Dear Dr Hughes,

Thank you for submitting your work to Cortex for consideration as a Registered Report (Stage 1). Your revision has now been reviewed by expert referees, whose comments are enclosed for your perusal.

Both reviewers are very happy with all your work in the revision, so thanks for your care and attention there. I think we are quite close to this being ready. They both have some final comments, and as these are about issues of design and analysis I wanted to give you a chance to reply to them and incorporate them or not. One way or the other, these are all things that need to be nailed down and seen by the reviewers and myself prior to this being given Stage 1 approval.

So, on the basis of these comments, we cannot accept your manuscript in its present form but would like to invite you to revise your paper, taking into account the issues raised by the reviewers. Please note that in-principle acceptance (IPA) 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.

If you would like to submit a revised manuscript, we request that you do so within three months. Please let us know if you will not be able to meet this deadline.

Comments from the Reviewers:

Reviewer #1: We want to thank the authors for their careful reply to our earlier review. They have done a good job addressing our comments and we are much more confident about the success of the project now. We only have two comments regarding the new methods.

Thank you again for your very helpful comments, we are very pleased that we have been able to address most of your concerns.

In the response letter, the authors say that participants will be instructed to "find the singleton", rather than being told the specific target to look for. We advise against those instructions (and against mixing the single feature trials from two different dimensions). The reason is that there is a literature regarding the differences between searching with a "feature" mode vs searching with a "singleton" mode. They can produce different phenomena (wrt attentional capture, for example) and more worrisome, singleton mode can produce RTs that decrease with set size, particularly at small set sizes (with an effect that may or may not recover... that is, RTs might never begin rising with set size).

We are actually working on two manuscripts where we deal with this issue. But, this observation was already published in Buetti et al. (2016). That paper includes a "singleton" experiment, with otherwise identical displays to the fixed-target-feature experiment, and indeed, RTs at small set sizes can be as much as 150ms slower. Co-author Lleras published a modeling paper that touched on this (Tseng et al., 2014), the idea being that in singleton mode, observers don't know what to attend to, and particularly at small set sizes, they must first determine what stimuli are repeated in the display (and thus have distractor status) before understanding what stimulus is the target in the current display. It is difficult to say whether this would be at play in the current experiment because the target feature variability is much smaller in the current set up (for instance, it never switches to a feature that has distractor status on other trials), but it might make the data a bit tighter if participants are always given a specific instruction as to what the target that they are looking for is. From a theoretical perspective, this is also important because TCS is all about having a well defined target template in mind (against which stimuli in the display are compared to). Anyway, this is a fairly minor change to the methods proposed by the authors. They simply need to group the "single feature" experiments temporally in the experiment and show the observers exactly what the target will look like in the current block of trials, so all search trials are completed under a "fixed-target" search mode.

Thank you for drawing our attention to this literature. We have updated our experimental protocol to take your advice into account (L418).

The second comment relates to the number of observations per condition. It's great that the authors were able to find that 12 observations per condition were enough to provide a good estimate of single-dimension D and that this estimate could be used to predict the bidimensional D parameter, with some accuracy. The only worry that we had was that the computational comparison between the three models compared was unsuccessful. The authors suggest it is simply a matter of power (i.e, having too few subjects in the pilot). That might very well be. But, the other possibility is that the estimated D parameters might be too noisy to produce precise enough Ds for the prediction part of the work, such that the predicted Ds cannot be differentiated by the three models being compared. Maybe this is something worth pointing out at some point in the registration report? That maybe, even though the design is within subjects (thus more powerful), the estimated parameters might be too noisy to produce bidimensional Ds that can be differentiated by the three models. That is, if the study is run and there is no way to separate the models (from a computational perspective), it might very well be because of noisy D estimates at the within subject level. Of course, the alternative, that more subjects will allow the authors to conclusively rule out some models or even find a decisively winning model, is totally possible. But is it worthy hedging their bets a priori and point out this possibility? The authors should perhaps more clearly highlight that this reduction in the number of observations per condition is a departure from our methods (we used 40 observations per set size, more than three times as many) and a potential source of noise in the prediction part of the work.

You raise some good points, and we agree that the discussion about the number of trials is often neglected in power analysis. We note that the collinear model does have the lowest mean absolute error assigned to it in our pilot analysis, as would be expected from your previous work, which is why we made the assumption that the issue lay with the small number of participants tested in our pilot study: but you are of course correct that measurement error could also be a factor. We have carried out some robustness checks (now documented in Section 4.1 of the main manuscript, L371), including re-running our computational verification of the original Buetti et al (2019) data with a subset of their trials and a simulation to estimate the confidence intervals on the mean when sampling from a log-normal distribution. Based on both of these, we think that using 20 repeats should be enough to keep noise to a reasonable level (in conjunction with the within-subjects design).

On a minor note, Figure 3.3 was really messy in the supplemental materials. The equations were partially cut off. The x axis numbers overlapped to the point where they were unreadable... and we wondered whether the titles on each panel were correct (best feature, collinear and orthogonal, in that order?) or is the order wrong? Something worth checking.

We apologise for the formatting issue. Hopefully, this now displays correctly.

Other than that, we are happy with the status of this registered report. Good luck with your data collection and we can't wait to see the results.

Signed,

Simona Buetti, Alejandro Lleras and Zoe Xu.

Thank you, we look forward to discussing them with you in the future!

Reviewer #2: When accepting this review, I informed the editor that, unfortunately, I cannot make the time to provide a full review, because I am currently organizing a conference. But when I started reading the paper, it kept me quite engaged, so that I did more than I had originally planned. However, I did not verify the code, which was made available with this revision, and I did not carefully look at the supplements. I trust that the authors have carefully validated these items themselves.

