**We thank the Editor and the reviewers for their helpful and constructive comments on the Stage 2 manuscript. All page and line numbers in our comments below refer to the PDF file (the submission system requires a Word file, but the manuscript was written in LaTeX and therefore formatting etc. may be slightly off in the Word file, which has been converted from a PDF).**

Both reviewers are generally happy, but have things they would like you to engage with for the final version of the paper. I will note that one reviewer spotted an analysis change from the accepted Stage 1 (Section 4.4., adding 'removing incorrect trials' to the data cleaning). A reminder that you cannot alter the pre-registered analyses or text, so please make sure that everything is correctly aligned. Any changes or additions to the pre-registered analysis plan need to be clearly labelled as 'exploratory'.

**Apologies for the confusion here. We always included removing incorrect trials in the supplementary material - see pre-registered OSF, file 3\_pilot\_analysis.html: *“We will import data from the pilot experiment (with pre-processing already in place, e.g. incorrect trials removed, 1st and 100th percentile reaction times removed)”*. We therefore think this doesn’t count as an analysis change. However, we agree that this detail was not included in the main text manuscript of the Stage 1 registered report. We have returned the wording in the main text to the Stage 1 wording, but have added a footnote to indicate that readers should look at the Supplementary Materials for further details of data cleaning. We hope this is an acceptable resolution of this issue, but do let us know if you’d prefer an alternative approach.**

Reviewer #1: This Stage 2 Registered Report is the outcome of a Stage 1 Registered Report that I had reviewed before. As far as I can tell after careful comparisons, the authors did exactly what was pre-registered and I therefore have no objections regarding publication of this manuscript. I've merely a few suggestions on how to improve the exposition (and thus the impact) of the paper and two suggestions for further exploratory analyses that might turn out to improve the interpretability of the present results:

Signed. Heinrich Liesefeld

1. It is relatively difficult to understand what exactly has been done and why - more detail on the analyses and the choices made would help to convince more readers about the quality of your work. Not everybody will try equally hard to understand each step and therefore might be relatively unimpressed in the end. Likewise, the supplemental material would benefit from a bit more commenting and clearer descriptive labeling. I'd suggest that you carefully go through all the material again and try to look at it from the perspective of the reader (without your specific background knowledge on the study and employed methods). I believe that this way you can still considerably improve the quality of the paper.

**We apologise - we have now tried to make both the main text and the supplementary materials clearer (e.g. the supplementary materials should now be more clearly commented throughout) and hope this improves the experience of reading the manuscript.**

2. From a comparison of the last version of the Stage 1-RR, it seems that Fig. 2 (rightmost panel) has changed and the data cleaning now includes "removing of incorrect trials"; this seems fine (even desirable) to me and these changes might have been done after I last saw the manuscript but before the Stage 1-RR was accepted. Nevertheless, I wanted to flag these changes here just to make sure this conforms with the journal's registered-report policy. As a suggestion (addressed at the editor or the respective responsible person at the journal) to improve the review procedure for this new format: You might want to make the Stage 1-RR available to the reviewers in the editorial system; I was unable to get hold of a copy (the OSF link in the paper led to a access-restricted page). Alternatively or in addition, you could ask authors to clearly mark any changes from the Stage 1-RR in the Stage 2-RR manuscript.

**Thank you for picking up on these. In reference to Fig 2: this is because the points plotted here are a random sample of the full number of empirical data points. In order to check everything ran as expected, we re-ran the Bayesian modelling and thus overrode the caching from the Stage 1 report and re-saving the figure. The code is identical and has not been changed since Stage 1.**

**In reference to the removal of incorrect trials: see above for our response to the Editor, but in brief, we always intended to do this. We appreciate you are not able to check the OSF at present, but hopefully the Editor will be able to make the materials available to you and/or confirm the removal of incorrect trials was documented in the Supplementary Materials.**

3. I think the biggest methodological difference that might explain why results differ from the original study is that (some) conditions were randomly intermixed. From the perspective of relative-coding and optimal-template theories, this means that a less ideal attentional set can be set up compared to when conditions are blocked or implemented as a between-subjects design. This should be discussed, I feel. For a great recent paper on these theories, including pointers to the relevant literature, see Yu et al. (2023). Good-enough attentional guidance. Trends in Cognitive Sciences.<https://doi.org/10.1016/j.tics.2023.01.007>

