Editor Comments (Yoonhee Jang)

Dear Dr. Maxwell:  
  
I have received two expert reviews of Manuscript MC-ORIG-24-064 entitled "Investigating the Effects of Item-Specific and Relational Encoding on Judgment of Learning Reactivity for Categorized, Uncategorized, and DRM Lists" that you submitted to Memory & Cognition. The reviewers suggest that in examining JOL reactivity, your study has the potential to contribute to our understanding of the topic in the literature. At the same time, they have identified several issues that they believe you may be able to address in a careful revision. I agree with the reviewers that substantial work is needed to realize its full potential. Thus, I am not accepting the manuscript but can offer you the option of revising it along the lines detailed below and resubmitting it for consideration.

**Comment 1:** It seems to us all that your results involved not only positive reactivity but also negative reactivity, but you focused heavily on the former only. Both of the observed findings should be thoroughly discussed not only within the item-order account (i.e., it can predict and explain both or only one of the findings [if so, which one]?) but also within other accounts. Specifically, full discussion about all the accounts regarding JOL reactivity should be clearly included in both the Intro and General Discussion, particularly as the item-order account is relatively new in the JOL literature. I had an extremely hard time following the rationale of the study and your claims given the current findings.

***Response:*** Please see my response to Reviewer 1 (comment X) regarding negative the reactivity that was observed in Experiment 1A with uncategorized word lists. The item-order account makes no specific claims regarding the presence or absence of negative JOL reactivity. Instead, this account posits that item-level JOLs should only improve memory when memory is assessed via recognition testing but not free-recall. Because this account makes predictions regarding positive reactivity only, the lack of positive reactivity on uncategorized word lists in Experiment 1A is still consistent with an item-order account, even though negative reactivity emerged for this comparison. I have clarified this point in the Introduction (pg. XX). Additionally, this negative reactivity pattern is consistent with other studies which have found that item-level JOLs can sometimes (but not always) produce negative reactivity when memory is assessed via recall and study items lack strong relatedness cues (e.g., unrelated cue-target word pairs). I now note this point on pg. XX of the Introduction. Finally, in my revision, the negative reactivity pattern observed in Experiment 1A is now discussed in the Experiment 1 Discussion (pg. XX) and General Discussion (pg. XX).

**Comment 2:** As pointed out by Reviewer 1, when reading, I was so confused about why/how familiarity was related to the item-order account and you cited the well-known paper of dual-process memory theories (Yonelinas, 2002). I am unsure how to make connections between item-level/relational processing and recollection/familiarity. You may want to refer to a recent study (Zheng et al., 2024) which investigated the JOL reactivity on recollection/familiarity (although you may already know it).

***Response:*** Thank you for suggesting this additional literature. This paper appears to have been published shortly after I submitted my initial manuscript for review. [UPDATED INTRO] [REMOVED THE REFERENCE TO YONELINAS WHICH WAS ON PAGE XX OF INITIAL]

**Comment 3:** The different results observed in free recall and recognition in Experiments 1A/B should be interpreted very carefully. As noted by Reviewer 1, there were substantial differences in the procedure, such as how to present items and retention interval (multiple study-test blocks vs. a single study-test session). It sounds far-fetched that the comparison of the findings across different experiments is considered for evidence of dissociation.    
  
Relatedly, I am unsure about the rationale of the item presentation method used (list-by-list) in 1A/B, which is different from typical in both fields of memory and metamemory (except for false memory studies). To avoid confusion (to be seen below), I would insert figures to clearly show how to present items. In addition, both the procedure and findings should be thoroughly presented/discussed along those of previous studies, as noted by Reviewer 1.

***Response:*** [WORDS HERE]

**Comment 4:** Self-paced learning is most problematic as pointed out by Reviewer 2. I am not convinced by the current findings from such an unusual procedure without a clear rationale (note I am aware that prior work used it, e.g., Mitchum et al., 2016, but this is a completely different case), nor am I sure about the replicability. There are a variety of confounding variables, which should have been controlled.

***Response:*** [CITE ALL THE THINGS, REFER BACK TO R2 RESPONSE]

**Comment 5:** The Method sections are quite confusing (please also refer to Reviewer 2’s comments). For example, there were 4 categorized and 4 uncategorized lists (8 in total) in Experiments 1A/B. Given “all four-study lists” (p.11) in 1A, I cautiously guess you actually created two sets of the materials: i.e., 2 categorized and 2 uncategorized lists per set: let’s say X and Y, then each participant received either X or Y, not both, as study lists. I believe you also applied the same thing in the remaining experiments. If this is correct (if incorrect, please just clarify to me), then the study lists for (about) half the participants were different from those for (about) the other half. This sounds like you actually ran two sub-experiments (e.g., 1A-1 and 1A-2; 1B-1 and 1B-2) while reporting only the combined data. Again, if that is the case, the list set effects should be examined and reported.  
  
