JP Comments — Not Addressed in Reply — 7/16/25

Editor's comments

In agreement with the reviewers, I find your work to be competently designed and carried out and the topic interesting and relevant for the journal. However, I remain unconvinced when it comes to the potential theoretical impact of this work which is of primary concern for this journal. Because of the methodological quality and current relevance of the work, I would like to invite you to revise your manuscript, refining your argumentation and potentially producing additional empirical evidence, to address this shortcoming, aiming to maximize the impact of your publication.

Thank you for the thorough reviews and the opportunity to revise the manuscript. Please find detailed responses to all the comments interspersed below in **blue**. As you will see, the biggest changes to the manuscript are:

- A **new experiment** with the words arranged horizontally rather than stacked vertically above/below fixation. The results are consistent with the other three experiments, thus resolving many of the reviewers' concerns. In that new experiment, we also increased the difficulty of the color task, to make sure that it still yielded a positive effect even when more difficult than the word recognition tasks.
- Discussion of the role of inter-stimulus **contingencies**, which do not account for any of our key results.
- Addition to the Appendix of model predictions for **RT variability**. In the main text we summarize these predictions and explain that our experiment was not designed to detect the subtle changes in RT distributions that the models predict.
- Throughout the manuscript, less focus on the issue of parallel vs. serial processing in natural reading, and more focus on the ~~more generalizable~~ literature on redundant target effects and how to model parallel and serial processing. in visual search
- Discussion of other evidence for **interference** between simultaneously presented words and models with **co-activation** or cross-talk between stimulus processes. In the General Discussion, we now propose that a more complex model with a stage of parallel orthographic integration (and interference) might be able to reconcile our results with other studies. Such a model has some difficulties however, to resolve with future research.

Let me start by highlighting the methodological issues raised by Reviewer #2, of which the one on contingency may be critical for interpretation (and may require the collection of additional data). Please carefully consider all of these constructive comments in planning your revision.

Reviewer 2's comments about interstimulus contingencies were astute. The most relevant contingencies concern the probability that a target is present at one location given what is presented at the other location. These inter-stimulus contingencies can modulate redundant target effects if participants learn to adapt their strategy. In the General Discussion we now emphasize this point.

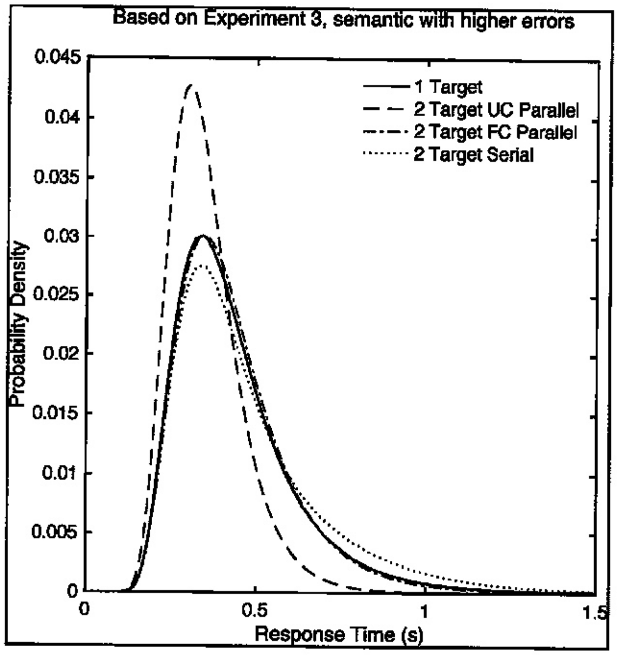
However, the contingencies cannot explain our key results that were consistent in all four experiments: a positive redundant target effect for the color task, and a negative effect for the semantic task. We say this for two reasons: (1) the contingencies in Experiments 1-2 could have made the redundant target effect more positive, while the contingencies in Experiment 3-4 could have done the opposite. Nonetheless the results were consistent across all experiments. (2) Within each experiment, the contingencies were the same for all three tasks (color, lexical and semantic) but the redundant target effects differed greatly across tasks. This is all made clear in the revised General Discussion (pg. **XX**), which also highlights the value of using several variations of the procedure and confirming the key result in multiple experiments.

Turning to the impact potential, the way I see it, the major contribution at the moment is the enrichment of the models with the addition of the potential for errors, and the perhaps counterintuitive implication of this improvement for the serial model. In my opinion this is a solid proposal. Turning to the evaluation of these modeling implications with respect to your data, I think it is possible to improve by digging a bit deeper into your RT distributions, given the clear implications from the modeling proposal: A target misidentification does not slighly prolong RT on the specific trial, but very greatly extends it by necessitating the processing of two stimuli instead of one. If this were the driving force leading to the RT patterns deemed consistent with the serial model, I am thinking that the distributional implications should arguably be discernible (a point noted by Reviewer #1); if not, some justification might be in order.

These are reasonable and interesting questions. We agree that *in theory*, our new serial model does make unique predictions about RT distributions. We begin by answering the reviewer's question narrowly and then consider it more broadly. Reviewer #1 asks if the standard serial model might predict enough of a change in the distribution to make it "visible". For our experiments, the short answer is no.