**Thank you, this is a very interesting suggestion and we have now included a paragraph in the discussion to talk more about this possibility (L698).**

4. To test the potential explanation "that because some participants had negative search slopes, the collinear contrast model predicts implausibly large reaction times, due to the mathematical formulation of this model, leading to worse predictions" (p. 31), you could exclude these data points from the analysis of the original dataset and see whether that changes the relative evidence for the various model versions.

**Reviewer 2 also had a very similar comment: please see below for details of exploratory analyses we have now conducted where we have excluded participants with negative search slopes (see Supplementary Materials: Suggestions from Reviewers, and discussion in the main manuscript from L608).**

5. To test your explanation why "search slopes were correlated within feature, but not between" (p. 32), you could try to examine how these correlations depend on the distance between blocks. For example, if this explanation holds, the correlation between Block 1 and 2 should be higher than between the first and the last block of the color part of the experiment.

**Thank you for this excellent suggestion. We have now done this analysis, which is detailed in ‘Supplementary Materials: Suggestions from Reviewers’, and discussed briefly in the main manuscript (see L758 onwards). In short, there still seem to be fairly good correlations between search slopes even when taking block into account, which we think argues against block/time being the only important factor. However, we acknowledge that to test this fully, we would need to design the experiment slightly differently.**

6. I think the last word in "Accepting this model of how the combination process works has theoretical assumptions…" (p. 31) must read "implications".

**Thank you for spotting this, this has now been corrected (L683).**

7. Maybe make clear that with "to consider only trials in which there are a large number of distractors" (p. 31) you still mean various set sizes (not just one), because otherwise you could no longer examine slopes. Furthermore, using many different set sizes, including rather small ones seems important to reveal the logarithmic nature of the relationship (which is of importance for the general model), so I'm not sure whether this is indeed good advice.

**We take both your points here. Reviewer 2 also had similar comments, and we have now removed this suggestion, instead discussing the possibility of altering the target set (see L797).**

Reviewer #2: First, we would like to congratulate the authors on a sophisticated paper. I do not have any qualms with the publication of this paper. Below, I note two relatively important issues that I would like the authors to think about/respond, along with a longer list of minor points. I realize it is a big commitment to embark on a process like this one and I wish it had only been able to produce more compelling results in terms of discriminability between different models (given that all three models ended up being about as likely). I do have some thoughts on this. At any rate, this is a fine piece and one that has definitely taught us some clever tricks.

Signed,

Alejandro Lleras.

Major comments:

1. Feature contrasts were very disparate across the two feature dimensions, with color having overwhelmingly larger contrast levels than shape (thus, stronger guidance).

Table 3: The log slopes reported for shape search are larger than the ones we reported. There is no longer an observable overlap in terms of values between the color and shape mean values. In Buetti et al., one color value was larger than one shape value, and comparable to another one. The stimuli that ended up being used in this experiment show a much larger processing efficiency in color (at all similarity levels) than shape. Perhaps this can be discussed both at the theoretical level, but also, perhaps, at the mathematical level. At the mathematical level this matters because feature contrasts are inversely related to slopes and all contrast combination formulas, thus, amount to sums of inverse values (raised to some power…. The max formula is essentially what happens when the exponent approaches infinity). If one does not have values that are in the general vicinity of one another, but, as is the case here, are almost an order of magnitude larger), then the inverse sums are going to be disproportionately determined by one set of values. To be more specific, 1/Dc values are going to be much larger for pink and purple (an order of magnitude larger, specifically these values are 23 and 66) than they will be for orange, triangle, diamond and circle (6, 3.9, 5, 5). As a result, the contribution of the shape term will be generally much smaller to the sum of contrasts (irrespective of the formula used). And in fact, this problem becomes more pronounced when the inverse values are squared (1/Dsquare). This means that perhaps the modeling approach might have been inconclusive because it becomes harder and harder to differentiate linear and orthogonal models when one subset of D values is almost an order of magnitude larger than the others. One possibility to test this idea might be to redo the predictions for individual RTs using only the D values that are generally in the same vicinity, so only use the D orange (in the color domain) and then use all three D values in the shape domain. It would be interesting to see if the results are more conclusive then.

