Dear Dr. Haaf,

Please accept our apologies for taking so long to send you our revised manuscript (and pass along our apologies to the reviewers). Although we had planned to complete the revisions almost two years ago, we realized early on in that process that a third study would help solidify our case. Then a combination of the pandemic and the start of graduate school for Victoria meant that we had to put the project on the backburner. We have finally completed the revisions, which include the new study 3 that addresses many of the original concerns. Below we provide detailed responses to your comments and those of the reviewers, and along with our submission we included a “change tracked” version of the manuscript showing everything differs from the previous version. We also have verified that the OSF pages are now all public and we simplified access by creating a single project page that links to the separate projects for each study. Thank you again for the helpful action letter and the insightful reviews—the comments and suggestions led to what we believe is a much stronger paper.

In the responses below, we have reproduced the entirety of the action letter and reviews and we have inserted our responses below each point in color. Please don’t hesitate to reach out if you have any questions.

Sincerely,

XXXXXXXXXX

Action letter:

1. The first important issue all reviewers and I encountered was that the preregistration was not accessible. Please ensure that, should you decide to submit a revision, the preregistration is accessible to the reviewers. Reviewers 1 and 2, for example, wondered about the relatively small sample sizes for a one-trial study and wanted to verify your rationale for it in the preregistration. Nevertheless, I suggest you also share the sample size rationale with us in the manuscript itself.

We thought that the preregistration was visible and that we had provided the view-only links, but apparently we forgot to do that. We’re sorry we couldn't provide that information for your review. We have now created an overarching OSF project that links
to the project for each study, and we have made all of the projects public so that they should all be accessible.

The preregistration included our sample size justification and stopping rule:

“We did a power analysis for subjects based on the original design. We will test 60 participants (95% power to detect the effect size of $z = 0.43$). We will continue collecting data and complete scheduled sessions until we have data for at least 60 subjects.”

We have added a power analysis subheading to each experiment that includes details about our calculations. For Experiment 1, the added section states:

“We preregistered a plan to test at least 60 participants, which a sensitivity analysis shows would give us 95% power to detect an effect size of $= 0.430$ (and 80% power to detect and effect of $= 0.325$) for a paired t-test comparing error rates for the attended and ignored colors (with alpha = 0.05). Eitam et al. (2013) reported observing a difference in accuracy for the attended and ignored feature corresponding to $= 4.36$ for study 1a and $= 2.86$ for study 1b in their forced-choice test. Given that we are unsure how their test was computed, we estimated their effect size for our planned test statistic (a paired t-test) by calculating the one-tailed p-value corresponding to their statistics and converted those to . The corresponding $= 4.598$ for their study 1a and $= 2.989$ for study 1b represent effect sizes of $= 0.467$ and $= 0.407$ respectively. Our target sample size of $= 60$ would have approximately 97% power to detect the effect in Study 1a and 93% power for the effect in Study 1b, not taking into account the increased precision of our continuous measure.”

The power analysis section for Experiment 2 states:

“We preregistered a target sample size of 60 or more participants based on the same sensitivity analysis as in Experiment 1. Note that our effect size in Study 1 corresponds to $= 0.501$, which is larger than the effect we calculated for Eitam et al.’s (2013) studies. With our target sample size of N=60, Study 2 would have approximately 99% power to detect the effect we observed in Study 1.”

The power analysis section for Experiment 3 states:

“A power analysis showed that a sample size of 175 participants would give us greater than 99% power to detect the effect size observed in Experiment 2 ( $= .34$)
with \( = 0.05 \) (using the same paired t-test). We preregistered a plan to continue scheduling participants until we had data from at least 175 participants.”

2. All reviewers suggest improvements for the statistical analysis of both experiments. For example, Reviewers 2 and 3 suggest reporting effect sizes in addition to sample means.

We added the effect size (Cohen’s d) for each between-groups statistical test and dz for each paired test.

