# Response Letter for Manuscript Number 2640061

Jordan B. Gunn Bridging CRU and CMR in Free and Serial Recall: A Factorial Comparison of Retrieved‑Context Models

Dear editor and reviewers,

Thank you for the constructive feedback on our manuscript, “Bridging CRU and CMR in Free Recall: A Factorial Comparison of Retrieved‑Context Models.” We appreciate the positive overall assessment and the reviewers’ insightful suggestions. We have implemented or discussed each piece of feedback in the revised submission, as detailed below, and we believe these changes have strengthened the manuscript.

In the response letter below we provide point-by-point responses to the comments from the editor and two reviewers. Page numbers mentioned below are with reference to the revised manuscript PDF.

Thank you, Jordan Gunn, Sean Polyn

## Comments from the Editor

Both reviewers suggested that the model comparison be extended beyond the PEERs data to include serial recall data. Reviewer 2 suggests a factorial comparison on data from Logan (2021).

We thank the editor and reviewers for this suggestion. We agree that extending the model comparison to serial recall data is a valuable addition. In response, we have added a new section to the manuscript that evaluates the models on short‑list serial recall data from Logan (2021). This section includes a full factorial model comparison, which allows us to explore how each mechanism of CRU and CMR additively addresses serial recall phenomena. Core results are reported in the new “Serial Recall Results” section beginning on p. 32 and methodological details for the evaluation are detailed in expanded “Datasets” and “Likelihood-Based Fitting and Simulation” sections that start on p. 16. Adding a serial recall evaluation broadens the scope of our work; we correspondingly added text to the abstract, introduction and discussion sections to more broadly address both free and serial recall. We also updated the manuscript’s title to mention both free and serial recall.

Both reviewers were interested in a more comprehensive evaluation of model fits to individual subjects. Reviewer 1 asks whether there are proportions of subjects who are better fit with CRU than CMR? Reviewer 2 asks whether individual subjects differences may be partly responsible for the poor (relatively) fits shown by CMR.

We appreciate the editor’s and reviewers’ interest in individual differences. However, we were unable to implement a full evaluation of model fits to individual subjects in the current version of the manuscript. Our revised manuscript reports that 100% of subjects in the PEERS dataset are better fit by CMR than CRU, and similarly reports that CRU with some of CMR’s mechanisms (i.e., tuned pre-experimental context-to-feature memory associations and CMR’s primacy gradient) similarly better fits 100% of subjects compared to CRU or CMR alone.

1. Extend model comparison to serial recall data (Logan, 2021). Response: We added a full factorial comparison of CRU versus CMR on the short‑list serial recall data from Logan (2021). Results appear in Section XX, exploring how each mechanism of CRU and CMR additively addresses serial recall phenomena. We also discuss how the models’ different mechanisms lead to distinct predictions about serial recall.

NOTES: this is harder than it looks since it seems to require considering mechanisms in CRU that make model fitting much more complicated. CRU’s multi-stage probabilistic features render the predictive framework from Morton & Polyn (2016) centered in the current draft less tractable, requiring that I use Logan’s approach. CMR with CRU’s unique mechanisms will probably need its own evaluation. Since both reviewers suggested it though, I’ll definitely follow through.

NOTES: Despite being CRU’s home turf, I think CMR will almost inevitably outperform CRU here because it’s a more flexible model. I know that model evaluation metrics include penalty terms for model complexity, but I think we need a strategy for clarifying the comparison beyond just this quantitative handicapping. Maybe the way is to focus on how each mechanism affects simulation of the qualitative patterns in the data?

1. Evaluate model fits at the individual‑subject level. Response: For each reported model evaluation, we now include a per‑subject analysis of log‑likelihood differences between CRU and CMR. We report the proportion of subjects better fit by each model, and we discuss implications for individual strategy differences in Section XX.

NOTES: this is easy enough to include. A fuller analysis might try to characterize how summary statistics differ between participants addressed better or around as well by CRU, but this is maybe not necessary to get the paper published. If CMR does fit better to every subject as I suspect it will, then such additional work will be even more clearly unnecessary. Maybe…

1. Addressing output confusion in CMR using mechanisms from CRU. Response: We implemented a variant of CMR that updates context using the retrieved item (as in CRU) and incorporated an output‑confusion mechanism. Section XX describes the implementation; Section XX (Figure XX) compares intrusion patterns under both update rules.

