January 9, 2026

Dear Professor Vespa,

Thank you for the opportunity to revise our manuscript titled *Preferences for Warning Signal Quality: Experimental Evidence* (Manuscript ID: YGAME-D-25-00181). We appreciate the thoughtful comments and suggestions provided by the AE and the reviewers. We have carefully addressed all points raised and believe the revisions have significantly improved the paper. Below, we provide a detailed response to each comment.

We hope these revisions meet the expectations of the reviewers and the journal. Thank you for considering our revised manuscript.

Sincerely,
Peter McGee
University of Arkansas

General Summary of Changes
We did stuff...

Detailed Responses

AE:

1. **Comment:** The introduction of preferences is currently confusing. The expected utility (EU) formulas simultaneously present both utility calculations and optimal actions. I recommend a more incremental approach. First, derive the EU of each possible strategy (e.g., following the signal, ignoring it, doing the opposite of what signal says etc.), then identify the signal structures under which following the signal is optimal. This approach, common in information economics, may clarify your presentation and address R1's concerns about Proposition 1.

   Response: [We...]

2. **Comment:** Can you strengthen Proposition 3 by deriving bounds for the dropped error terms?

   Response: [We...]

3. **Comment:** Reviewer 1 raises important concerns about the pooling of priors and the fact that the willingness to pay (WTP) for some signals gets too close to the bounds of the range, which are echoed by Reviewer 2. If these concerns are indeed valid, that makes your results questionable. If feasible, consider running robustness checks by changing priors to make WTP amounts more interior or varying the elicitation method of WTP. If your lab access is limited, you may want to switch to online platforms. If you pursue this, I recommend randomizing the order of priors and incorporating Reviewer 2s comment #9 into the design as well.

   Response: [We...]

4. **Comment:** Your individual-level analysis needs to be improved. In addition to the reviewers suggestions in this regard, you may consider classifying participants based on the consistency of their behavior across the four parts, and test your hypotheses for each group.

   Response: [We...]

5. **Comment:** Tables 2 and 3 are difficult to interpret. Please consider a more transparent way of presenting the data across treatments. The reviewers offer constructive suggestions here.

   Response: [We...]

6. **Comment:** I agree with Reviewer 1 regarding the terminology. Consider using information source or information structure for each treatment, and reserve signal for the realized signal.

   Response: [We...]

7. **Comment:** Introduce the definition of imprudent in the main text rather than only in the proof.

   Response: [We...]

8. **Comment:** Please cite the papers suggested by the reviewers.

   Response: [We...]

Reviewer 1:

1. **Comment:** First, the term signal used by the authors is somewhat confusing. When it is ex-post, the signal is the realization from an information source, while when it is ex-ante, it is the source itself. Given that the experimental design contains elicitations regarding both expost signals and ex-ante sources, I would like to see a clearer exposition of this terminology. In general in the literature, signal or news are often referred to as signal realizations, and ex-ante it is called the information source. I will follow this custom in this report.

   Response: [We...]

2. **Comment:** On page 6, the assumption that access to a signal can increase expected utility if the signal affects her protection decisions influences the calculation of the WTP b in equation (1), as well as the partial derivatives for sensitivity. I do not think this assumption is proper here. The assumption should be more fundamental and made on the properties of the information source (P_{ij}), not on behaviors. In fact, in the set of parameters chosen for the experiment, four tasks did not satisfy this assumption (non-instrumental information sources). Another suggestion is to verify whether subjects choices indeed follow this assumption. It provides a measurement at the individual level for contingent reasoning. If a significant number of subjects put a positive WTP for the non-instrumental information sources, then the theoretical analyses based on this assumption become questionable.

   Response: [We...]

3. **Comment:** Also in Appendix B.1 Proposition 1, the last equation is not correct. It should be $\leq$ not $=$. The statement Because expected utility with a signal is a decreasing function of b_0 we obtain $b > b_0$ should be $b < b_0$.

   Response: [We...]

4. **Comment:** On page 7, I do not understand the prediction that the model of a risk-neutral agent suggests that subjects WTP should have equal sensitivity to costs from false-positive and false-negative signals (equation 2). In equations (2)-(4), the partial derivatives of the WTP b with respect to FP and FN are different and only equal when the prior is 0.25. But the experiment tasks have varying priors from 0.1-0.5. As a result, Hypothesis 2 is not correct and the analyses for Result 2 should be restructured.

   Response: [We...]

5. **Comment:** Propositions 2 and 3 rely on quite restrictive assumptions and simplifications. I would like to see discussion of the generalizability of these propositions, perhaps with simulations or numeric methods.

