Dear Dr Marieke Schiffer,

Thank you for your positive assessment of our manuscript, for providing these two helpful reviews and for inviting us to resubmit revisions in response to these reviews.

As described below, we have submitted a considerably revised manuscript, which has been improved in response to these reviews. We apologise for the extended time frame in returning these revisions. We hope it will be apparent that the revisions entailed a substantial amount of new research.

Below we address point by point the editor’s and reviewers’ comments.  
  
  
*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.*

We have added numerous new analyses (including fitting and comparison of theoretical computational models to behaviour) and three studies worth of extra data. The result is that our narrative has changed to reflect our new knowledge. The new manuscript will claim that undersampling bias is susceptible to payoff scheme and sequence length, in part because the optimality ideal observer model is more sensitive to these factors than participants are. We find that participants’ limited willingness to adjust their sampling rate (relative to that of the ideal observer) is explained (via computational modelling) by biased expectations / belief about the distribution of future option values (i.e., our biased prior model). Our new analyses now actually find that the use of subjective and objective values in modelling the ideal observer and theoretical models produces similar biases.

We thank the editor and reviewers for encouraging us to drill deeper in our data set and to bring in new data, as (in our opinion) we stand on more rigorous ground in the current report and believe the revised manuscript can make a more substantive contribution, especially in terms of theory. We were asked to provide extra data by Reviewer 2 and a theoretical model comparison by both reviewers. Having done this, the new manuscript we think is much improved.

*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.*

Our new pre-registered Study 3 shows that sequence length can modulate participants’ sampling rate. This finding, in fact, is itself a replication of Costa & Averbeck (2015), upon which our study’s design is principally based.

We contend that if participants’ show a rigid sampling rates, that would be a legitimate and novel finding that would be surprising enough to warrant reporting. This would be true even if we had not been able to find a manipulation that successful modulates participant sampling rates.

We would not expect an interesting and novel finding that participants are reluctant to change their sampling rates to deserve rejection on a purely *prima facie* basis. Neither the editor nor the reviewer identified in their reviews any flaw in our design that would prevent measurement of changes in participant sampling rate.

*Please also note Reviewer #1's critique regarding the effect of payoff schemes as you address this point.*

Given our new data and analysis, we now agree with Reviewer 1 that the payoff scheme has important influence over the ideal observer’s sampling rates and our narrative has changed throughout to reflect this.

*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.*

We believe that our new model comparison provides this “insight into why human behaviour differs from the optimal solution”. We tested a number of theoretical computational models, each of which implemented a bias term that specifies how normally-optimal performance might be systematically skewed. We fitted and compared these models in eleven conditions / datasets and obtained remarkably replicable results, which implicated prior expectations of option quality as the computational factor that leads to bias.

*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.*

Thanks to the editor and this reviewer for encouraging us to take a deeper look at our data. We now print the results of Bayesian pairwise tests on our plots throughout the manuscript. However, we suspect it is the Bayesian tests in the new Figure 2 that might most directly address the reviewer’s concern. Here, we see that (with the exception of the full condition, which uses a different payoff scheme), participants’ mean sampling rates were statistically equal (greater evidence for the null model) for all other pairwise tests between conditions. This suggests positive evidence that none of the methods feature we tested in these conditions changed participants’ sampling rates.

*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.*

We appreciate the reviewer’s positive response and constructive attitude towards improving the manuscript. Please note that the narrative has changed from what is described here, as a result of new analyses and new datasets that have been introduced in the response to the reviewers’ and editor’s comments.

*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.*

Our new data and analyses show that the reviewer’s perspective here turned out to be quite prescient. Indeed, our new narrative now reflects a more crucial role for the payoff scheme. It appears from our new analyses that using a payoff scheme that rewards only the top three ranked choices drives up the ideal observer sampling rate (relative to the “full” condition that rewards all chosen ranks by the magnitude of the chosen option value) and seems to be the main factor that modulates the undersampling bias, as participants are not as sensitive to the payoff scheme as the ideal observer is.

*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.).*

In light of the changed results, these comparisons of subjective versus objective values are now moot and so no longer reported in our manuscript. Our newly analysed results (e.g., the newly-added Studies 2 and 3) suggest that the use of subjective and objective values in the ideal observer model produces relative small and inconsistent differences, compared to factors like the payoff scheme or sequence length.

