We thank the editor and the reviewers for their comments, which have helped us clarify the theoretical and empirical contributions of our manuscript. 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 have revised the title of the manuscript to include a subtitle to clarify the main theoretical questions of the study. The new title is: "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. We have added a sentence in the manuscript to explain this decision:

A small-*N* design was chosen for this study because by collecting 720 trials over the course of the four sessions, experimental power was concentrated at the level of the individual participants, rather than at the level of the experimental sample (Smith & Little, 2018). (p. 19)

**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. 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’?**

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 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. Participants for whom model recovery failed were the same participants who exhibited uniform or near-uniform level responding, and when we conditioned the analysis on participants who exhibited strong memory-based performance, there were almost no misfitting cases. Regarding the flexibility of the hybrid and threshold models, it is worth noting that these models do make predictions about the RT data in that response accuracy and RT are linked in the model. When comparing the threshold and hybrid models to each other, very few participants required the additional flexibility the hybrid model provides. Overall, we believe that this demonstrates that the models are sufficiently distinguishable from each other, and that the results of the model comparison reported in the main text reflect meaningful differences between the models.

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

**R1C3: 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.**

We appreciate the theoretical elegance of the MDL approach to model complexity but the technical challenges of applying it to diffusion models are formidable. The MDL penalty term depends on the integral of the determinant of the Fisher information matrix over the parameter space, where the elements of the matrix are the second partial derivatives of the likelihood function with respect to its parameters. For diffusion models, in which the likelihood function depends on a multiple integral of an infinite series, the technical challenges in rendering these kinds of computations tractable are significant. Even for simple multinomial signal detection models (e.g., Klauer & Kellen, 2011), the expressions for the Fisher information matrix are fairly complex. We suspect that this is the main reason why MDL has been confined to simple algebraic models and has not been extended to sequential-sampling models. Despite the acknowledged limitations of penalized likelihood statistics like the AIC and BIC, we believe they provide a principled way of selecting between models. In the absence of a tractable scheme for MDL calculations for diffusion processes and mixtures of diffusion processes, we really cannot speculate on how MDL model selection might behave. However, regarding the results changing with a different fit statistic, it is worth noting that the qualitative difference between the models’ performance is large, particularly in terms of model fits to the distribution of response errors, which is reflected in the quantitative difference in the BIC. 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.

**R1C4: 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 (extract below) to acknowledge that changes in overall performance could have an influence on the relative performance of the models. However, given that the length of the distracter task places retrieval clearly in the scope of what many researchers would define as long-term memory, we know of no theoretical reason why we would expect longer lists or longer study-test lags to qualitatively affect the pattern of results.

Our experiment was designed to be similar to the short-delay condition in the Harlow and Donaldson (2013) paradigm, but not the long-delay condition. With longer study-test lags, changes in overall performance and levels of interference in memory may affect the relative performance of the models presented in this study. As such, it is unclear how our results generalize to other continuous-outcome source memory paradigms and making broader conclusions will require future work. (p. 46)

**R1C5: 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:

Dual-process models propose that memory in an item recognition task is supported by a mixture of familiarity and recollection processes, as both types of information retrieved can inform whether an item has been previously encountered or not. Dual-process models and SDT can make similar predictions about item recognition, particularly when the contribution of familiarity in the dual-process model is high. (p. 3)

**R1C6: 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.

**R1C7: 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…**

The results of the Zhang and Luck (2008) mixture model fits to the response accuracy data, as originally fit to overlapping categories of recognition confidence (Table 1), showed that high-confidence recognition is associated with a higher proportion of memory-based responses relative to all recognized items. High-confidence recognition was not associated with an increase in memory precision, but a decrease. When we revised the mixture model results to fit mutually exclusive confidence categories (Table 2), as suggested in Comment 3, this counter-intuitive finding is further pronounced, as items recognized with low confidence are associated a very high proportion of retrieval failures, comparable to items that were unrecognized, and high precision for the few cases that were memory-driven.