For instance, Proposition 2 relies on the imprudence assumption that might be inconsistent with standard utility specifications and the authors own empirical findings. The condition $U'''(w) < 0$ fails to hold for the most commonly used utility functions in economics: CRRA utility has $U'''(w) > 0$ for reasonable risk aversion parameters $\gamma > 0$, while CARA utility has $U'''(w) > 0$ for any $\alpha > 0$. In the analyses for BP, the authors acknowledge that many of our subjects (24) start protecting at lower priors (0.05-0.15), indicating strict risk aversion. A smaller group of subjects makes choices consistent with risk loving by protecting at a probability of 0.3. There should be further discussion of the validity of the imprudence condition.

Response: [We...]

6. **Comment:** Proposition 3s derivation relies on a first-order Taylor approximation that drops higher-order terms without establishing error bounds or validity conditions. The proof provides no justification for why the neglected terms $o(b)$ and $o(b + c)$ are small enough relative to the retained terms. For meaningful values of b and c (5protectioncost; upto5 WTP), the neglected terms could be substantial. I would like to see error bounds demonstrating when the approximation is valid.

Response: [We...]

7. **Comment:** 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: [We...]

8. **Comment:** My main concern about the experiment is the parameter designs. As we observe from my calculations, many conditions produce extreme values. I am worried that some key findings appear to be driven by systematic measurement artifacts created by the experimental parameter choices rather than genuine behavioral patterns. 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...]

9. **Comment:** This may also drive the key conclusion about asymmetric sensitivity. For low prior conditions, the low base rate ($\pi = 0.1$) means theoretical WTP is already near zero for most signals. Adding false positive costs should theoretically drive WTP even lower toward zero, but there is a boundary effect where subjects cannot bid negative amounts, so any positive bid appears as insufficient reduction. This creates a measurement artifact where the failure to reduce WTP enough is simply because observed WTP is bounded below at zero. For high prior conditions, the high base rate means theoretical WTP could be substantial ($3-5), and adding false negative costs should theoretically reduce WTP from these high levels. However, subjects may be reluctant to bid close to their $30 endowment or have budget constraints, creating a measurement artifact where the underreaction is because observed WTP is practically bounded above.

   Response: [We...]

10. **Comment:** The analyses and conclusions of this paper might be stronger if the FN and FP combinations also induced some intermediate posteriors. Otherwise, I would suggest the authors have a discussion about how the parameter design may create artifactual results.

    Response: [We...]

11. **Comment:** The data analyses are not well-structured for readers to grasp the take-aways. Most results use aggregate statistics that mask individual inconsistencies. They often pool subjects across tasks rather than tracking individual patterns and focus on population averages rather than individual coherence.

    Response: [We...]

12. **Comment:** A missing piece of the data analyses is cross-task consistency checks. The authors experimental design contains four parts, yet they analyze each component in isolation without examining individual-level consistency across tasks. This represents a significant missed opportunity for validation, as their theoretical framework explicitly requires subjects to integrate these components coherently.
    For instance, a BE-IP consistency check could examine whether, given a threshold posterior of 0.25, a consistent decision maker who reports posterior ¿ 0.25 in the BE task also chooses Protect in the IP task. Across the six tasks generating 12 posteriors, the authors could identify decision thresholds for different subjects and inconsistency measures at the individual level.

    Response: [We...]

13. **Comment:** Similarly, an IP-WTP consistency check could examine whether participants who chose to protect or not protect regardless of possible signals assign zero WTP to information structures, as predicted by the assumption that access to a signal can increase expected utility if the signal affects her protection decisions. These analyses would likely reveal that high-consistency subjects show clearer false positive/false negative sensitivity patterns, while low-consistency subjects drive the null aggregate results, explaining why their current pooled analysis fails to detect the theoretically predicted effects.

    Response: [We...]

14. **Comment:** In the key section WTP and Signal Characteristics, the authors use the difference between individual WTP and risk-neutral WTP for regressions. With the decisions in other tasks, I think it makes more sense to use the difference between observed WTP and predicted WTP, which should be a function of individual risk attitude from BP, individual beliefs from BE, and individual action plans from IP. The function could be a simple OLS but could also incorporate behavioral elements that account for different biases in each component of the tasks.

    Response: [We...]

15. **Comment:** For section 6, the evidence provided for the key conclusion is not convincing enough. The claim that people cannot distinguish false positive from false negative rates relies mainly on failing to find a significant difference in how people respond to FP versus FN costs, but failing to find a difference does not prove that no difference exists. It could simply mean their test lacks sufficient power or that their data is too noisy. 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. By pooling everyone together, they make it impossible to detect the very differences their theory expects to find. The authors also never directly test whether people actually use a total error rate strategy but simply assume this explains their results without considering other possibilities like measurement problems or task confusion.

    Response: [We...]