To illustrate the predicted distributions, consider the predictions of all three standard models for a case based on the semantic task of experiment 3. For this illustration, we increase the probability of an error in the one-target condition from the empirical mean of 2.2% to 10%. That increases the distinctiveness of the prediction of the standard serial model, which depends on initial misidentifications of one target before processing the other.

The following figure shows 4 distributions. We start with a distribution ~~fit to~~ predicted by the linear ballistic accumulator (LBA) model as described in the Appendix, fit to the one-target condition. This distribution for the one-target condition is the solid black curve in the figure. It matches the data in both mean and standard deviation, but its exact shape is from the LBA model. Given this starting point, the three standard models predict specific distributions for the two-target condition. The prediction for the unlimited-capacity parallel model is the dashed blue curve. It shows the usual positive redundant target effect: a shift to faster response times. Notice the reduction in the tail of the distribution. The prediction for the fixed-capacity parallel model is the dot-dashed green curve. It shows a tiny negative redundant target effect but, for this example, mostly falls near the one-target distribution. The prediction for the standard serial model is the dotted red curve. It shows a somewhat larger shift to slower response times. Its unique feature is the tail of the distribution which is heavier on the right than the other models' predictions.



Unfortunately, even with the error rate exaggerated at 10%, the serial model's prediction of a heavier tail is **a small effect.** It is nothing like a response time distribution with two modes. This is because the variability of response time is large relative to the duration of the additional processing time in the redundant target condition. Detecting the heavier tail predicted by the standard serial model would require a large and careful experiment and is beyond the current experiments.

To consider response time distributions more broadly, the next step is to investigate the **variability** of the distributions. First, we attempted to derive predictions for the standard deviation (SD) of responses times, for each of the three standard models. The standard serial model predicts that the SD in the redundant target condition is larger than in the single target condition. The fixed-capacity parallel model can predict either a larger or smaller SD in the redundant target condition. For the unlimited capacity parallel model, we have not been able to formally derive a prediction. But for all the special cases we have investigated, the SD in the redundant target condition is smaller than in the single target condition. Thus, for our three standard models, the predicted effects on the SD are analogous to the predicted effects on the mean. We have added these new predictions to the Appendix (pg **XX**) and summarized them in the Introduction (pg. **16**).

Second, we analyzed the SDs of the response times in all of our experiments. The results generally went in the predicted directions, but only two of the statistical tests were significant. For the color task, the standard deviation for the redundant target condition was significantly smaller than for the single target condition only in Experiment 1. This is consistent with the prediction of the unlimited-capacity parallel model. For the semantic task, the effect went in the opposite direction significantly only in Experiment 2. This is consistent with the prediction of the standard serial model. Because these results were so noisy, and the analyses were unplanned, we did not include them in the revised paper. To measure these effects, a future experiment needs to have more trials/participant and ideally tested in the lab, rather than remotely online as we did, in order to get more precise estimates of across-trial RT variability.

In summary, we have dug into the response time distributions and examined through ~~of~~ models how they are affected by redundant targets. The distributions could in theory confirm the results in the RT means, but **the current experiments were not designed to reveal these effects.** Moreover, the predictions for RT variability exactly mirror the predictions for the means, for all three classes of models. Therefore, **they would not add any unique ways to distinguish the models.** We now explain this rationale in a new section of the Introduction (pg. **16**).

In agreement with Reviewer #3, I admit that I am less enthusiastic about the parallel models as implemented. I do realize and appreciate the theoretical usefulness of simplicity, so I am not in principle against "standard" models. However, I am not aware of any claims of "unlimited-capacity parallel" processing of words in the reading literature; this model is more like a strawman than a usefully rejected alternative, and I do not see much information contributed by its consideration. I do not imply that it should be entirely removed from the manuscript, but it seems to me that its implausibility could be more forcefully acknowledged. Your "fixed-capacity parallel" is not trivial in this sense, as it may be relatable to reading models positing spatially distributed attention as a fixed-amount resource; however, your results cannot say much about this model because -- as you note -- it is already too flexible in terms of predictions you can test with your experiments.

We understand the concern here. First, let us clear up any possible misunderstanding of **what "unlimited capacity" means.** It does not mean that an unlimited *number* of stimuli (e.g., words) could be identified in parallel. Rather, in this context it means that there is not a capacity limit that degrades the processing of *two* stimuli that are presented simultaneously as compared to just one.

More generally, allow us to explain what makes our models are "simple". First, all of our models assume "selective influence"; in other words, that each processing channel is spatially selective. That is, the specific features of one stimulus do not influence the features represented for another stimulus. This assumption is now made clearer on pg. [redacted]. In contrast, interactive models assume that the processing of stimulus is directly influenced by neighboring stimuli.

These kinds of simple models have **a central place in the redundant target literature**, where they are often called "race" and "coactivation" models. The "race model" is in fact identical to our unlimited capacity parallel model. Importantly, *coactivation models are also unlimited capacity.* That simply means that the *quality* of processing one stimulus is not affected by the number of other stimuli being simultaneously processed. By 'quality' we mean the speed and accuracy of each stimulus process. We also address the importance of considering coactivation models in the response to Reviewer 2 below.