This relationship between precision and confidence is inconsistent with the underlying dual-process theory, and reflects a limitation in the data, specifically the way in which participants use the confidence ratings. Given that participants did recognize an item, they tended to use the maximum rating of 6 and rarely used ratings of 4-5. When participants did use ratings of 4-5, the associated source responses were mostly guesses. We added a new paragraph in the manuscript to explain the implications of this on the mixture model parameter estimates:

…As the proportion of memory-based responding approaches zero, the precision parameter becomes unidentifiable as there is little memory information to constrain it. Overall, that the precision parameter is higher for items recognized with low confidence than items recognized with high confidence is much more a product of the idiosyncratic way that participants use confidence ratings rather than anything more substantive about source memory retrieval. (p. 26)

With the sharp difference in responding with different confidence ratings at the level of source accuracy, it is difficult to draw strong conclusions about source accuracy across different levels of confidence. With that said, when we used the same confidence categories to fit the circular diffusion model in response to Comment 2, we found that the difference between confidence categories was primarily reflected in the mixing proportion instead of precision (drift variability). This could be due to the additional constraints imposed on the diffusion model by fitting RTs as well as source accuracy.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 1 | | | | |
| *Overlapping Confidence Categories* | | | | |
| Recognition Rating | Low Imageability | | High Imageability | |
|  | κ | *π* | κ | *π* |
| All (1-6) | 19.22 | 0.51 | 23.89 | 0.51 |
| Recognized (4-6) | 18.24 | 0.50 | 23.79 | 0.54 |
| Highly Recognized (6) | 18.03 | 0.50 | 19.81 | 0.56 |

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Table 2 | | | | |
| *Mutually Exclusive Confidence Categories* | | | | |
| Recognition Rating | Low Imageability | | High Imageability | |
|  | κ | *π* | κ | *π* |
| High (6) | 22.43 | 0.48 | 22.89 | 0.51 |
| Low (4-5) | 42.43 | 0.09 | 49.73 | 0.07 |
| Unrecognized (1-3) | 51.97 | 0.06 | 11.30 | 0.07 |

**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 guesses 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 memory-driven responding does not, supporting a purely thresholded view:

If the change in the source response distribution across recognition confidence is the product of a pure threshold process then we would expect that allowing the mixing proportion to vary as a function of confidence would improve the fit of the threshold model. In comparison, allowing drift rates to vary across recognition confidence should not improve the fit of the model, as the quality of evidence retrieved should not differ across recognition confidence. Table 9 compares the joint fit of the threshold model to high-confidence and low-confidence data, and while allowing for different proportions of guessesacross confidence improves the fit of the threshold model compared to the original parameterization, the addition of different values of does not similarly improve the fit of the model enough to outweigh the BIC penalty. (p. 38)

**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.**

In addition to revised fits of the mixture model detailed in response to Comment 1, Figure 4 and Table 4 have been updated to report mutually exclusive confidence ranges. The same confidence ranges were used to jointly fit the threshold circular diffusion model to high and low 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?**

For the reasons outlined in response to Comment 1, it is not clear from the current data whether both the proportion of memory-based responses and memory precision improve as recognition confidence increases. We recognize that a unified model of recognition and source memory would provide a clearer account of how source responding changes across recognition confidence, and acknowledge that an approach like the bounded bivariate Gaussian model in the continuous-outcome paradigm is a potential research direction for the circular diffusion model:

…An example of a model that provides a joint account of recognition and source confidence in two-choice tasks is the bounded bivariate Gaussian model presented by Starns, Rotello, and Hautus (2014), which is a continuous model that represents evidence in memory as a bivariate Gaussian distribution with recognition and source evidence as its two dimensions. When compared with a bivariate dual-process model that incorporated a threshold recollection process like the threshold variant of the circular diffusion model in the present study, the authors found that the continuous bounded bivariate model successfully predicted a range of qualitative patterns in the joint recognition and source data, while the dual-process model did not. As the data and modeling presented in the current study precludes drawing strong conclusions about the relationship between recognition and source memory, a natural direction for future research in the continuous-outcome domain is to extend the circular diffusion model to joint recognition and source data. (p. 42)

**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 emphasized in the following extract, along with an acknowledgement that these assumptions may not be generalizable to other decision model:

It is important to note that the source distributional assumptions that are suitable for the circular diffusion model may not be generalizable to other decision models. The circular diffusion model, like Ratcliff's diffusion model for two-choice decisions, assumes the evidence entering the decision process is normally distributed across trials. Unlike his model, the evidence has a bivariate rather than a univariate normal distribution, one component of which represents variability in evidence strength and the other of which represents variability in the retrieved stimulus identity. We found that most of the across-trial variability was in evidence strength and that there was little variability in retrieved identity. These findings replicate those of Smith et al. (2020) who applied the circular diffusion model to decisions about the hues of noisy color patches. As our models were all versions of the circular diffusion model, we are not able to say whether this finding depends on the properties of this particular decision model or whether it is a feature of source-memory representations more generally. (p. 46)