**We have carried out an exploratory analysis (documented in ‘Supplementary Materials: Suggestions from Reviewers’, and also referenced in the main text at L608) where we use only data from orange targets (as well as some other conditions, to reduce negative slopes, related to some of your later suggestions) to make reaction time predictions. We don’t see any major overall differences in the findings: again, the orthogonal contrast model seems to do best, but broadly all models are similar.**

As speculation, the authors might also suggest that in situations where observers experience a feature as being substantially more informative than the other, than they might change their search strategy and attend preferentially to that feature. Maybe there is a subset of participants that simply search by color, as a result. Can the authors test for this hypothesis?

**We agree that this could certainly be happening - in the supplementary material (4\_planned\_explanations) we plot the predictions for each participant individually, and at least some participants do seem to be best explained by the ‘best feature’ model which may mean they are searching only by colour. We have now highlighted this more in the manuscript at L745. We can imagine devising a new ‘colour only’ model that uses only the colour information to make predictions about the double feature search slopes, and it would be possible to test whether this does a better job overall across all participants (and indeed, we would be very happy for people to use our data/code base to test this type of prediction if they are interested).**

At any rate, I would like to point out that in the second round of review, we raise with the authors the issue of how to best select slope values to have experimental conditions that could easily discriminate between these models. Part of the relevant comment at the time was:

"It is important to understand that the more the log values are disparate in magnitude, the less

discriminating the model predictions are, mathematically speaking." It was a longer paragraph, of course. I am only pointing this out just to make sure the authors realize this was an anticipated shortcoming of their study design, rather than some posthoc comment. The authors say that they used our same materials, but the search slopes they obtained were much different than ours, with some very fast color slopes, and overly long search slopes. This should have been caught earlier. Perhaps the solution would have been to simply use larger shapes (to increase processing efficiency and thus reduce slopes).

**We remember this discussion and we did try to adjust our paradigm accordingly e.g. we changed the colours that we were using, and included pilot data using our proposed experimental design. It is definitely a good lesson in the fact that even straightforward seeming replication-type studies can have many seemingly insignificant decision points that can change the overall findings! We have included more discussion of the general point you are making at L785.**

2. Negative search slopes (an issue raised by the authors): In contrast, this is not an issue that we had anticipated during the review process for the proposed study, but it is an issue that we have faced in previous studies. The relevant section is on Line 597-600 (Very hard to parse the formulas, by the way. Initially, I could not follow the argument being made). More generally, the issue of negative search slopes can sometimes be addressed by restricting the set size over which the slope is computed (see Xu, Lleras, Shao & Buetti, 2021). When set size increases, display and stimuli factors can produce inter-item interactions that are so strong as to facilitate RTs at larger set sizes (call it a textural facilitation effect), yielding negative search slopes. It is unclear whether the authors included the target only condition in the computation of the slope (in general, one should). But if they didn't, there might be a remedy for these negative values. The key is that, if strong textural facilitation effects are present (and are the cause of the negative values), those effects only appear at large set sizes, so one can overcome the negative slopes by simply computing the slopes over small set sizes (i.e., before those textural facilitation effects arise). This is an easy fix and the authors are invited to try it.

**We do include the target only condition in the computation of the slope (see Supplementary Materials: Registered Analysis - the summary statistics for d1 and d2 (near the top of the document) show that the minimum number of distractors is zero, i.e. we include the target only condition). We would argue that the textural facilitation effects are real effects we see in the data, and therefore our models of how the cognitive processing might work should be able to account for them (or at least we should be very explicit about the boundary conditions!) As we discuss above, we do try removing negative search slopes (via a slightly different method) and don’t see any major changes in the overall conclusions.**

