We thank the editor and the reviewers for their comments, which have helped us clarify the contributions of our manuscript. Below, we address each of the editor’s and reviewers’ comments point-by-point, and where a change has been made to the manuscript in response to a comment, a page reference to the change has been provided and the change itself has been highlighted in the main text of the manuscript.

# Comments from the Editor

## My one big concern is whether your manuscript makes a sufficiently large theoretical contribution to warrant publication in Cognitive Psychology, given that you already published a closely related article (Zhou et al., 2021). It looks as if the present work primarily addresses a limitation of your earlier work by now accounting for intrusion errors.

While it is true that the work presented in this manuscript does follow directly on from Zhou et al.(2021), we see the contribution of the present work as more than simply addressing a limitation of our previous publication. The distinction between intrusion errors and guesses relates to an important question in the wider memory literature: whether memory retrieval fails in a discrete manner (i.e. if guessing in a no-information state occurs)

## Variable-precision models of visual WM are able to account for the fat tails of the error distribution without assuming a mixture, whereas your model (with the circular diffusion model as decision model, and variability in drift rate) does not… it could also be that your way of introducing trial-by-trial variability is not a faithful implementation of the variable-precision idea. My worry that it might be the latter – or some other mis-specification in the model – is reinforced by the fact, which Reviewer 2 pointed out, that the circular-diffusion model fits the error distribution much more poorly than the simpler models fitting only errors. One way forward would be to analyze in more detail what causes the systematic mis-predictions of the error distribution, and to demonstrate more convincingly that the trial-by-trial variability in drift rate is actually equivalent to the precision variability in the van den Berg et al. (2012) model, and related models

Show that the variable-precision model is able to produce the sorts of response error distributions we are talking about, but has RT implications.

Cross-fit simulated data from variable precision and dual-process models, and show that variable precision is unable to account for dual process simulated data.

## Concerning the trial-by-trial variability in drift rate: Whereas in the initial description of the circular diffusion model, the drift rate was a vector of two mu parameters, when you introduce the variability (p. 39), you refer to a single mu parameter. At this point it is not clear to me how the variability is applied: Is the mu parameter on each dimension drawn independently from a Gaussian distribution, or is the same Gaussian error term added to both of them on each trial? The latter would introduce variability only in the norm, but not the angle; the former would introduce variability in both. Either way, there appears to be no variability in the angle that’s independent of the variability in the norm – and I’m wondering whether that could constrain the variability in the angle in a way that makes it difficult for the model to fit the error distributions as well as a variable-precision model fit only to the errors.

True that we assume eta components in x and y are the same (eta1 = eta2), so same Gaussian error added to both dimensions.

## Is the response time measured until the mouse hits the circumference, or until the person confirms their response with a mouse click (assuming they did that)?

There was no confirmation on mouse click, the trial ended immediately upon the mouse reaching the circumference of the response circle, and the response time was measured until this point i.e. no further input was required from the participant. We have added text clarifying this point on p. 21.

## The means given for absolute error on p. 23 (M = 0.06 and 0.08) can’t be correct, unless I completely misunderstand the scale on which they are measured (certainly not radians or degrees).

Instead of mean absolute error, we mistakenly reported values of mean error. These values have been corrected. The results of the statistical test (comparing error between simultaneous and sequential display conditions) remain nonsignificant, and so the surrounding commentary has not been altered.

## You treat space and time as if they were equivalent, which they often are in memory, but in your case, spatial location is the to-be-retrieved information, whereas time is incidental information. They probably play different roles. The temporal context of an item could be used as a mediating retrieval cue (if people use the word to retrieve its temporal context, then use the temporal context to retrieve the associated location), but spatial location can’t have that role. Therefore, it is not clear to me through which mechanism spatial proximity could affect the prevalence of intrusion errors.

Why does it have to be linked to a retrieval mechanism/mediating retrieval cue? Can’t items just overlap in spatial representation even if space itself is the to-be-retrieved information? Q doesn’t make sense to me.

## Your explanation for the absence of a semantic-similarity effect did not convince me. When words are used as retrieval cues, increasing their similarity should lead to more confusions, simply because more similar cues lead to more confusions. That prediction does not hinge on semantic associations between the words being formed.

