We would like to thank the editor and the reviewers for their comments. Below, we address each of the 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. Additionally, we wish to revise the title of the manuscript to include a subtitle to clarify the main theoretical questions of the study. The updated title reads: "A circular diffusion model of continuous outcome source memory retrieval: Contrasting continuous and threshold accounts".

**Reviewer 1**

**R1C1: I do not see any confounds that would limit the ability of draw conclusions from the data. Although the sample size might be considered small (n=20), I believe it is likely sufficient for the issue under investigation. I am not aware of any method to compute statistical power on model fit statistics as those reported here.**

We believe that a small-*N* design is appropriate for this study. Each participant in our sample completed 720 trials in each phrase of our experimental paradigm. With a large number of observations for each participant in the sample, each observation can be thought of as a replication of an effect for that participant (Smith & Little, 2018). In this sense, experimental power in our study is concentrated at the level of the individual, rather than at the level of the experimental sample.

**R1C2: Model recovery simulations should be included in supplemental materials. The conclusions reported here rest on evaluation of model fits. This is particularly important because (at least to my knowledge) this is the first report using the threshold and hybrid variants of the diffusion model. Such simulations will help give readers a sense of how much one model can mimic another.**

Model recovery simulations have been included as supplementary material attached to this manuscript. We used the variable-precision and threshold models to generate a single simulated dataset using the parameter values that provided the best fit of each model to the empirical data from each participant. We excluded participants whose source response distributions were at or near uniform levels, as that data would not be informative in distinguishing between the models. We then cross-fit the simulated data with the two models, to assess how often the model that provided the best fit matched the model that generated the data. Our finding was that when the continuous and threshold models were fit to simulated data generated by the other, the appropriate model was recovered for most participants both when the Variable-Precision model and the Threshold model were used to generate data. We believe that this demonstrates that the models are sufficiently distinguishable from each other.

|  |  |  |  |
| --- | --- | --- | --- |
|  |  | Generating Model | |
|  |  | Variable-Precision | Threshold |
| Fitting Model | Variable-Precision | 12 | 0 |
| Threshold | 2 | 14 |

**R1C3: Relatedly, does the implication that the threshold and hybrid models make no specific prediction for RT data suggest that these models might be more ‘theoretically flexible’?**

* What are the a priori predictions? Do we need a simulation spanning a range of parameter values?
* The threshold model a priori predicts a heavy tailed response error distribution, but what are the a priori predictions of the threshold model in the RT data? Is its ability to capture the RT data just due to model flexibility
* Different boundary settings for the guess state and the memory state does contribute to flexibility
* What do the QQ plots look like when you simulate the threshold model?

**R1C4: BIC (and other related fit statistics such as AIC) makes the assumption that each parameter is weighted equally in terms of complexity, at least given how it imposes a penalty on the data. Other measures, such as normalized maximum likelihood/minimum description length, apply penalties based on the functional form of a model parameter. Is there any reason to believe that the results would change with a different fit statistic such as NML/MDL? That is, is the addition of diffusion process in the threshold and hybrid models introducing more flexibility, and thus overfitting, the data? I don’t know if this can be addressed or if these other fit measures have been developed for diffusion models. Nonetheless, I think it is important to state and discussion the assumptions made with the fit measure used.**

While we recognize the advantage of alternative ways of quantifying model complexity compared to the BIC/AIC approach of penalizing the number of parameters, applying penalties based on the functional form of parameters in the circular diffusion model is difficult to implement.

Regarding the results changing with a different fit statistic, it is worth noting that the difference between the models’ performance is large, which is evident in simply observing the graphical fits of the models to data. We could triple the penalty for the number of parameters in our models and the threshold models would still be favored over the variable-precision model.

**R1C5: The experiment was designed to be similar to the short-delay condition in Harlow & Donaldson (2013), but not the long-delay condition in the same study. Many other paradigms that have implemented similar continuous source measures use much longer lists, and thus study-test lags and interference levels. These factors might be critical for model fits as they will influence the overall level of performance. Thus, the discussion should make clear that it is unclear how these results generalize to other paradigms, and such conclusions require future work. That being said, I do not believe this limitation detracts at all from the importance of this study.**

A note has been added in the manuscript (page 45) to acknowledge that changes in overall performance could have an influence on the relative performance of the models.

**R1C6: In the Introduction (page 3), it is stated that “SDT and dual-process models make similar predictions about item recognition because they both assume that recognition relies on familiarity, which is continuously distributed”. I do not think this is an accurate representation about dual-process models which propose that even on item recognition tests memory can be supported by a mixture of recollection and familiarity (though, I do acknowledge that item recognition can be supported by familiarity even if recollection fails). Part of this assumption depends on which dual-process model is being adopted. The sentenced should be updated.**

This language on page 3 has been updated to reflect that the predictions of purely continuous and dual-process models can resemble each other when the relative contribution of familiarity is high in supporting recognition memory.