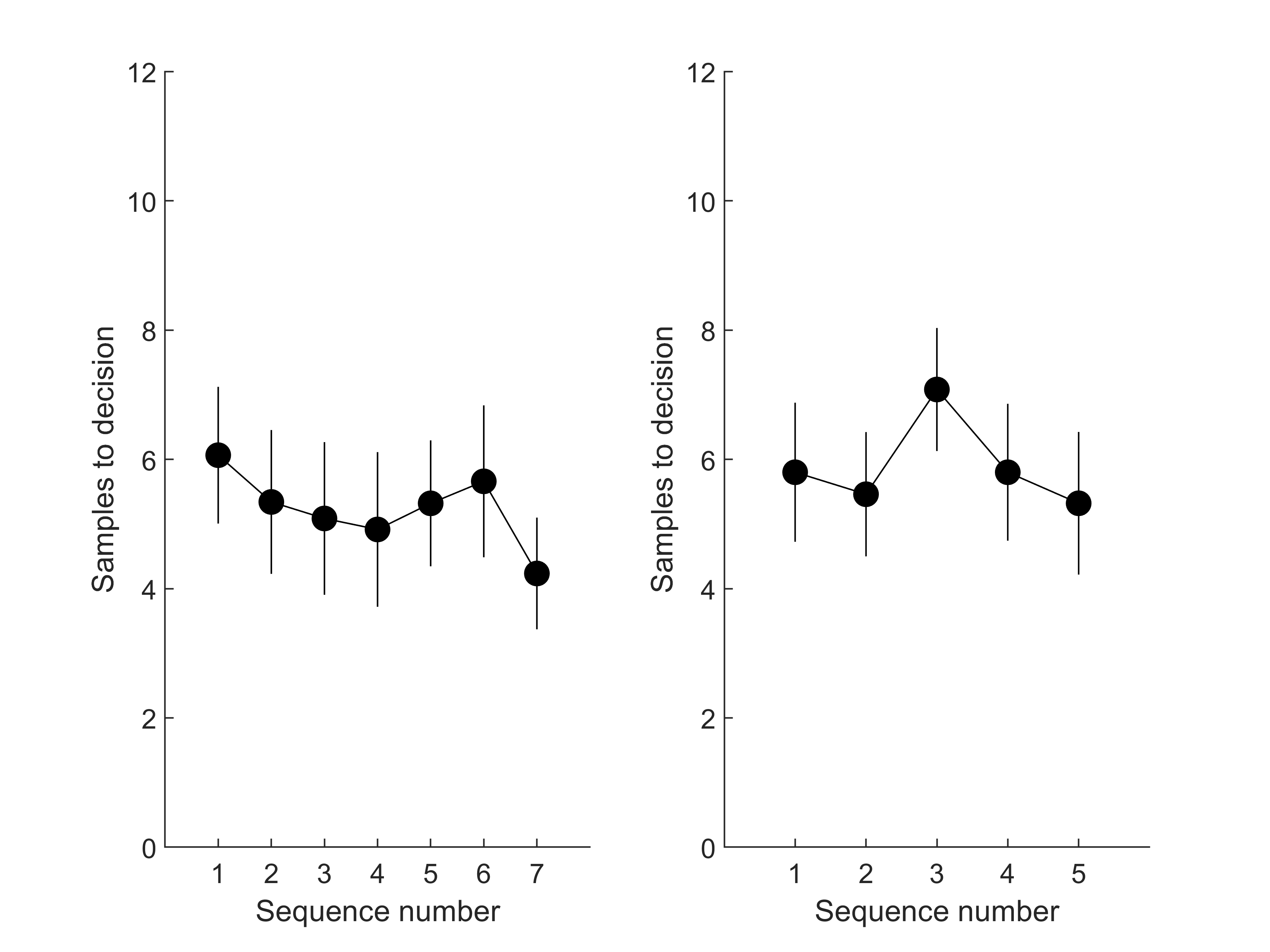
*3. Additional Analyses/discussions – In my opinion, several statistical analyses/discussions could be added to better understand the data and strengthen the manuscript: Learning effects – Did the participants show any learning effects during the studies? For example, did they move from undersampling to oversampling across trials (or vice versa)?*

As with all of our studies using full information problem paradigms, and as has been previous reported in the literature (Lee, 2006), there are no compelling learning effects to report in our data. We have added some content to the Discussion stating this.