Of course, this only makes sense if the negative slope values arise because of facilitatory effects at the larger set sizes. If the effect is present from set size 1, that is problematic. That would be very unusual in the visual search literature (that RT with no distractors is larger than RT with even the smallest distractor set size). This might be a red flag. If the authors were using the target only condition to fit the slopes and they obtain negative slopes even over small set sizes, this might suggest that the anchor point for the slope function is incorrect. Why might that be the case? This might happen because of different levels of practice with the target-only condition overall compared to the single-feature search task. It is my understanding (I might be wrong) that there were no target-only trials run inside the single-feature block of trials (at least not in the color block?), or maybe if those trials were run, it might be the case that the RTs for the target-only condition in the color blocks were mixed with the target-only RTs that were interspersed across the rest of the experiment. Can the authors clarify if that was the case?

**The target-only trials were intermixed throughout the experiment. The shape-only task (Experiment 1b, in Buetti et al 2019) was run as a single block (due to a different template for search being required) and thus the target-only trials for this condition (white semicircle only) were found only in this block.**

The point here is that, if the RTs for the red semi-circle were combined across the experiment, but the trials for the color trials were only measured at one point in the experiment, it might be that for some subjects, the average RT for the target-only condition across the entire experiment might be an overly inflated estimate of how long it would have taken Ss to respond to those trials during the color-only search block. Let's take an extreme example. Suppose the purple block occurred at the end of the experiment. By that time, participants will have lots of experience in the task and RTs will be overall much faster than if those RTs had been measured during the first block of trials in the experiment. If it is possible to do, I recommend the authors try to solve the issue of negative slopes (even at small set sizes) by making sure they are using an appropriate target-only condition as baseline for those search slopes. Maybe only use target-only trials in the near vicinity of the problematic color-only blocks. Or, if it is determined that this is not feasible and that the culprit for the negative slopes is mostly an over-inflated target-only condition, the other option of course might be to simply compute the search slopes for those conditions excluding the target-only datum.

More generally, it might be noted that the same conditions that are causing trouble for the modeling with the negative slope values (pink and purple, see Figure 5, by causing implausible values when a positive inverse is added to a negative inverse value, an issue that only occurs in the linear combination model) are the ones also causing trouble with the mathematical modeling (point above, because their inverse values are so much larger than the inverse values for the other four conditions). So, again, it might be worth trying out the modeling excluding these two conditions, as exploratory work. Or try to fix the negative slope estimates one way or another. Actually, looking at Figure 5, just excluding purple as a condition might end up helping in both regards (most negative slopes come from that condition and it is also a disproportionally smaller slope value than most other slopes). So, maybe re-try the exact same analysis, but leave out the purple condition? I would say that it might be worth adding this as a supplementary analysis, particularly if the discriminability between the models improves (I am honestly interested in finding that out).

**The target only trials were interspersed across the whole experiment, so we think that the problem described shouldn’t occur. It is possible that the shape-only trials differed based on when that particular block occurred in the experiment, but this position was randomised across participants, so we think it should even out (i.e. some participants might have been slower with the white semi circle target only trials because they encountered that block first, but others would have encountered that block last, balancing out any learning effects).**

**Alongside removing purple and pink targets from the exploratory analysis we carried out above, we also removed ring 1 targets, and then additionally any participants who had still had any negative slopes: as before, this didn’t seem to make any difference to the overall main conclusions regarding reaction time predictions.**

Minor comments:

Table 2: Reminding the reader in the Table legend what the features of the target are might help them understand how target-distractor similarity impacts the slope value. Better yet, the table can be edited to say "contrast" instead of "feature" and on each row say what the target distractor condition was: cyan-blue, cyan-yellow, cyan-orange; and semicircle - triangle, semicircle-diamond, etc.. This places the emphasis on the fact that the slope value is not associated with a feature per se, but with a target-distractor feature difference. Just a suggestion.

**This table is in the Stage 1 part of the manuscript so we are not sure if we are able to change it at this point. If the Editor is happy for us to make these modifications, we are of course happy to do so.**

Formula 3: the square root symbol is off

**We think this is an issue with how the figure is represented in the Word version of the manuscript – please see P.10 of the PDF version where we think it should appear properly.**

Line292: increase

**Thanks for the correction here, we have edited this (hopefully as this is a typographical error it’s okay to fix it, though it’s in the Stage 1 part of the paper).**

Figure 3. Something appears off on the right figure. There are not three different shaded regions, but three are mentioned in the legend.