16. **Comment:** The contribution is not clearly pointed out within the field of experimental studies for the demand for information sources. For instance, the failure of distinguishing the value of FP and FN is also implied in Xu (2022) while also accounting for the effects of risk attitudes and belief updating biases. I want to see a clearer comparison and innovation of this paper. Some other key references are missing, for instance, Masatlioglu, Orhun, and Raymond (2023), Charness, Oprea, and Yuksel (2021), Montanari and Nunnari (2023), which all study skewed information sources, and some of their skewness conditions have the same structure as the FP and FN rates studied in this paper.

    Response: [We...]

17. **Comment:** Another strand of the literature missing is ROC (Receiver Operating Characteristic) analysis, which is a statistical method used to evaluate the performance of binary classification models, particularly in radiology, diagnostic testing and machine learning. The structure is relevant. I hope there would be more exploration of how the results in this paper relate to findings in this literature. Please see Fawcett (2006) for a review.

    Response: [We...]

18. **Comment:** Please indicate significance of the estimated parameters in your regression tables for better reading experience.

    Response: [We...]

19. **Comment:** On page 1, the sentence While the optimal threshold depends on user preference over the costs of these probabilistic errors, currently there is no guidance on what this threshold might be beyond assuming that decision-makers weigh false-positive and false-negative costs equally is not consistent and not true. The optimal threshold depends on the costs of the false-positive and false-negative probabilistic errors, which in most cases are not weighted equally, such as in medical tests where the consequence of missed detection is weighted more than the consequence of false alarm.

    Response: [We...]

20. **Comment:** On page 8, the IP task description is not clear. It states subjects must make a protection decision given the prior probability of drawing a black ball. Subjects learn a prior and signals accuracy, but subjects should make two decisions about protection, each for a possible hint.

    Response: [We...]

21. **Comment:** On page 9, there is a concerning inconsistency between the belief elicitation mechanism described in the paper and the actual experimental instructions given to subjects. The paper claims to follow "the stochastic version of the Becker-DeGroot-Marshak mechanism developed by Grether (1992) and Holt and Smith (2009)" with a specific mathematical formulation involving uniform random draws and conditional losses. However, the experimental instructions describe a completely different mechanism that simply tells subjects "you make more money if your guess is closer to the actual probability of the event" with examples using different probability values and payoff calculations that bear no resemblance to the theoretical mechanism described in the paper. This discrepancy raises serious concerns about the validity of the belief elicitation data. Moreover, the instructions themselves appear potentially confusing to subjects, particularly the statement about "the actual probability of the ball being black is 10 percent" when each individual ball is deterministically either black or white. If the goal is incentive compatible belief elicitation where accuracy is rewarded, it would be much clearer to implement a proper scoring rule that directly penalizes prediction errors rather than using this convoluted mechanism that appears to differ between the papers description and the actual implementation.

    Response: [We...]

22. **Comment:** Given the central finding on asymmetric sensitivity depending on the prior probability, I think the comparative statics of the sensitivity for different priors is missing in this paper. It is easy for the risk-neutral benchmark but is not straightforward with risk attitudes.

    Response: [We...]

23. **Comment:** On page 14, We define updating errors as the difference between the subjects elicited belief and the actual posterior probability of drawing a black ball for a given signal will be understood by most people as Updating Errors = Elicited Belief - Actual Posterior Probability, which is not consistent with your figures and follow-up descriptions.

    Response: [We...]

24. **Comment:** There are also some ambiguous pronoun references throughout the paper. On page 15, This suggests that subjects understand the signal structure but have difficulty implementing the optimal strategy leaves unclear what This refers to, since the preceding sentence discusses multiple results.

    Response: [We...]

25. **Comment:** Similarly, on page 17, These results are consistent with our hypothesis but contradict previous findings does not specify which results or which previous findings. Page 18 references low WTP conditions and high WTP conditions without clearly defining the threshold or providing descriptive statistics.

    Response: [We...]

Reviewer 2:

1. **Comment:** *(1a)* I believe that, compared to existing studies on the demand for information, the most novel element of this paper is the inclusion of protection actions. As the authors mentioned, the existing literature employs prediction games; thus, states are usually symmetric, and guessing the underlying state is relatively trivial. Here, the protection action is asymmetric, and protection decisions become non-trivial, as we have known in the literature on choice under uncertainty. As future protection actions affect individuals WTP for information, the presence of protection action would have a non-trivial effect on subjects WTP.

    Response: [We...]

2. **Comment:** *(1b)* 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 subjects 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.

    Response: [We...]

3. **Comment:** *(2a)* Result 2 on page 18 is interesting. However, it is a comparison of subjects WTP with the risk-neutral benchmark, and it is unclear what implications we can draw from it. One could also question whether risk neutrality is a meaningful benchmark. I understand the novelty and importance of the results only until the end of the paper, where the authors discuss the subjects failure to distinguish FP and FN rates. Figure 5 on page 22 is particularly clear and striking. I suggest that the authors move the discussion of subjects failure to distinguish FP and FN rates to right after result 2, and frame it as an analysis on the underlying mechanism. The rest of the discussion could be framed as a section discussing/ruling out alternative explanations.