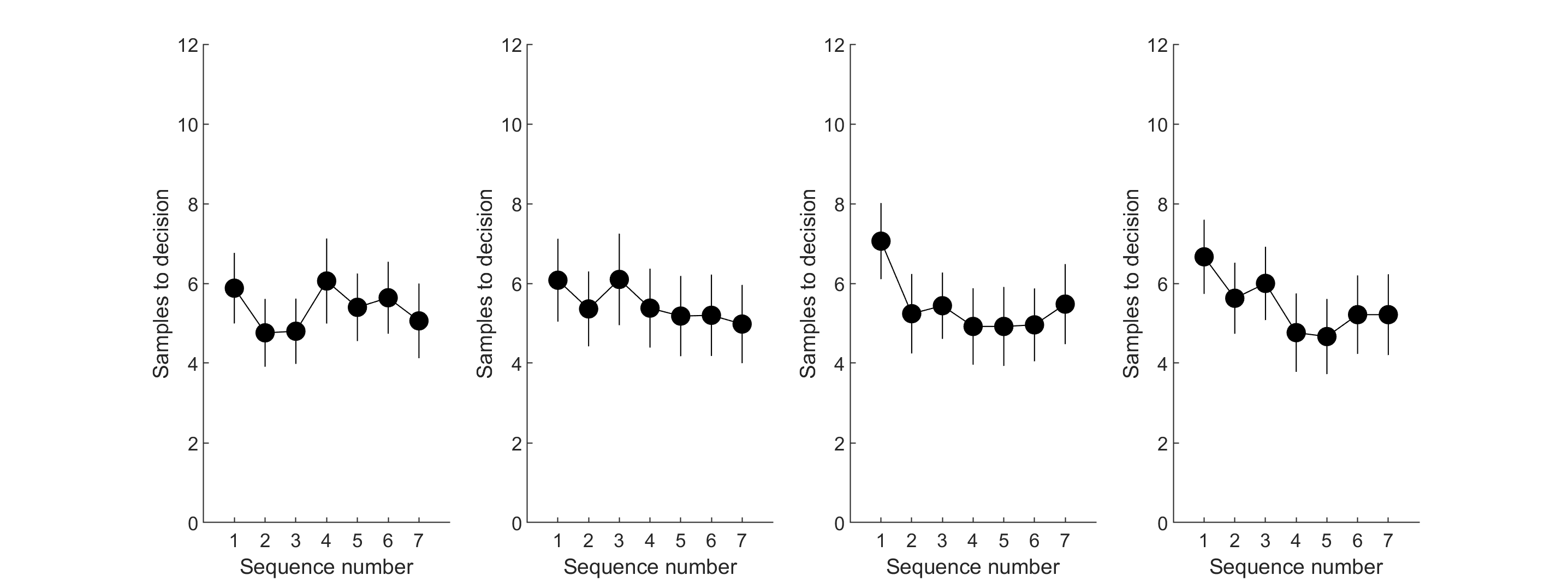
We show below plots of participants’ sampling rates in each of our studies as a function of sequence number (As the ideal observer treats each sequence independently, there can be no ideal observer differences in sampling rate). In some of these plots, there appears to be a hint of a slight decrease in sampling rate, but any such effect appears small (mainly within the error bars, which are 95% confidence intervals of each mean) and do not replicate well over studies. We feel it would be a risk to attempt to draw an inference on the basis of these results.

Moreover, while learning effects are certainly an interesting research question in their own right (and presumably the reason they have already received attention in the literature before this), it not clear to what extent learning effects would directly support or deny our main conclusions.