For now, it suffices to say that the concept of an unlimited capacity parallel model is clearly relevant to the redundant target literature. No one has previously distinguished unlimited capacity from fixed-capacity parallel models for redundant target effects. Most simply assume unlimited capacity. Therefore, this **unlimited capacity model is not a straw man;** it is in fact the most common model that has much support from the relevant literature. We believe we are making a contribution by distinguishing it from specific fixed-capacity parallel models and the standard serial model ~~that make~~ different predictions. We now make this statement clearly on **pg. 10**, and we added **throughout the Introduction** more clarity on the assumptions of each model and their position in the wider literature of redundant target studies.

Lastly, you asked whether our models, in particular the unlimited capacity parallel model, is relevant to the study of **natural reading**. We do not wish to make ~~strong~~ any claims about natural reading based on these experiments, and in the revised manuscript we are more cautious. The diversity of phenomena in reading requires that these simple models be embedded within a large set of processes.

Nonetheless, one can see the simple models within the heart of existing models of reading. The simplest case is the E-Z Reader model (Reichle, et al, 2006). It has a standard serial model at its heart. A more complex case is the SWIFT model (Engbert, et al, 2005). It was the model

formulated to encompass both serial and parallel models of the processing of words using a parameter that can vary from zero to infinity (see their equation 21). At one extreme, this part of the model is a standard serial process. At intermediate values, it is a particular kind of limited-capacity parallel process. And the other extreme, this part of the model is an unlimited capacity parallel process which they call "fully parallel". Their larger model includes other dependencies that might make their fully parallel version have limited capacity when the entire model is taken into account. But from examples in their paper, it looks like the activation of neighboring words is largely independent and thus effectively has unlimited capacity.

Other reading models are more different from our simple models. For example, the OB1 Reader (Snell, et al, 2018) emphasizes interactions between words. That is not consistent with our simple models that assume selective influence (i.e., spatial selectivity). But even this model has at its core a simpler model that does not assume a capacity limit.

In summary, we argue that simple models are useful. For the redundant target literature, they are the bases of the alternative theories of the effect of redundant targets. For the reading literature, they provide a context to understand how the processing of one word affects the next word. The distinction among our three simple models is useful to understand the differences among the existing reading models.

This needs more explan? Also is it 2 or 3 things?

Moreover, in both of your parallel models, stimuli are assumed to be processed independently; the possibility that stimuli might actually interfere with each other -- which you note on p. 48; and is also arguably the "something" to add to the model to make it nonstandard (p. 46) -- is not taken into account, despite what seems to me to be mounting evidence in this direction (which might render your contribution obsolete if it proves robust and applicable to natural reading situations). Still, I agree with your carefully--and in my opinion appropriately--phrased conclusion that limited-capacity parallel models cannot be rejected; in other words, it is not clear to me what exactly we could learn about the possibility of parallel processing from this work.

As we strove to make clear in the Abstract and conclusion, from this work **we learn two main things** that are generalizable to many domains (not just reading): First, performance in search tasks is not always enhanced by the presentation of a second target, as is widely assumed. In fact, the opposite can occur. Second, the most popular models of redundant target effects previously ignored the possibility of errors. But when errors are accounted for, negative effects of redundant targets are readily predicted.

We also have gone beyond any previous studies in this area by working out the math for several flavors of parallel models, and finding that ~~a~~ specific fixed-capacity parallel model can predict negative redundant target effects. That is novel.

Regarding the possibility of **interference between stimuli**: we agree that is an important aspect to the word recognition literature that might be consistent with some of our findings. We added a section to the general discussion about that (see pg. **XX)**, in which we acknowledge the potential benefit of more complex models that incorporate a stage of parallel orthographic integration across neighboring words. We review evidence from the flanker effect literature and the visual world paradigm (Ziaka et al., 2025).

It is notable that such models assume that there can be deleterious effects of neighboring words, but they also often argue for parallel processing of high-level lexical and semantic information. Therefore, it is not clear what they would predict for our lexical and semantic tasks in terms of redundant target effects (they could have been positive or negative). We suggest avenues for more complex models to account for the full range of findings across the literature.

As you acknowledge, some of your methodological choices render your experimental setup far removed from reading situations. This leads me to question the extent to which your discussion of word recognition and natural reading is warranted. I am not aware of any reading model positing parallel recognition of words two lines above and two lines below fixation. In addition, as you note, there is evidence that, in reading, attention is differentially allocated over the horizontal and vertical axis. Moreover, your tasks involving processing of the words demand metalinguistic judgments about the words rather than simply activating the corresponding lexical representations. In all, I remain unconvinced of the relevance of the present findings (as well as related previous findings with very similar limitations) for the current debates on serial vs. parallel word processing in reading. Perhaps you could downplay the implications for reading and instead use the opportunity to further discuss these limitations of the applicability of such findings in that debate. Or you might want to run an additional experiment directly addressing these shortcomings (which would necessarily be subject to its own limitations, of course) in order to support a more concrete link to reading.

We took this concern to heart and ran **an additional experiment** with two words arranged horizontally with a single letter space between them, as in natural reading. That is now **Experiment 4.** The results were consistent with the previous three experiments. We therefore feel more confident making claims about capacity limits for word recognition. At the same time, we removed from the Introduction some of the commentary about parallel vs serial processing in natural reading, as that is not what our study is directly investigating. Those issues seem to have distracted from the main motivation of our study. We return to the many questions about natural reading in the General Discussion.