I would also like to point you to Reviewer 1’s suggestions. Both Reviewer 1 and I would like to see a more careful approach to the statistical analysis and interpretation thereof. Just to give an example, the fact that participants can report the unattended color fairly well when the attended color has already been reported seems to be an important feature of experiment 1 that perhaps warrants further exploration.

We conducted a follow-up analysis and addressed this point in greater detail in the discussion. Our experimental design addresses the confound in the original paper: some of the error stems from participants misreading the question. We suggest that the unattended information is still encoded but with less precision than the attended color. See our response to Reviewer 1 below for details on the changes we have made. Our new Experiment 3 addresses these concerns more completely.

Another example is for experiment 2 where participants responded to two color wheels in one display, yet it is unclear which they chose to respond to first. Are these data available? If so, I suggest an additional exploratory analysis to see what the effect of that choice is.

Unfortunately, we do not have data to analyze responses based on which color participants reported first in Experiment 2. Participants clicked a button to continue after having selected both colors, so we just have their answers for each and not the sequence of responses.

Reviewer 3 also suggests inspecting performance differences between the inner and outer ring.
We added a test of the difference in performance for the inner or outer ring in order to examine whether eccentricity mattered (as suggested by reviewer 3). It didn’t ($t_{[66]} = 0.38$, $p = .708$, $d = 0.06$). Given that there was no requirement to fixate and that the display was relatively small and centered on the screen, the lack of an effect of eccentricity is not surprising. See our response to Reviewers 1 and 3 below for more details on the analyses we added to the paper.

3. Reviewers 2 and 3 criticize the lack of engagement with the wider literature on attention and working memory. That seems to be a valid critique, and you find many useful suggestions in both reviewers’ comments.

We have tried to better situate our study in the broader literature. We tried to balance that goal with our goal of keeping the paper concise and focused. See our detailed responses to Reviewer 1 & 2 below.

4. I would also like to point you to the last suggestion by Reviewer 3. This seems to be a good solution for the issues that participants seemingly have with this experimental design. Perhaps this suggestion could be included in the discussion.

We have conducted a third study based on their suggestion.

**REVIEWER 1:**

In the rationale for using a precision measure, as opposed to the previously used forced choice task, the introduction says: “a small difference in the mental representation favoring one option could yield a large difference in choice accuracy” (p.4). I found this rationale slightly confusing as usually the rationale for using more fine-grained, ‘precision’ measures is that choice measures may be less sensitive to subtle changes in memory fidelity.

Forced choice might yield great performance even if the memory representation were imprecise. When given a forced choice, the smallest leaning toward one of the options can lead to a large difference in the likelihood of selecting it. A 51:49 difference in representation can result in a 100/0 selection if participants are forced to choose one option or another based on judging which one is closer to the representation. But, in a
recall task like the one we used, we can measure whether the representation is perfect (as the 100/0 would suggest) or fuzzier as we'd expect if it's 51:49).

In essence: A color wheel allows for greater generalizability of color perception, since other colors not tested in Eitam’s paradigm, may fall within a continuum.

Also on p.4, discussing Eitam et al: “Participants likely did not expect a memory test for the unattended feature, which could lead them to mistakenly report the attended color when asked to report the unexpected color first.” When I first read this I thought “yes this would be the case for the first trial but not necessarily after that, once participants knew they would be asked about both”. It was only later that I found out that these are one-trial experiments - this could be made clear earlier on in the introduction.

We have clarified in both the introduction and method section that both the original study and our replication are single-trial designs. In the introduction, we added the following:

“Eitam et al. (2013) used a single-trial design to prevent participants from paying more attention to the unattended feature in subsequent trials.”

We also now discuss related studies with multiple trials as a comparison.