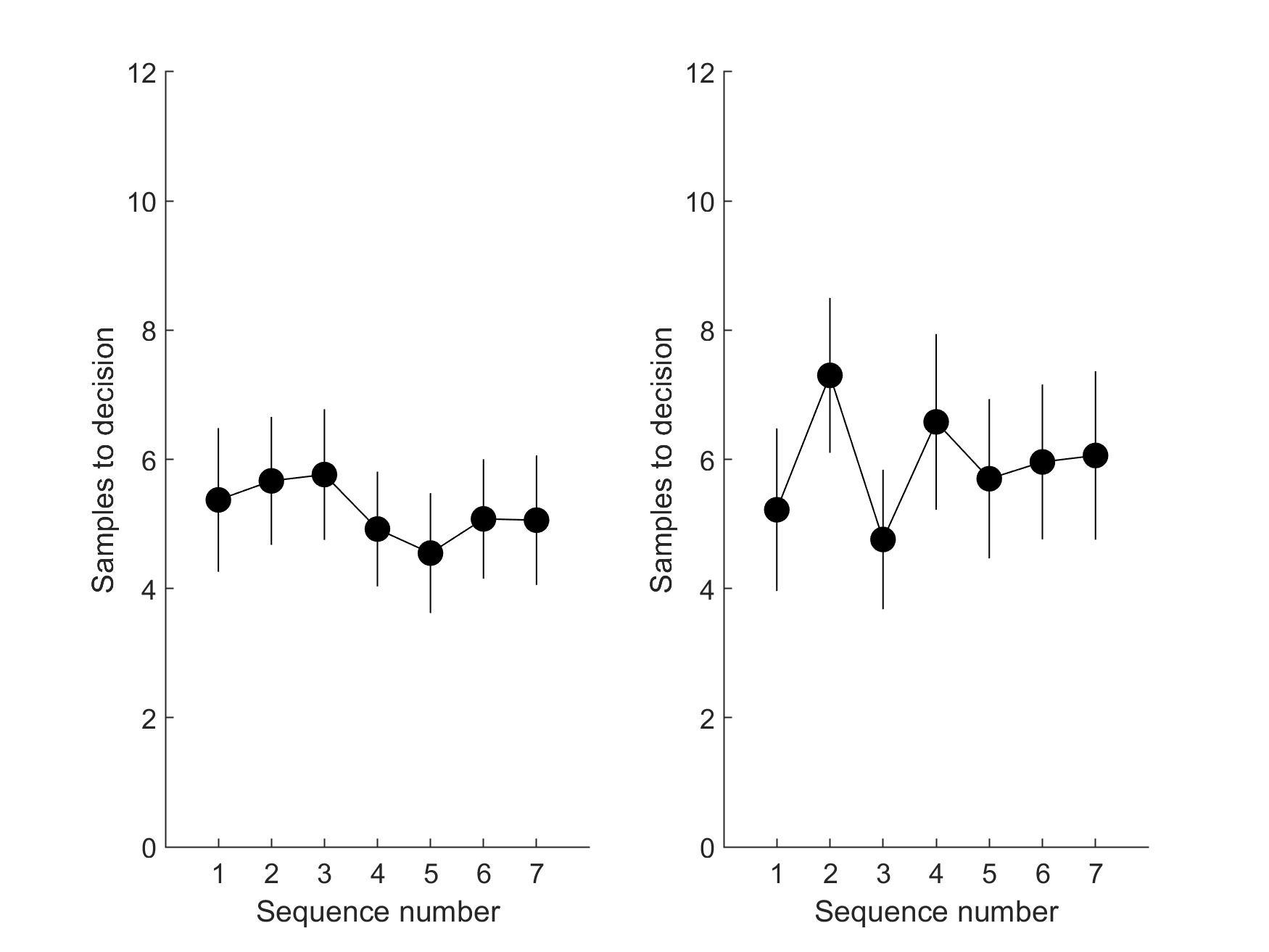
Pilot baseline and Pilot full



