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 comments and questions raised, and where a change has been made to the manuscript, a page reference to the change has been provided and the change itself has been highlighted in the main text of the manuscript. Where the comments overlap with each other, particularly in relation to the fit of the circular diffusion model, to avoid repetition we have pointed to our response to other comments which are relevant.

# 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.

Whether guessing occurs in memory tasks is a hugely consequential question, and our results emphasize that source intrusion effects cannot be overlooked in this debate. Our analysis of intrusion errors in the continuous-outcome source memory task, guided by the work by Bays and colleagues on swap errors in VWM, links the debate around threshold and continuous models of source memory to the work on capacity limitations and interference effects in the VWM literature in a straightforward but novel way. We found that the estimated guessing rates were substantially reduced by the inclusion of intrusions in the model. Through this estimate was further reduced by elaborations of the intrusion component to capture systematic effects of similarity on intrusion probability, using several important variables from the episodic memory literature, the contribution of guessing always remained. As an important caveat to this, reviewers raise excellent points about whether the uniform component of the model can be said to reflect pure guesses, which we have acknowledged and discussed in our revision. We wish to highlight that while intrusions and swap errors have been of interest to the field, RT models of intrusion effects like the one we present are novel especially with continuous responses. On top of that, our modelling of similarity gradients represents an early step towards a more comprehensive understanding of how interactions between items impact the response error and RTs observed in continuous-outcome tasks.

In summary, while it is true that the present study does build upon our findings in Zhou et al. (2021), we believe it does so in a way that relates to important theoretical questions in the memory literature at large, and the characterization of the manuscript as addressing a limitation of our own work minimizes its theoretical contributions.

## 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

We believe trial-to-trial variability in drift rate in the circular diffusion model is an appropriate implementation the variable precision idea as employed in VWM, and the circular diffusion model has been successfully applied in more typical VWM contexts. For example, Smith et al. (Psych Review, 2020) used the circular diffusion model to model a color discrimination task with three levels of difficulty (chromatic noise). At lower levels of noise (i.e. when the task was easier) the circular diffusion model provides a good fit of the distribution of response errors without resorting to any uniform component, in accordance with the variable precision idea. However, at higher levels of noise, the tails of the response error distribution are heavier, similar to what we see in the source memory task. As stated in Smith et al. (2020), any distribution of response errors that can be produced by the van den Berg (2012) variable precision model, which assumes a mixture of von Mises components, can in theory also be produced by the circular diffusion model since the circular diffusion model predicts a von Mises distribution of decision outcomes for any given values of drift rate and decision criterion. Naturally, the challenge of jointly fitting response times is that the parameters that provide the best account of response times can simultaneously result in a failure on the RTs (see response to Reviewer #3). We have attempted, in our responses to the specific questions of Reviewers #2 and #3 below, to analyze the cause of the misfit in more detail. Ultimately, our position is that while the circular diffusion model certainly misses some important structure the data, visible in both the marginal distribution of response error and in the joint distribution of RTs and errors, we can still illustrate that a substantial proportion the heavy tails seen in the former are explained by intrusion effects.

## 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.

Thank you for this very helpful suggestion, which we have implemented and found to improve the fit of the model. Because the effect of increased variability in the drift angle is to smear the central peak of the predicted distribution of errors, allowing this variability to be estimated independently of variability in the norm does substantially improve of the model by allowing for smaller values of trial-to-trial variability in the tangential direction and larger values in the radial direction. We have refit the circular diffusion model with this modification to both Experiment 1 and 2 data, and updated the quantitative results and figures. Because a similar degree of improvement was seen across all variants of the circular diffusion model, no change was made to the results of the model comparison, or the broader narrative of the manuscript as it relates to the intrusion effect. However, some of the systematic misfits pointed out by Reviewer 2 remain, and we offer a separate reply to those concerns below.

