We thank the editor and reviewers for their extremely detailed comments. We have substantially revised the manuscript in light of the reviews, and therefore we begin our response with a high level overview of our approach to some of the key issues that came up throughout the reviewer comments. As such, in some places during our more detailed responses, we refer back to these high level topics. Similarly, as the manuscript has changed quite extensively, there are some places where the comments are no longer relevant as those sections have been removed. We have attempted to clearly indicate where this is the case.

# Editors Letters

*Dear Dr Hughes,*

*First, thank you for your submission of a Registered Report. I have now seen two detailed reviews and read the paper myself. Please note I will email you separately with Reviewer 1’s comments in the original pdf in addition, the copy-paste below created formatting issues.*

*Both the reviewers and myself were generally in favour of the project, but there are a large number of issues that need to be addressed. I’ll summarise the major ones below. The scale of the required revision is very large; however, all of the changes are doable. There are no intractable problems, and so I am entering a decision of ‘major revision required’. Let me emphasise that there are a great deal of things to address, and that I agree that they need to be addressed properly for the paper to be considered for in principle acceptance. The benefit of the RR process is that these issues can be dealt with at this stage; of course, please note that in-principle acceptance and progression to Stage 2 is not guaranteed and any revision is likely to be sent back to the referees for further review. You should, therefore, include a point-by-point response to the reviewers' comments, outlining each change made in your manuscript or providing a suitable rebuttal.*

*Review 1 was a joint submission from the 3 authors of the original paper you are replicating and expanding on. They were exceptionally engaged, and have provided a lot of important things that need to be addressed. They propose some pilot testing of the modified design, and raise some not unreasonable concerns about your power analysis and sampling plan. They also list several areas of concern about your presentation and interpretation of their work and their model; given this is a replication study these issues are particularly important. Finally, they were not convinced that the new approach you are proposing to test is actually doing anything beyond what they did; your model fits and discriminations did indeed seem worse than theirs.*

*Overall, I would suggest pilot data is a good idea.*

*I also think modifying your analysis plan to be a more explicit DV by DV comparison would help; analyse your data with their DVs and yours to see if you can show your analysis is adding the value you claim.*

*In line with this, these reviewers suggested that you refocus the introduction to emphasise the methods and measures vs the underlying theories and I suspect that will be required once other revisions are made.*

*See their reviews (attached) for more specific details. It is very extensive! Note it is repetitive in places, and I there may be some issues they raise that matter more to them as the authors being replicated than to you. However, please make sure that in any revision, you address all the points.*

*Reviewer 2 had some similar issues, but was overall more optimistic about the value of moving to your methods. Again the review is very detailed so I would like to see you address all the points raised.*

*As I say, the required revisions are a substantial overhaul of the manuscript. However, I believe the overhaul is certainly possible and that overall it will result in a better Registered Report. I’m happy to answer any questions you have as you go through the reviews, etc. Please note, given the scale of the revision I have given you 120 days to resubmit, just in case.*

Overview comments

**Computational verification**

We have revised our computational verification to further align our presentation of the original results with those presented in Buetti et al (2019).Previous differences were largely cosmetic (i.e., number of decimal places, limits on graph scales etc). We have also now included Experiments 3 and 4 in this reanalysis. We hope the original authors can agree that we have precisely matched their implementation, and so fully understand their original analysis.

We have chosen to split the supplementary material into two files; the first fully details the computational verification so that it is clear that our methods completely replicate the findings from the original manuscript. We have also added a Bayesian implementation (not multi-level) to provide further stepping stones between the two approaches.

**Modelling approach**

One issue that came up during the reviews was why we took the modelling approach we did. We firstly want to emphasise that our aim with this paper is not to say that previous work is in any way wrong. We hope that by taking a slightly different approach, we can provide more support for the TCS model: we think it is valuable for us to attempt to reproduce key findings from the model, in the spirit of strengthening scientific evidence via independent replications. We think our modelling choices also allow extension of the method to, for example, show uncertainty around parameter estimates and explicitly model variance, which we think may be useful tools for further development of the theory.

We hope that the ‘computational verification’ section now included (see above) provides reassurance that the choice of Bayesian vs. frequentist statistics is to some extent simply a matter of preference: we can essentially replicate the original findings within both frameworks. We prefer Bayesian approaches as they force us to make our assumptions explicit (and because for complex models, they are often easier to fit than frequentist models) but we appreciate that others may have different views.

The other novel aspect of our modelling is in the use of multi-level models, which we think are useful in this case as they allow us to explicitly model the non-independence of data, aiding correct inferences about the factors we are really interested in, while not leading to the loss of variance associated with averaging across observations. Humans rarely behave in an ‘average’ way so we think it is important to try to build models that allow us to capture trial-by-trial behaviour. We have now included more motivation for the use of multi-level models in the manuscript (L255-258).

**Modelling Comparison**

To compare our models, we are calculating the marginal likelihood of the relevant models, and then calculating the posterior probabilities, giving us a probability for each model that represents the likelihood that the model gives us the best prediction. This approach is now generally preferred over AIC e.g. McElreath has commented that “AIC is of mainly historical interest now. Newer and more general approximations exist that dominate AIC in every context” (Statistical Rethinking, 2020). Furthermore, AIC is only valid when flat priors are used (which is not possible with multi-level models) and is not appropriate for Bayesian models as it is based on point estimates (see ‘Understanding predictive information criteria for Bayesian models’ by Gelman et al, 2013). WAIC statistics could be used, but these do not provide any additional information to the LOO statistics already included. In addition, failure of WAIC computation is more difficult to diagnose (see personal communication with Aki Vehtari, Stan developer - <https://twitter.com/avehtari/status/1406990096574955526>). We have now included some extra information in the manuscript to explain why we are not using AIC (L320-323). For further discussion of some of these issues, see Richard McElreath. Statistical Rethinking : A Bayesian Course with Examples in R and STAN. Vol. Second edition, Chapman and Hall/CRC, 2020 (Chapter 7).

.

**The intercept**

One of the points of contention between us and the original authors appears to arise from different terminology around our modelling procedures.

We can think of the pairs of experiments as being training and test sets: Experiments 1a and 1b/3a and 3b are used to “train the model” which is then evaluated in its ability to predict RT in Experiments 2/4x (the “test data”). However, in Buetti et al (2019) this is slightly complicated in that they take the training data to be Experiments 1 and 3 (for calculating D) and the no-distractor conditions from Experiments 2/4x (for calculating *a*). While Buetti et al (2019) describe *a* as the observed reaction time for N\_T = 0, we have described this as a calculation (e.g. mean of RTs): however, this is clearly just a semantic difference which we don’t think is of any major significance for the conclusions.

We agree that Buetti et al.’s method may be appropriate when different participants carry out each experiment: as they identify, the differences in intercept in this case may be due to factors such as participant motivation, which are not interesting to include in future predictions. As such, we have used their methods when carrying out the computational verifications.

An alternative way to tackle these issues is to use a within-subjects experimental design. Thus, one of our proposed modifications to Buetti et al’s workflow is to fit both parameters (*a* and D) per subject on the training data., taking the single feature and no distractor conditions as training data. We will then evaluate our models using the compound feature data. The within-subjects design means that the between-participant differences in motivation etc. should be controlled for (i.e. will be the same across both test and training experiments). Our design also allows each participant to have a unique intercept, (estimated from the training data). This is powerful as it allows us to account for variance (rather than having to ‘average it away’). It also allows us to see the variance between participants in performance (increasingly recognised as an important topic in psychology) while still also being able to identify trends that occur across all participants. Finally, we can also add in extra parameters to our model (e.g. which ‘ring’ the target is on), allowing us to account for other factors that may influence reaction times while still being able to draw conclusions about the primary variables of interest. We demonstrate the use of this within-subjects design using our pilot data.

**New design**

Both reviewers had very valid concerns about running the experiment online. We therefore have made the decision to run the experiment in-person, which will allow for greater control over factors such as screen calibration & light levels, and will enable us to run more trials per participant.

We have also simplified the experimental design, removing the additional ring condition, as suggested by both reviewers, and have run pilot experiments in order to verify that our choices of target and distractor colour and shape (which we have modified in line with the reviewers’ helpful suggestions and comments) will produce useful data that will allow us to distinguish between the different contrast models.

# Reviewer 1

*We are the original authors of Buetti et al. (2019) (Simona Buetti, Zoe Xu and Alejandro Lleras). We want to be up front about the fact that we knew the authors were preparing this registered report and we helped them make sense of our data and equations in the lead up to their modeling efforts. We were flattered and very enthusiastic about the idea of doing more sophisticated analyses on our data and models and also, to have an independent validation of our methodology, and theoretical conclusions.*

*Below, we provided a detailed response to the authors’ proposal. There are several important points that we wish the authors to consider. Perhaps most importantly, we think the authors ought to do some piloting before committing to collecting data from 150 subjects. For instance, it would be best if the authors pick a set of three color values that best help differentiate the models (see comments below), as well as demonstrate that the data can be trusted when samples with as few as 6 observations per condition are used to model individual differences. See our simulations below. We are very skeptical that this will be a useful approach, but we remain open to be convinced otherwise (with pilot data).*

Taking these comments (and also Reviewer 2’s comments) into consideration, we have decided to run the experiment in person, in a lab space. This removes the requirement to have only a small number of trials per condition (as we will be able to make the experiment longer) and will likely also reduce the number of participants required.

See the Supplementary Materials for information about our pilot experiment and L346-351 for information about power analysis.