In the method section (under “Stimuli and procedure”), we now state: “In this one-trial experiment…”

Another, more important point, on the above quote from p.4: In Eitam et al’s task it was not possible to report the attended color if asked about the unattended color first, and vice versa (see their figure 1). Therefore, participants mistakenly reporting the attended color first when asked about the unattended color cannot account for the greater error in their experiment.

In the Eitam et al. (2013) study, the question order was counterbalanced across participants. Figure 1 in their paper just illustrates one counterbalancing order. Half of the participants focused on the inner ring and half focused on the outer ring. And, for each group, half were asked first about the “relevant” ring and half were asked first about the “irrelevant” one. To quote the method section from their Experiment 1a:

- “The participants were first asked to focus on the inner or outer (with equal probability) of two to-be-presented concentric circles.”
● “...participants were presented with a recognition test asking them to choose the color of one of the 2 stimuli (either the relevant or irrelevant stimulus, counterbalanced across participants)”

Because the focus of attention and the order of questions was counterbalanced, it was possible to “report the attended color if asked about the unattended color first, and vice versa.” If participants did not carefully read that they were being asked about the unattended feature, they might give a response based on the attended color. The recognition task order was counterbalanced, so for half of the participants, the first question was about the color they did not focus on.

It is also worth noting that in Eitam et al. there was an additional cue for the color being asked about, as the color choices were presented either as disks or outer rings (see their figure 1). In the present figures the color wheel looks very much like the outer ring - is this just because things aren’t drawn to scale?

While the Eitam et al.’s cues were shaped as either a disk or ring, subjects might still inadvertently report the color they were attending. Only one stimulus was presented at a time, so participants might not have realized that the shape cue was informative. For example, the disk could have been interpreted as a color swatch.

The reviewer is correct that our study lacked this shape cue because the color wheel was shaped like a ring. Study 3 addresses this concern by having participants reproduce the actual stimulus. We also conducted a follow-up analysis to see if subjects used this ring shape as a cue to what they should report. We checked whether participants had better performance when reporting the outer ring’s color given the match in shape. They didn’t. We added the following to the results section:

“An exploratory analysis suggested by a reviewer confirmed that there was little difference in the average error for the inner disk (m=45°, sd=52°) and the outer ring (m=41°, sd=56°; t[66] = 0.38, p = .708, d = 0.06; where \( d = \frac{(X\bar{1} - X\bar{2})}{\sqrt{\sigma_1^2 + \sigma_2^2}/2} \), suggesting that the visual eccentricity difference between the ring and disk likely was unimportant to task performance.”

The method does not appear to mention the size of the color wheel (or the confidence slider).

We added the following more complete description of the color wheel’s size:
“For each, a ring-shaped, RGB color wheel with the same dimensions as the original stimulus appeared at the center of the screen…”

We also note that the original stimulus consisted of a “...colored disk (1.87° diameter) surrounded by a differently colored ring (1.87° thick) against a gray background.”

Confidence slider (already included on pg. 2): “After selecting the remembered color, participants rated their confidence in their judgment on a 1-100 slider (cursor started at 50).” We don’t think the physical size of the slider is important given that it was sufficient to permit participants to select any number from 1-100.

p.5. Both of the osf links are private so I couldn’t access the pre-registrations. See here for how to make a view-only link to share with reviewers: https://help.osf.io/hc/en-us/articles/360019930333-Create-a-View-only-Link-for-a-Project

As noted in response to the editor, we had thought these were viewable, but apparently they weren’t. We wish we could have remedied this before you completed your review. We have now consolidated links to each study within an overarching osf project, and all of the components are now public (https://osf.io/pzavd/).

p.10: “For the attended color, participants were equally accurate when tested first or second. For the ignored color, participants were inaccurate when tested first and relatively accurate when tested second. That difference in pattern is consistent with an alternative explanation”. Firstly, no evidence is presented that participants are equally accurate for first/second tested attended items - only that they do not significantly differ. Secondly, this interpretation appears to be based on the difference between significant and non-significant. As I cannot access the pre-registration it is not clear why separate t-tests were chosen to analyze the data. An ANOVA with attended vs. unattended and attended first vs. second would better address the question of whether order effects are specific to unattended colors.

