Dear Dr van de Wouw,  
  
Thank you for your patience during the peer-review process. Your manuscript titled "How we model prior belief is crucial for predicting decision biases in realistic contexts" has now been seen by 2 reviewers, whose comments are appended below. You will see that they find your work of some potential interest. However, they have raised quite substantial concerns that must be addressed. In light of these comments, we cannot accept the manuscript for publication, but would be interested in considering a revised version that fully addresses these serious concerns.  
  
We hope you will find the Reviewers' comments useful as you decide how to proceed. Should additional work allow you to address these criticisms, we would be happy to look at a substantially revised manuscript. If you choose to take up this option, please highlight all changes in the manuscript text file, and provide a detailed point-by-point reply to the reviewers.  
  
Please bear in mind that we will be reluctant to approach the reviewers again in the absence of substantial revisions.  
  
Editorially, we consider three aspects key: the evidence put forward in support of your interpretation must be strengthened through additional empirical data and further analyses, the key advance must be demonstrated more convincingly and explained more clearly and the use of appropriate statistics and improved statistics reporting is required. Please note that the editorial requests incorporate advice we received from Reviewer #2 in an additional email exchange in which we enquired about ways to address their key criticism as listed below.  
  
First, as Reviewer #2 highlights, it is presently not evident whether the absence of measurable change in behaviour is a result of manipulations that are genuinely without an effect, or a feature of the paradigm. You will need to conduct additional work that demonstrates that the task allows manipulation of human behaviour. This work should be preregistered and powered a priori to detect subtle effects. Please also note Reviewer #1's critique regarding the effect of payoff schemes as you address this point.  
  
Second, as likewise mentioned by Reviewer #2, a key issue is clarifying and strengthening the insights that arise from the work. The effects demonstrated here arise from a comparison between computational models of human behaviour, with little insight into why human behaviour differs from the optimal solution; at a minimum, the goal of revision should be to convincingly demonstrate that commonly used implementations of the model produced results that are artificially interpreted as over/undersampling.  
  
Finally, Reviewer #1 provides a number of constructive suggestions for how additional analyses would strengthen the evidence and generate a more complete understanding of human behaviour in the task. We ask you to address these suggestions, and at the same time, provide Bayesian statistics or equivalence tests for all null-results, which can otherwise not be interpreted. You will find more information about our guidelines for statistics in the PS.  
  
If the revision process takes significantly longer than six months, we will be happy to reconsider your paper at a later date, provided it still presents a significant contribution to the literature at that stage.  
  
We understand that due to the current global situation, the time required for revision may be longer than usual. We would appreciate it if you could keep us informed about an estimated timescale for resubmission, to facilitate our planning. Of course, if you are unable to estimate, we are happy to accommodate necessary extensions nevertheless.

Reviewer #1 (Remarks to the Author):  
  
In this manuscript the authors examined sampling behavior in full information stopping problems. First, in a pilot study, they showed that the participants’ sampling rate is lower (numerically) than that of a Bayesian ideal observer model that relied on the objective price values (undersampling bias), but higher than that of an ideal observer model that relied on the subjective values (oversampling bias). Then, in their main study, they showed that the sampling rate of the participants did not significantly change across different task features. However, the classification of the participants’ performance (undersampling or oversampling) changed as a result of the model it was compared to: a comparison to model 1 (objective values) resulted in undersampling bias (or no bias at all), while a comparison to model 2 (subjective values) flipped the results. The authors suggest that this pattern of results stems from differences in the prior distributions of the objective (model 1) and the subjective values (model 2).  
  
The results are novel and interesting and the manuscript is well-written and easy to follow. I have a few suggestions that I hope would help to improve it.  
  
1. Payoff schemes – I found the different payoff schemes (Reward 1 & Reward) a bit problematic. The authors mention that “payoff schemes at their most potent cannot switch between under- versus over-sampling biases” (Supplementary Text C). However, the different payoff schemes can almost completely eliminate the bias. That is, to change the classification of behavior from under/oversampling bias to no bias at all.  
  
2. Statistical backup – In several places the authors interpret the results based on visual inspection without backing up their claims with statistical analyses. For example, in Figure 4A the subjective values are presented as a function of the objective prices, and the relationship between them is described as ‘sigmoidal’. However, this relationship should be examined more quantitatively. For instance, the authors could fit a mixed model sigmoidal function to the data and compare its fit to other models (e.g., linear model, decaying exponential function, etc.). In addition, the authors mention that the prior distributions in Figure 4B are different, but do not show a significant difference between them (e.g., using a Kolmogorov–Smirnov test). Finally, in Supplementary Text C, the authors compare different conditions without performing any statistical analyses (for example, is the mean sampling rate in Model 1 – reward 1significantly higher than the mean sampling rate in Model 1 – reward 2? etc.).  
  
3. Additional Analyses/discussions – In my opinion, several statistical analyses/discussions could be added to better understand the data and strengthen the manuscript:  
  
· Learning effects – Did the participants show any learning effects during the studies? For example, did they move from undersampling to oversampling across trials (or vice versa)?  
  
· Bayesian models – Models 1 & 2 provide benchmarks to which participants' sampling rate is compared. Could the authors add a discussion (or analysis) about whether these models can also account for the cognitive mechanisms underlying the behavior of the participants.  
  
· Sampling efficiency – It would be interesting to compare the mean payoff obtained by the participants to that obtained by the optimal models.  
  
· Rating consistency – Each price was presented twice at the rating phase. What was the correlation between the two ratings?  
  