*There are a couple of places where the authors misunderstood our method, in particular, it is not the case that we ever used the intercepts of the log functions to make any predictions. We always used the target-only condition as the anchor to all of our search functions because it is a good index of all non-search processes that impact responses (including motivational differences between groups).*

We think that some confusion may be arising here, possibly from using different terminology to describe the same thing. It is our understanding (please correct us if we have misunderstood) that *a* in Equation 1 of Buetti et al (2019) “represents the reaction time when the target is alone in the display”. This is the same as in Equation 1 in our original manuscript (which simplified to Equation 2: though please note we have now removed the first equation, as discussed later in this response). As *a* is on the right hand side of the equals sign, it (along with *D*) is used to predict reaction times, within the context of that equation. However, there is a discussion about which values are used for *a* in Experiments 2 and 4. In the section above (entitled ‘Intercept’) we discuss this question in more detail(and why we think our approach may be beneficial with a within-subjects design).

We have added in a sentence to clarify the theoretical role of *a*: “as such, it should be independent of both shape and colour, and can be thought of as the role of non-search processes (such as motivation, motor preparation etc.)” (L238).

Please also see the section above (entitled ‘Intercept’) where we discuss these issues in more detail.

*Most critically, we were not convinced by the superiority of their methodology. The authors did not do a very good job presenting their new methods. For instance, they criticize the Rsq approach, but do not convince the reader that their bridge sampling approach is better (we are not familiar with that technique and we hope that one of the reviewers is). The authors also do not discuss the validity (or lack thereof) of using an AIC metric to compare the models. That’s one of the measures we used and the authors could have also used it to provide a comparison to our own results. Note that the AIC metric is commonly used across many disciplines to compare models. If the authors want to move the field towards better methods, they should provide an explicit argument of why the new method is superior, which was lacking here.*

We have included a short section discussing approaches to model comparison, and provide references to explain our approach (see ‘Modelling Comparison’ section above).

*Also, the authors did not present the other metric we used (the average mean prediction error), which is a measure that helps one measure how closely the predictions are to the observed means. The authors criticized our paper for using Rsq because they argued it was not a good measure of prediction accuracy. But they did not mention that we did include a metric to do precisely that (the average mean prediction error). Ideally, the authors would have used the same metrics as in Buetti et al. (2019) to compare the models. And ideally, the new approach would show better fits to the data, as well as better discriminability between the models.*

We have now included mean absolute error (see supplementary material for both the computational replication, and for the new analysis of the pilot data).

*Unfortunately, we are presented with predictions that the authors acknowledge have a poorer fit to the data and that has worse discriminability between models. To be honest, this is less promising than what one would have hoped for.*

Prediction accuracy between the two approaches is not directly comparable. The original models presented in your paper only account for the mean of means, whereas a mixed effects model accounts for the whole distribution of response times. Direct comparison is also hindered by the different approaches to setting the *a* parameter (see discussion elsewhere in this response - see ‘Intercept’ section above). As we hope is clear from our analysis of our pilot data, this will not be as large an issue when we analyse the real experiment data, as we will be able to estimate a on a participant-by-participant basis, and thus take into account individual differences in non-search behaviour (e.g. motivation) rather than on a group basis, which we think will allow for better model fits.

*Finally, we were a bit disappointed by the manner in which the authors presented our data. It seemed like they went out of their way to always present the worst examples in our data, without presenting any of the successes.*

We apologise for this: we in no way intended to be disrespectful of your original paper, and indeed, we very much like the paper, which was the main motivation for working one the current manuscript. Our original manuscript was focused on motivating our proposed improvements, while avoiding excessive use of figures, but we acknowledge this may have led to the overall tone being more negative than we originally intended. We have edited the new draft to hopefully give a more balanced reflection of your earlier work.

*We ran a total of 10 experiments, with 18 different dual-feature distractor conditions. But a naïve reader who is not familiar with our paper might not realize that (see as an example Figure 3). Although it is valuable to highlight the conditions that did not work out and that they might want to improve upon, it seems more transparent to present the entirety of the data.*

We focused on your first set of experiments (in your terminology: Experiments 1A, 1B, 2A, 2B, 2C) forming 50% of your dataset. We concentrated on replicating this study as it was the first one you presented in your paper, with the other study being described as a replication. However, we have now included Experiments 3 & 4 as well.

*Overall, there are some limitations to the proposed approach. At this point, we are not convinced that the study is ready for data collection (see comments below regarding the proposed methods for the new experiment).*

*Introduction: The current introduction focuses on the theoretical literature of visual search. The present paper being more methodologically focused would benefit from a more direct introduction to the key methods/approaches that the authors propose to use, along with a discussion of the finer points regarding why this approach is better than traditional Rsq and AIC based model comparison approaches. If the goal is to promote better practices for data analysis and modeling, then this should be presented more up front, with the pros and cons. This will allow the reader to better understand the goal of the paper and to better understand the motivation behind the analysis. For example, the authors propose to use a bridge sampling technique, but say nothing about it, no details about how it works, why it is preferred, etc… Given that this seems to be a key contribution of the paper for model comparisons, it would be preferable if readers had all the information needed to make sense of this method within this paper.*

Yes, you are correct: as well as building upon your work, one of the goals of our manuscript is an attempt to present state-of-the-art best practice (to the best of our ability), particularly focusing on ways to translate specialist methods papers into useable analytical procedures for cognitive psychology papers. Many of these tools are unapproachable for people as there are few good examples of them being applied outside of these specialised methods papers. We therefore agree that we should highlight this approach more in our introduction, and have included further rationale and citations for e.g. Bayesian mixed effects models and our approach to model comparison (see ‘modelling approach’ and ‘modelling comparison’ sections above).

*It is also important to note that the authors present the paper as a test of sorts of TCS. TCS was not developed with the goal of predicting new slopes. TCS is a general framework theory of parallel processing, that follows the initial one proposed in 2016, with the goal of explaining performance in a specific subset of visual search tasks: tasks with a fixed target and relatively large levels of target-distractor similarity, such that the search can be accomplished in parallel, through peripheral vision. The RT prediction approach was first presented in Wang et al. (2017) for predicting RT in heterogeneous displays. Lleras et al. (2019) provides additional insights regarding the interpretation of some of the factors used in this approach, specifically, how spatial configurations impact predictions. The first application for predicting search slopes of combined features was developed by Buetti et al. (2019). In this latter paper, we introduced the idea that the contrast between a target template and a distractor is inversely related to the observed logarithmic slope. TCS presents the general architecture that allowed us to develop this mathematical relationship between contrast and log slope, but is agnostic as to how the visual system computes multi-dimensional contrasts.*

We apologise for misrepresenting TCS, and agree we are only looking at a special case of predicting search slopes of combined features. We have rephrased to make it clear that our focus is on replicating your work on the contrast combination method, rather than on the basic principles of TCS (L167). However, we do note that at least some of our proposed modifications (e.g. multi-level modelling, using a lognormal reaction time distribution) could be more generally applicable to the TCS framework.

*Suppose the authors find a different contrast combination formula for color and shape. The key thing to remember here is that such result will not go “against TCS” because the contrast combination formula is not a part of TCS per se. What this finding will mean is that feature contrast for shape and color combine in a different way than what was proposed by Buetti et al. (2019). And this has implications for our understanding of the visual system, but does not directly challenge TCS. On the contrary, it will continue to validate it because as long as feature contrasts are the measures that drive attention.*

*For example, Xu, Lleras and Buetti (2021) explored how shape and texture contrasts combine. The results showed that those two sets of features combine orthogonally (in agreement with other studies in the literature that used different experimental tasks), but unlike how color and shape combine. More generally, these sets of findings add to the understanding of how contrast is computed in the visual system and how it goes on to direct attention.*

As above, we agree and did not mean to imply otherwise. We hope that the changes we have made to address the above comment make this clear. We hope that our manuscript will improve the empirical support for TCS by building further knowledge of the process of contrast combination.

*P 4: Equation 1: Please note that in all our equations we use the natural log (i.e., ln ) not log function (which is based 10).*

Yes, we are using the natural log (which is the default behaviour in R). We have made this explicit in L105 and L215.

*Also, the indexing function is over cases where j=2 or larger, (not L), so the published equation is correct, not the authors’ version here. Apologies for the previous miscommunication. Take as an example the case where L=2. RT= a + (D2-D1)\*ln(Nt - N1 + 1) + D1\*ln(Nt + 1) The index function makes the sum inside the parenthesis zero in the last term of the equation. If the authors were right, even in this simple case, there would be a sum term inside the last term of the equation that would add from i=1 to… 0. That does not make sense. It is to prevent the sum term from not making any sense when j=1 that the indexing function is used.*

*Mechanistically speaking, this equation means that all distractors (Nt + 1) are being processed at the rate dictated by D1 and then, once the distractors N1 have been rejected, the remaining distractors (Nt – N1 +1) continue to be processed at a rate dictated by (D2-D1). At any rate, given that the authors do not use this formula at all in their paper, it might be simpler to move to the one distractor case [RT = a + D ln (Nt +1) ]. Also, a minor typo: the bracket on the infinite sign is by convention always open as infinity cannot be contained.*

We agree that removing this formula will help streamline this manuscript, and therefore we have removed this formula and concentrated on the simpler one that is relevant to our paper.

*Table 1. Perhaps mention that*

*- Wang et al. used images of real-life objects*

*- for Lleras et al. (2019) add: “distractor-distractor interactions can facilitate processing when nearby items are similar to each other”.*

*- Xu, Lleras & Buetti (2021). Extended Buetti et al. (2019) to shape and texture. Shape and texture combine according to a Euclidian metric (orthogonal contrast integration model).*

Thank you for these suggestions; we have implemented these changes (see P5).

*P 6 line 23-26. Missing citation for oriented line study. As stated, it almost reads as if it were one of the TCS experiments.*

This has been fixed (L126).