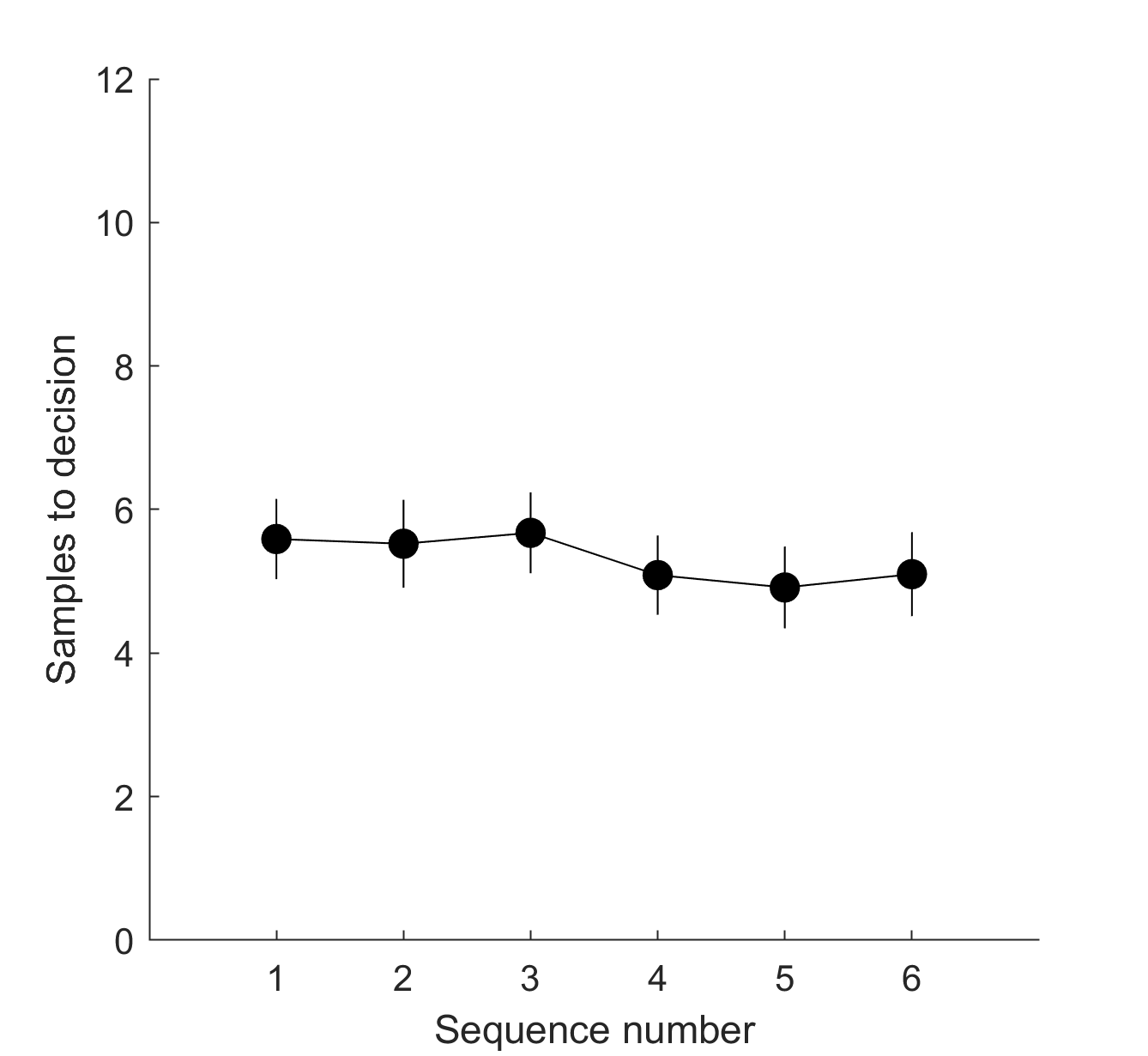
Study 1 baseline, squares, timing and payoff