I don’t think I argued that semantic associations necessarily need to be formed (consciously?), I’m really just talking about similarity. Semantic similarity is just low across the board, so there’s not a strong semantic effect. Increasing the amount of intrusion probability based on the semantic similarity between target and non-targets doesn’t help the model because most targets aren’t really semantically similar at all to the non-targets because of how we have constructed the lists. Maybe I am totally missing the point of this question? just can’t get into his head here

# Reviewer’s Responses to Questions

### If applicable, is the application/theory/method/study reported in sufficient detail to allow for its replicability and/or reproducibility?

#### Reviewer #2

##### The authors need to describe how the models were fit to data. What steps did the authors take to avoid local optima? Given the scale of the models (the largest has 14 parameters) I'd say it's essential to perform model recovery analyses on simulated data to validate the model comparison results.

The models were fit using maximum likelihood estimation (specifically a combination of the Nelder-Mead simplex and Differential Evolution algorithms), using starting parameters sampled between boundaries obtained through exploratory simulation of the models. Using a range of starting points with multiple iterations for each model fit was our countermeasure against local maxima. Text has been added to the manuscript on pp. 32, 41.

Regarding model recovery, we conducted a model recovery exercise but because of the number of candidate models we restricted this exercise to the most complex models (the spatiotemporal, spatiotemporal-orthographic, and spatiotemporal-semantic models), which we believed would be the most likely to have issues with identifiability. The results of this analysis were reported on pp. 46-47 of the originally submitted manuscript, but was only briefly described. To highlight the model recovery analyses, we have added more detailed description of the procedure we followed, and sectioned this text under its own subheading so it more easily distinguishable from the rest of the Experiment 2 results section.

##### p21 There was no time limit, but were participants given any instructions about how to weigh speed vs accuracy?

Participants were not given any instructions about prioritizing speed or accuracy. This detail has been made explicit on p.22

#### Reviewer #3:

##### Page 45. Methods. Could help to say under procedure how many trials total, or range of trials, across the many sessions

The number of trials each participant completed, and the range of trials analyzed included after data exclusion, have been added on p. 46.

##### Page 45. Results section. It could help to orient the reader if you present some brief descriptive text regarding the performance of the different participants, making note, for example, of the range of performance. With P5 performing more poorly and P2 performing particularly well.

We have prefaced the modelling in Experiment 2 on p. 46 with descriptions of the range of performance shown across the participants, making reference to the distributions of response error, as well as the responses recentered on non-target angles (i.e. intrusions).

### Could the manuscript benefit from additional tables or figures, or from improving or removing (some of the) existing ones?

#### Reviewer #1: It’s very hard to see the temporal and spatiotemporal models in Figures 9 & 10. Maybe they could be shown in a duplicate figure, or the authors could note that they overlap almost entirely with the predictions of other models to make it clear that they are there (just hardly visible)?

We have added text clarifying that the predictions of the temporal and spatiotemporal models overlap almost entirely on pp. 41 & 47. We did this rather than duplicating the figures to save space.

#### Reviewer #3: Page 31. Table 1 gives a number and descriptive label to each model. There are some points later in the paper where models 6, 7, and 8 are referred to, but their descriptive labels aren't handy, and I had to come back many pages to remind myself which one was which. It would help the reader a lot to make these labels more locally accessible in the paper. There are a few ways to do this. One could involve giving each model variant a letter-based code, like Intrusion+Guessing: IG; Temporal: T; Spatiotemporal: ST; Spatiotemporal-Orthographic: STO, Spatiotemporal-Semantic: STS, Four-factor: 4F. That does add extra acronyms though. But it would make Table 4, 5, 6, 8 easier to read by putting the short code next to the model number. Alternatively, you could just have some table captions, where the text just lists the relevant model numbers and their full labels, for that specific table.

I opted to include captions to Tables 4, 5, 6, and 8. The letter-based code is a good suggestion, but when I tried implementing that in the text, that I found that I mostly referred to the models by their number and full descriptive label in parentheses, which made the letter codes seem redundant. Later on, when the tables appear, I’m not sure how useful the letter code would be to readers, given I don’t use them consistently in the text itself, so having the full descriptive labels in the caption of each table seemed the best way to make sure that information was locally accessible.