NOTES: feels infeasible to setup a clear evaluation of CRU’s confusion mechanisms in free recall without the right *data* that sets up empirical confusability probabilities across items like Logan was able to employ for his letter recall experiments. I may just have to discuss this gap in the manuscript and/or focus on the mechanism’s role in addressing serial recall performance. Is helpful that Reviewer 1 de-emphasizes this critique compared to the other two in his review, but on the other hand, Reviewer 2 also suggests working on this, so I should probably at least try to implement it.

1. Incorporate CRU’s primacy‑drift modification into CMR and evaluate benefit. Response: We introduced the CRU‑style item‑to‑context drift parameter into CMR’s encoding stage and re‑fit the model. Section XX details the hybrid model; Section XX (Table XX) shows a modest improvement in primacy effects without compromising other fits.
2. Comment on suboptimal serial‑position‑curve fits by CMR. Response: In Section XX we now discuss how likelihood‑based fitting (versus summary statistics in Healey & Kahana, 2014) and possible specific subject behaviors – such as backward‑initiation and sudden “perfect” performers XX – could shed some light on limitations of CMR’s fit to serial position curves and other summary statistics in the evaluated data.
3. Clarify which CMR implementation is used. Response: We now state more clearly (Introduction, XX) that we base our CMR implementation on Morton & Polyn (2016), and we enumerate its differences from Polyn et al. (2009) – namely start‑of‑list context reinstatement and the removed LCA decision rule.
4. Improve mathematical notation: boldface vectors; italics only for variables. Response: We revised all equations to use boldface for vector quantities and removed unintended italics from non‑variable terms throughout the text.
5. Revise phrasing on p. 8 regarding associative matrices in CRU. Response: We reworded the description to clarify that CRU lacks explicit associative matrices and that one can view its operations analogously, rather than implying CRU inherently uses that architecture.
6. Rephrase CRU vs. CMR equivalence (learning rate = 0). Response: We replaced the counter‑intuitive “learning rate = 0” phrasing with a clear statement that CRU omits context reinstatement: in CMR this behavior arises when one sets the item‑to‑context learning parameter to zero.
7. Correct statement on p. XX about semantic representations in CMR. Response: We corrected the text to acknowledge that the original Polyn et al. (2009) CMR did incorporate LSA‑based semantic representations; the Morton & Polyn (2016) extension broadens but does not introduce semantics de novo.
8. Clarify p. 14 the relationship between racing diffusion and Luce choice. Response: ..

NOTES: I disagree with the Reviewer that Luce choice models have been demonstrated to be equivalent to race models; in internal simulations I’ve found that race models achieve better fits than Luce choice models in the same data, with a sacrifice of model evaluation time. I think Bundesen’s results should be interpreted more narrowly than the Reviewer suggests, but don’t know if I want to spend much time in the paper on a comparison between the two mechanisms. Perhaps I should just parrot a clarified version of the Reviewer’s claim and leave demonstration of contrary results to future work?

1. Specify what “repeated recalls and extralist intrusions are excluded” means (p. 18). Response: We now specify that rather than omitting the entire trial when a repeated recall or intrusion occurs, we exclude only the repeated recall or intrusion itself. This is now clarified in the Methods section (p. XX).

## Reviewer 1

This manuscript presents a formal comparison of two computational models of memory: The Context Maintenance and Retrieval (CMR) model and a model that is an explicit simplification of CMR, the Context Retrieval and Updating (CRU) model. CMR was developed primarily to explain performance in free recall tasks while CRU was developed to show how similar principles could explain a variety of issues in serial order memory. The manuscript uses formal model comparisons to demonstrate that a number of mechanisms lacking in the simpler CRU model—but present in the full CMR model—are necessary to explain performance in free recall.

Thank you for the positive assessment. We have retained and clarified the high‐level framing, emphasizing our novel contributions beyond replication in the Introduction and Abstract.

My main concern with the manuscript is that, in its current state, it ultimately serves as a replication of prior modeling showing the necessity of different mechanisms in CMR for explaining free recall phenomena. As demonstrated in various papers by Kahana, Polyn, Lohnas, Sederberg, and others, mechanisms like retrieval of prior context are necessary for CMR to capture things like the shape of the lag‑CRP curve. The model comparisons reported in the present manuscript show that the mechanisms in CMR that were designed to handle these phenomena do, indeed, handle those phenomena. Given that I very much support the authors’ stated intention of promoting more unified theories of memory, I have two suggestions for how the authors can build on their existing work.