· Subjective vs. Objective values – The authors speculate that participants sampling rate would be affected by the distribution of subjective perception of prices (e.g., £550 is roughly equal to £400, p. 10/ first paragraph). This can be empirically examined by comparing participants the sampling rate of participants with a relatively linear relation between the subjective values and objective prices, to participants with a more ‘curvy’ relation.  
  
· Effect Size – The authors reported only significance levels, but not effect sizes.  
  
4. Code availability – As this manuscript compares human sampling behavior to that of ideal observer models, it would be great if the authors could make the code used to implement the ideal observers publically available.  
  
  
Reviewer #2 (Remarks to the Author):  
  
## Summary  
  
The paper investigates full information optimal stopping problems, specifically when people oversample/undersample in this scenario. The author first hypothesized that the number-based tasks led to undersampling and picture-based tasks led to oversampling (as reported by previous studies). But oversampling was observed in the number-based task in the Pilot study, which indicates that pictorial stimuli may not be the only reason causing oversampling. The goal of the Main study is to isolate which task feature leads to oversampling in number-based tasks. They found the human sampling rate is unchanged in all six conditions (i.e., Baseline, Full, Squares, Payoff, Timing, and Ratings). The conclusion about over versus undersampling is determined by the implementation of the model. The paper implemented two models with different prior generating distributions for the Bayesian optimality model. Model 1 uses objective prices as the prior generating distribution, and Model 2 uses subjective evaluations of prices.  
  
  
  
## Review  
  
The potential contributions of the paper are to show that some task features are insignificant in affecting human sampling biases and that the conclusions about under/oversampling are completely model-based. However, the paper falls short in several places. It is not clear that the study itself was sensitive enough to detect any changes in behavior so it gives low confidence in the conclusion that human sampling is unaffected by task features. The paper does not give any insight as to why there is a difference between the Bayesian optimal models and human behavior; the only conclusion here is that participants under sample the Bayesian optimal model, but even that conclusion, as the paper establishes, is completely model-based. Finally, overall the paper was very difficult to parse. Thus, I cannot recommend the paper for publication.  
  
  
  
  
  
## Comments  
  
1. \*\*Sensitivity to task features.\*\* The author state, “Our Pilot Study shows that oversampling is not limited to tasks that present option values as pictures but can also occur for some tasks using numeric stimuli to communicate option value. Our Main Study then attempted to systematically isolate which task feature leads to oversampling on number-based tasks.” The idea here was that there were several features in the experimental protocol that apparently differed between the current numeric version and the past numeric versions that established undersampling. The Main study then changed these features to see if that could explain the difference between the current result and the past result. But, one reasonable issue here is a sensitivity issue. Could it be that in this particular study, participants are just relatively insensitive to these task features? That is, they should show that the study could actually prompt changes in behavior and that the study can detect it. That would give confidence in the conclusion that participants are insensitive to task changes.  
  
2. \*\*Model 1 and Model 2\*\* A real struggle in this paper is the use of Model 1 and Model 2 and the definition of undersampling and oversampling. Over many, many reads. Here is what I understand. Model 1 is the Bayesian Optimal model with objective values. Model 2 is the Bayesian Optimal Model with subjective values. Many papers showed in these optimal search tasks that people under sample with references to the Bayesian Optimal Model with objective values (Model 1). A couple of papers (Furl et al., 2019; van de Wouw et al., 2022) came out using images instead of numeric values, which meant researchers needed to collect subjective values of the images to run the Bayesian Optimal Model. But, comparing behavior to this model led to the conclusion that people oversample. This paper set up a situation where both Model 1 and Model 2 could be used, establishing that the conclusion of undersampling (with reference to Model 1) and oversampling (with reference to Model 2) is completely model-based. This is useful information, but honestly, it takes a lot of effort for the reader to figure this conclusion out. The writing obscures this because it flips between Model 1 and Model 2. But, there could be a lot more work done to make that clear. For instance, in the Pilot study, it is really confusing to read that “Our Pilot Study shows that oversampling is not limited to tasks that present option values as pictures but can also occur for some tasks using numeric stimuli to communicate option value. Our Main Study then attempted to systematically isolate which task feature leads to oversampling on number-based tasks.” But, as a reader, you are looking at Figure 1 and Model 1 and saying there is no oversampling.  
  
  
  
A deeper issue here is that, in the end, the main conclusion of the paper is that the conclusion of over vs undersampling is completely model-based. But, this is not that interesting because it is true by definition. There is really little to no insight as to why human behavior differs from the optimal math solution. The paper's opening suggests it is going to investigate this, but the studies do not do this. The main study seems to have been designed to do this, but really all we learn is that changes in the experimental task features did not have any impact. And as mentioned above, there is a real issue here that it could be that the experiment itself just can’t show or detect any changes in human behavior.  
  
  
  
The paper could also generate models that closely replicate human stopping behavior and find the best model by comparing the models' results with human behavior. The authors seemed to try to do this with, for instance, the subjective value version of the Bayesian model. But, then, the experiments give us no information about why there are differences between human behavior and the models. Moreover, it isn’t really clear if the Bayesian optimal models would have predicted any changes based on the manipulations in the main study.  
  
  
  
  
  
### Minor Comments  
  
1. The difference between 1 and 10 is not the same as between 1000 and 1010.  
  
  
  
2. "Although full information problems lack the many restrictive assumptions of the secretary problem, they cannot employ such a simple rule to derive optimal performance for comparison with human performance. "  
  
  
  
3. The paper states that secretary problem is a simpler problem of full information problems, and the simple mathematical rule of the sceretary problem cannot used to derive optimal performance for comparison with hunman performance. But the sceretary problem is known as the "no-information game", the assumptions of the no information game are very different from those of the full information game. Thus, the optimal solution of no information game is of course not applicable to the full information game.