Similarly for Experiment 2, a t-test is reported for attended vs. unattended then separate t-tests comparing left and right for each color. Why? An ANOVA with attended vs. unattended and attended left vs. attended right as factors would be appropriate here. As additional exploratory analyses it may be interesting to look at the order in which participants recalled the colors (assuming this information is available). Did they tend to report attended then unattended or did they go left to right?
An ANOVA would not provide a better test of whether the order effects are specific to the unattended colors. The ANOVA would give three distinct tests: A main effect of order (combining across attended/unattended), a main effect of attended/unattended (combining across first/second), and an interaction between attended/unattended and first/second. The main effect of order is uninteresting, so it’s an unnecessary test. The main effect of attended/unattended is functionally equivalent to the t-test we reported. The interaction term in an ANOVA tests the fit to a focused contrast in which the effect of order is opposite for the attended and ignored objects (i.e., a cross-over). We had no reason to predict a cross-over interaction, so that test would not be a precise hypothesis test of interest. We’re instead interested in the simple effects, which is why a set of 3 t-tests was a more direct test of the hypotheses we cared about than was the set of 3 tests provided by an ANOVA. Our preregistered plan explained our rationale for using a set of tests that address our specific hypotheses.

We also added an additional, exploratory t-test to compare accuracy for attended items when tested first or second:

“For the attended color, participants were equally accurate whether that color was tested first (m=29°, sd=49°) or second (m=27°, sd=39°), t(52.54)=0.12, p=.901, d = 0.03. In contrast, for the ignored color, participants were highly inaccurate when it was tested first (m=79°, sd=64°) and relatively accurate when it was tested second (m=30°, sd=37°), t(61.13)=-3.99, p<.001, d = 0.95.”

Why only 67 and 69 participants in these single-trial experiments? I assume the pre-registration covers this.

Based on a sensitivity analysis, those sample sizes give us greater than 95% power with N=60 (alpha=.05) to observe a difference of the size reported by Eitam et al. (2013) (see the more detailed discussion now added to the paper and documented above). These are all repeated-measure effects, and they appear to be fairly large. We planned to continue scheduling sessions until we had at least 60 participants in each experiment, but we tested all of the participants who came for the scheduled sessions. We had higher turnout than expected for the scheduled sessions, which is why our final samples had more than 60 participants. See also our response to the editor above.

The finding from Experiment 1, that participants can report the unattended color fairly well when the attended color has already been reported, is kind of glossed over in the discussion. Doesn’t this suggest an important boundary condition on this effect and that the unattended information is still reportable under the right test conditions (i.e., when
participants no longer have to retain the attended color)? Looking at the order of recall in Experiment 2 may provide some more relevant data on this issue.

We added the following prose to the Discussion:

“In Experiment 1, we observed more errors when the questions were asked sequentially, suggesting that task confusion might have amplified the difference in precision for the attended and ignored color. In Experiment 2, where the response wheels were presented simultaneously to the left and right, it seems plausible that participants would respond using the left wheel first, even if it were the one for the unattended feature. If so, that might make the subsequent report for the attended color less accessible. Unfortunately, Experiment 2 did not record which wheel they used first, so we cannot be certain whether response order affected performance.”

It occurred to me when reading this paper and Eitam et al. (2013) that a possible driver of the attended > unattended effect could be a demand characteristic of the experiment and the instructions to participants. Participants are told to focus on one color in particular and therefore may be less concerned with the accuracy of recalling the unattended color (even if they could possibly do better) as it is disincentivized by the instructions. Or they may assume that they are supposed to do more poorly for the unattended color and make it so. Obviously this is an oversimplified interpretation (e.g., it doesn’t account for the order effect in Exp 1) but it could be true of a subset of participants, and could contribute to the mean difference. Perhaps this is worth discussing, if only to show that it couldn’t possibly work as an account (if that’s the case).