**R1C7: There is a Figure 8 referenced in the main text (pg. 36) that I could not find. This might have been referring to Figure 6.**

The reference was intended to refer to Figure 5 and has now been corrected.

**R1C8: Were source memory judgments elicited for both old and new items, or just for old items? I was not able to determine this from the text. If new items were given source responses, is there any benefit to modeling source memory responses to false alarms to get measures of bias? This last point likely won’t need any modification to the manuscript and can be ignored if the authors wish.**

Source memory judgements were only elicited for old items. A line has been added on page 20 to clarify that source judgements were only made for old items. Collecting source memory responses to false alarms is an interesting idea and would give a measurement in source response bias, but it does not seem like there is a systematic bias in the source responses we did collect.

**Reviewer 2**

**R2C1: The most important thing I would like to see in a revision is model fitting for source accuracy across the different levels of recognition confidence. I believe that the current version fits models to a single pool of items, which ignores what seems to me to be the most theoretically meaningful aspect of the data: the change in the source accuracy distribution across recognition confidence levels. The results showed that high-confidence recognition is associated with two changes in the source distributions: the proportion of retrieval failures decreases and the precision of retrieval successes increases. At a conceptual level, this seems like evidence against a pure threshold model. Wouldn’t such a model predict that changes across a predictor variable, like recognition confidence, should only influence the proportion of failed trials? The manuscript should say more about the implications of this result and…**

* I don’t know if it's true that precision actually increases with recognition confidence if we look at mutually exclusive confidence levels. I need to code up the Zhang and Luck mixture model myself and fit it to each confidence level individually to see if the precision parameter actually changes.

**R2C2: ...ideally, fit the models jointly to source accuracy distributions for high-confidence and lower-confidence recognized items. Can the threshold-style models continue to match the data if they can only change the proportion of failed trials for the two different confidence ranges?**

We jointly fit the threshold and hybrid circular diffusion models to items recognised with high confidence and low confidence. For the threshold model, allowing the proportion of failed trial to differ across confidence ranges improved the fit of the threshold model relative to the original parameterization, which constrained all parameters to be equal across the confidence ranges. We also compared a version of the threshold model that allowed both the proportion of failed trials as well as mean drift on successful trials to differ, and found that the improvement of fit associated with the additional parameter did not outweigh the BIC penalty. In the manuscript, this is used as the basis to state that the proportion of guesses changes across confidence ranges, while the precision of non-guesses does not, supporting a purely thresholded view.

**R2C3: By the way, I think it would be better to have mutually exclusive categories of recognition confidence, like recognized with high confidence, recognized with lower confidence, and unrecognized. Currently, results are reported for overlapping categories, like all recognized versus specific confidence ranges.**

* TO DO: Code up the Zhang and Luck Mixture model and re-do those fits
* The Zhang and Luck mixture model as originally reported showed a change in precision and not mixing proportion across confidence ranges. We did not allow for different parameter values across confidence ranges.

**R2C4: A revision should also discuss the implications of joint recognition and source modeling for the continuous report paradigm/results. I realize that adapting this complex model to joint recognition and source data would be an additional major theoretical achievement, so I’m not asking for any actual implementation here, just discussion. Specifically, I wondered how the results might mesh with the idea of bounded bivariate distributions in a continuous source/recognition model (Starns, Rotello, Hautus, 2014). Of course a correlation between source and item memory has always been a part of continuous models, and the bounded model just implements the reasonable assumption that low recognition strength should be associated with chance-level source memory, not “reversed” source memory whereby the participant gets strong evidence for Source B when the test item is from Source A, and vice versa. Could this version of a continuous model better account for the heavy tailed distributions of source errors than the standard continuous model, with the idea being that the “below threshold” guesses are items whose recognition strength falls below the point where the source evidence distributions have converged? Compared to the threshold-style models, perhaps this account better explains the fact that both precision and the proportion of retrieval failures are improving as one moves up on the recognition confidence scale?**

* By conditioning analysis of source responses on items that were recognized, we account for chance-level source memory for items with low recognition strength. For a model like this to fit our data, it would need to say that items that were above a threshold in the recognition task are then below another threshold where source evidence “converges”. Are these thresholds at different levels? If not, the recognition and source tasks are in very close proximity to each other in the experiment, so its unlikely that memory strength decays between tests.
* How does a continuous model that is bounded in this way differ from a some-or-none model, like the hybrid circular diffusion model in the present study?
* The continuous model does not do that well for only highly recognised data.
* I don’t know if it is true that source precision increases with recognition confidence in our dataset. This is based on the results of the Zhang and Luck mixture model that was fit to overlapping confidence bands. Hard to say with the sparsity of data for items recognised but with lower confidence (rated 4 or 5). At a group level, it basically looks fit until ratings of 6.

**R2C5: In the interest of caution, it is probably a good idea to acknowledge that the source distributional assumptions that work best within the circular diffusion model might not be the ones that work best if other decision models are assumed.**

The distributional assumptions of the circular diffusion model are emphasised in a new line on page 37, along with an acknowledgement that these assumptions may not be generalizable to other decision models.