As you are well aware, the rejection of serial search for colors is a rather trivial finding. I do not think that the fact that color (a single simple visual feature) can be processed across the locations your stimuli are displayed at has any bearing on whether complex visual forms can also be processed there. I have the impression that the very structure of the visual system suggests that the claim implied on p. 42 does not carry much weight in the present context, even though I can certainly imagine a less trivial processing distinction between simple visual forms and linguistic properties (but also between orthographic and semantic properties, and between lexical activation and metalinguistic judgment, which diminish the relevance of the contrast).

As we had previously stated in the Introduction, the color task serves as ***a positive control*** to verify that of our experiments were capable of measuring the usual redundancy gain, but with words as stimuli. Some authors in fact argue for serial processing of simple features such as color. For example, Jeremy Wolfe's Guided Search model is a hybrid model with initial parallel processing but then a serial stage that would also apply to simple color target. Nonetheless we also do not find it *surprising* that the color task yielded a positive redundant target effect. But it provides an important comparison to the lexical and semantic tasks, with the same stimuli at the same locations and with the same experimental designs.

Regarding the section on pg. 42 (now pg **45**): we stand by the claim that there should be no **crowding** between words on opposite sides of fixation. Crowding operates within relatively small zones within each hemifield (Pelli et al. *Nature Neuroscience* 2008).

But as noted above, we have re-written a new section about the possibility of **orthographic interference** between neighboring words (pg., **XX**). That is distinct from crowding but is perhaps what you were referring to. In short, we think that our results might be revealing a similar effect as in those other studies that use the flanker paradigm or visual world paradigm. But more tests and modeling would be required to distinguish limited capacity or serial processing from a more complex form of interference between ~~stimuli.~~ words

In all, what I am saying is that, although I appreciate your measured formulation of the conclusions, and without disagreeing with any of your statements or conclusions, I would like to see a richer contribution with higher impact potential. I am leaving it to you to consider whether this might best be achieved with additional analyses, the collection of additional data, different or additional lines of argumentation, or a combination of all of the above. The specific comments made by the reviewers can help guide you in your way forward.

In the full replies to the reviewers below, you will find a detailed explanation **of additional data, analyses, and lines of argumentation.** We hope you find them satisfactory. Thank you again for the thoughtful comments and suggestions and opportunity to resubmit.

Reviewer #1: Review for XHP-2024-0756R1

Summary
The authors introduce a new model of the redundancy effect including response time as well as accuracy as outcome variables. Critically, this model predicts negative effects of redundant stimuli in a semantic categorization task (in contrast to a color task). The authors observe such negative redundancy effects across different tasks go/no-go and choice. This finding shows that serial vs. parallel processing likely differs between tasks. Whereas parallel processing appears to occur in the color task, the semantic tasks most likely includes capacity limitations (if not serial).

Restriction in Expertise
I cannot review the modelling part in the appendix, as I do not do modelling work myself (yet).

Evaluation
This is a very good paper. I enjoyed reading it and learnt new things while doing so. The idea of combining RT and accuracy within one approach is very interesting. The corresponding modelling as well as the empirical finding of negative redundancy effects promotes the field of redundancy effects. This manuscript eventually should be published. I have a couple of comments which I will list below. As you will see, they are not very severe and I am optimistic that the authors will have good answers to them.

Thank you for the positive evaluation and helpful suggestions!

Comments

I was wondering whether the idea of prolonged response latencies due misidentifications of the first word would become visible in the response time distributions. It seems like the argument is that there are actually two distributions (one with and one without initial misidentification). Are there enough misidentifications to make this visible in the combined distribution?

This is an excellent point and we agree that in theory that our new serial model should predict a different RT distribution for two-target than the parallel models predict. However, by modeling this effect we conclude that it would be too small for us to detect with the present dataset.

Please see a full explanation in the response to the Editor's comments above (pg. **2** of this document). There we plot predicted RT distributions. The serial model predicts a surprisingly small increase of the rightwards tail, even when the error rate is much higher than we observed in the data. As explained in that section above, we also analyzed RT variability but found the results to be quite noisy.

We also added new sections on predictions for RT variability to the Introduction (pg. **16**) and to the Appendix (pg. **XX**). The important point for now is that the predictions for RT variability do not provide a unique way to distinguish the models; they ~~exactly~~ mirror the predictions for RT means. Nonetheless it would be worthwhile in the future to conduct a larger experiment in a controlled laboratory setting with more trials per condition, to test these predictions. Thank you for bringing it up!

Lastly, in examining these RT distributions (especially for the new experiment), we realized that our trial exclusion criteria were insufficient. In the first submission, we were excluding trials with RTs more than 4 SDs above each participant's mean RT. But that still included a few very long RTs, because some participants apparently took long breaks between trials.

In the revised manuscript we adapt a **simpler strategy for excluding trials** that reduced noise in the data. It is explained in the new *"General methods: Participants and sample size"* section on **pg. 19**. We simply include trials with RTs between 250 ms and 3s. This is an approach recommended by Miller (2023, *JEP:G*). We also then exclude individual participants for whom more than 10% of trials had RTs outside the acceptable range. These new exclusion criteria had negligible influences on the overall results (compared to the last version of the manuscript), but they did narrow the RT distributions. Nonetheless, the effects of redundant targets on the RT SDs were not detectable. consistently

I wonder why the authors used a 1 s rt cut-off criterion in the go/no-go task when their preceding numerical example for trials with initial misidentifications clearly exceeds that one second. Given the fast rts in the other experiment, the numerical example might need some values in the range of the data.

give citation maybe Mullen & Egeth?