We briefly mention this concern in the constraints section at the end of the discussion:

"Finally, participants might have expected that they should be less accurate for the unattended color, leading them to intentionally perform worse. Although we cannot eliminate such a “demand characteristic” explanation, we find it unlikely given that our results held across multiple types of responses (sequential presentation, simultaneous presentation with separate responses, and simultaneous presentation with feedback)."

REVIEWER 2:
In two experiments, the authors replicate and extend a task from Eitam et al. to test whether there are differences in precision for an attended vs. ignored color held in
working memory. Each experiment consisted of a single trial per participant. Before viewing the memory item, participants were instructed that they should attend only 1 of the 2 colored parts of an object in the display. At test, they were actually tested on both of the colors, not just the cued colors (i.e., the cue for this 1-trial experiment was 0% valid). As expected, memory precision for the cued color was better than memory for the uncued color.

I think that, overall, the results are clearly presented, and I appreciate that the authors pre-registered their studies. However, I think a significant weakness of the manuscript is inadequate engagement with and citation of prior work. Since this is a replication study, I understand that the authors might want to keep the paper fairly tightly focused on following up that specific paper. However, in my view, the narrow focus of the framing is too extreme, and I think the manuscript in its current state fails to appropriately contextualize the work within the literature.

We have broadened the discussion and framing in the introduction (see detailed response below point 1).

In addition, I have some questions about how the authors’ chose their sample sizes: with only a single trial per participant, my hunch is that the sample size is perhaps a bit lower than ideal. Perhaps the issue of sample size was addressed in the pre-registration, but the pre-registration link was set to private, so I was unable to view It during the review process. In either case, I think it would be beneficial to the literature to include a discussion of sample size and to report effect sizes for the main statistical tests.

We’re sorry that the preregistration was not available for your review. The preregistration included a sensitivity analysis showing that N=60 gives 95% power to detect an effect of dz=0.43. We expected a large effect based on the original study, and we fully expected to replicate it (and did). We now report details of that analysis in the method section for each experiment (see also our response to the editor above).

These and other minor concerns are detailed below.

1. In my view, the biggest issue with the paper is insufficient citation of prior work. The authors state in the abstract that “this study is one of the few to demonstrate an effect of attentional selection on memory using such a simple display”. I think this framing is faulty given the vast amount of work on pre-cue and retro-cue effects in working memory. I’ve added some pointers to references
below (these are not exhaustive lists but should provide a good starting point). As there have been many, many papers looking at pre-cue and retro-cue effects in working memory, I don’t think the authors need cite every single paper showing an effect of prioritization in working memory. However, I think they do need to engage with this literature in the paper in order to discuss how their study fits in. I also think the authors do need to cite and discuss the most highly relevant papers (e.g., at least a few other papers to my knowledge have used an extremely similar “surprise test” paradigm with 2 features from a single object).

- The effects of prioritization on visual working memory performance (pre-cues and retro-cues), for example: (Berryhill et al., 2012; Griffin & Nobre, 2003; Myers et al., 2017; Palmer, 1990; Souza & Oberauer, 2016)

- Neural evidence for prioritization of one simple featural object over another in working memory, for example: (Bocincova & Johnson, 2019; Hajonides et al., 2020; Harrison & Tong, 2009; LaRocque et al., 2016; Serences et al., 2009; Yu & Shim, 2017)

- Papers using a similar paradigm whereby a ‘surprise trial’ is used to test whether participants remembered the non-cued feature (with some minor differences, e.g., 1 color & 1 orientation, rather than 2 colors), for example: (Born et al., 2020; Chen et al., 2018; Chen & Wyble, 2015b, 2015a, 2016; Shin & Ma, 2016; Swan et al., 2016)

We have updated the introduction with the following prose based on your suggestions:

“Eitam et al. (2013) used a single-trial design to prevent participants from paying more attention to the unattended feature in subsequent trials. Other studies have addressed this concern by only asking about the unattended feature on a final, “surprise” trial. For example, Chen and Wyble (2015a) asked subjects to report a letter on 11 trials and only asked about the letter’s color on the 12th trial. Only 25% of participants could recall the color even though they had focused on the letter. Unlike Eitam et al. (2013), the letters were presented only briefly (200ms), making the representations more susceptible to proactive interference or rapid forgetting (Nee & Jonides, 2013; Oberauer, 2002).”
In a similar “surprise trial” design, participants were asked to report an arrow’s color across multiple trials and then were unexpectedly asked to report its orientation (Swan et al., 2016). People were able to recall the orientation, but recalling the orientation impaired their memory for the color when they reported it immediately after. In a larger study, participants who completed 30 trials followed by a surprise question about either orientation or color showed worse precision for the previously irrelevant features (Shin & Ma, 2016). Collectively, these findings suggest that we encode both relevant and irrelevant features, but with varying degrees of precision.”

2. I have concerns about statistical power given that this is only a 1-trial experiment (i.e., no practice trials or experience with the stimuli prior to the critical test). How did the authors determine their sample size? What were the observed effect sizes? It would be very useful for others planning future studies if the authors provided a measure such as Cohen’s d or Hedge’s g with each statistical test. I was not able to see if the authors pre-registered the sample size, because the pre-registration links were set to ‘private’. Since the authors did find an effect, this might be a testament to how strong cuing effects are in working memory. However, I think the authors need to acknowledge potential limitations in their study design and power (e.g., maybe they got a bit lucky, and might actually need more subjects to be well-powered, particularly for the between-subjects analyses that further split up the data).

As noted above, these are large effects measured using repeated measures. We fully expected to replicate them even if we were testing an alternative explanation and using recall as a measure. We have added a discussion of power and our plan to the method section.

3. The authors touch on some weaknesses of a true ‘1-trial design’ on p.18. In addition to the level of noise, I think another downside of a single-trial design is that it likely exaggerates participant confusion and doesn’t allow the participant to learn to trust that the cue is actually valid. In my experience, the average naïve participant doesn’t understand just by reading the instructions what the stimuli will actually look like or what it will be like to make a response on a color wheel. In some ways, the fact that the authors still find an effect with no practice and a true
1-trial design is probably a testament to the robustness of cuing effects! However, the authors might also want to compare this approach to prior work that has first given participants some experience with the valid cue (e.g., ~50 trials) before the surprise trial (see above references).

We agree with the Reviewer. That is what we thought might have been happening in the Eitam experiment: Participants didn't know what to expect and potentially responded with the wrong color. The fact that we still replicated their result after removing that confound suggests that the effect itself is robust and not purely due to confusion. We also now discuss prior research that used multiple trials in the introduction.

4. Unless I missed something, I think the authors forgot to provide information about whether the study was approved by the local IRB and the participants provided informed consent.

We have added this information to the method sections. In study 1, we state:

“The protocol (#09441) was approved by the Institutional Review Board at the University of Illinois at Urbana–Champaign, and participants provided signed, informed consent.”

Study 2 followed the same procedures under the same protocol.

For the online Study 3 the manuscript specifies the waiver of signed consent:

“Participants read an online description and consented to participate by entering the study (under University of Illinois IRB protocol #09441 which includes a waiver of signed consent for online studies).”

5. Did the participants come into the lab only for this 1-trial experiment alone, or was this 1-trial experiment ‘attached’ as part of other ongoing studies? If the latter, it would be good to include this information for completeness.

Participants completed this 1-trial task before doing an unrelated study. We have added this information to the method section:
“Since this was a one-trial experiment, subjects completed the task prior to participating in an unrelated study that examined individual differences in pseudoscientific beliefs. Participants were informed that they would complete this brief task before participating in a longer study.”