For the materials of 1A/B, I found descriptive data in Table A1. Were the three variables (concreteness, length, and frequency) controlled? I cautiously guess yes for frequency but maybe not (just unsure) for the remaining two (e.g., categorized words would be more concrete than uncategorized words?). Also, if you used two sets of the materials (as noted just above), please provide the outcomes of the same analysis for each set.

***Response:*** For each experiment, two sets of study lists were generated. This was done so that control items in Experiments 1B and 2 could be counterbalanced across participants so as to provide further confidence that any potential memory changes between groups resulted from the JOL manipulations rather than the materials being studied. The Experiment 1B methods section has been updated to more accurately reflect this (pg. xx; see my response to Reviewer X, comment X).

[ANALYZE RESULTS WITH VERSION AS A FACTOR AND ADD A FOOTNOTE]

**Comment 6:** From my understanding of the procedure/results in 1B, out of 96 words in total (= 8 x 12, as seen in 1A), there were 48 targets and 48 distractors (p.13). Given “from the counterbalanced lists” (p.13, though unclear, as noted by Reviewer 2), 24 targets and 24 distractors were used for each of the categorized and uncategorized word lists (i.e., four-study lists as in 1A, and each contained 24 words?). If this is the case, there should be two distractor types, categorized and uncategorized word distractors. However, I found only one type of new (distractors) in Figure 1. Relatedly, how were d’ and c calculated (as I am unsure how false alarm rate was calculated for each list type)? Again, if use of the two sets is the case, d’ is the index of recognition accuracy (not just hit rate) as both hits and false alarms depend on list type.

***Response:*** [FALSE ALARMS WERE COLLAPSED ACROSS CATEGORY TYPE. DISTRACTORS WERE NOT CATEGORICALLY RELATED TO PRESENTED ITEMS] [ANY OTHER STUDIES THAT HAVE DONE IT THIS WAY?]

**Comment 7:** BTW, in Experiment 2, either hit rate or d’ can be used as accuracy index, as there was only one type of new (distractors), except for the critical lures. Although you classified two distractor types (list item controls and critical item controls although the terms sound odd [i.e., what are controls in this situation?]), this is true to you, experimenter, but both were new from the participant’s point of view (from each of the two unseen, categorized word lists, 3+1 distractors were semantically associated, though). I think it is ok to report separately false alarm rate (as seen in Table A3) as some readers may expect the (numerically or significantly) higher false alarm rate for the critical item controls than for the list item controls (which could be referred to as false memory involved during the retrieval). However, to calculate d’, I think the same false alarms should be used for the targets and critical lures. Is this what you did?

***Response:*** [WORDS HERE]

**Comment 8:** Why did you use word generation for the filler task (which is also unusual)? It is tricky as retrieval involved in such a task can change context as well known, which affects (either improves or impairs) memory performance and potentially in different ways depending on the groups, lists, procedure, etc..

***Response:*** The filler task was modeled after Huff, Maxwell, and Mitchell (2022; https://link.springer.com/article/10.1186/s41235-022-00448-9) who similarly utilized a list word generation filler task for participants who studied a set of DRM lists. Although Huff et al., were not interested in the effects of JOLs on free-recall and hit rates, they still assessed whether encoding-based manipulations would modify both correct and false memory for study lists. Thus, there is a precedent for using generation filler tasks within the context of DRM lists.

**Comment 9:** In 1A, p.9, “Based on Senkova and Otani’s findings, JOLs were expected to benefit free-recall of categorized lists. However, the item-order account makes diverging predictions for item-level and global JOLs. First, this account predicts that item-level JOLs would not be reactive, as this JOL type should inhibit relational processes that facilitate free-recall. However, global JOLs would be expected to improve recall, as this task encourages participants to reflect on list-wise relations which are beneficial to recall. Thus, the inclusion of item-level and global JOL encoding groups allowed for a comparison between item-specific and relational oriented JOLs. Finally, because categorized lists contain pre-exiting relations, any benefits of global JOLs on this list type were expected to be greater than uncategorized lists.”

In 1B, p.12, “because this [item-order] account makes no claims regarding the effects of list relatedness on reactivity, item-level JOLs were also expected to benefit recognition memory for both categorized and uncategorized lists.”

I am very confused. Please clarify. Note this point was made BEFORE I read the Reviewer 1’s comment.