For go-no/go tasks, a 1 second cutoff is standard. As shown in Figure 5, mean RTs were well under 1 s in all three tasks. The numerical example that relates to Figure 2 has been updated in light of our new experiments and analyses (see below). But we wish to emphasize that even if there is an initial misidentification that requires processing a second stimulus, the time required to process both stimuli before preparing a response can be well under 1s. Moreover, in the other three experiments, we allow RTs as long as 3 s and the results are consistent.

The authors report F-values with super large degrees of freedom; and I am not really sure were they come from. There is no ANOVA here right, but you mention t-tests (where is the t?). Without looking in the data, my intuition would be that there was an ANOVA being calculated as the results read a bit like it, but with raw data rather than aggregated data. Also some F-values only have one degree of freedom. I'd recommend double checking all F-values and their analyses.

We agree that this was not clear previously. We now explain these *F* statistics in the Analysis section on pg. **25**. These statistics come from fitting linear mixed-effect models to single-trial data, and examining the interactions between the fixed effects of task and the number of targets. The degrees of freedom reflect the total number of trials included in each analysis. And you were right, in a few cases we made a typo that left out the first degree of freedom (1), which we have added back in.

I am not really convinced by the attempts to generalize the findings to natural reading (even with the caution the authors do it). Naturalistic reading is of course different from processing to unrelated nouns at unrelated locations in parallel. I do not doubt that the present experiments are important, but in natural reading, words are neither unrelated not sequences of nouns. I assume that also chunking processes play a role in natural reading and might bring back parallel processing (simply because some word combinations appear together more likely than others). I am not an expert on reading though... just some thoughts...

This is a fair point. The biggest change in the revision is that we added a fourth experiment with the words to the left and right of fixation, which is closer to the conditions of natural reading. The results were consistent with the previous three experiments. We also removed from the Introduction mention of the debate about parallel vs serial processing in natural reading, and are now more cautious about the implications in the Discussion.

We believe the most important contributions of this study ~~concern~~ are models of processing capacity in redundant target experiments, which can inform us about isolated word recognition and many other visual tasks.

that include error

the continuing measurement of negative redundant target effects

Reviewer #2:

I suggest bringing up the issue of whether single-target trials include a non-target in the second location much earlier. (The paper called these "filler" items, which is OK, but I don't see why we need a new label for these.) As I was reading the comparison of the simple parallel and serial models on pages 4-6, I couldn't help but think that serial models have no problem explain an

and general

advantage of two targets over one target and one distractor. I know that this has been addressed, but leaving this question in this reader's mind was not helpful.

Point taken. You are correct that the serial model predicts faster responses to two targets than one target and one distractor, as do all other models. In the first paragraph of the section *Fundamentals of redundant target effects* on pg. 4, we now say, "On some trials, one target is presented alone or with one or more non-target stimuli." One paragraph below, we explain that the standard serial model makes the same prediction as the standard parallel model, with regard to the contrast of two targets vs. one target and one non-target. Thus, those trials are not useful for testing the serial.

With regard to the term "fillers," we use this for the special case of Shepherdson & Miller (2014). Targets were animal names, distractors were words that refer to non-animal objects, and the "fillers" were pseudowords. Thus, the fillers were neither targets nor distractors, as now explained on pg. **7.** We further explain why this comparison cannot reject the serial model in the Discussion section titled "Relation to previous redundant target studies of word recognition" on pg. **48-49**.

who introduced the term

It's also worth noting that the comparison between go/no-go and forced-choice, which is discussed in this section was often confounded with whether distractors were included in single-target displays in the early literature. This matches the current finding of no real effect of response type across Expts 1 and 2.

We agree that there is a broad correlation across studies between the use of go/no-go versus choice tasks and whether mixed trials were included or excluded. Early studies tended to use the choice task and include mixed trials while later studies tended to use the go/no-go task and exclude mixed trials. But for letter tasks, there are two studies that break this correlation. Both studies contrasted go/no-go and choice while holding the rest of the design constant. Van der Heijden, et al. (1983) compared these tasks while excluding mixed trials, and Grice and Reed (1992) compared these tasks while including mixed trials. Both studies concluded that the redundant target effects were more convincing for the go/no-go task than for the choice task. Thus, there is reasonable evidence for this task effect using letters. On the other hand, we do not know of any studies that directly compared designs including mixed trials versus those excluded mixed trials. Our own view is that it is likely that both the details of the task and inclusion of mixed trials can sometimes modulate the redundant target effect. However, ~~these modulations are~~ modest compared to the differences due to using simple features versus words.

we find that

Given your comment and the state of the literature, we added the following sentences to our review on **pg 7**: "In some of the redundant target literature, variation in the task procedure (go-no/go or choice) was confounded with the inclusion or exclusion of mixed target-distractor trials. But two studies using letters do show differences between go/no-go and choice tasks while holding the rest of the design constant (Van der Heijden, et al., 1983; Grice and Reed, 1992). Thus, for letters, there is reasonable evidence that positive redundant target effects are more robust for the go/no-go task procedure."