*P 9 line 11-16. There is a slight misrepresentation here. We do not use a different target-only condition for different distractor types. On the contrary, we fit all of our logarithmic slope functions anchored on the target-only condition, as correctly described on p 4, line 40 of the current manuscript. That is to say, we do not “calculate” a. We simply use the observed RT in the target-only condition. So, it is not true that the RT at set size zero depends on the features of distractors that are not present. In Table 2, a is the same for the color experiment, and it is different for each feature in the shape experiment. That is incorrect. We had only one value for the target only condition in that shape experiment and it is the one that anchored all three search functions.*

We think this is a semantic difference: *a* is not directly observed (it is generated by taking a mean of RTs) so it is indeed a calculation, albeit a simple one. However, we have reworded this section to highlight that *a* reflects group differences, as discussed (L241). Our computational verification shows that we have exactly matched the original methods (as we get exactly the same numbers).

*Did the authors use these intercept values in any of their predictions? Note that, in our predictions, these a values (intercepts of the fitted log functions in the single feature search) are not used to make any predictions. The D values are used, along with the RT0 of the experiment being predicted (because different targets are used in different experiments and the group of subjects is different, one expects different speeds in overall response times, which are not of interest to TCS theory).*

*The misunderstanding might be coming from a distinction between the RT at set size zero (which is included in the fitting of the best fitting log slope) and the intercept value that is reported once the log function is fitted. These are two different values. But, we always used the RT0 as the true anchor of all of our search functions. It is worth noting that the RT0 in the color experiment is allowed to be different from the one in the shape experiment because the targets are different in the two experiments, and because they arise from different groups of subjects. Clarifying this might help the reader as well.*

As above, we have tried to clarify our approach re. the intercept in the ‘Intercept’ section above. We apologise that this was not sufficiently clear in our original manuscript.

*So, in terms of the first proposal : “As such, in our proposed new version of the TCS theory, we will estimate a at the level of an experiment (e.g. Experiments 1 and 2).” That is already the case in TCS.*

In our within-subjects version, we will use *a* from the single feature experiments (i.e. Experiment 1) to predict *a* in the compound feature experiments (i.e. Experiment 2). Hopefully the section about the Intercept (above) makes it clearer how our new methods will differ from yours.

*P 9: Equation 4 is wrong. If Cc,s 2 = Cc 2 + Cs 2 (sum of vector magnitudes, as correctly explained in the text) and D = constant/C Then Dc,s = 1 / sqrt ((1/Dc) 2 + (1/Ds) 2 ) This is a small but important difference. If we focus on the denominator of the equations, the term in our model is sqrt[ (Dc 2 +Ds 2 )/ (Dc 2 \* Ds 2 ) ] Whereas the authors’ denominator is: sqrt [ 1 / (Dc 2 +Ds 2 ) ] Can the authors please redo their model predictions and model comparisons using the correct formula? (if indeed they used the wrong formula).*

Thank you for picking up on this. We accidentally have given the same incorrect equation as you do in Buetti et al (2019) (equation 4 page 9). We see that you submitted a correction to your published paper in April, after we had submitted. We have fixed our equation to match your correction and removed the footnote accordingly.

*P 9: footnote: We communicated the typo in the original Buetti et al. (2019) to the journal and they are in the process of correcting the online version, as well as publishing a correction. When this manuscript is published, this error might no longer be in print. Importantly, the calculations reported in the Buetti et al. paper were based on the correct formula. It was just a publisher error. It would be nice if the authors alerted the readers that the correct formula was used in the computations in the original formula.*

As above: this correction was not available when we submitted, but we have now removed the footnote as it is no longer relevant. We would like to thank you again for the help resolving this query, and in no way meant to imply any mistakes on your part or wrongdoing.

*P 9 ln 52. It might be worth noting that the R2 = 0.915 is for all the experiments reported in the paper, not just the one based on the values reported in Table 2. Or the authors might want to increase Table 2 to include the two sets of color and shape slopes (depending on the target template).*

Apologies for the lack of clarity here: we have now carried out the computational replication using all of your experiments (see tables in computational verification supplementary material).

*P 10: line 46-48: The right panel of Figure 2 has not been presented/explained. So, it is difficult to follow the authors’ line of reasoning. The authors direct the reader to the supplemental materials. There is also an important aspect of the data that the authors are not presenting. In our experiments, we acknowledge that there are multiple factors that can influence a specific RT, other than target-distractor similarity, like target eccentricity (see Buetti et al., 2016 and Ng et al., 2018 for examples of eccentricity analysis), and spatial arrangement of the distractors. So, the goal of averaging among many observations and among many observers in a given level of target-distractor similarity is also to average out all these sources of variability. The Mean Predictions are therefore not aimed at being representative of any single trial RT. Instead, they represent the mean tendency of trials in that condition when these factors average out across multiple observations. To make an example, imagine a trial with only two stimuli (one target, one distractor). RTs will be different if the target is near fixation and the distractor far from fixation than in the opposite configuration (target far from fixation, distractor near fixation). And RTs will also be different if these two items appear near each other than far from each other. All of this variability is averaged in the corresponding group mean for “set size 2”. So, there are reasons to expect that randomly sampling RTs and comparing those individual RTs to the group mean will not work very well (as illustrated in Figure 2, right panel). This might be worth mentioning. One can imagine developing an approach that includes variables that tag all of these factors (eccentricity, inter-item distance, etc…), but that was beyond the scope of the original theory (and would require a lot more observations per subject and per condition).*

You raise some good points about trial-to-trial variability and indeed the estimated sigma parameter in our multi-level model will represent variance due to the factors that you mention (eccentricity, inter-item distance etc.) As you mention, it would certainly be possible to extend the model to incorporate some of these other ‘nuisance’ factors, and one benefit of the multi-level approach is that it makes partitioning out these different sources of variance easier. As an example, we demonstrate in the supplementary materials how it would be possible to include a factor for ‘which ring the target was in’ as a proxy for distance from the centre of the screen. This reduces the model’s residual variance and as such offers a better fit to the data. However, this also adds an additional level of complexity to the analysis which is outside the scope of the analysis plan we wish to formally register in this report, although we have added it as a possible factor to explore further in our exploratory analyses section (beginning L324).

*P 10 line 50: The way the Discussion is written it makes it sound like Buetti et al (2019) only used Rsq as a measure of how well each model was performing. We also presented a measure that is doing what the authors here are proposing: a measure of how close each prediction was to the observed measure. We called it the mean average prediction error, computed by averaging the absolute value of the difference between each specific prediction and its corresponding observed measure. In this metric, the Collinear contrast integration model also won handedly when prediction slopes (mean deviation of 3.41 log units/ms compared to 11.37 and 7.27 for the Best Guidance and Orthogonal models, respectively). We also reported the mean deviation for the collinear model: it was 13 ms, which is a fairly small error in cognitive psychology (and about 5% of the prediction range).*

Yes, we agree that your model makes good quantitative predictions (this is one of the main reasons we wanted to explore it further). Apologies if this did not clearly come across in the previous draft of the manuscript. We have rewritten the relevant section of the manuscript (L152 onwards) to hopefully be clearer about our aims.

We have added the mean absolute error to our computational replication, and in the supplementary material you can see that we obtain the same values as you did in your original manuscript. We also include mean absolute error in our pilot analysis in the supplementary material.

*Figure 3: Top: How did the authors pick these three conditions? In all, we predicted RTs in 18 conditions. Are these meant to be representative of all the 18? Or did they pick the three conditions where the prediction was the worst? It is understandable if they want to do the latter, but this figure is really conveying the idea that our model was awfully wrong. Maybe the authors can present the best three and the worst three conditions? Or present all 18, organized from worst to best fit? This comes back to the issue that they have not yet listed all the conditions we tested. Anyway, a naïve reader will look at this panel and conclude that our paper did much worse than it actually did.*

We have now removed this figure.

*Note that there is one aspect of the data that we have commented on previous papers: the smaller the slope of an RT log function is, the harder it is to fit it properly (the noisier the estimate). 6 Indeed, a quick look at even our first paper (Buetti et al., 2016) reveals that the Rsq for the shallowest log functions are also always the weakest Rsq. This matters because the noisier the estimate is, the worst the predictions it will make.*

We hope that our increased emphasis on Experiments 3 & 4 (where the slopes are less shallow) addresses this point. We have also followed the advice you give elsewhere in your response regarding selecting appropriate colours and shapes for the targets and distractors.

*P 11 line 12: “for example, the intercept, a, is set at the level of each sub-experiment and thus would be expected to be perfect.” Unclear what the authors are suggesting here.*

This has now been removed, as the discussion of *a* has been substantially rewritten.

*P11 line 22-24: “assess model fit by computing the log marginal likelihoods via bridge sampling. This directly assesses how well our values of Dc;s are able to predict RTs, avoiding the issues with the R2 measure”. Please say more about this method as many readers will surely be unfamiliar with it. It might be useful to for example say why it is better than the measure we used (mean average prediction error).*

We have included more details about our modelling approach (see ‘modelling approach’ and ‘modelling comparison’ sections above).

*P 11. Line 28: The authors reference the bottom of Figure 3, which is not a great visualization. For instance, it is difficult to understand why the categorical variables on the x axis are (1) joined by lines in the graph as if a continuous variable,*

*(2) why circle and diamond are close together and triangle is so far away?,*

*(3) why the scale was chosen to minimize differences in the middle and left panel, and finally,*

*(4) why is the y-axis truncated at an arbitrary value ? Why not start at zero?*

*And (5) perhaps more importantly: why are the authors only showing half of the data? The authors are only presenting one set of predictions (from experiment 2? Or 4? It is hard to tell, since the authors do not specify which one they are presenting).*

This figure no longer features in the manuscript.