### If applicable, are the interpretation of results and study conclusions supported by the data?

#### Reviewer #2: The mixture model results are consistent with typical findings for VWM in that some - but not all - of the trials a normal+uniform fit ascribes to the uniform component are in fact better described as intrusions/swap errors. However, describing the remaining uniform component as "guessing" is contentious because, as the authors are aware, the all-or-nothing nature of the normal+uniform mixture has come in for a great deal of criticism in the VWM literature, and in particular compared unfavourably to models in which precision varies continuously. Without evaluating this alternative, the claim that "a purely continuous view of source memory retrieval is incompatible with the data" (p43) is premature. The authors may respond that the diffusion model, which forms the second element of the analysis, does allow for variability in precision. However (while mathematically elegant) the circular diffusion model provides a qualitatively poor fit to the data - at both group and individual level - with or without the addition of a uniform component. While I agree with the authors' argument that we should expect a poorer fit to the marginal distributions of error and RT compared to the mixture model (which models only error), this is insufficient explanation for the consistent qualitative failure of the diffusion model and its variants to capture key elements of these distributions (visible in e.g. almost every panel of Fig 11). I don't see how strong conclusions can be drawn from model comparison under these circumstances. On a more minor note, Schneegans & Bays (J Neurosci, 2018) previously proposed a continuous model that at least qualitatively captured error and RT in a VWM task. It would be interesting to know how this compares to the circular diffusion model both conceptually and in terms of fit.

Smith et al. (2020), the circular diffusion model applied to color judgements. There’s something about heavy tails.

Regarding Schneegans & Bays, I don’t think that is a good comparison:

Quote from Schneegans & Bays (2018):

“To focus our analysis on trials in which saccades were directed toward the target item, we fit the three-component mixture model of [**Bays et al. (2009)**](https://www.jneurosci.org/content/38/21/4859?etoc=&utm_source=TrendMD&utm_medium=cpc&utm_campaign=JNeurosci_TrendMD_0#ref-11), which distinguishes between responses directed toward the target, responses directed toward one of the nontarget items, and random responses. Using the method of [**Schneegans and Bays (2016)**](https://www.jneurosci.org/content/38/21/4859?etoc=&utm_source=TrendMD&utm_medium=cpc&utm_campaign=JNeurosci_TrendMD_0#ref-42), we classified as target trials those with a probability exceeding 75% of arising from the target component of the model. Only these trials were included in the analysis of response measures.”

So it’s not really a fair comparison, because Schneegans & Bays remove a lot of the mass in the tails and focus on the central mass which corresponds to targets likely directed toward the target. We could, for the sake of comparison, try doing a similar treatment on our data or trying to fit the full Schneegans & Bays dataset with our the circular diffusion model (or even their model?), but I’m not sure that’s worth the effort since it’s all a bit tangential to the focus of this paper, where what is happening in the tails is the point of interest.

#### Reviewer #3: Page 49. Regarding how in both cases the joint response and RT model does worse at capturing the distribution of response errors. This could make a useful discussion point, some discussion of the worse joint fit. Specifically, I'd be interested to hear a bit more about the misfit in terms of the model mechanisms. Like is it that the model can't fit a distribution with the observed shape, and this is it doing the best it can? Or like if the model could fit the response distribution but this would cause a specific worsening of the RT fit. Or whether there are any promising modifications or additional parameters for future work that could increase the model's flexibility and help it improve the misfit.

### Have the authors clearly emphasized the strengths of their study/theory/methods/argument?

#### Reviewer #2: This manuscript provides an important caveat to previous similar studies that have overlooked the possibility of intrusion errors. The careful consideration of spatial, temporal, orthographic etc influences on intrusion errors is also a strength. I found it a little unclear how exactly the intrusions were incorporated into the circular diffusion model (was there a separate drift process for each item?) and also how successful the model was at capturing the intrusion element of the data - perhaps this could be developed in more detail.

The intrusion component of the circular diffusion model was implemented as a secondary drift process (relative to the primary on-target drift process) with a different mean drift rate and trial-to-trial drift variability (μ2, η2). Instead of having a separate drift process for each non-target item, we circularly shift the distribution of likelihoods of response outcomes (angles) by the distance between the target and each non-target angles. To calculate the overall likelihood of the intrusion component of the model on each trial, we take a weighted average of all the shifted distributions, where the weight for each non-target is determined by the similarity calculation of the particular model. To illustrate how successful the model was at capturing the intrusions, we relied on the recentering method described by Bays et al. (2009): the models predict the degree of central tendency observed in the data when expressed as offset between response and non-target angles.