My guess it that the new parallel model continues to predict a redundancy gain, even with errors, unless and until accuracy drops to chance. If I'm correct, this ought to be mentioned.

Yes, that is correct. And not only that – the redundancy gain continues all the way down to 0% hits. But keep in mind that this prediction is just for hits, considering only target-present trials. Responses to non-targets are their own story. Because of that complication, we prefer not to dive into it in the main text; it is already implicit in the Introduction and the Appendix.

Given the state of the redundancy-gain literature in (pure) response time, do we need a coactive model with errors? If so, what happens when one channel on a two-target trial misperceives a target? Could such a model produce a negative effect?

This is an interesting question, especially in light of the flanker effect studies with word recognition that we now review more thoroughly in the General Discussion (starting pg. **52**). We then expand upon how those results and ours might inspire future development of a model that includes some degree of interactive processing (see pg. **54**). Such a model would be similar to the 'co-activation' models of redundant target effects in that there is a stage of integration across processing channels. However, co-activation models were generally developed to predict large positive redundant target effects, while our goal is to explain *negative* redundant target effects.

To answer your question more directly – we suspect that it is possible to build a co-activation model with errors that predicts a negative redundant target effect. Donkin, Little and Houpt (2014) discussed co-activation parallel models with limited capacity. Perhaps one of those could produce a negative effect, if there is partial pooling or partial co-activation. But those would be complex models with many free parameters that we cannot determine in this study. We already show that one specific form of a fixed-capacity parallel model can also predict a negative effect (depending on the evidence accumulation process it assumes).

Second

More generally, we have chosen not to add detailed co-activation models to the current paper for two reasons. First, coactivation models ~~are not trivial~~ to develop and test, in contrast to the straightforward and distribution-~~free~~ models that we use to account for both negative and positive redundant target effects. The empirical literature on coactivation models depends upon the "race model inequality" (Miller, 1982) and its variations. Exactly how to test this property is debated in the literature, as reviewed in Gondan and Minakata (2016). Furthermore, tests of this property have been criticized because it can fail due to a lack of context invariance (e.g. Lin and Otto, 2020). Reviewing and addressing these issues is nontrivial.

Moreover, the theory for existing coactivation models is quite different from the theory developed in our paper. The relevant work that includes errors is based on specific stochastic processes such as the diffusion model (Schwarz, 1994; Diederich, 1995) with a recent example from Blurton, Greenlee and Gondan (2014). It is not obvious how to generalize this work to both include errors and to be distribution free, as our simpler models managed to do.

Correction: the method of dealing with **favored locations** suggested by Miller and Lopes is actually less conservative than that used by Mullin et al. (which was taken from Biederman and Checkosky). The method used here is the fixed (across all subjects) version, which is not recommended. Either keep the faster location on a subject-by-subject basis or only when it's faster at p < .10. Avoid forcing all subjects to have the same favored location.

This is indeed an important issue and we have clarified our explanation of it on pg **25-26**. The most important question is whether the calculation of the redundant target effect biases the estimate to be positive or negative. All of the prior discussions of alternate baselines (e.g., using the one-target performance at the fastest location) were concerned with the potential bias of *overestimating* the redundant target effect, which was usually assumed to be positive.

Here we are most interested in detecting the negative redundant target effect predicted by our new models. If we use as the one-target baseline the mean of RTs at each subject's fastest location, we risk overestimating how *negative* the redundant target effect is. Therefore, the more conservative approach is to compare two-target responses to the mean of one-target responses at the two locations. That is the approach we use in all our experiments (although for Expt. 1 we also report the results with the fastest-baseline).

The design of Expt 1 appears to have a **contingency** issue of the sort discussed by Mordkoff and Egeth or Yantis. The overall probability of a target in a given position is 37.5%. The conditional probability of a target in a given location when a target also appears in the other location is 66.7%. Is this a problem? Does this bias the results in favor of unlimited capacity parallel models? Note that this is not at all the same as the way that "correlated" and "uncorrelated" are being used in this paper, which contrasts with what was said in the paragraph crossing from page 34 to 35.

...(Continuing a point from above...) the design used for Expt 3 is quite different from that used for Expts 1 and 2. Now the overall probability of a target in a given location is 45% while the conditional probability of a target in a given location when a target appears in the other location is only 33.3%. So, instead of having targets being positively related, now they are negatively related. So, it might not be the inclusion of distractors (or filler) on some single-target trials that caused the change in the pattern; it could be this. Therefore, any and all conclusions with regard to the lexical-decision task must be taken with some salt. You can try to argue that the lack of change for the color and semantic tasks makes this unlikely, but it still should be addressed.

Thank you for bringing up the issue of contingencies. We now raise the issue when introducing the design of Experiment 3 (pg. **35-36**), and we added a section to the General Discussion about it (see pg. **47**). In short, we agree that the interstimulus contingencies differ between the experiments. In Experiment 1-2, contingency learning could have made the redundant target effects more positive / less negative; Experiment 3-4, contingency learning could have done the reverse. It was therefore important that we tested both designs in this study.

The most important point is that within each experiment, the contingencies were the same for all three tasks, but the redundant target effects differed qualitatively across the tasks. Also, while the contingencies differed across experiments, the color task yielded consistently positive effects and the semantic task yielded consistently negative effects. Thus, the contingencies cannot explain why one effect was positive and the other was negative.