*A reader would be forgiven for thinking that all the data are presented and that the only differentiating conditions are in the left panel. But that is not the case. We ran two sets of experiments, and when they are all considered, there were more than just three conditions that differentiated the three models. Even if the authors are planning on only testing one set of stimuli (with the cyan target), it would be best if the authors could present the entirety of our data. Note that the issue the authors have with orange in Experiment 2 is not an issue in Experiment 4.*

*At any rate, as a more general comment, the authors seem to be following no systematic rule for re-presenting our data, they seem to pick data that make their argument more convincing, and they are not clearly disclosing which data they are presenting (and which they are omitting).*

As requested above, we have now presented the data for all experiments, including experiments 3 & 4.

*Page 11, Line 11: In retrospect, one thing we learned from our experiments is that in order to have more differentiating conditions, it is important that the D values are somewhat comparable along the two dimensions and as far away as possible from being too shallow (10 log units/ms or smaller). So, rather than a priori saying that yellow, green and blue distractors “should” work, what would be much more useful is to present data with several different colors, data from different shapes and pick the colors and shapes that work best. We encourage the authors to do this prior to starting on the feature combination experiments. This is all the more important given that the experiments will be conducted online, so it is harder to control stimuli characteristics. Moreover, if the authors are planning on using the stimuli of Experiment 2, we would suggest staying away from the orange (9.8, too small) and even the yellow (16 sort of small). The Blue is fine (77) and hopefully a carefully chosen green will have a larger than 20 value. And pick a third color that falls somewhere in between. So, as mentioned, it might be better to run an 7 experiment with several colors (5 or 6) and pick the best three values among them (not too small and not too close to one another). This is one of the lessons we have learned after this first paper and what we are now trying to do as we move forward in similar projects.*

*Alternatively, they could try to follow Experiment 3, which used slightly better color values (15, 22, and 51). 22 and 51 are good values. Maybe add a color that has an even larger log slope (a near red pink or purple)? With regards to shape, perhaps drop the semicircle shape (152) which is perhaps too large to meaningfully contribute to model predictions, so perhaps replace with a shape that is slightly more different from the target (i.e., with a smaller log slope). In Buetti et al. (2019) we were limited to the colors and shapes we could use because we wanted to keep as much as possible the same exact features and distractors in the two sets of experiments (Experiments 1-2 and 3-4) because we wanted to demonstrate these distractors change performance as a function of their contrast to target templates (as opposed to their physical properties/salience). The authors are not restricted in that regard, so have more freedom to pick the best possible sets of features.*

*It is important to understand that the more the log values are disparate in magnitude, the less discriminating the model predictions are, mathematically speaking. As mentioned above, we are also moving in this direction.*

Thank you for this invaluable advice. We propose the following changes, guided by our pilot data, and the above suggestions:

* We will use the colours from Experiment 3 i.e. we will replace our cyan target with a red one.
* We will use the shapes from Experiment 1 i.e. our target will be a red semicircle, and the distractors will be diamonds, circles and triangles.

*P 12 ln 39: It is unclear what the authors mean by “ We therefore tested a linear transform model alongside the three original models for predicting D for compound colour shape stimuli.” What is the model doing? Is this the same as the “log normal with linear Nt” model from the supplementary materials? Because this model was not doing a good job at fitting the data in Figure 3.2 S. See more on this below.*

Given that the manuscript is already fairly complex, we have decided to remove the linear transform model for the sake of clarity.

*P 12. Line 47: why were the values of 53% and 97% chosen. Is that common practice?*

As the reviewers are no doubt aware, the commonly used 95% threshold is ad hoc and increasingly people are using 99% or even smaller thresholds (e.g. Benjamin et al, 2017). The wider the percentile interval is, the harder it is to estimate the boundaries as the computational methods rely on sampling from the tails of the distribution. Therefore, we chose 97% as a compromise to give a wide, conservative estimate for where we can be confident that the vast majority of the posterior mass lies. We chose 53% as a secondary interval as it gives approximately 50% chance of the value lying within the interval. We avoided using 50% as perceptual biases mean that people generally interpret this interval as being more likely to contain the real value than not. We are happy to change to any other arbitrary values if you prefer.

*p. 14 l.38: Typo: easiest*

This figure no longer features in the paper.

*P 14 line 43: Can the authors please provide more details as to how the posterior density distributions for Dc and Ds are obtained? It was not clear to us, since we are not familiar with these analyses techniques. Are they sampling from the distributions shown in Figure 3.3 multiple times, and then fitting a log function to these sampled RTs to produce these D values? One thing that is also not clear to us is what role does the “participant” variable play in these analyses. Are these distributions basically analyzed as if all the data came from a single participant (ie., is subject variability ignored?) Why would that be appropriate? Can the authors discuss this choice?*

We have done our best to provide a high level overview, while also providing pointers towards relevant textbooks for readers who want to understand the methodology in more detail (e.g. MCMC sampling for Bayesian analysis) - see paragraph beginning L138. As our paper is not a tutorial paper, we think that further details would make the main message less clear. We hope that we have struck a suitable balance in this version.

*P 14: line 55: “it consistently predicts values that are too low”. Perhaps direct the reader to Figure 5 here?*

This is no longer relevant in the new version of the manuscript.

*P14-bottom: It might be helpful to compare this Figure 5 predictions with the predictions from our model, to illustrate the differences between the two approaches and how using a different underlying distribution leads to predictions from the collinear model that do not fall on the y=x line (whereas they did fall along this line in Buetti et al., 2019, see Figure 4 as well as Buetti et al.’s Supplemental materials Figure S.3).*

We think that with the updates to our model, our re-analysis of your data now looks much closer to your model fit (see computational verification supplementary material).

*8 p. 15 line 10: “Interestingly, the linear combination model also has a high R2 value” It is surprising to us that the authors are presenting a model that, yes, predicts well here, but does not model correctly the observed data. What is the theoretical value of this model if it cannot account properly for the real RT data and distribution? Put another way, how can they justify pursuing this model given the awful fit to the data in Figures 3.2 and 3.3? Also, we still don’t understand just “what” is this model. How is it computed? What is the underlying theory and architecture behind it? If we had looked at figure 3.2, we probably would have rejected any attempts at pursuing that model given its poor fit and not pursued it any further to make any predictions.*

Apologies for the lack of clarity here: we hope that the new way of showing the data in the supplementary material, and our pilot data analysis, is clearer than the previous way we were analysing this data.

*P 15. Table 3. What exactly is the bridge sampling method? Why is it preferable to say AIC model comparison? Since that is what we used in our paper, perhaps it would be nice for the reader if this method is better explained and any advantages over the AIC model comparison approach is explained. Note that the AIC method is widely used for model comparison across many fields, so readers will really benefit from learning a bit more about its shortcomings.*

*The paper would be improved if the authors presented an AIC comparison approach between the results from their three models. From a naïve readers perspective, it would seem like the Bridge method is unable to detect a likelihood difference between the models, where the AIC method might. So, why should we trust what appears to be a less sensitive method?*

*We took the liberty to attempt an AIC model comparison on their results by computing the relative likelihood using exp((AIC i -AIC min )/2) and the data from Table 3. The results suggested that the Collinear contrast integration model (AIC = 786.86) was 1.67 x 10^18 and 2.43 x 10^7 times more likely than the Best Feature Guidance model (AIC = 870.79) and the Orthogonal contrast combination model (AIC = 820.87), respectively, in accounting for the variance of the observed data. The authors should explain why these results should be trusted and why their method is preferrable, so that the readers will want to adopt their approach.*

See our section on ‘modelling comparison’ above for further discussion of these points.

*P15 line 48: “over a number of simulated new observers” What does this mean? Why are the authors not predicting the real observers? Also, it is still unclear to us how “participants” are modeled in this approach. Where is the between-subject variability modeled? Or are the authors assuming all data comes from the same distribution irrespective of observer? That would be a really strong oversimplification to make, right? Or is this usually how these Bayesian analyses are done? (ignoring subject variability).*

Bayesian multi-level models allow for generation of simulated new observers: as we have estimated the variability across the real observers, we are able to use this information to generate simulated observers following the same distribution.

*P 15 line 55: It is not “at first glance” evident that TCS does better. Where is this visible for the reader?*

This was based on a comparison of Figure 3 with Figure 6, and most of the discrepancy was due to our decision to estimate a from the Experiment 1 and 3 data, rather than from each sub-experiment. In our new version of the manuscript, we have followed your methods for estimating *a* and thus this issue does not occur. Hence, we have removed this section from the manuscript.

*P 16. Line 32-34. The authors present the fitting we did in Buetti et al. (2019) as if we had improperly used a different constant for each experiment that we were predicting. This reveals a critical misunderstanding on the part of the authors. TCS is a “visual search” theory. It is not a theory of (1) stimulus-response compatibility, (2) stimulus affordances nor (3) inter-group motivational differences. Why do we say this? Each sub-experiment in our paper (Experiment 2a, 2b and 2c, 4a, 4b and 4c) has a different group of participants. These groups were run at 9 different times and therefore likely vary in terms of motivation and other non-search related factors that end up determining just how quickly each member of that group ends up pressing the response button, in general. TCS and our combination models are moot on these points. And the important part is that our predictions do not rely on (are agnostic with respect to) these differences. What matters to us is: how does the presence of distractors in the display delay the simple response time to the target? That is why we compute search slopes and why all of our predictions are based on search slopes (not intercepts). Note too that we used two different targets between Experiments 2 and 4, so, there are bound to be different stimulus-response compatibility factors at play with one target perhaps leading to overall faster responses than the other. And the targets are also different between Experiments 1a and 1b and 2, as we they are between Experiments 3a, 3b and 4. Etc… So, this is why TCS does not use intercept values from Experiments 1a and 1b to predict performance in Experiment 2. This is theoretically out of scope for the theory. That said, the easy solution is simply to assume that the reaction time for a given experiment in the target-only condition contains all these sources of variance. This is why our predictions for each sub-experiment are anchored on the target-only RT for that sub-experiment.*

We apologise, we did not intend to imply improper methodologies. We are grateful for the opportunity to discuss in more detail some of these modelling choices that often do not receive much airtime but can make important differences to results and interpretation. We have done our best to align our choices with yours where appropriate e.g. the multi-level Bayesian replication of your experiments now uses *a* set at the level of each sub-experiment, as you do. Thank you for the clarifications which has made this process much easier.