#### Reviewer #3:

##### Page 25, introduction of the Pure Intrusions model. Towards the end of the discussion, there was a description of how the current modeling framework relates to decision models with multiple competing choices. I think the idea was that the current model can be thought of as representing multiple accumulators. I think the idea was to get across how an equation like Eq 4, which has a term for each of the list items, could involve giving each item an accumulator, and cases in which the participant makes an error that's due to a different list item than the one being probed, that would be like if that item's accumulator crossed threshold prior to the accumulator assigned to the target item. It could be that I'm mangling the details here, but that is ok, because my suggestion is to add a bit of extra description here (or earlier in the intro I suppose?) that foreshadows that point from the discussion. Here, the description of the equations and how they produce model behavior was clear, but I could have used a bit more detail regarding the broader cognitive theory. Specifically, just setting up some of the ideas that will be important in the discussion. I think I'm just asking for a bit of description of the corresponding process model involving multiple racing accumulators. I would leave it to the authors whether and how to best accomplish this.

I don’t have a very clear sense of how to provide the detail this reviewer is asking for, they seem to have described the idea of the multiple accumulator version of the model pretty much how I thought about it. Not sure what detail I can provide, might need help addressing this one in a satisfactory way.

##### Page 27. As someone who cares about free recall, I thought Figure 5 was very interesting. Fig 5A shows examples of how this lag-based intrusion probability function is affected by the model parameters. It would really pay off nicely if there was a panel here, or another figure that shows the best-fit version of this function for one or both of the experiments.

Figure 7 shows lag-conditioned recentered response error predictions of the models in Experiment 1. We have added a panel to this figure which shows the lag-based intrusion probability using the fitted parameters of the temporal and spatiotemporal models to data from Experiment 1.

### Have the authors clearly stated the limitations of their study/theory/methods/argument?

#### Reviewer #2:

##### The authors significantly understate the challenge of obtaining reliable RT measures in online experiments. They cite a paper that found an ~25ms lag was introducing by using JavaScript to PsychToolbox (on one specific model of PC), but do not consider the considerable and machine-specific display lag (>100ms is not unusual on an LCD display) or response lag (which can also be significant).

Machine specific lag. Maybe? As long as it is consistent within a participant and across the participants’ trials I’m not sure if this really matters? A consistent machine-related lag would be absorbed by non-decision time I guess. What is response lag? Measurement taken by javascript seem pretty good, but maybe I need to provide more evidence, maybe some papers by Roger?

##### Relevant to this, did the experimenters obtain any information about specs of the machines the experiments were run on? As a case in point, the Methods say "mouse", but is it possible some participants used trackpads?

We collected information about the browser, including the reported screen resolution and browser size, but did not collect any information about the actual hardware participants used to access the experiment. The only specification was that people complete the task on a computer (and not say, a tablet or phone), but participants were not explicitly instructed to use a computer mouse, so it is possible that some participants completed the task on a laptop and used a trackpad or other peripheral devices instead of a computer mouse. It is possible that a degree of the variability between participants is due to this lack of control over how participants interacted with the experiment. To reflect this uncertainty, we have replaced all mention of “mouse” with “cursor”. We do not expect this to fundamentally affect the results and interpretation of our analyses, since our models were all fit on an individual- participant level, and participants were instructed to keep their setup consistent across sessions.

### Does the manuscript structure, flow or writing need improving (e.g., the addition of subheadings, shortening of text, reorganization of sections, or moving details from one section to another?

#### Reviewer #2:

##### p5 It seems odd to provide references for VWM models that combine slot and resource concepts, but not for the slot (e.g. Luck, Vogel, Pashler, Cowan) or resource (e.g. Palmer, Bays, Ma) models themselves.

##### I found the first full sentence on p4 hard to parse - maybe split into two sentences?

Dual-process models? Split into two sentences as suggested.

# **Reviewer Comments**

#### Reviewer #1:

*In this paper, the authors take a look at the patterns of responses in a continuous-response memory task to understand the roles of intrusions and guessing. The authors do a good job of demonstrating how "pure" guesses alone cannot account for the patterns of errors observed in the experiment (e.g., Figure 6B) and quantitatively showing that this is the case in model comparisons, illustrating the importance of intrusions in the decision process. They also develop and test several dynamic models - using the circular diffusion model - instantiating hypotheses about intrusions due to spatial, temporal, orthographic, and semantic similarities among stimuli. While the results with respect to semantic intrusions seem somewhat inconclusive (given they were not manipulated), the authors were able to show that intrusions constitute a significant portion of responses and that they are influenced by spatiotemporal relationships among stimuli during the experiment.  
  