We agree that the fluctuations in the lexical decision task's effect across experiments could be related to changes in the contingencies, as we now describe on pg. **48.** That is one reason why we base our general conclusion on the semantic task and the color task.

It's probably obvious to most, but you might want to be clear that the Bayes Factors reported are BF10 not BF01.

Yes, thank you for catching that. We have clarified the terminology on pg. **24** (in the Analysis section).

In Expt 1, was there a relationship between speed and accuracy as a function of location? In other words, if a given subject was slower on single-target-upper trials, did they also miss more upper-only targets?

For single-target trials, we analyzed the correlation between the side asymmetry for RTs (right – left), and the side asymmetry for error rates. There was a significantly positive correlation for some tasks is some experiments but not others (and never a significantly negative correlation).

We are not sure what to make of this, except that there is **not a speed-accuracy tradeoff.** In other words, a participant who responds more quickly to words at the at the top location is also more likely to respond correctly to words at the top location.

Maybe report Expts 1 and 2 together (to dispense with the response-type issue ... see above)...?

To keep a consistent organization across all four experiments (especially now that we have added a fourth) we prefer to keep Experiments 1 and 2 separate.

With all that said, the paper was very nicely written. The Introduction did its job, the Discussion (other than any reliance on Expt 3 vs Expts 1 and 2) was fair and appropriate, and the Appendix was very good.

Thank you very much for the positive evaluation!

Reviewer #3:

I've seen a previous version of this paper, and have to say that the current version is a nice improvement. I think this is worth publishing, if the authors can integrate answers to the questions below in their manuscript.

Thank you for the overall positive comments and helpful suggestions!

- In the Intro, the authors provide a brief justification for presenting the two words above and below fixation, rather than left and right of fixation. The arguments that are provided there are alright; but Snell et al. (2018, Scientific Reports) found that attention is deployed to horizontally aligned flankers but not vertically aligned flankers. In your 'standard' models, is there room to accommodate the fact that specific tasks and stimuli engender specific deployments of attention? I can imagine that when the task is to detect colors, attention isn't strictly deployed along the horizontal meridian, but that things are different in the case of linguistic stimuli.

We agree that the position of the words is an important issue, and we took your concern to heart and ran an **additional experiment with the words arranged horizontally.** The words were 1 letter space apart, with nothing between them. This new experiment, now number 4, used otherwise the same design as the preceding Experiment 3 (which had words above/below fixation). The results are consistent: for both arrangements of the words, there is a positive redundant target effect in the color task and negative effects in the lexical and semantic tasks.

We therefore conclude that the negative effects we found initially are not due to the unnatural or unfamiliar arrangement of stimuli, nor to an inability to attend to the words' locations.

- Relatedly, it is said that "we focus on parallel processing of two non-identical targets, given a long-term goal of understanding natural reading, when neighboring words are not identical." If the goal is really to understand natural reading, one would've expected horizontally aligned stimuli. And earlier the authors do claim that they're interested in the question of whether readers CAN attend to multiple words (more so than the question whether readers DO attend to multiple words). If that is so, then the case of repeated target words is interesting, because serial and parallel models do make different predictions here.

We agree that the case of repeated word targets is interesting, and it might reveal parallel processing at earlier visual or orthographic stages of analyses. This is now discussed more in a new section on flanker effects in the General Discussion on pg **54-55**. Our goals were to understand lexical and semantic processing. That is why we used lexical decision and semantic task categorization tasks. The question was whether two words could be processed to that level simultaneously and independently, so we made the two words different from each other.

In a future study we would love to use repeated identical words and also parametrically vary the amount of orthographic overlap between them.

- In the General Discussion you point to the importance of parsimony; e.g., a serial model can account for your data without having to incorporate task-specific attentional deployments or mechanisms. That's true - but it could also be argued that your set of experimental conditions is "too parsimonious". I think there are some crucial experimental conditions that we're not seeing in the present study but where models would make quite different predictions; namely, conditions that would allow you to test for redundant target congruency effects. With the "mixed" trials in Experiment 3 you came close, and there indeed seems to be a congruency effect, but this cannot be interpreted because participants didn't know where the target would appear. The crucial prediction of a parallel model is that the congruency of the redundant word impacts the response for the attended target. The serial model doesn't predict such an effect. As the authors well know, such effects have been widely reported with the closely related flanker paradigm, where the participant is pretty much guaranteed to attend the target (and ignoring flankers is beneficial).

We hope that changes throughout the revised manuscript have made the models more clear, and why the key test is between one-target and two-target trials (not comparing to trials of the standard That comparison is

with 1 target and 1 distractor; see pg. **4-3**). All models, parallel and serial, predict faster responses to two-target displays than mixed target-distractor displays. ~~Those are~~ therefore not informative. Moreover, they are not quite the same as a flanker congruency effect because in our tasks, both stimuli are task-relevant and either one could trigger a response.

But regarding the flanker paradigm, we agree that it is theoretically related and we added a significant section about to the General Discussion (starting on pg. **52**). There you will see that we reviewed three findings from that literature: (1) that in some contexts recognition of a single target word is impaired by the presentation of other words at the same time, which is similar to our negative redundant target effect; (2) that less interference is caused by non-target words that contain similar letters or bigrams (or are identical to the target); (3) that ~~nonetheless~~ there are arguments for parallel lexical, semantic or syntactic activations for multiple words.