*But yes, a within-subject design can eliminate some of these issues (between group differences) but not all (because different targets are used in the simple feature experiments compared to the double-feature experiments, and it is entirely foreseeable that some of these targets will induce faster responses than others, even in the same participants). This is something the authors should spend some more time thinking about. But it is \*not\* a shortcoming of our prediction methodology. It is a feature, not a bug. (see comment on page 17 “allows a to be calculated in a more principled manner”, again, a should not be calculated/predicted as it is not part of the model, so please refrain from suggesting we did something “unprincipled”).*

As discussed elsewhere in the response, we have made several changes that hopefully clarify these issues (and have removed the phrasing of ‘unprincipled’, which we agree was unnecessary).

In our implementation of the experiment (i.e. the proposed RR), we do think it will be possible to estimate the intercepts on a participant by participant basis (because of our within-subject design) which will allow calculation of variability. We hope that this provides an advance in that it will offer a way to model ‘unknown participants’, and allow us to explore individual differences in this paradigm more closely (which has generally been a relatively underexplored area of visual search research).

*P 16: The model comparison was not visually presented for the RT predictions, nor were the models compared in a similar table as with the slopes in Table 3. Why where they not presented? It is difficult to reach a conclusion from this part of the predictions/analyses when they are actually not presented. The authors argue our method is better but they do not present a common metric where this conclusion can be reached. Maybe they can present all the predicted RTs for each model in a similar way as we did in our paper and present the corresponding statistics for each model.*

We do not necessarily claim that our models offer a better ‘fit’ on all possible metrics, but instead argue that they allow us to predict the full distribution of reaction times over multiple trials and participants - perhaps a slightly different goal to the original TCS (which focuses on predicting mean values) but we think a valuable one when trying to understand the range of human behaviour.

*P 16 line 43: The authors include again a discussion of the new linear transform model, but again, we have questioned the theoretical and empirical validity of that model (above).*

As above: we have now removed the linear transform model.

*P 16. Line 51: “However, our analysis suggests that there is no strong evidence that any of the models tested is superior to the others.” See comment above about AIC. Suppose you had two methods to measure the length of a pencil and you want to compare the length of three different pencils. One method uses is a 12-inch ruler with 1/32-inch tick marks. The other method uses a yard stick with tick marks every foot. If the three pencils are close in length, using the yard stick method, all three pencils might score the same length on the yard stick and thus, one might not be able to conclude which is the longest pencil. Whereas with the 12-inch ruler, one might confidently conclude that one pencil is longer than the other two, by perhaps 1/16 of an inch. The authors argument is that their bridge method is the best tool for the job here, comparing the three 10 models. But, we are left unconvinced by their presentation since we were not given a reason to understand why their method is better. As far as we can tell, they are using the yard stick and the AIC is the 12-inch ruler, that is, a more sensitive method for detecting the best model. And by that metric, the collinear model might still come out as the winning model (see our attempted AIC comparison above). And, as a reminder, in Buetti et al. (2019), we also compared the magnitude of the average mean error for each model and using that metric, collinear model also wins. Can the authors do something similar? That is, use the same metrics that we used to see how their new model predictions score on these metrics?*

See our discussion of model comparison metrics at the top.

*P.16-17, lines 55 and 9-10. Our work was never aimed at describing the RT distributions and it is nice that the authors aim at doing that. However, by their own account, the resulting estimates from that analysis produce poorer predictive fits than our approach that instead uses the group means to fit the logarithmic slopes. We would also point out that we always thought it was a strength of our approach that we could predict performance across different groups of subjects (i.e., use performance from Groups A and B to predict performance in Groups C, D and E), which is precisely where the current method is falling short. Aiming to test everything within subjects might also introduce other problems (see response p. 19 line 44) and also takes some of the “generalizability oomph” away from the results. And finally, just a reminder that we also compared the average mean prediction error (not just Rsq) in our paper, a measure the authors have not grappled with here. Is that a bad method? Worse than their method? Why?*

We agree that this is a strength of your approach, and one of the reasons that attracted us to working on this project. We believe the weakness in our approach is due to the way we were estimating the intercept in the previous version of our manuscript (see above discussions) In the revised manuscript, we have further aligned our methodology with yours which leads to better estimates. Please note that when analysing the planned experiments we will have intercept estimates on a participant by participant basis, and many of these complications will no longer apply.

*2.3 DISCUSSION Page 17, line 19: “Moving from a normal to a shifted-lognormal distribution allows us to accurately model the skew seen in the empirical data, and avoid predicting negative response times.” It is true that the authors have demonstrated an advantage in modeling the distributions of RTs, yet, their approach falls short when it is time to predict new RTs. The onus remains on the authors to explain why using this approach allows for a better understanding of the predicted data. For instance, when comparing the three theoretically-motivated models (best feature, collinear, orthogonal contrast), the collinear model makes the best predictions in terms of slopes when using the Rsq metric. However, their model produces slope values that deviate significantly from the y-x line (the slope is 1.8). In contrast, our modeling effort also identified the collinear model as the best model (i.e., with a higher Rsq value) and the predicted slope values were much more accurate (i.e., they fell on the y=x line). So, again, it is difficult to understand what is the utility of their approach. Finally, we still question the value of incorporating the “linear transform” model because (a) it is not theoretically motivated (the authors did not present a cognitive architecture that produces that behavior) and (b) it does a terrible job in fitting the simple experiments. So, unless the authors better motivate this model, we don’t see the value added to the paper by including it. It seems like it is a “curve fitting” exercise, where they are trying to maximize the fit of the predicted data (while ignoring the fit of the single-feature data).*

We hope that our new approach reduces these issues. As above, we have now removed the linear transform method of cue combination.  
  
*Page 17, line 25: Again, a is never calculated in our approach. If the authors are referring to the intercept of the logarithmic function, it is critical that they realize we never use this intercept in 11 any of our modeling efforts. We always use the RT in the target only condition as the anchor for our models (as described above). There is no need for a “principled” approach here.*

See discussion of *a*, as before.

*Why modeling distributions might not be the best approach to predict slopes. As we mentioned before, when running these visual search experiments, a simplifying assumption is that set size and distractor type are the two most important variables (certainly, the only ones we tend to control as independent variables). However, these are most definitely not the only variables influencing RT. Take the simple example of placing two stimuli (one target and one distractor) in the display. There are 36 possible locations in the display used by Buetti et al (2019), so there are C2(36) = 630 different spatial configurations possible for this “simple” search condition. For set size 4, there are even more possible configurations. Similarly, stimulus eccentricity is known to have a major impact on RTs. There might be also processing asymmetries between the upper and lower visual field, and so forth. The point then is that a single participant views only a small subset (~40-50) of all possible configurations where these “hidden” variables are impacting his/her performance. A second participant might not have a single exact configuration in common with the first. So, the solution for us is similar to what is done in ERPs: averaging a lot of observations with the hopes that with big numbers, variability due to these factors becomes more or less comparable across the experimental conditions of interest (set size and distractor type). In our paper, we used 40 observations per condition and upwards of 20 participants per experiment, allowing us to use about 800 total observations to compute one mean per condition.*

See new section on modelling approach in the introduction. We agree that there are strengths and weaknesses of both aggregate and distributional models, and that there is room for both in different contexts.

*Page 17, line 47: Rather than a priori deciding that “green” will be a better color, as we mentioned above, it would be better if the authors piloted a few different colors and picked the best set of three colors that produce (a) sufficiently different values between each other and (b) sufficient overlap with the values in the shape dimension.*

Please see above for further discussion of this point, and the colours we intend to use, based on your advice and our pilot work.

*Page 18, Line 9: “The TCS model, if it is to be a useful predictor of human behaviour, should generalize beyond the conditions originally tested.” As a reminder, the proposed work is not a test of TCS because the manner in which features combine to produce overall contrasts is not embedded in the processing architecture of TCS. The proposed work is challenging the proposal by Buetti et al. (2019) regarding how color and shape contrasts combine to guide attention, and if another contrast combination works better that would still be in line with TCS. Now, if the authors discovered that contrast was \*not\* driving attention, that would indeed be a blow to TCS. Finally, we could argue that Buetti et al.’s results did generalize beyond the conditions originally tested since we used performance in single-feature tasks to predict performance in dual-feature tasks (in a different group of subjects). And we tested the winning model on a separate set of experiments (Experiments 3a,b, and 4a-c). But, yes, we understand what the authors are trying to say.*

We appreciate that the cue combination methods are really an extension of the TCS model. But we do agree that there have been many tests of TCS model and its generalisability, and we didn’t mean to imply that this hadn’t been done before - just simply that there would be value in testing this with a pre-registered hypothesis, given that TCS is able to make predictions about what should happen in this case.