In general, I think the experiments and modeling are carried out very carefully, and the results clearly and logically correspond to the conclusions that the authors have drawn. I also think the work addresses a clearly defined issue in the memory literature regarding intrusions versus guessing. The justifications for the experiments (and particularly the small-N design in Experiment 2) are convincing, although in retrospect I imagine the authors would agree it would have been good to explicitly manipulate semantic similarity. Otherwise, I have relatively few remarks, mainly about contrast effects and the nature of "pure" guesses, that I think could be easily addressed in a revision. I outline these below.  
  
1. This is maybe just my eye / mind playing tricks on me, but it almost looks like there is a suppression or maybe contrast effect in the data , shown by "dips" in response frequencies at angles around +/- pi/2 (Figure 7). Given the assimilation effects resulting from the spatiotemporal gradient, is it also possible that there is a contrast effect in the spatial and temporal domain? For example, is a participant less likely to respond consistent with an intrusion from a distractor placed at 90 degrees versus one placed at 45 degrees? It also seems possible that this could happen with semantic similarity. For example, if a very different word to the target was recently presented at a nearby location, do participants shy away from that contrast-word location and become more likely to respond toward the opposing side of the target? If either of these types of contrasts appear in the data, it might be worth incorporating these kinds of effects into an additional model or showing that the present models can capture them.  
  
1b. It might be difficult to detect the semantic effects, since (as the authors note) semantic similarity wasn't intentionally manipulated. So I could see these effects not showing up. If semantic contrast effects are of interest, I could certainly see the authors testing these effects by deliberately manipulating them in an additional experiment, although I would understand if they prefer to save this for future work.  
  
2. Maybe it's just how I've thought about it, but "guessing" has always seemed like a blanket term covering a bunch of processes including intrusions, anchor effects, and other idiosyncratic processes that we don't feel like modeling (possibly because doing so would be totally impractical!). In that sense, I am glad the authors are examining the distinction between "pure" guessing and intrusions and demonstrating that intrusions are a core part of the data. At the same time, it may be worth a mention that the guessing component of the model almost certainly covers some other idiosyncratic elements that aren't completely stochastic and uniformly distributed (which is what I think of as "pure" guesses) but participants doing things like repeating previous responses, making responses in cardinal directions because memory strength is too low, or something else that is inconvenient to try and model at present. In that sense, the intrusion + guessing model might be thought of as intrusions + [we aren't sure yet, but let's assume it follows a uniform distribution]. Calling it pure guessing makes it sound like there isn't any more theoretical work to be done to understand that latter component, which seems (to me, at least) like it isn't the case.*

It is true that the guessing component of the model probably catches non-uniformly distributed responses of the kind the reviewer mentions, and we agree it is somewhat misleading to refer to this as pure guessing. We required a useful shorthand that distinguishes between responses guided by some information from memory (i.e. target and intrusion trials) and those that are not. The “pure” part of the pure guess model label is meant to reflect the fact that the model does not include any intrusion mechanism as part of its explanation for errors, to parallel the terminology of the “pure intrusion” model which has only intrusions as a mechanism to explain non-target directed responses. The intention is not to claim that all guesses are homogenous or “pure” in that sense of the word, but again I can see how it is misleading, especially since the original manuscript never acknowledges the different things participants could be doing when not guided by the trial target and non-target items.

To strike a compromise between keeping the focus of the paper on the distinction between intrusions and errors covered by the guessing descriptor, while acknowledging the idiosyncratic nature of the latter, we have added text to the general discussion paper on p. 57.  
 *3. Minor notes:  
Abstract: "circular diffusion [model] of decision making"*

Fixed. *It's very hard to see the temporal and spatiotemporal models in Figures 9 & 10. Maybe they could be shown in a duplicate figure, or the authors could note that they overlap almost entirely with the predictions of other models to make it clear that they are there (just hardly visible)?*

Noted, rather than duplicate (addressed earlier).

#### Reviewer #3:

*The paper is very impressive in its technical achievement, and well organized in terms of getting the message of the modeling work across to the reader. It uses an important modeling framework (the circular diffusion model for continuous judgments), and demonstrates how the model can be constrained by both continuous responses and their latencies. The paper makes a solid contribution, and is relevant to several ongoing theoretical debates regarding memory for source characteristics of studied materials. It will certainly be of interest to a wide group of cognitive modelers of memory, attention, and decision making. It will also be of interest to researchers in the visual working memory community, and the general community of psychologists interested in behavioral analysis of list-learning tasks (recognition, cued recall, free recall, and other tasks).  
  
Minor and typographical  
  
Page 36, line 5, missing close parenthesis after "1 and 2"  
Page 36, line 10, "that this" -> "that is"* Fixed these typographical errors.