These are all related phenomena. As you noted above, they highlight the importance of prior redundant target studies that found positive effects for two copies of the same word, suggesting a pre-lexical stage of pooling of letter identities. The new section on "Interactive processing" (pg. **XX**) discusses how a more complex model could unify many lines of evidence.

Lastly, given the mixture of past studies on the flanker effect, we would not find it easy to predict how our semantic task should have turned out: a positive redundant target effect, consistent with parallel semantic processing (as most previous parallel models would predict), or a negative effect, due to the interference caused by neighboring words with different letters? It will take more theoretical modeling to bring these two areas of study together. ~~Perhaps the balance of orthographic integration and simultaneous semantic activation for multiple parallel models yields predictions something like the fixed-capacity parallel models that we developed here. We now raise that possibility on pg. XX.~~ Cut!

The authors correctly note that the winning model should be the one that can explain all effects (across different paradigms). Can the standard serial model really do this in its current parsimonious state? Consider here that the authors' serial account for flanker congruency effects is quite non-parsimonious (because one would have to assume that participants on a portion of trials base their response on a flanker instead of the target, and that response would be incorrect when the flanker was incongruent, and incorrect responses tend to be accompanied by longer RTs, hence a congruency effect in RTs - but how did the participant know that the response was incorrect if the target wasn't processed? Long story short: you need to concoct quite a non-parsimonious scenario to save the serial model here).

We're willing to concede this point and have simply removed that line of argument from the Discussion. As just mentioned, we now have a much more thorough discussion of flanker effects (and related phenomena), in which we are open to a common cause for these effects, including some initial parallel integration of information across words.

- It is claimed that "with a diffusion process of sensory evidence accumulation (Palmer et al., 2005), [fixed capacity parallel models] yield positive redundant target effects on correct response times, for all relevant parameter values (as well as a positive effect on accuracy)." This is counter-intuitive to me. How would diffuse evidence accumulation across two words be beneficial relative to simply processing one word? Is it assumed in this type of model that attention is diffusely distributed across two word locations even if there is only one word in the visual field? Would that be plausible at all?

This is a good question and we have expanded the relevant section (on **pg. 15**) to explain more. First, a point of terminological clarification: a "diffusion process of sensory evidence accumulation" does not refer to the distribution of attention or processing resources "diffusely" manner across words. Rather, it is a type of model of that specifies how evidence about any one stimulus accumulates over time towards a decision. In a diffusion process, the evidence goes up and down in a stochastic fashion over time. That is distinguished from a linear ballistic accumulator model in which the rate of evidence accumulation is determined immediately and that rate is then constant until a decision is reached.

Here is how a fixed-capacity parallel process could yield a positive redundant target effect: the capacity limit means that the SNR of each stimulus process is reduced by a sqrt(2) when there are 2 targets as opposed to 1. That SNR applies both to the speed and accuracy of the process. A second target can still make RTs faster because the two processes are *racing* independently to trigger the response. Becuase any one processing time is quite variable, there is a statistical benefit, on average, to taking the minimum of two processes' completion times.

To make a long story short, the diffusion model predicts a positive effect because the processing times are quite variable. The linear ballistic accumulator model predicts the opposite because the processing times ~~are~~ less variable, and the race benefit is outweighed by the loss of SNR from fixed capacity. can be made

We have added that explanation to page **16,** when we introduce the fixed-capacity parallel models.

- In the paragraph that starts with "One last consideration is the possibility of parallel orthographic processing across multiple words, as is assumed by some models of reading", you could cite a few studies that evidence parallel orthographic processing: Dare & Shillcock 2013 (QJEP), Grainger et al. 2014 (Acta Psychologica), Angele et al. 2013 (JEP:HPP).

Thank you for pointing out these papers. We now cite them in the new discussion section on "interactive processing" that starts on **pg. 52.**

- "Such parallel orthographic processing could also result in interference between multiple words that contain different letters, perhaps explaining the negative effects of redundant targets in the experiments above. This explanation is essentially a non-standard parallel model with dependencies between each stimulus process." Isn't it awkward that, in the context of reading, the acknowledgement of the existence of orthographic processing is 'non-standard'? Can we really make progress in the serial-vs-parallel debate if we deliberately abstract away from a process that is undoubtedly so fundamental to reading?

We understand where you are coming from. But we must emphasize that the word "standard" means something very specific in this context. In particular, it refers to class of models with a specific set of assumptions, as we explain on pg. **5.** These are well established models for visual search and redundant target experiments, which is why we have focused on improving them in a few specific ways for this redundant target study. They are not "standard" in the literature on natural reading.

Applying these models to natural reading is not trivial, but we feel there would be much to gain by incorporating this style of quantitative capacity limits into reading models. But to avoid this confusion we have rewritten the sentence that you highlighted (on **pg. 54**) so there can be no confusion about what we mean by "standard."

We also focus less on natural reading throughout the manuscript, and as noted above, added an section on orthographic integration and flanker effects to the General Discussion.

- I'd like to see a few words on what the authors would predict if the current setup is repeated but with horizontally aligned words (separated by a space, like in normal reading, which the authors claim to be interested in).

As noted above, we do not have to just make predictions, we now have data! **See the new Experiment 4.**