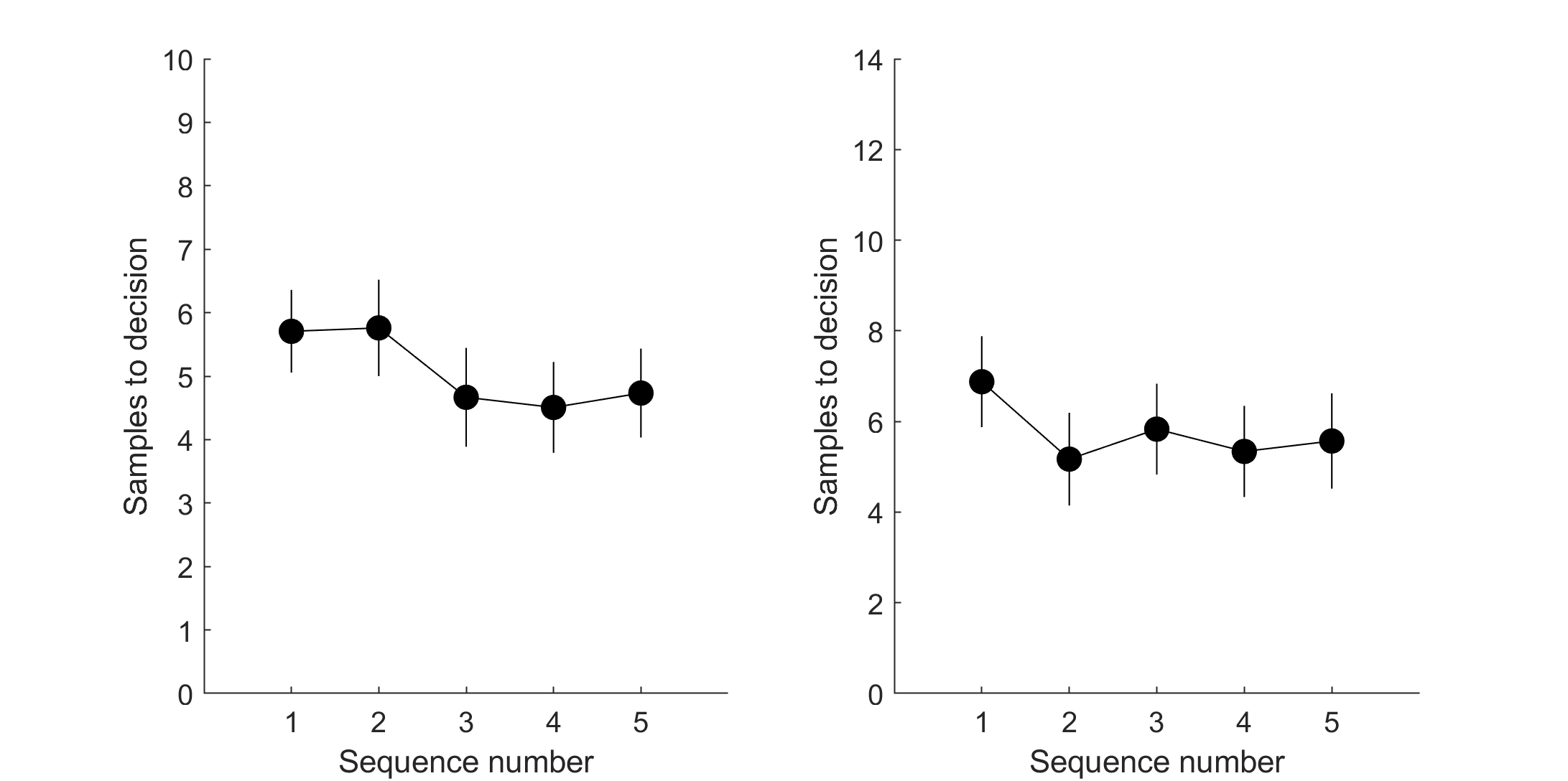
Study 1 ratings and full



Study 2



Study 3, sequence length 10 and sequence length 14



*·*

*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.*

The largest change to the new manuscript is the introduction of theoretical computational models. In response to this request by both reviewers and the editor, we have included some of our new work, which was indeed designed exactly to “*account for the cognitive mechanisms underlying the behavior of the participants*”. We show that errors on this task are theoretically accounted for by biased expectations of future option values.

*·Sampling efficiency – It would be interesting to compare the mean payoff obtained by the participants to that obtained by the optimal models.*

In the Supplementary Materials, we have added plots of the mean rank of the chosen prices for each condition / study. This new analysis adds an important dimension to our narrative, as it can be seen that participants perform close to optimally, despite sampling less than optimally.

*· Rating consistency – Each price was presented twice at the rating phase. What was the correlation between the two ratings?*

The correlations (now reported in the respective Methods sections) are Pilot full: 0.83, Study 1 full: 0.87, Study 1 ratings: 0.81, Study 2: 0.85, Study 3 10 options: 0.88, Study 3 14 options: 0.84.

*· 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.*

As we have changed our claims with respect to subjective versus objective values, this comparison would no longer be key to our conclusions.

*· Effect Size – The authors reported only significance levels, but not effect sizes.*

We report effect sizes for all comparisons involving human behaviour including a new Figure S4 in the Supplementary materials that visualises the effect sizes for all pairwise comparisons in Study 1 (i.e., the null effects, in which participants appear to sample at nearly the same rate over conditions).