To explain why, we introduce drift as a vector of two elements, but only estimate one parameter for mean drift rate. In our implementation of the circular diffusion model, the stimulus angle is “canonically oriented”, meaning the target is always positioned along the positive x-axis. Drift is then expressed in Cartesian, rather than polar, co-ordinates. In canonical orientation, x corresponds to the radial component (norm) of the drift rate, and y corresponds to the tangential component. There is only a single μ parameter because we don’t estimate the parameter corresponding to the mean value of the angular or tangential component of the drift vector (μy), it is fixed at zero because we do not see a bias or shift in the center of the distribution of response errors. We implemented the intrusion component of the model by generating predictions in the canonical orientation and then circularly shifting the response angle by the offset between the stimulus and non-target angle. The omission of these details which may also contribute to Reviewer #2’s comment that they were unsure how the intrusion component was actually implemented, and we have added some more detailed description to the main text of the revised version to address this on pp. 42-43.

## 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.

We agree that space and time may not play equal roles, given that memory for location is the goal of the task, while memory for time is incidental. In our models with spatiotemporal similarity gradients, we allow for different degrees of contribution of temporal and spatial similarity values to the overall probability of an intrusion error occurring. This is the *ρ* parameter, which determines the weight of spatial similarity compared to temporal similarity (1- *ρ*). In the best-fitting spatiotemporal circular diffusion model to small-n dataset (experiment 2), the average parameter estimate was 0.86, which suggests that spatial proximity had a comparatively larger effect on intrusion errors compared to temporal similarity (in experiment 1 this value was 0.56).

It is possible that time and space might play different roles given the importance of the latter to the demands of the task, while temporal information is incidental, but if they are different it is unclear to me in what way they should be different. Consequently, I think it is reasonable to treat them similarly in the model.

## 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.

We agree that more similar cues should lead to more confusions. Our explanation for the lack of a semantic similarity effect was that the similarity between words was so low that any effect of semantics was dwarfed by comparatively larger effects of spatiotemporal context and even orthography. As a result, the improvement to model fit when incorporating semantics into the model was too small to support the parameter penalty incurred by its inclusion. Our point regarding the formation of semantic associations was not necessarily to claim that semantic associations are not formed, but to reconcile this explanation with findings in the free recall literature, where semantic similarity in even unrelated lists *do* lead to more errors. We have attempted to clarify our position in the revised manuscript (pp. 60-61):  
  
“Our explanation for the absence of a semantic similarity effect, that overall similarity was too low to exert a noticeable effect, would appear to be inconsistent with previous findings that semantic similarity can exert effects even in lists of unrelated words, including transitions between list words in a free recall task (Howard & Kahana, 2002b; Morton & Polyn, 2016) and predicts false alarms in recognition memory (Osth et al., 2020). This may be due to differences in the particular demands of the source task. When items are presented individually on a study list, they are associated with the list context and with other items on the list (e.g., Gillund & Shiffrin, 1984). Semantic similarity can exert a large effect on recall transitions because each recalled item is used as a cue for further retrievals – semantically similar items facilitate this process. In recognition memory, items are matched against all of the other items on the list to produce an index of global similarity that is the basis of the recognition decision (e.g., global matching: Clark & Gronlund, 1996; Osth & Dennis, 2022). Thus, semantically similar items on the list will contribute to the global similarity and increase the likelihood of a false alarm. In a source task, in contrast, items are associated with the source of their occurrence, and it is not necessarily beneficial to associate items with other list items. At retrieval, both the list context and the item cue are used to retrieve the specific source location. Given the lack of associations formed between items, semantic similarity between the items may exert less of an influence in source judgements than in a task such as free recall.”

# 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.

These details have been added to the manuscript on p. 32, which describes the models of response error:

“To find parameter values that optimized the fit of each model to the observed data, a maximum likelihood estimation approach was taken using the differential evolution algorithm as implemented in the R package DEoptim (Mullen et al., 2011). To avoid local optima, we fit each dataset five times, using starting parameter values sampled randomly between upper and lower boundaries”

And on p. 41, describing the circular diffusion models:

“To fit the circular diffusion models to data we used maximum likelihood estimation using the Nelder-Mead simplex algorithm, and like our approach to fitting the response error models, we attempted to avoid local optima by fitting each participants’ data five times, sampling starting points for each iteration between boundaries obtained through exploratory simulation of the models.”