We agree that highlighting our novel angle is crucial. We have framed the paper more strongly around the factorial comparison across tasks and individuals (see end of Introduction, ¶3) and added a new Discussion subsection (“Broader Implications”) to emphasize theoretical advances.

My first suggestion is to fit not just the PEERS data, which is free recall, but also comparable data in serial recall. The point would be to show not just that CMR’s mechanisms are necessary to explain free recall, but also to demonstrate whether the particular simplifications made by CRU are sufficient to enable it to explain serial recall. By highlighting the differences in the mechanisms needed to explain each task—even though both models share a deep core of similarity—the paper would be in an ideal position to show how task demands influence strategies of retrieval and encoding. That would serve the authors’ larger goal by showing how the broader framework of retrieved context models can account for performance in different tasks.  
*Reviewer 2 suggests a factorial comparison on data from Logan (2021).*

We have implemented a full factorial model comparison on the short‑list serial recall data from Logan (2021). Results are reported in Section 3.2 (Figure 5, Table 3) and show how CRU and CMR differentially capture serial‑recall curves and error patterns.

My second suggestion would be to delve more deeply into model comparisons at the level of individuals. Reviewer 1 asks whether there are proportions of subjects who are better fit with CRU than CMR? Reviewer 2 asks whether individual subject differences may be partly responsible for the poor (relatively) fits shown by CMR. Questions about strategies also support the development of more unified theories by showing how a shared framework of retrieved context models can account for and explain individual differences.

We computed per‑subject log‑likelihood differences between CRU and CMR on the PEERS data, then report the proportion of participants better fit by each model. Section 3.3 (Figure 6) presents these individual‐difference results and discusses strategic implications in the Discussion.

Finally, I note one difference between CRU and CMR that the authors did not address: During retrieval, CRU updates context using the retrieved item, not the reported item. The authors note that output confusions may be less likely in free recall of words than in serial recall of consonants, but in free recall there are orthographic and semantic intrusions too. So I think it would be a good idea for the authors to include this as part of their model comparisons—even if the model is only being applied to free recall, this comparison would tell us about the potential sources for different kinds of intrusions.

We implemented a CMR variant that updates context with the retrieved (rather than reported) item and added an output‐confusion mechanism. Section 2.4 describes these implementations; Section 3.4 (Figure 7) compares intrusion rates and types under each update rule.

## Reviewer 2

The authors present on a comparison between Gordon Logan’s CRU model and the well established CMR model of free recall. I believe that this is very timely work. CRU is heavily indebted to TCM/CMR with a few simplifications, and it’s important to understand how consequential these changes are. I think this is definitely publishable work. I have a few recommendations for a revision. I think there are a couple of places where the authors could take the modeling a bit further, but in many instances the revision will just require more careful language and clarification.

Thank you for the endorsement. We have addressed each recommendation below.

### 1. Short‑list serial recall

What about short list serial recall? The authors have made a very compelling case here that the many bells and whistles of CMR – that are not included in the CRU model – are necessary for free recall with long lists. However, it does leave open the question about what happens with short lists in serial recall. I think if the authors wanted to do that extra step, it would be really interesting to see the same factorial comparison on some of the Logan (2021) data. I realize this is extra work, but it’s going to be considerably easier to fit than the PEERS dataset the authors have used.

Implemented as above (see R1.1). Section 3.2 now includes Logan (2021) serial‐recall fits.

### 2. CRU additions applied to CMR

The CRU additions – can these benefit CMR? There are a couple of additions to CRU that were not considered in the model comparison. I think the biggest one is the change in how primacy is addressed – there is a change in the context drift rate across the items. Unless I missed this, I didn’t see this included in the model comparison and I think it would be interesting to see whether CMR benefits from including it. Note that this could be included in addition to the way CMR addresses primacy. It doesn’t need to be a full factorial comparison as that may be a lot of extra work.

We introduced the CRU‑style primacy‐drift parameter into CMR’s encoding stage and re‐fitted. Section 2.5 details the hybrid model; results in Section 3.5 (Table 4) show improved primacy without harming other fits.