Basically, all my questions have been answered and all potential issues that I could see have been fixed. Here are merely a few follow-up thoughts:

1. Moving this to the lab indeed resolves many of the problems I saw and will generally strongly improve the quality of the data, I think (even beyond what could have been achieved in an ideally designed online study). One further potential improvement that has become possible due to this decision regards the precision of measuring response times: depending on the polling rate of your input device/port, there might be a considerable amount of across-trial jitter induced by when exactly the key presses are made on a given trial. I'd recommend measuring how strong that jitter is in your specific case and gauge whether it contributes a significant amount of variance to your data. This is more relevant here than in most RT studies, because you plan to analyze RT distributions rather than just means.

Thank you for your comments here: we have decided to use a button box to record responses, as this should give higher precision (L412).

And as you were asking: yes, I'd recommend using a chin rest for this type of experiments; in particular, if you plan to include eccentricity and inter-item distance as predictors at some point (as mentioned in several places; e.g., p. 13).

This is a good point, and we will ensure that participants use a chin rest, as suggested (L408).

2. When discussing the Wald (inverse Gaussian), it might be worth mentioning that it is psychologically somewhat more plausible than the log-normal (if you agree) and how well the log-normal approximates the (ex-)Wald.

We agree that there is a case to be made that the Wald distribution is more plausible than a shifted-lognormal distribution. However, the issue does not appear to be straightforward, for example, see Matzke & Wagenmakers (2009) who state “we conclude that researchers should resist the temptation to interpret changes in the ex-Gaussian and shifted Wald parameters in terms of cognitive processes.”

As already stated in the paper, our main reason for selecting a shifted-lognormal is that it is easier to fit computationally with our toolchain, while still offering a good fit to the data (especially when compared to lognormal and normal). We feel that our analysis pipeline is already relatively complicated compared to the original implementation of the TCS model and we do not wish to further complicate things and run the risk of hitting technical issues with our planned analysis.

We may explore fitting these distributions in the exploratory analysis. Alternatively, we are planning on sharing all of our data and would be delighted if researchers with stronger modelling skills were to build upon our work!

We have added a short section on L283-285 to discuss the issue of psychological plausibility in more detail.

3. Yes, I'd like to further explore the question of whether it is the template-distractor contrast or the actual target-distractor contrast that determines performance in these tasks and would be happy to discuss ideas on how to do this! I will definitely consider working with the final data set once this is published.

4. Regarding a concern by the other reviewers: They pointed out that one impressive aspect of their modeling was that they were able to predict performance across participants. I agree that this is impressive and would therefore like to point out that this is also possible with the data set that will be collected in this project: you can simply split the sample (in various ways) if you like to replicate that aspect of the original work.

Thank you very much for your kind comments. We would be very happy to help with any further modelling you would like to do with the data!

5. I think that the extensive work by Krummenacher and Müller is still quite underrepresented in this manuscript and also in the manuscripts by the original authors. They have quite a number of studies on what you call "double-feature search", which they would refer to as "redundant-signals paradigm" or the like. Again: I highly recommend reading the pertinent review by Krummenacher & Müller (2012). Mortkoff & Yantis (1991) also seems to be an essential reference here. In the context of that literature, it might be interesting to relate the models tested here (best feature guidance vs. orthogonal contrast combination vs. collinear contrast integration) to the models discussed in that older work (independent-race models vs. interactive-race models vs. coactivation models).

Thank you for these suggestions. We have included a short sentence into the introduction to highlight the Dimensional Weighting Account (L73-78), and have also referenced ‘redundant feature search’ (L33-35). We think it will make sense to discuss the results in comparison to other models in more detail in the discussion section, as our methods won’t directly test independent/interactive race models and coactivation models.

6. Re. "however, the main issue with this family of models remains their limited usability in complex real-life search arrays (Tatler et al., 2011; Koehler et al., 2014)" (p. 3), I would add: "and in abstract laboratory search arrays (Kotseruba et al., 2021)". The recent Kotseruba results seem to be of even higher relevance for the present study with abstract laboratory stimuli.

Thank you for this reference, we have now included it as suggested (L52).

7. A few typos:

p. 12, last line: ingratiation

p. 13, l. 20/315: poster

p. 15, l. 39/397: "will be included if their search accuracy \*is\* over 90%"

Thank you for spotting these, they have now all been fixed.

Kotseruba, I., Wloka, C., Rasouli, A., & Tsotsos, J. K. (2021). Do Saliency Models Detect Odd-One-Out Targets? New Datasets and Evaluations. ArXiv:2005.06583 [Cs]. <http://arxiv.org/abs/2005.06583> Krummenacher, J., & Mueller, H. J. (2012). Dynamic weighting of feature dimensions in visual search: Behavioral and psychophysiological evidence. Frontiers in Psychology, 3, 221.

Mordkoff, J. T., & Yantis, S. (1991). An interactive race model of divided attention. Journal of Experimental Psychology. Human Perception and Performance, 17(2), 520-538. <https://linkprotect.cudasvc.com/url?a=https%3a%2f%2fdoi.org%2f10.1037%2f%2f0096-1523.17.2.520&c=E,1,0liBqewbj5OTygN9giQPyteUmh99o9XOjMFAE22KBo2x5xJgp_PJ3PgS9HQD9QDISZC0tprAZtC1wPcLKHHtZF79hF7jks4nE6glwYw2kj9wfvbRhBGOAQ,,&typo=1>