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:

“One concern in model comparison is the diagnosticity of the results when the models can make similar predictions. To evaluate the extent to which our models mimic each other, we conducted a model recovery exercise, focusing specifically on the spatiotemporal, orthographic, and semantic similarity models. We restricted this exercise to the most complex models as these were the ones most likely to lead to parameter tradeoffs and therefore be most difficult to identify. The parameter values for each model that resulted in the best fit to each participants data was used to generate five simulated datasets for each participant, each with the same number of observations as the empirical dataset for that participant. Each simulated dataset was then cross-fit with the same set of models, and using the AIC as the fit statistic, we observed the number of times that the generating model was recovered as the best fitting model. Across all of the simulated data, the spatiotemporal and orthographic models were successfully recovered in 80% and 84% of cases respectively. However, the semantic model was not recovered in any of the simulated data, for which the spatiotemporal model was preferred in all cases. The likely reason for this failure was because the semantic similarity of the stimuli was not explicitly manipulated when study lists were constructed and the average similarity between items was consequently low. We elaborate this point in the discussion section to follow. Because the differences in semantic similarity are minimal in these data, the estimated value of χ was so low that simulated data generated from the fitted parameters could not be distinguished from the spatiotemporal model”

##### 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:

“There was no time limit on the decision task, and participants were not instructed to prioritize either speed or accuracy in their responses. A schematic for one trial in each of the phases is shown in..”

#### 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:

“Each participant completed a total of 1200 trials over the course of 10 sessions. Trials with a response time of faster than 300 ms or slower than 7000 ms were excluded, resulting in a range between 1121 to 1197 trials remaining for each participant, or the omission of 1.73% of total data.”

##### 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. 49 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):  
  
“Participants varied in their overall level of performance in the task, as well as the extent to which intrusions appear to contribute to errors in source responding (Figure 10). Participants 1 and 2 responded with less error, and when responses are recentered on non-target angles there is little indication of intrusions. In contrast, Participant 5 performed more poorly, and the recentered error plot shows much stronger evidence for intrusion errors.”

### 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.

We have opted to include captions to Tables 4, 5, 6, and 8. The letter-based code is a good suggestion, but when each model is introduced, they are referred to by their number and full descriptive label in parentheses, which made introducing the letter codes seem redundant. Later on, when the tables appear, the letter code is unlikely to be as helpful as they are not used consistently in the text itself. 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.

We acknowledge that the circular diffusion model systematically misses elements of the data, particularly in terms of the steepness of the central peak of the marginal distribution of response errors (i.e. the precision of target responses), and in terms of the joint distribution, the models also consistently predict faster tails in the response times for the most accurate responses (i.e. there are a proportion of very slow but very accurate responses that the model misses). In our response to Reviewer #3 immediately below this comment, we explore in more detail why these misses occur with reference to the specific constraints imposed by the response times. It is possible that future refinements to the model, or an alternative modelling framework entirely, may attribute more responses to another source of variability in the memory process. We concede that it is probably premature to rule out all continuous views of source memory and have softened the language accordingly (now on p. 46):  
  
“Experiment 1 suggests that previous threshold models may have similarly overestimated guessing rates (Harlow & Donaldson, 2013; Zhou et al., 2021). However, the poor fit of the pure intrusion models, both in terms of error and joint error and RT data, also do not support a purely continuous view of source retrieval.”  
  
With this said, we maintain that the circular diffusion model is an appropriate tool in distinguishing between errors due to non-target items and other sources of error, which is the primary goal of our study. We would also like to mention that this is possibly the first investigation focused on modeling targets and intrusions in continuous report while simultaneously considering continuous responses and response times at an individual trial level.