*Although it is intuitively appealing to propose to test more distractors and more eccentricities, as we mentioned above, this will increase the noise in the observations of all conditions (see simulation presented below, p. 19 line 44). So, we do not recommend the inclusion of this condition (note that, if an additional eccentricity is used, the authors should remember to use the cortical magnification equation we used to compute the different eccentricities, to minimize the crowding effects). Furthermore, the shapes will be less and less distinguishable in the periphery, perhaps forcing participants to make eye movements and introducing linearities into the RTs. At any rate, the authors should consider carefully what advantages there will be from including these new conditions and weigh them against the downsides.*

Thank you for raising these points. We have removed this condition, as suggested.

*P18, line 18: It is wonderful that the authors want to use online data collection. We have done so on other projects as well. However, with online data collection come some limitations in terms of stimulus control. In fact, we ended up dropping the farthest eccentricity for our many of our online experiments. This is because we have no control over the dimensions of the displays and therefore the size of the stimuli. This is something important to consider. Will participants be able to fit four different levels of eccentricities in their displays? With stimuli large enough to be distinguishable in parallel? Maybe piloting some of this is necessary as well.*

In light of this concern, along with R2’s comments, we have decided that we should switch to lab based testing, which should alleviate these issues.

*P 18. Line 28: Point 2: It is unclear what the authors are planning on doing when they say that they will test the TCS model assumption that Nt has a log linear effect… We have presented data in a multitude of papers (reviewed by the authors here) showing that the log approximation are better than the linear approximations, when there is parallel processing at play. We don’t understand just what they propose on testing here.*

We very confidently expect to replicate your results, and don’t wish to cast doubt on them. Given that the vast majority of visual search work is based on a linear rt x set size slope, we believe that there is no harm in verifying again that log is better, in the format of a formal registered hypothesis. We have tried to clarify this in our registered hypothesis section (L280 - 281).

*P 18. Line 34. Point 3: The “linear transform model” has not been presented appropriately, as we mentioned before. Unless it is justified theoretically, this will come across as a curve fitting exercise with little theoretical value. The Best Guidance model is inspired directly by the theory of Guided Search and other feature-centric theories of attention that use a single threshold to reject dissimilar distractors. The Orthogonal and Collinear models have been used since the 70s as metrics for evaluating perceptual similarity (see Garner’s work). Note that the underling mechanism that are captured by our log functions have also a mathematical and theoretical justification, going back to the work of Townsend and Ashby. So, we have tried to develop a theory from the ground up in a principled manner. In contrast, we don’t understand what the linear transform model is or where it comes from.*

As above, we have now removed the linear transform model.

*Line 48-50: What posterior probability? Are they referring to the weights in the bridge sampling method? Maybe the authors could help the naïve readers a bit here.*

See the ‘modelling comparisons’ section above.

*P 18, line 52: Planned Explorations: We really like this approach, but as we mentioned above, including an individual level variable without really taking into consideration the differences in the spatial configuration of stimuli across participants could lead to very noisy and different to fit data. Even within a participant, what are the chances that the conditions seen in the single-feature experiments will be representative enough or overlap enough with the conditions in the dual feature experiment to warrant extrapolation from one to the other. As an extreme example, suppose that in the single feature conditions, a participant see overwhelmingly far eccentricity targets but, by chance, they see near targets in the dual-feature condition…*

As we have decided to carry out the experiments in the laboratory, rather than online, we are now able to run more trials per condition, which we hope will alleviate these issues. Our pilot data suggests that even a relatively small number of observers allows us to verify hypotheses 1 and 2 with the number of trials we are collecting from each participant.

*p. 19 line 44. We are very surprised by the authors proposal to reduce the number of observations per condition from 40 to 6, in an online experiment (which will increase the noise in the data, introduce response time-outs, etc…), but this seems to be associated to the choice of running everything within-subjects. We do not fully understand how this value was estimated using the synthetic data (Figure 7), but we are very skeptical that this will prove to be a successful strategy. Note that what the authors are proposing is in the general ballpark of what we have done (we estimated 40 observations in about 20 subjects, for a total of about 800 per condition, the authors are proposing 150 \*6 observations per condition, so 900 total). But, whereas we were aiming to average out all the variability across these 800 observations, the authors are proposing to estimate a slope for each participant? Based on 5 set sizes? Looking at our data, estimating a slope from these few number trials will not lead to stable estimates of the mean. So, if the goal of the study was to use the entire data to compute a group mean to estimate slopes, we might think that this approach could work (ignoring the fact that between subject variability tends to be much larger than within subject variability). But we are entirely unconvinced that computing 150 slopes based on 6 observations per set size will be successful. Indeed, it sounds like a recipe for decreasing (no increasing) power.*

There should be a trade off between number of participants and number of trials However, as suggested above, we are moving from online to lab based presentation, which gives us the opportunity to present more repeats of each condition to each participant.

A within subjects design in the mixed model framework also allows ‘partial pooling’, which tries to share information between data points in an optimal way, permitting a trade-off between global (mean) performance and individual performance. In the case where an individual’s data is very uncertain (i.e. as in a case where there are a small number of data points for each condition, and thus any individual subject’s data might be quite influenced by a small number of outliers) the estimates for this individual would be ‘shrunk’ towards the global mean. In this sense, we anticipate that the modelling framework should be fairly robust to the issues described, at least once data is collected from a large number of participants.

*To illustrate these points, we ran a few simulations. In the simulation, we took performance from one subject in one of the Experiments in Buetti et al. (2019) who had a good Rsq when the log slope was fitted to the RT by set size function, using the entirety of the observations (40 observations per set size). We only analyzed one target-distractor pairing (searching for the cyan semi-circle among orange diamonds). In this simulation, we randomly picked 6 values from each set size (6 out of 40) and computed the slope and r-sq for that logarithmic function. We repeated this procedure 1000 times. In the first panel, we see a histogram of all 1000 log slope values. As can be readily seen, there is huge variability in the log estimate when only six observations are sampled per set size. The variability here suggests that any attempts at making predictions based on any one of these slopes is likely to fail more often than not. This is without even considering that the authors are proposing to use additional stimulus locations, which will only add to the number of possible configural conditions in each set size condition. The Rsq histogram for this slope values is also shown below. It suggests that most often than not, a log fit through these data is a widely inaccurate representation of the data. The authors have all of our data and can do further simulations to estimate the robustness of slope estimates based on just 6 observations per condition. Initially, we thought that there might be some “savings” in the authors’ proposal to use only 6 observations per condition. Our simulations suggest this will be unwise. And there are not many savings, in fact. The authors propose to use 150 subjects, each with noisy slope parameters. In contrast, to test the same idea, we used only 100 subjects (20 subjects or so per experiment, in five experiments, two single feature and three dual-feature experiments). So, it is not clear what the advantage of the proposed method is compared to ours.*

See above: as we now plan to run the experiment in the laboratory, we are able to run more trials per participant, which should solve these issues.

*P. 20, line 34: We believe the authors should plan to run all the conditions tested in Buetti et al (2019), that is, the conditions with the blue target and the conditions with the red target. If that is not possible, then, the conditions with the red target are preferrable (see comment above). In addition, it would be preferrable if the authors piloted a few different colors and picked the best set of three colors that produce (a) sufficiently different values between each other and (b) sufficient overlap with the values in the shape dimension.*

See above for further discussion of how we have adjusted the methods in line with the reviewer’s recommendations.

*P. 21, Procedure: the instructions for the different conditions are not specified and it is not clear how the different conditions (six single-feature conditions and nine double-feature conditions) will be ran. They authors only mentioned the following: “Participants will be told to search for the target among distractors and report if the semicircle target points to the left or right, by pressing either the ‘1’ or ‘2’ key respectively on their keyboard. They will first complete 20 practice trials where they will receive feedback immediately after completing each trial. In the real experimental trials, participants will receive feedback on their average accuracy and reaction time after each block of 30 trials. Participants will complete 19 blocks of trials (570 trials overall)” .*

*Details about the experiment design are missing. Will the participants first complete the single feature conditions first (like in Buetti et al.), will they be blocked? or will all the conditions be intermixed?*

Apologies; all conditions will be run intermixed. This has now been clarified.

*In the instructions the authors talk about “the target” but they should remember that there are three different target stimuli in the proposed experiment. (1) The target for the color experiment has the same shape as the distractors. (2) The target for the shape experiment should be gray (as are the distractors). (3) The target for the dual-feature experiment is a blue semicircle (or a red triangle target). How will the participants be instructed to react to the different targets?*

Participants will be told to find the unique singleton. The training trials will feature all possible combinations, so they will know what to expect in the main experimental blocks. We have done this because we assume that the target contrast is a saliency based, bottom up metric that should not be affected by contextual priming. However, we admit that this is an assumption, and we would be happy to switch to a block based design if you think that top down influences are important.

*In our experience, participants in online studies tend to move quickly past the instruction. We ended up adding example displays to visualize what the task will be and to improve the changed that participants understand the task. So, having 3 sets of instructions in the same experimental session might proof challenging. If the authors end up lengthening the duration of each experiment (to increase the number of trials), they will also have to think about practice effects and motivation, which is difficult to maintain in online experiments.*

We are now planning to run the experiment in the lab (see elsewhere). This should allow us to deal with any issues regarding not understanding instructions properly etc.

*P 21, line 27: The math does not add up regarding the number of trials. In the text above, the authors mentioned there would be 6 trials per condition. When we do the math, we don’t come up with the 570 trials they report. Can they please break it down?*

Each experiment (1a, 1b, 2a, 2b, 2c) consists of 192 trials - this can be broken down into 3 distractor types (i.e. the three colours in Expt 1a) x 5 set sizes x 12 repeats = 180 trials. We then also include 12 zero distractor trials in each experiment, giving a total of 192 trials per experiment, which gives 960 in total across 5 different experiments. Note that the number of repeats has now been increased, this gives a different total number of trials: we have updated the manuscript accordingly (L385 onwards).

