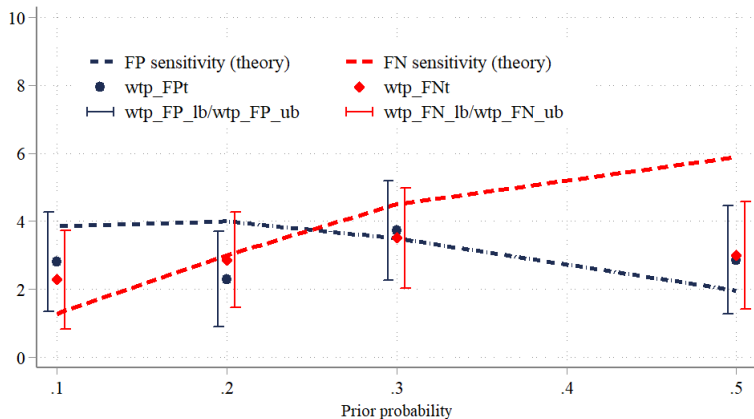
# WTP Sensitivity

- We can never reject the hypotheses that empirical FP/FN sensitivities are the same, but sometimes cannot reject that they are equal to theoretical ones.



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

# WTP Sensitivity (Tobit)



Estimates of sensitivity to FP and FN rates by prior probability of a black ball (tobit)

# Comment: Cross-Task Consistency Checks

*1) However, the current version of the paper does not explore much of the relationship between protection actions and WTP. Therefore, I encourage the authors to investigate how the protection actions observed in the experiment affect WTP, which sets the paper apart from the existing literature. The analysis would also lead to new implications. Intuitively, WTP for signals is roughly affected by two factors: 1) the understanding/preference (etc.) over information 2) anticipated protection action taken by the subject's future self. As shown in Table 7, subjects exhibit failure to distinguish FP and FN even when choosing their protection actions. Is the equal sensitivity of WTP w.r.t. FP and FN driven by rational anticipation of the "bias" in protective choice? Or is it driven by the heuristic when subjects compute the expected benefit of getting the signal? Depending on the answer, the result will have different policy implications for encouraging the acquisition of warning signals. ""*

# What Expected Bias Entails?

- Easy to imagine different potential protection biases with different implications:
  - ① I know that there are some signal structures for which I err in certain direction ("Always hide for that stupid tornado siren test on Wednesday"). Why not adjust the protection decision if knowing the bias direction?
  - ② I know that I do not account for particular information when making the decision (fail to distinguish FP and FN rates), better theoretical grounds, easier to model.
  - ③ I poorly understand particular signal structures (we already do observe that subjects with good quiz have higher WTP). Less assumptions to make and to use, but plausibly belief errors can indicate understanding of signal structures.
- We imply that the second type of bias exists for WTP in which subjects do not differentiate FP and FN costs  $\implies$  What if subjects are aware of that bias in IP too and pay less for signal structures known to suffer from this bias?

# Narrowing on Beliefs

- I am against studying biases of the first type - too internally inconsistent.
- Both types 2 and 3 imply poor use of information. Possible test: do they pay less for signals in which they have large BE errors?
- We can write a simple model for type 2 bias by rewriting the signal purchase problem in terms of variables they do account for:  $\pi$ , proportion of dishonest gremlins.
  - Still not clear how we test it? Bias correction significance is one option, but inconclusive because as a function of signal characteristics it can proxy for something else (preference non-linearity).
- Note: these biases cannot explain paying for unused signals, and as average WTP roughly equals the risk-neutral value, either the bias has to be small, or risk preferences increase baseline WTP



# Do errors correlate with discounting WTP?

- Belief error in round indeed strongly correlates with WTP (with subject and prior FE)

	(1)	(2)	(3)
Subject-round-specific belief error	.294 (0.3)	-.795*** (0.3)	-.723** (0.3)
FP costs		.482*** (0.1)	.571*** (0.1)
FN costs		.342*** (0.0)	.129 (0.1)
$p > 0.2 \times \text{FN costs}$			.234*** (0.1)
$p > 0.2 \times \text{FP costs}$			-.321*** (0.1)
Prior dummies	Yes	Yes	Yes
Observations	1230	1230	1230
Adjusted $R^2$	0.50	0.56	0.56

Standard errors in parentheses