## Reply to Reviews

### Reviewer #1:

*“My concerns about the manuscript are rather minor. Most importantly, I believe that the authors could do a better job at explaining and characterizing the behavior of the ideal observer model. Figure 8 shows that the model tends to fixate over and over again near the centers of the four display quadrants. Why does this happen despite the information update that is achieved with every fixation? It would be helpful if the authors expanded their discussion of this topic, explaining exactly how their target detectability function leads to this behavior. I think this is of central importance to the paper, as it shows the deviation of human observers from "ideal" ones. Najemnik and Geisler avoided such obvious discrepancies by using round search displays, but in the present study, the square shape allows to investigate this point.”*

We have attempted to clarify the description of the model. In particular, we have elaborated on why the model fixated the centre of each quadrant (para1, page 13). However, as R2 points out, there are a number of important differences between our model and the optimal model presented by Najemnik & Geisler, so one must be careful in making direct comparisons. However, we have also collected a small amount of data (three participants) using circular search stimuli, and find that the characteristic “doughnut shape” cited by Najemnik & Geisler as evidence of optimal behaviour in human search is not present in our data (Figure 10).

*Regarding the target detectability function, it took me a long time to at least guess what the authors did. This may be just me, but I believe the authors could explain their computation more clearly. I now believe that the authors included the factors r and x^2 in their linear model to model differential detectability with greater eccentricity in the x- and y-directions (i.e., ellipses rather than circles in Figure 6). Is that correct? If so, is this x then the horizontal distance between target and fixation, i.e., is it different from the x in d\_x,y? Then the notation should be changed, otherwise Equation (1) would be confusing. At any rate, all of this should be better explained.*

We apologise for the confusion. We have decided to change the target detection model slightly, to hopefully make it easier to understand. Now the model is (leaving out the surface roughness factor) p~x^2+y^2, where x and y are relative to the current fixation location. We have changed x,y in equation 1 to (x-x\_f) and (y-y\_f).

*Speaking of Equation (1), regardless whether my interpretation of the authors' computation is correct, it contains some mistakes. Why are beta, b1, b2, and b3 shown as vectors? They are introduced as scalars. Also, I am assuming that the last term should include b3 instead of b2. These problems should be fixed, and there should be more explanation of the variables near that equation.*

b1 and b2 are now vectors, and they are multiplied with beta to give a scalar. That is, for the first factor level of surface roughness, we have beta [1 0 0] and b1 = [-0.17, 1.27, 2.39], so b1xbeta = -0.17. We apologise if there were mistakes or if this was poorly explained in the previous manuscript.

*Furthermore, the fact that only two observers were used to establish the target detectability function weakens the conclusions drawn from the study. While the authors (rightfully) criticize Najemnik and Geisler for only using two subjects, the same criticism applies to their own study. Ideally, they could have obtained data for each of the 7 subjects in their main experiment. Of course this would be difficult to do at this point, but the authors should consider running additional subjects in the preliminary experiment and re-running their models using the resulting, more reliable target detectability function. Otherwise, they should at least discuss and justify this small number of subjects.*

As the reviewer suspects, it is not possible to collect more data from the Clarke et al 2009 study. However, we have taken your advice onboard and collected a whole new dataset (nine observers, with each one carrying out both the Target Detection and Visual Search study). This gives us over twice as much data compared to the original manuscript. Furthermore, it demonstrates that the results (mainly based on the number of fixations required to find the target) are not particular to any one eye tracker or fixation filter. The new data are presented in Experiment 1, while original analysis has been moved to Experiment 2.

*Finally, I think that the title of the manuscript should be changed. The current title does not reflect at all what this study is about. I will not make any suggestions but leave it to the authors to find a more descriptive and accurate title.*

The title has been changed.

*Minor points: Generally, the manuscript could benefit from another round of proofreading. Here are some mistakes and typos that I noticed:*

* *p. 2, first para.: "Hwang, Higgens, and Pomplun, 2009" -> "Higgins" (also in references)*

We apologise for misspelling the name. This has been corrected.

* *p. 3, first para.: "…and to then make a judgement about a dot was that could appear… " -> delete "was"*

Corrected.

*- same para.: "…observers failed adapt their…" -> \*to\* adapt*

Corrected.

*- p. 3, second para." "Over, Hooge, Vlaskamp, and Erklens (2007)" -> "Erkelens" (also in references)*

Corrected.

* *Figures 3a and 4: Units of measurement need to be added to charts.*

This graph has been removed as we deemed it unnecessary as Figure 9 shows the same information

* *p. 8, last line: "Model coeﬃcents are showin in Table 2." -> "shown in"*

Fixed

* *p. 12, first para.: "… both models are both very similar to human ﬁxation data" -> "both models are very similar…"*

Fixed

*- p. 15, second para.: "Our focus in this study is on developing the the stochastic model…" -> delete one "the"*

We have deleted the second “The.”