*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.*

<https://github.com/nicholasfurl/Model_fitting_hybrid_study>

*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.*

Our two new studies, reanalysis of our existing data and the addition of our theoretical model building and comparison has shifted our narrative and conclusions considerably.

Nevertheless, we still contend that participants adopt a relatively rigid sampling strategy, where many methods and modelling techniques that were attempted did modulate participants’ sampling rate. And, where participant sampling rate was modulated experimentally (i.e., by sequence length), it was still modulated to a lesser extent than ideal observer sampling.

We have added Bayesian tests of the null model to better bolster conclusions based on the null effects of the various task feature manipulations on participants’ sampling that we observed in Study 2. We have replicated a previous report that sequence length can modulate participants’ sampling rate (So our paradigm can successfully modulate participant sampling). And our new model building and comparison provides a possible theoretical and computational explanation for participants’ relatively rigid sample rates.

## 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.*

As requested by the editor, we now report Bayesian tests that show positive evidence for equivalent means between these various methods conditions. Our design was sensitive enough to detect changes in ideal observer behaviour (at least for the different payoff schemes).

*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.*

Our revised manuscript now gives “*insight as to why there is a difference between the Bayesian optimal models and human behavior* “. We have built a number of models that could explain undersampling and, in the new manuscript, we fit them to participants’ decisions. The ensuing model comparison replicates highly across the studies and conditions in suggesting that participants do not increase their sampling rates in the same conditions as the ideal observer does because they subjectively perceive sample costs that the ideal observer does not. Possible reasons why are now discussed in the Discussion.

*Finally, overall the paper was very difficult to parse.*

We are submitting effectively an entirely new paper. Nevertheless, we are of course happy to entertain any specific or constructive suggestions for rewriting.  
  
  
  
  
## 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.*

Costa & Averbeck (2015) using the methods we adapted for our study, already showed that sequence length increases participants’ sampling rates. We now replicate this finding in Study 3. Therefore, some manipulations using this paradigm can indeed affect participants’ behaviour. We have pre-registered and a priori powered Study 3, as requested by the editor.

Our Study 2 is sufficiently sensitive to detect positive evidence for null models where means are equal, using Bayes Factors (Figure 2).

We note that our paradigm methods did suffice to modulate the ideal observer’s sampling rate in response to payoff scheme and sequence length, even if participants showed no modulation (payoff scheme) or less modulation (sequence length), suggesting the design elements capable of modulating sampling for an ideal observer are nevertheless present.

We note that the reviewer has not asserted any specific problem with our paradigm that would prevent detection of participant shifts in sampling. The claim that “*in this particular study, participants are just relatively insensitive to these task features”* appears, at least from this comment, to be made on the basis of the results alone.

*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.*  
  
We suspect that the new manuscript may not pose the same challenges, as we have abandoned the Model 1 / Model 2 notation and, more generally, the new narrative is no longer rooted in comparisons of these two model types.

*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.*

Our narrative has changed to the point where this comment is now moot.

Nevertheless, we stress that findings are not *“true by definition”* whenever they are demonstrated using model simulations. Indeed, our new empirical data, observed in the new Studies 2 and 3 suggest that models using subjective versus objective values actually need not behave in this way at all.

*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 new manuscript is now centred on theoretical model fits and comparisons that give insight into “*why human behavior differs from the optimal math solution*”.

*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.*

We have followed this reviewer’s advice and, indeed, our new model comparison has highlighted a possible computational mechanism for explaining undersampling bias, which replicates well across a number of studies.

As this reviewer prefers, the new manuscript is now centred on theoretical model fits and comparisons that give insight into “*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.*  
  
The ideal observer sampling rates for the different manipulations in the main study (now labelled Study 1 in the new manuscript) are visible in Figures 4 and 5.  
  
  
  
### Minor Comments  
  
*1. The difference between 1 and 10 is not the same as between 1000 and 1010.*

This sentence is no longer in the manuscript.  
  
  
  
*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. "*

It is not clear to us what is being recommended with this quote.

*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.*

We do not claim that the secretary problem is a type of full information problem. To avoid any misunderstanding, the revised manuscript provides more thorough comparison of these problems in the Introduction.

Also, in the Methods, when we introduce the cut off (CO) model, which is derived from the solution to the secretary problem, we critically discuss the merits of Peter Todd’s and Geoffrey Miller’s (1999) claim that the CO model should be applied beyond the secretary problem.