The comparison of our model results with those of Schneegans and Bays (2018) is not a straightforward one. The main issue is that for Schneegans and Bays (2018), the key point of interest was investigating variability in responses directed to the target. To achieve this, they identified trials likely driven by the target by fitting the Bays et al. (2009) three-component mixture model (analogous to Model 3 in our study), and filtered for responses with greater than 75% probability of arising from the target component of the model. The full RT model was fit only to the filtered data. The same sort of treatment, applied to our data, would effectively slice off the distribution of errors around the shoulders, removing much of the mass around the tails, which is where the focus of our study is: to distinguish between item and non-item related responding in the tails of the distribution. It is not clear how well the Schneegans and Bays model would fit more typical heavy-tailed data, but as we describe in our response to the editor’s comment #2, Smith et al. (2020) showed that the circular diffusion model is also able to provide a good fit of data without heavy tails.

Ultimately, we believe the current model, though it is clearly missing important structure in the data, still gives sufficient support to our central claim, which is that a substantial proportion of errors previously ascribed to uniform guessing by prior work is explained by intrusions, which as you say is consistent with the work on swap errors in the VWM literature.

#### 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.

The model can produce the observed response distribution through a variety of mechanisms. The most straight forward way to do this, relative to the fitted estimates we report, is to simply increase the mean drift rate of the memory process. Figure 1 below shows simulated data using the group-level estimates of the data from Experiment 2.

**Figure 1**

Diagram

Description automatically generated

The leftmost column shows the original parameterization of the spatiotemporal model. In the middle column, we allow the tangential and radial components of drift variability to be different in response to the editor’s suggestion, which slightly improves the fit of the model by reducing the degree to which tangential variability “smears” the central peak of the distribution. However, the key misses of the original model remain. To prioritize the fit to the distribution of response errors, we can increase the mean drift rate (i.e. the precision of responses), but this causes a worsening of the RT predictions, specifically in the prediction of much faster responses in the leading edge of response times for more accurate responses than we observe in the data.

The other systematic misprediction we see in the model is in the top left area of the joint quantile plot. Simply put, there are a proportion of accurate responses that occur with much slower RTs than the model predicts, which is shown in the plot in the 0.9 RT quantile (triangle) being further up the y-axis (slower) than the model (Figure 2). We explored the minimal modification required to improve this aspect of the model, and the most straightforward way to do this is to assume that targets responses are not homogenous, but that on a proportion of trials people require more information than usual, while working with the same quality of evidence. This was implemented in the model as a probability mixture with a secondary diffusion process centered on the target angle, with the same drift rate but a larger decision criterion value. This modification introduces two additional parameters: one for the proportion of “second criterion” target trials, and another for the value of the second criterion parameter. When the proportion of trials is relatively high (15%), the RT quantile predictions are improved across the entire range of errors but particularly for accurate responses, and the marginal response error predictions are also better. However, the downside is that the second proportion of memory responses results in a clear bimodality in the predicted response time distribution, which is not present in the data. At a more moderate proportion (5%), we get some improvement in the joint quantiles without compromising the marginal RT predictions.

**Figure 2**

Diagram

Description automatically generated

While these post hoc explorations offer potential recourse for the misfit of the current model, with plausible psychological interpretations, it’s worth noting that implementing these changes in the modelling fitting routine may not be straight forward. For example, by adding a discrete mixture of decision criteria to the model, we complicate an already fairly complex model, and this would likely come at the cost of parameter identifiability. Given we are primarily interested in estimating the proportion of intrusions, we offer these simulations as a proof of concept that illustrates what is missing with the current model, rather than a definitive next step.

### 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, which is mixed with the primary on-target drift process, with a different mean drift rate and trial-to-trial drift variability (μ2, η2). This is identical to how intrusions are implemented in the models of response error alone. 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. We have clarified this by extending our introduction of the circular diffusion model on p. 42. 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.