## Reviewer #2:

*The authors carried out many steps in pulling off this study.  In my opinion, each of these steps has some modest weaknesses. Unfortunately, the cumulative effect of these modest weaknesses is to greatly reduce the value of study. One weakness is that the modeling and analysis is not done within the framework of signal detection theory.  Rather they use the old-style high-threshold modeling approach.  Thus, there is not a proper representation of uncertainty or neural variability, and no possibility of false positives.  This is a departure form human behavior as well as the behavior of signal detection models.*

We discussed this issue at length while working on the manuscript, and in response to this comment we have elaborated on our justification in the manuscript for deciding to use this approach both in the original experiment and in the replication that is now Experiment 1 (see page 4-5). We believe there are advantages of both approaches. As our main objective is to compare human performance to a simple stochastic strategy, we believe our methods are the best choice for our study, as we discuss in more detail below.

*A related weakness is that the authors chose to use a single-fixation search task to measure the visibility maps.  They suggest that this approach is superior to earlier approaches (e.g., Najemnik & Geisler) that use a standard detection task where the target location is known/cued. They seem to miss the fact that modeling search is all about modeling the effects of location uncertainty. The reason for measuring detection limitations without uncertainty is so that the search models can be directed at modeling the effects of position uncertainty.  If they had measured detection for cued locations they would undoubtedly have obtained higher sensitivities. Thus, it is quite possible that they are underestimating the performance of the ideal searcher in this task (also, as mentioned below their model is not a realistic ideal searcher), which calls into question the results in Fig. 7.*

We agree that we may be underestimating the performance of the ideal searcher compared to the model used by Najemnik & Geisler, but we would like to argue that this is not a major concern for two reasons. Firstly, if we have underestimated the target detection function, then this will affect both the ideal and stochastic strategies, and so we would expect the relative performance to stay equal. Furthermore, if there is a difference between the stochastic and ideal strategies it will be most evident on trials requiring a larger number of fixations. Increasing the probabilities of target detection will reduce the number of fixations required to find the target across all conditions, and hence further reduce the non-significant difference between the two models.

We ultimately decided to use an uncertain target location to map visual sensitivity largely because the cued paradigm allows for the observer to deploy covert attention away from fixation, to the target’s potential location. This will also inflate performance in the target detection task and is orthogonal to effect of uncertainty during search to which you refer.

Presuming both methods inflate visual sensitivity to some degree, then the results of the two approaches can be taken together as a lower bound on the number of fixations required to find the target, that is, we could imagine that both strategies should actually find the target in fewer fixations than a human during search. However, if this were the case, then our results would suggest that human observers fail to match the performance of a random walk - a result that seems implausible. Finally, the current performance of our ideal strategy is similar to human performance. This suggests that despite the fundamental differences in approach, our results are consistent with those presented in Najemnik & Geisler (2005, 2008).

*Also, because of their choice of detection task, they were compelled to increase the number of target-present trials to five times the number of target-absent trials (subjects then gave about equal "present" and "absent" responses).  From a signal-detection framework it seems likely that this would cause subjects to shift their decision criteria.  I am not sure what the consequences of this odd asymmetry are for generating predictions for the fixation search task, but there must be some.*

In the new experiment, we have taken a slightly different approach. In the target detection part of the experiment, observers are instructed that there is a target on nearly every trial, and they are just to answer if they can see it or not. We also included 10% catch (target absent) trials. We find that the median false positive rate is 3.88% in the Target Detection study. Both approaches produce models that match human search relatively well, and to a similar extent.

*Another weakness derives from using the high-threshold model in the ideal searcher.  The net effect is that up until the fixation where the target is "detected," the sequence of fixations will be identical given a display with the same target location, and a background having the same roughness.  This is because there is no noise in the model (except for "detect" or "no detect" at the target location; all locations with no target are "no detect") and after each no-detect fixation the posterior probability that the target is at each location is reduced in a deterministic way.  Humans do not do this.  Also signal-detection models (like the earlier ideal search models) have substantial statistically independent noise at each potential target location, and in general have variable scan paths for the same display.  Double pass threshold experiments for the textures in this study would undoubtedly reveal substantial internal noise.  Thus, real ideal searchers would behave considerably different from the authors' ideal searcher, making it difficult to make much of Fig. 9.*

We agree that this is an important difference that we failed to highlight in our accounting of the differences between models in the discussion (page 18-20), and we have added this to paragraph 1 on page 20. As with the other differences, the fact that we find results that are broadly consistent with those of Najemnik and Geisler (2005, 2008) suggests they are worthy of note but not critical to our conclusions.

The main aim of Figure 9 is to show that the two-step saccade amplitudes and directions are similar between the stochastic model and human subjects. That is, there is no need to add an IOR or saccade momentum component to the model beyond what is implicitly included in the distribution of saccades the model samples from. We have made the caption for this figure more explicit.

*The authors acknowledge some of these weaknesses, but suggest that fixing them would not have much affect on the conclusions.  I think this is far from certain.*

We thank the reviewer for their critical comments and believe that we have acknowledged the key points of this discussion in the manuscript (see the introduction to Section 2, Experiment 1, page 4-5, and throughout the general discussion).

As a final point, we would like to again emphasize that the main aim of our paper is a comparison between the stochastic search model and human behaviour, as we believe that Najemnik & Geisler’s original work failed to use a rigorous baseline. Despite any shortcomings in our own optimal model, we believe our findings that a stochastic model matches human search performance could make a useful contribution to the literature, and reconciles previous findings of optimality in search nicely with other recent findings suggesting non-optimality (e.g. Morvan & Maloney 2012).