   Response: [We...]

4. **Comment:** *(2b)* 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.

   Response: [We...]

5. **Comment:** *(2c)* Figure 5 shows the sensitivity to FP and FN are similar and unchanged as prior changes. Does that mean fixing FP and FN, WTP is unchanged in prior? That would violate the intuition that the value of information decreases as the prior becomes more extreme. As I suggest the authors increase the spotlight on Figure 5, it would be a good opportunity to present some analysis on the reported WTP instead of the difference between the reported WTP and the risk-neutral benchmark.

   Response: [We...]

6. **Comment:** *(3)* Another potential explanation of the results is probability weighting. If somehow subjects exhibit probability weighting to the extent that probability compresses towards 0.25 for all priors, then sensitivity to FP and FN would be equal and is invariant in prior. I dont think this explanation is realistic, but some discussions (to rule it out) would be nice.

   Response: [We...]

7. **Comment:** *(4)* I find it unclear how subjects are allocated to different treatments. The authors write in the 2nd paragraph page 10: Subjects go through two different priors and three types of signals. But on Table 1 page 10, there are total 4 types of priors and 7 types of signals. Are subjects randomly allocated to 2 out of the 4 types of prior and 3 out of the 7 types of signals? But the authors write in the 2nd paragraph, page 10: Subjects go consecutively over all three signal types starting from the honest one for each prior.. That means one of the three signals is from the honest-only composition (3-0-0) for all subjects, and subjects are not randomly assigned to 3 out of the 7 types of signals? What are the sample sizes for each treatments?

   Response: [We...]

8. **Comment:** *(5)* I am also confused about Table 2 and 3, especially, column 4, the posterior. It seems that the authors have pooled the different 4*7 treatments into 8 different groups, based on the presence of FP and FN, and the hints received by the subjects. The authors write on page 12 that Column 4 shows the posterior probability of a black ball averaged across all the treatment within a group How are the posteriors calculated? Is it based on the empirical data, or is it some weighted average of Bayesian posteriors? Similarly, are the share optimal calculated based on the empirical data or the Bayesian posteriors?

   Response: [We...]

9. **Comment:** *(6)* 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.

   Response: [We...]

10. **Comment:** *(7)* The authors cited two laboratory experimental papers on the demand for information. I believe there are many more. I do not think the missing literature diminishes the contribution of this paper, but citing them would highlight the novelty of the papers setup and better position the paper in the existing literature. Here are some examples:

    (a) Intrinsic Information Preferences and Skewness by Yusufcan Masatlioglu, Yeşim Orhun, and Collin Raymond

    (b) How Do People Choose Between Biased Information Sources? Evidence from a Laboratory Experiment by Gary Charness, Ryan Oprea, Sevgi Yuksel

(c) Beyond Value: on the Role of Symmetry in Demand for Information by Aniol Llorente-Saguer, Santiago Oliveros, Roi Zultan

Response: [We...]

11. **Comment:** *(8)* I think one of the low-hanging fruits to improve the papers contribution is to provide more discussion on policy implications related to warning signals. At the moment, the authors only discuss the papers contribution in the introduction by saying that currently there is no guidance on what this threshold might be beyond assuming that decision-makers weigh false-positive and false-negative costs equally. Throughout the rest of paper, the authors did not discuss any implications/guidance on the design of warning signals systems.

    Response: [We...]

12. **Comment:** *(9)* The use of multiple gremlins frames both FP and FN as dishonest. Thus, one would imagine that subjects refrain from trusting/paying dishonest gremlins, leading to equal sensitivity to FP and FN. The authors could consider running new sessions that frame signals as generated from one imperfect gremlin/expert that has asymmetric tendencies to send false positive and false negative signals.

    Response: [We...]

13. **Comment:** *(10a)* There are no stars in Table 5 onwards. Is that intentional? I find it impossible to calculate the p-values for every coefficient.

    Response: [We...]

14. **Comment:** *(10b)* In various parts of the paper, the authors have used the terms true posterior and accurate posteriors (for example Table 2 and 3, the bottom of page 14, and many others.). Do they refer to Bayesian posterior or empirical posterior?

    Response: [We...]

15. **Comment:** *(10c)* Page 18 after the regression equation: how exactly are FP cost and FN cost calculated? It is confusing to see FP and FN sometimes refer to rate and sometimes refer to cost.

    Response: [We...]

16. **Comment:** *(11a)* Extra = sign in Equation (1) page 5.

    Response: [We...]

17. **Comment:** *(11b)* Missing a fullstop at the end of footnote 7 page 15.

    Response: [We...]