*P.21, line 33: In our online experiments (not Buetti et al., 2019, which were all in lab experiments) we added a couple of other inclusion criteria due to the nature of online data collection. We advise the authors to think about the added peculiarities to online data collection (much higher rates of time-out trials and/or trials where no response is given, unrelated to the difficulty of the trial, but related to other factors, such as distractions, that are common in online data collection. These errors should not be included in assessments of performance. The reason why we mention this is because the authors are planning on only collecting six trials per condition. Compared to in-lab data collection the proportion on time-outs is larger in online data collection. What will the authors criterion be regarding the minimum number of trials per condition?*

We are now planning to run the experiment in the lab (see elsewhere).

*p. 19, line 40: typo. “from BETWEEN-subjects”*

This has been corrected.

*Supplementary materials: In general, the supplementary materials could be improved in clarity by adding some more text to explain what is being presented.*

We hope the revised supplementary materials are now more readable.

*Figure 3.1: Why are the scales so large (up to 10) if the largest values don’t seem to be higher than 2.5?*

*3.1.2: “The plots before show our estimates for average participant to lie.” Not clear. Also, which plots?*

*Figure 3.2: It is difficult to assess whether any of the top three rows are different from one another, aside from the fact that there is a slightly larger shaded area (confidence interval?) in the shifted-log normal model. The fourth model is weird and not well presented. It is really not fitting the data well. The data fall outside the confidence intervals of the model in four of the six panels. So, why is it justified to use a model that so poorly fits the observed data (compared to the top three which all account much better for the data)?*

*Figure 3.4: It might be useful to also visualize the estimate for D that we had used for each of these conditions so that readers can see where the differences exist. Of course, those are point predictions, and not distributions, but it would be helpful nonetheless.*

The supplementary materials have now been changed extensively to reflect our new approach - as above, we hope that this new version is now more readable and avoids the problems highlighted by the reviewer.

*We realize that this is a very lengthy review but we thought it would be best to provide the authors with our best feedback to improve the likelihood of success of this proposed registered report. Signed, Simona Buetti, Zoe Xu and Alejandro Lleras.*

# Reviewer 2

*The project described in is Stage 1 Registered Report attempts to replicate and extend the modeling results from Buetti et al. (2019) within the Target-Signal-Contrast theory (TCS) by Lleras and colleagues who found that search times for a target redundantly defined by two features can be predicted from search times for targets defined by the constituting individual features.*

*Overall, this is a commendable project that might yield interesting novel insights on a topical question. In particular the modeling work promises to provide an advance for theories of visual search. Notable extensions of the modeling work with respect to the original Buetti et al. (2019) paper are: (a) a more sensitive index of model fit, (b) modeling of response-time distributions rather than mean response times only, and (c) taking individual differences into account (exploratorily). Notable changes in the proposed replication are: (a) within- instead of between-subjects design, (b) using a more diagnostic target-distractor contrast. I have a few concerns that I would like to see addressed before recommending this Stage 1 Registered Report.*

*1. I think the paper somewhat misrepresents TCS and - at least at times - seems to reduce it to Eq. 1 (which is the same in the present paper, Buetti et al. , 2019, and Lleras et al., 2020). For example, "The original version of the TCS model is given below: [followed by Eq. 1]" (p. 4); "The Target Signal Contrast theory is built around a linear model for predicting mean reaction times from the logarithm of the number of distractors (see Equation 2). In particular, the TCS theory allows us to predict the value of the logarithmic slope, Dc,s, in this condition based on the corresponding Di in the single feature search experiments." (p. 8). Notably, Eq. 1 merely specifies the relationship between individual contrast signals and their co-active effect when the target stands out by two features. The core of TCS - as far as I can tell - are the simulations of individual diffusors towards a non-target decision boundary (highly similar to Moran et al., 2016, by the way). The current writing, including formulation of the research goals, must be revised accordingly. What the present project aims at (among other interesting goals) is to verify Buetti et al. (2019)'s conclusion that Eq. 1 is the best characterization of contrast signal combinations in visual search within the TCS framework.*

Thank you for these comments: in discussion with Reviewer 1 (the original authors of the manuscript) we have decided to remove Equation 1, in order to streamline the manuscript (as we don’t actually use it for our implementation, as we focus on the simpler case where all the distractors are the same). However, we have tried to make our aims in this project clearer (“in our registered replication, we will attempt to verify the conclusions of Buetti et al (2019), that the collinear contrast integration model does indeed offer the best characterisation of contrast signal combinations in visual search within the TCS framework” - L167), as well as trying to represent the core principles of TCS more clearly, in accordance with your suggestions. Finally, we also now reference Moran et al (2016) when talking about categories of model that share similarities with TCS (“TCS also has features in common with other models that propose parallel identification of all items in a scene, with diffusion based mechanisms for identifying targets from distractors” - L82).

*2. I agree that accounting for RT distributions is an important goal as, for example, suggested by Wolfe et al. (2010).*

*However, "switching from linear regression to a multi-level model" (p. 11) is not the only (or main) reason why taking search-time distributions into account is of interest for theories of visual search. Foremost, the hope is that taking these distributions into account will help us to distinguish models that mimic each other on the level of average response times (see Moran et al., 2016, 2017). For detailed discussions of the issues with basing theories of visual search on search slopes derived from average-RT data, see also Liesefeld and Müller (2020).*

Thank you, we are glad that you agree. Thank you for the suggested references, we have expanded the relevant section of the introduction to now read:  
  
“Taking search time distributions into account is also important for constraining theories of visual search (Wolfe et al, 2010; Liesefeld & Mueller, 2020): for example, they have been used to help distinguish between models that make similar predictions at the level of average reaction times (Moran et al, 2016, 2017). Including subject and trial level data into our implementation of the TCS will therefore further aid model development and assumption testing.” (L132 onwards).

*3. Assessing model ﬁt by computing the log marginal likelihoods via bridge sampling is also important progress. How about also considering more traditional (chi-square-based) measures, such as BIC, if only to provide some model-fit indices that are more familiar to most readers?*

Thank you for your suggestion. However, it is our understanding that many of these more traditional methods are superseded by LOO, not appropriate for Bayesian models, or only appropriate in special asymptotic cases. We have included discussion on this in the manuscript (please see reply to R1 and the “modelling comparison” section at the top of our response for more details).

*4. Given that TCS is based on a diffusion process, which can yield (ex-)Wald distributed decision times (Schwarz, 2001), a Wald distribution might be the more psychologically valid starting point (rather than a normal or a lognormal distribution; see Moran et al., 2013, 2016, 2017). Or is there a justification for a lognormal that is equally psychologically convincing?*

We believe that the Wald distribution is the same as the inverse Gaussian distribution. We did attempt to fit an inverse Gaussian model to this data, but ran into computational problems. Indeed Paul Brukner (the author of the BRMS package) has stated that “Inverse.gaussian is extremely hard to initialize and I would argue not using this distribution in most cases (there is a note about this somewhere in the doc).” on the Stan forums.

[<https://discourse.mc-stan.org/t/model-fails-to-converge-when-using-brms/9062>].

Hence, we opted to use the shifted lognormal distribution which is easier to initialize. While the Wald distribution may give an improved fit, we believe that this trade-off is worthwhile as it will make it easier for others to reproduce our results and rerun our code.

We have added the sentence below that outlines this reasoning:

“Note that a Wald, or inverse Gaussian distribution, would also be a reasonable distribution choice for this data given that TCS is based on a diffusion process e.g. Moran et al (2013): we chose not to use this distribution as it often leads to computational issues, which would make it harder for others to reproduce or build on our approach later.” L273).

*Relatedly, "We demonstrate that moving from using a normal distribution to a shifted-lognormal" (Abstract) misreads as if the original modeling used a normal distribution.*

We have now reworded this sentence as follows: “We investigate normal and shifted-lognormal distributions, and show that the latter allows for a better fit to previously published data.”

*5. A few notes on your statement (which had implications for your modeling): "Unfortunately, this leads to a situation in which the model's prediction of how long participants take to ﬁnd a target when NT = 0 depends on the features of distractors that are not present in the stimulus!" (p. 9): this assumption is not so implausible after all and could be achieved by adaptations of decision thresholds (e.g., in the Moran et al., 2016, model; in TCS this effect must probably be deferred to the second stage). To give another example where features not present in the current display affect search times: The model in Liesefeld et al. (2016) explains variation in physically identical target-absent displays by variation in the propensity to quit search. Thus, such variation in parameters depending on features that are not present in the currently presented search display can be explained by previous experiences made by the observer within the experimental session or block of trials.*  
  
*As Liesefeld et al. also varied the crucial target contrast within participants (but between blocks) and still found "contrast" effects on target-absent trials, I am not so sure whether the proposed within-subject design will really avoid the "issue" of condition-specific intercepts (as suggested on p. 16 of the present manuscript). If this question is close to the authors' heart, it might be advisable to also run a condition with various targets randomly intermixed so that such condition-specific preparation is not possible.*

Thank you for bringing this to our attention. We apologise that the design of our experiment was not clear: we intend to run all conditions completely intermixed (i.e. in the manner that you suggest), to try to ensure that condition-specific preparation is indeed not possible. We have written a section on the intercept above which hopefully clarifies our position. We have also tried to clarify the experimental design in the main manuscript.

*6. I think changes 1 (within-subject design) and 2 (more diagnostic contrast) with respect to the Buetti et al. (2019) design (p. 17) are well justified, but I could not find any strong motivation for changes 3 and 4 (p. 18). Why is adding an extra ring a particularly relevant generalization? As it stands it appears quite trivial to me.*

Our original idea was that this manipulation could provide better evidence for the log relationship of distractors. However, we have now removed this change, based on both your advice and that of Reviewer 1.