**As above, we think this is an issue with how the figure is represented in the Word version of the manuscript – please see P.15 of the PDF version where we think it should appear properly. All Figures should also be viewable in our GitHub (**[**https://github.com/Riadsala/single\_double\_feature\_search/tree/main/plots**](https://github.com/Riadsala/single_double_feature_search/tree/main/plots) **- this one is n\_trials.png). We note that this is also in the Stage 1 part of the manuscript, so don’t know how much we are able to change at this point (although if the Editor would be happy for us to e.g. change the colour scheme to make it clearer, we can of course do that).**

Line 447: I tried to access the registered report through the OSF link, but access was blocked. I also tried to access the registered report via the submission portal, but was unable to. I was asked to compare the two manuscripts, but I was not able to do this because I lacked access to the initial one. I don't think this is a big issue at this point, but the authors shouldn't forget to open OSF access to readers.

**We apologise for this - unfortunately, we don’t have control of this OSF (it’s the one owned by the publisher/journal) and can’t set it to public ourselves.**

Line 561: consider citing Wang et al. (2018) here as it uses a very similar methodology and also links eccentricity effects directly to cortical magnification.

**Thank you, we have now included this citation (L567).**

Line 725: We never used "odd one out" in our paper. Please do not use that phrase. It is misleading as it relates to an entirely different search task (oddball task), where the target is not known ahead of time. Processing in that task is quite different (see Buetti et al., 2016 and other papers on the topic). We are also working on a manuscript where we address this task in the color domain. Anyway, not a big deal, but fixed-target search tasks should not be labeled "odd one out". Please change. The authors might be interested to know that Tsotsos' lab, for instance, showed that no AI deepnet system can actually predict human performance in the oddball tasks. So, yes, very different task indeed.

**Apologies, we have changed this phrasing as requested (L767).**

Line 740: This is a misquoting of our theory and paper. That paper actually includes eye-tracking data where the effects of eccentricity (and its interaction with target-distractor similarity) are demonstrated. Not only can it accommodate conditions where peripheral information is insufficient and eye movements are triggered, we even identify the parameter (T0) that is responsible for "guess saccades" (i.e., eye movements that occur prior to the end of peripheral processing). We go on to discuss how this relates to the concept of functional viewing field, etc. So, please edit or take out this section that is mis-quoting our theory and results. Thus, sure, suggesting that contingent displays can improve methods/results in this literature is a totally fine thing to say.

**Apologies, we meant only to refer to the particular experimental settings in Buetti et al (2019) here (where it is indeed argued that this task should be able to be completed with only parallel processing). We have modified this paragraph to make it clear that the broader theory does include eccentricity effects (see paragraph beginning L767).**

Line 775: We appreciate the recommendation to focus on trials with larger number of distractors, but that can lead into the problem of textural effects (as mentioned earlier, this was a challenge we faced in Xu et al., 2021). So, it is not immediately clear that large number of distractors should be a recommendation. A recommendation that could be added is that one should aim to find experimental conditions that best differentiate between the models (major point 1). That might happen (mathematically speaking) when the slope values across the two dimensions have some overlap and/or are generally of the same magnitude. Having a confound (like in the current report) where one feature dimension (shape) is overly much longer to process than the other (color) makes mathematical differentiation of the models difficult (which is why the three models end of doing reasonably well - line 755) but also might create a psychological environment where participants may choose to preferentially attend to the more discriminating feature (color). I think something along these lines should be added as a caveat/limitation to the current study. Maybe if the shapes had been made larger, the shape contrasts would have been easier to process and the shape slopes would have been smaller/more in the vicinity of the color ones.

**Thank you for your comment here, we completely agree that colour may be a very ‘strong’ feature which may lead to different results compared to other feature dimensions (e.g. orientation). We have modified the final suggestions accordingly (see L830) and have also discussed this idea in more depth in a new section (see L785).**