We have added some discussion of the multiple accumulator framework, which you have summarized accurately here, to the manuscript on p. 43 (quoted below). This is later on in the paper than perhaps desired, but it seemed clearest to discuss how intrusions can be implemented in the circular diffusion framework after the circular diffusion model had itself been introduced. On p. 25 when we provide the equation for the Pure Intrusions model, we are using the simple case of a model of just response error to set up the idea of intrusions, and the worry is that introducing the idea of multiple accumulators operating over time would confuse the reader at this stage. With that said, we hope the description on p. 43 better sets up the idea that we expand upon at the end of the discussion section:  
  
“In this study, we take a mixture modelling approach to estimate the relative contribution of memory, intrusion, and guess responses, an alternative way of conceptualizing how these processes interact would be to associate each item in the study list with a separate evidence accumulator, which compete over time to reach a decision criterion in a timed race structure (Hawkins & Heathcote, 2021). Because our goal in this study was primarily to estimate the proportion of intrusions and distinguish them from guesses, we have restricted our modelling to the mixture approach for simplicity, but we discuss the process interpretation implied by multiple accumulators in the general discussion.”

##### 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, which we have copied here:

Diagram

Description automatically generated

### 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).

Ratcliff and Hendrickson (2021) considered the potential issues with obtaining RT measurements in online experiments, with specific reference to estimating diffusion model parameters, and found that when fast outliers are filtered out (as we do in the present study), results are consistent with in-person data collection. We have included a citation to this paper as well as some context to better justify our methodology on p. 19:

“…biases of this magnitude are negligible for the purposes of the inferences we wish to draw about RTs in our task (for a comparison of the reliability RT measurements in online data collection, particularly with regard to diffusion model parameter estimation, see Ratcliff & Hendrickson, 2021).”

Furthermore, because we analyzed data at the level of individual participants, the reliability of measurements within each participant is more important than the absolute difference between running the experiment online compared to typical laboratory conditions. Consistent machine-related differences, in terms of display and response lag, would be captured by the non-decision time component of the diffusion model.

##### 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 mentions 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.

Added the suggested citations and a few others on p. 5

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

Split into two sentences as suggested. Excerpt:

“These alternative retrieval processes are not mutually exclusive but may instead co-exist. For example, in dual-process models, recognition is thought to be supported by two processes: familiarity, which yields a continuous measure of strength for an item in memory, and recollection, which yields rich information about the study event itself when memory strength exceeds a threshold but fails absolutely when it does not.”

# **Reviewer General 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.*Histogram

Description automatically generated with low confidence

The model doesn’t have a specific mechanism to capture an effect like this.

#### 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.

We agree that a natural extension of the findings we report is to directly manipulate the semantic similarity of the word lists and seeing if there is a stronger contribution of semantic similarity on intrusions when semantic similarity within the list is higher. We are currently investigating this in a follow-up experiment, but we believe this investigation falls out of the intended scope of the present study. Our goal in this manuscript was to evaluate the extent to which responses which previous studies have attributed to uniform guessing are systematically related to non-targets information. We discuss this on p. 61 of the revised manuscript:  
  
“Because this study was designed to reassess previous inferences about discrete retrieval failures (Harlow and Donaldson, 2013; Zhou et al. 2021), and none of those studies manipulated orthographic or semantic similarity, our experiments were designed to best reflect the conditions of the previous experiments used to argue for the contribution of a guessing process. An effect of semantic and orthographic similarity on intrusion probability may be observed in a future replication of the current paradigm with word lists that are constructed specifically to maximize similarity along these dimensions.” *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.*

We agree that the guessing component of the model almost certainly 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. Describing the model without intrusions as a “pure guessing” model was intended to serve as a shorthand that distinguishes between responses guided by some information from memory (i.e. target and intrusion trials) and those that are not. The intention is not to claim that all guesses are homogenous or “pure” in that sense of the word, but we acknowledge the terminology may be 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 other sources of error, while acknowledging the idiosyncratic nature of the latter, we have added discussion of this point to the general discussion paper on p. 57:

“Although we refer to all responses not directed towards items on the study list (i.e. targets and non-target intrusions) as guesses, it bears mentioning that such responses are not necessarily pure guesses, in the strict sense that guesses are completely uniformly distributed. For example, in cases where participants resort to responding in cardinal directions, repeating previous guesses, or other idiosyncratic elements when memory strength is low, these responses would be captured in the uniform component of our model. In the same way that intrusions appear uniform relative to the target, but can be distinguished from guesses when recentered on non-target angles, the remaining guesses may in turn be further decomposed into other processes.”

*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.