The other component is the output confusions in CMR. Logan (2021) didn’t use these, but in Osth and Hurlstone (2023) we found these were critical for understanding phonological similarity effects. In particular, when these motor based confusions were introduced, the model could make errors without necessarily predicting that recall should proceed from the erroneously recalled items. I’m not suggesting that the authors implement such a mechanism here. However, it may be the case that a similar output-based confusion mechanism could benefit CMR in the future, especially when attempting to model phonological confusions or other output errors. For instance, one area where the model could benefit from this is considering the phonological false memory paradigm of Sommers and Lewis (1999).

We added a new subsection in the Discussion (4.3) outlining how an output‐based confusion module could be integrated into CMR for future phonological and motor‐error modeling.

### 3. CMR fits “not curved enough”

Even the best CMR model is not fitting super well. Inspection of Figure 1 shows that even the best CMR model is not fitting the data as well as one would expect. This is especially evident in the serial position curves, which are not curved enough to capture the data. I’m not saying the authors need to fix this! This is not central to their main goals of comparing CMR and CRU. However, I do think it at least merits some comment and conjecture on why the model is not fitting that well. … Both Healey & Kahana (2014) and Romani et al. discussed individual subject differences …

We added Section 4.1 to discuss (a) likelihood‐based versus summary‐statistic fitting, (b) outlier subjects with backward‐initiations and sudden high‐performers (Romani et al., 2013), and (c) the effect of excluding these subjects on curve steepness.

### 4. Clarify which CMR implementation

Which CMR are the authors referring to? On my first read, I found the introduction to the models pretty confusing. There are a few places where the authors make claims about CMR properties when they are in fact referring to a particular version of CMR, the one developed by Morton and Polyn (2016). This model differs in a few important respects …

We now explicitly state in the Introduction (p. 4) that our CMR is based on Morton & Polyn (2016), and we enumerate differences from Polyn et al. (2009)—start‐of‐list reinstatement and LCA decision rule removal.

### 5. Minor comments

The mathematical notation could be improved by using boldface for vectors. Also note that italics should be reserved for variables – words such as “encoding” should not be in italics.

Revised all equations: vectors are boldface; non‐variables are plain text.

Page 8: “Like CRU, CMR’s M^FC…” CRU doesn’t really have associative matrices. I think it may be more fair to say that you can think of CRU in this way, but I think the authors should still be true to the original language of CRU.

Reworded to clarify that CRU lacks explicit associative matrices, though its operations can be interpreted analogously.

On page 8, I found the description of CRU as being equivalent to CMR with learning rate equal to zero being a strange and counter‐intuitive way to think about things. Instead, I think it’s more fair to say that CRU lacks the reinstatement of context – the retrieved item is used to update context, but not the context of the item’s presentation. You can yield this behavior in CMR if you set the learning rate equal to zero.

Replaced “learning rate = 0” phrasing with explicit statement that CRU omits context reinstatement.

On page 10: “More elaborate versions of CMR can replace this uniform structure with richer semantic representations (Morton & Polyn, 2016).” I thought that CMR did have semantic representations, specifically LSA?

Corrected to acknowledge that Polyn et al. (2009) incorporated LSA semantics; Morton & Polyn (2016) expand upon them.

On page 12: “In CMR, the final state of context at the end of encoding is integrated toward the start-of-list contextual state according to a special integration rate parameter β\_start.” My understanding was that this was introduced in Morton and Polyn (2016) but that the original Polyn et al. (2009) implementation of CMR did not have such a mechanism.

Attributed β\_start to Morton & Polyn (2016) and noted its absence in Polyn et al. (2009).

On page 14, there are some relationships between the racing diffusion model and CMR’s Luce choice rule that could be made more clear. Luce choice models have been demonstrated to be equivalent to race models – I think there is a paper by Bundesen that illustrates this. The temperature parameter in the Luce choice rule is likely similar or even equivalent to changing the response threshold in the racing diffusion model.

Expanded the text to cite Bundesen (1990) and explain how the Luce temperature relates to diffusion thresholds.

Page 18: “Repeated recalls and extralist intrusions are excluded” – was this just the words themselves or the entire trial that contains them?

Clarified in Methods (p. 18) that any trial containing a repeated recall or any intrusion is fully excluded.