*Testing online is more a concession to the pandemic rather than a methodological advance, isn't it?*

In accordance with the concerns of Reviewer 1, we will now complete the experiment in the lab.

*Related to the latter, maybe it would be important to also replicate Buetti et al. (2019) exactly (most importantly: between participants and with the same number of trials) in an online setup to make sure the online-testing as such does not affect the results and the conclusions drawn from modeling. Relatedly, it might be important to show that the inclusion criterion used in the lab does not produce a different bias in participant selection when employed in an online study. I could well imagine that more conscientiousness is required to perform with high accuracy in an online setting. For further potential issues with online testing for this particular study, see Points 9a and 9b.*

Only in part: online experiments also allow for a larger number of (potentially) more diverse participants. As we are interested in individual differences, we initially thought that an online study would be suitable. However, we take your points about colour calibration, etc. Furthermore, Reviewer 1 would like a larger number of trials per participant, which will make the experiment too long for online (where we have found attention spans are extremely limited). Therefore, we agree with your suggestion to carry out the experiment in a lab.

*7. Statistical power depends on both the size of the effect and the amount of noise and, thus, the sentence "we believe that this is due to the feature values used in their study, rather than low statistical power from sampling" (p. 19) does not make much sense to me. To base your "sensitivity analysis around the width of the 97% HPDI when estimating Dblue" (p. 19), might be overly optimistic as you don't know yet how well the new (green) contrast works.*

We have now carried out a pilot study, which allows us to base our sample size off conditions more similar to those we will use in the full experiment.

*8. The script in the supplement contains many custom functions that I was unable to access. Are these contained in the GitHub project? If so, I'd need pointers.*

We apologise if we failed to set the correct permissions on our GitHub project: this should now be fixed. We have also included additional comments in the code to clarify how the helper functions work: the “analysis\_revised” folder contains code for the computational verification, pilot analysis and a ‘scripts’ folder that includes the functions used. We are very happy to try to help with any further specific questions you may have!

*9. Many details on the planned experiments are missing; please revise carefully. Most importantly:*

*9a. As the exact color is so important here (determining the contrast to the target), it is necessary to define color exactly along a commonly used color space (e.g,. CIE-Lab or XYZ). Do you have access to the necessary equipment to measure color (i.e., a good spectrometer)?*

We have access to a spectrometer, and have now included details of the colours used in the Supplementary Material.

*How do you handle this issue for online experiments where you don't know the hardware? At the very least you should ask participants to turn off any special settings. Maybe you can include a brief color calibration procedure in the beginning?*

We are now planning to carry the experiment out in the lab, and have included details of the experimental set up that will be used.

*9b. The items' locations are not sufficiently specified in the methods section. For the online testing, use of a "virtual chin rest" (Li et al., 2020) is necessary to make sure that the locations in terms of degree of visual angle are known. (Relatively) precise location is of high relevance here as TCS (and other theories of visual search) assumes a reduction in processing speed towards the periphery.*

As we are now intending to carry out the experiment in a laboratory setting, we will be able to introduce further control measures e.g. participants will be sitting at a given distance from the monitor. We think that this, along with observation from an experimenter, means that we will be able to get good enough accuracy in terms of target distance/degrees of visual angle without the use of a chin rest (given that we are not recording eye movements). However, if the reviewer would still prefer us to use a chin rest, this would be possible.

*9c. What is randomized within (set size?) and across blocks (target-distractor contrast?) needs to be made explicit (see my Point 5).*

See above - we will randomise completely, so each block can potentially contain trials from any condition. We have attempted to make the experimental design clearer in the manuscript (L388).

*10. In case the prospective advance is considered too incremental, one option could be to compare the model against a saliency-based model. Saliency models assume an integration of target-contrast signals at a search guiding overall (priority) map as well and as such would - given the correct implementation - also predict that search times from single-feature targets can predict search times from dual-feature targets (see the work reviewed in Krummenacher and Müller, 2012). The crucial (and highly relevant!) question would be what the contrast is: between distractors and the search template (TCS) or between the target and surrounding distractors (saliency models). If a top-down influence is desired, the model by Liesefeld & Müller (2021) might be suited as it assumes a (dimensional) weighting of saliency signals with color being a multidimensional feature (i.e. producing contrast signals on at least two color-saliency maps).*

This sounds like a very interesting hypothesis, but we think it is outside of the scope of the current manuscript (which is already quite long given that it is a relatively simple replication study!) However, we will include these ideas in the final discussion for the paper as suggestions for future work (and we would be interested in future collaborations to tackle these questions if the reviewer is keen!)

*Minor issues (in no particular order)*

*11. As you point out that "Contrast Signal (TCS) Theory is a rare example of quantitative model that attempts to predict search slopes for efﬁcient visual search", it might be worth pointing the reader to the attempts by Liesefeld et al. (2016) to account for variation in search slopes at the transition from inefficient to efficient searches.*

Thank you for this reference, we have now included it in the manuscript (see L19).

*12. typo: "As we are moving from the \*between\*-subjects design used by Buetti et al. (2019) to a within-subjects design"*

This sentence has been removed during rewriting of the manuscript.

*13. "usually positive, but occasionally negative" (p. 2) needs a citation; e.g., Rangelov et al. (2017)*

Thank you for this citation, yes, this is exactly what we had in mind. We have updated the manuscript to include this reference (L10).

*14. Other papers that come to mind with regard to "with a range of different search slopes being seen for different types of targets and distractors" are Duncan & Humphreys (1989), Wolfe (1998), and our Liesefeld et al. (2016).*

Thank you, we have now included these references as suggested (L16-17).

*15. Using the '1' and '2' keys for response collection is not necessarily ideal, because the hand position is uncomfortable - "f" and "j" or the two shift keys are better, for example.*

We agree, we have switched to ‘f’ and ‘j’ as suggested (L381).

*16. Instead of referring to "stimuli" as in "corresponds to stimuli in which only the target is present" (p. 4) or "depends on the features of distractors that are not present in the stimulus" (p. 9), it is usually less ambiguous to refer to "search displays", because the individual items within each search display are also often referred to as "stimuli".*

Thank you, we agree that this was not optimal phrasing. We have reworded to display where appropriate, although the specific examples highlighted have now been removed from the manuscript due to our change in approach.

*17. "If we want to distinguish between these three models, we will have to combinations of features that lead to larger target-distractor combinations." (p. 11): a verb is missing and the last word ("combinations") seems not to be the intended word.*

Apologies for this word salad! This section has now been removed, but the same sentiment is discussed in L289.

*18. Footnotes 1 and 2 read very harsh (at least to someone from the visual-search community) - maybe explain why the original versions were incorrect instead of the sentence "The (current) version (given here) is correct." That would also be more informative.*

We apologise if our wording came across this way, we did not intend to be rude - we discussed with the original authors who confirmed that they are typesetting issues (which are correctly reproduced elsewhere!) We merely wanted to be as clear as possible to prevent future confusion. We have now removed the footnote as a correction is now available for the manuscript.

Signed. Heinrich Liesefeld

Additional References

Duncan, J., & Humphreys, G. W. (1989). Visual search and stimulus similarity. Psychological Review, 96, 433-458.

Krummenacher, J., & Mueller, H. J. (2012). Dynamic weighting of feature dimensions in visual search: Behavioral and psychophysiological evidence. Frontiers in Psychology, 3, 221.

Li, Q., Joo, S. J., Yeatman, J. D., & Reinecke, K. (2020). Controlling for participants' viewing distance in large-scale, psychophysical online experiments using a virtual chinrest. Scientific Reports, 10, 904.

Liesefeld, H. R., Moran, R., Usher, M., Müller, H. J., & Zehetleitner, M. (2016). Search efficiency as a function of target saliency: The transition from inefficient to efficient search and beyond. Journal of Experimental Psychology: Human Perception and Performance, 42, 821-836.

Liesefeld, H.R., & Müller, H.J. (2020). A theoretical attempt to revive the serial/parallel-search dichotomy. Attention, Perception, & Psychophysics, 82, 228-245.

Liesefeld, H. R., & Müller, H. J. (2021). Modulations of saliency signals at two hierarchical levels of priority computation revealed by spatial statistical distractor learning. Journal of Experimental Psychology. General, 150, 710-728.

Moran, R., Liesefeld, H.R., Usher, M., & Müller, H.J. (2017). An appeal against the item's death sentence: Accounting for diagnostic data patterns with an item-based model of visual search. Behavioral and Brain Sciences, 40, e148.

Moran, R., Zehetleitner, M., Liesefeld, H.R., Müller, H.J., & Usher, M. (2016). Serial vs. parallel models of attention in visual search: accounting for benchmark RT-distributions. Psychonomic Bulletin & Review, 23, 1300-1315.

Moran, R., Zehetleitner, M., Müller, H. J., & Usher, M. (2013). Competitive guided search: Meeting the challenge of benchmark RT distributions. Journal of Vision, 13(8), 24.

Rangelov, D., Müller, H. J., & Zehetleitner, M. (2017). Failure to pop out: Feature singletons do not capture attention under low signal-to-noise ratio conditions. Journal of Experimental Psychology: General, 146, 651-671.

Schwarz, W. (2001). The ex-Wald distribution as a descriptive model of response times. Behavior Research Methods, Instruments, & Computers, 33, 457-469.

Wolfe, J. M. (1998). What can 1 million trials tell us about visual search? Psychological Science, 9, 33-39.

Wolfe, J. M., Palmer, E. M., & Horowitz, T. S. (2010). Reaction time distributions constrain models of visual search. Vision Research, 50, 1304-1311.