6. In the abstract the authors allude to ‘a potential task demand confound’ but don’t specifically describe the confound until later in the paper. It would be useful to come up with a way to succinctly describe the confound in the abstract, so that interested readers can better understand what was done just from reading the abstract.

We added the following text to our abstract:

“Our goal was to replicate this finding while addressing a potential task demand that could have contributed to the results. Specifically, participants might have misread the instructions and mistakenly reported the attended color when asked to report the ignored color first.”

**REVIEWER 3:**

I’m making these exact comments available for both the authors and the editor. Overall, the manuscript tackles an interesting issue and has the potential to become publishable. It is also highly appreciated that the authors went for replicating the original results and then aim to solve the issues of the original design. However, below you can find the issues that need to tackle before resubmitting the paper.

1. In the results of the first experiment is mentioned that performance differences between the inner circle and outer ring won’t be reported because that’s not the interest of the study. However, as mentioned by Eitam et al., the original article the authors aim to replicate, a difference between the inner circle and the outer ring could be cause by differences in the eccentricity of the stimulus. As increased stimulus eccentricity has been associated with poorer performance is necessary to report whether there is a difference or not, as done in the original paper by Eitam et al.

We added a t-test comparing overall accuracy for the inner and outer color as a reviewer-suggested exploratory analysis to test for effects of eccentricity. There
was no meaningful difference, which is not too surprising given that fixation was not constrained and the object was presented at the center of the display (not much of an eccentricity difference between the two color regions). Here’s the added text:

“An exploratory analysis suggested by a reviewer confirmed that there was little difference in the average error for the inner disk (m=45°, sd=52°) and the outer ring (m=41°, sd=56°; t[66]=0.38, p=.708, d = 0.06; where d=(X‾1-X‾2 )/√((σ1^2+σ2^2 )/2)), suggesting that the visual eccentricity difference between the ring and disk likely was unimportant to task performance.”

2. In the introduction is not mentioned why confidence is being measured, nor what are the expected results for this measure. This needs further explanation.

Our preregistration contained the following prose:

"Finally, we will explore confidence ratings for the responses. We expect confidence will be higher for match responses than mismatch responses (repeated measures t-test)"

We added the following to the introduction:

“Finally, Eitam et al. (2013) also found that participants’ subjective perceptions aligned with their objective performance, so we included confidence measures to replicate that aspect of their study as well.”

3. I recommend adding to the figure 3 an equivalent plot but for answers when the ignored color was asked secondly. Right now, is not clear how the distributions of answers differ when recalculating the error for attended instead of the ignored color.

Thank you for this helpful suggestion. We added that plot to Figure 3 and added more discussion of it in the text. We also revised the prose to further clarify the pattern we would expect if participants were and were not making this sort of error. Here is the new figure and surrounding text:
“... That would lead to a bimodal distribution, with some responses clustered near zero and others clustered near 135. Figure 3 plots the difference between the reported color and the attended hue when participants were asked to report the ignored color.

The pattern shown in the figure confirms the possibility that some participants were reporting the incorrect color when first asked to report the ignored color. When asked to report the ignored hue first, some responded with a hue close to that of the attended color and others reported a hue close to what would have been the ignored color. The bimodal pattern of errors suggests that some participants mistakenly reported the attended hue rather than the ignored one when they were asked first about the ignored color. In contrast, when asked about the ignored color second, responses were centered around what would have been the ignored color; when they first reported the attended color and only then reported the ignored one, they reported the intended hue for both.”

4. In the exploratory analysis of experiment 1 it is said that 18 participants responded with a HUE closer to the attended color than the ignored color. However, in the figure 3 there are 19 dots with an error smaller than 90 degrees. Maybe I’m miss interpreting the plot, but if a median split is being apply here,
there should be 19, or 17 (if error < 50 degrees) participants. Given this, is not clear to me what a HUE closer to the attended color means.

Our approach was to check whether the error was greater with respect to the attended or ignored item. Even if the reported hue was far from both, it could have been closer to the attended or unattended color. The reviewer is right that there would be an issue if the colors were always 180 degrees apart, however, they ranged between 90 and 180 degrees apart. Therefore, it is not a median split.

For example, let's assume they differed by 100°, so attended = 0° and ignored = 100°. If asked to report the ignored first, any error that was closer to 0° than to 100° (e.g., 49°) would count as a confusion. It's also possible that a number could be greater than 90° and still closer to the attended. For example, if the attended were 0° and the ignored were 100° and the person responded with 235°, they would be 135° from the ignored and 125° from the attended. That would still count as closer to the wrong hue.

5. The general discussion centers too much on Eitam et al. paper and fails to bring the results to a more general discussion. On the one hand, in the abstract is mentioned that the study of Eitam et al. is one of the few studies to show an effect of attentional selection on memory using simple displays. However, there is no mention to what the characteristics of these other studies are. Are this similar enough to be relevant for these results? Can the results of this manuscript offer insight on the results of those other papers? Secondly, Eitam mentions how, for example, the perceptual load theory states that when capacity limitations are not meet there should not be an effect of attention selection. How can we reconcile the results of the manuscript with theories like the one just mentioned? In general, the discussion fails to integrate the results in a broader theoretical framework.

We have broadened the introduction and framing. Given that the study is intended as a replication/extension of an original work, the theoretical implications are constrained by those of the original work. And, for this sort of empirical replication, we prefer not to greatly expand the theoretical claims.

We have updated the introduction with the following prose:
“Eitam et al. (2013) used a single-trial design to prevent participants from paying more attention to the unattended feature in subsequent trials. Other studies have addressed this concern by only asking about the unattended feature on a final, “surprise” trial. For example, Chen and Wyble (2015a) asked subjects to report a letter on 11 trials and only asked about the letter’s color on the 12th trial. Only 25% of participants could recall the color even though they had focused on the letter. Unlike Eitam et al. (2013), the letters were presented only briefly (200ms), making the representations more susceptible to proactive interference or rapid forgetting (Nee & Jonides, 2013; Oberauer, 2002).

In a similar “surprise trial” design, participants were asked to report an arrow’s color across multiple trials and then were unexpectedly asked to report its orientation (Swan et al., 2016). People were able to recall the orientation, but recalling the orientation impaired their memory for the color when they reported it immediately after. In a larger study, participants who completed 30 trials followed by a surprise question about either orientation or color showed worse precision for the previously irrelevant features (Shin & Ma, 2016). Collectively, these findings suggest that we encode both relevant and irrelevant features, but with varying degrees of precision.”

6. Standardize effect sizes should be reported along the main comparisons (i.e. attended versus ignored and question order.

We have added Cohen’s d for all of the reported independent-groups t-tests and dz for all paired t-tests.

7. Lastly, given the above-mentioned modifications are implemented I would recommend the article to be published after access is granted to the preregistration form and to the data to rerun some of the analysis. If this was possible to be asked before this submission I apologize as I could only work on this review one day before the deadline. Anyway, the manuscript should be checked against the preregistration form.
Again, we apologize that the preregistration was not available for your review. We had intended for it to be, and we have now added details to the method section. The osf projects are now public (https://osf.io/pzavd/).

Finally, maybe more of a comment on the side that an issue to be solved. It looks like both the sequential and the simultaneous settings result in problems for some subjects. Wouldn’t it be a simpler approach to present again a circle and a ring and ask participants to reproduce the stimulus by adjusting the colors using two difference color-wheels? In this way it’s unlikely that a participant will color the outer ring with the color of the inner circle if asked to reproduce the stimulus presented before.

This excellent suggestion motivated the new Study 3 which eliminates such issues.