***Response:*** [WORDS HERE]

**Comment 10:** Several comments regarding the statistical analyses presented in the results sections

***Response:*** Please see my response to Reviewer X, Comment X.

**Comment 11:** The reason that some (so-called) Bayesian statisticians are in favor of using Bayesian factors or pBICs is to help avoid making an erroneous, binary decision based solely on the p value. That is similar to why researchers prefer to refer to effect size (although they are referred to differently). Please provide effect size (and pBICs if you’d like), considering its popularity, regardless of significant or non-significant results.       
  
Several pBICs do indicate week evidence, such as .74, .71, and .62 (.50 < pBIC < .75, as suggested by Raftery, 1995) and even support H1 (e.g., .46) for the non-significant results (e.g., p = .12, .12, .07, and .09, respectively). What is the effect size, once again?

***Response:*** [WORDS HERE]

**Comment 12:** In 1B, which method, either 1(1/2N) or log-linear rule, did you use, following Hautus (1995)? It will be great to let readers know so that they can save time, not looking for or guessing what you did.

***Response:*** The *psycho* package (Makowski, 2018) uses the log-linear rule. This detail has been added on pg. x.

**Comment 13:** In Experiment 2, two subheadings of the Results are not accurate as all of the indices, hits, false alarms, d’, and c are the terms under Signal Detection Theory (SDT).  
  
For recognition studies, I would provide descriptive stats of d’ and c in tables.  
  
T. O. Nelson and D. L. Nelson should be distinguished.

***Response***: Thank you for your attention to detail. These minor concerns have been addressed accordingly.

Table A1. Please provide inferential stats for each variable (please see above, too, if needed).  
In addition, please provide a summary for the materials of Experiment 2 (e.g., BAS) so that readers can understand how they served as counterbalances.  
  
Table A2. How about false alarms? Only reporting hits does not make sense (please see above, too, if needed). Also, please make all the typos correct here (as well as in the text, as pointed out by Reviewer 2).  
  
Table A3. Mean recognition is inaccurate – p(yes or old). In note, do parentheses really indicate 95% CIs?   
  
***Response:*** The Appendix tables have been updated accordingly. In addition, I now report inferential statistics for Table A1 in the Appendix. Appendix Table AX (Table A3 in the initial submission) reports mean hit rates, xxx, and xxx. And yes, you the values in parentheses reflect 95% CIs for each comparison (computed as (sd / sqrt(n)) \*1.96).

**Reviewer: 1**  
  
The present paper reports three experiments. The first two experiments examined the reactivity of item-level and global-level JOLs on categorized and uncategorized lists in recall (Experiment 1A) and in recognition (Experiment 1B).  Additionally, the author also examined the reactivity of these two types of JOLs on DRM lists in recognition (Experiment 2).  Exp 1A showed that only global-level JOLs but not item-level JOLs enhanced recall for categorized lists and neither type of JOLs enhanced recall for uncategorized lists. Exp 1B showed that for both categorized and uncategorized lists, only item-level JOLs but not global-level enhanced recognition. Exp 2 showed that for DRM lists, item-level JOLs enhanced true recognition while global-level JOLs increased false recognition.  The author thus concluded that these results showed that item-level JOLs primarily enhance item-specific processing, which support the item-order account for item-level JOL reactivity as item-level JOLs produced reactivity only in recognition but not recall. I think the studies are carefully designed and results are interesting and informative. However, I have some concerns and questions that prevent me from recommending the paper to be published in its current form.  
  
**Comment 1:** I understand that the study focuses on the item-order account of JOL reactivity, but I think the author should still also discuss the other alternative theoretical accounts as background.

***Response:*** In the initial submission, I focused primarily on the item-order account of JOL reactivity given that this set of experiments was designed to test this account. However, I agree that a broader discussion of other reactivity theories would be beneficial to the reader. As such, I have expanded the discussion of other accounts in the Introduction and, specifically, discuss both the cue-strengthening (Soderstrom et al., 2015) and changed-goals accounts (Mitchum et al., 2016) (see pgs. XX)

**Comment 2:** The author ties the current study to item-order account of JOL reactivity, which states that item-level JOLs enhance item-specific processing and impair relational processing. In terms of theory testing, I think the findings for item-level JOLs only replicate prior studies by Zhao et al. but did not add much new insights.  Instead, I think one novel contribution of the current study is the dissociative reactivity patterns between item-level and global-level JOLs, which “are likely determined by the stimuli participants study and the method of testing.” This notion seems to be in line with the cue-strengthening hypothesis, which emphasizes the cue consistency between study and test. Thus, related to my first point, I would suggest the author elaborate on the item vs. global JOL reactivity and discuss the implications of the current findings for other alternative theories.

***Response:*** This is an excellent point. On pg. xx of the General Discussion, I now discuss how the dissociative reactivity patterns between item-level and global-level JOLs relate to Soderstrom et al.’s cue strengthening account. Specifically, [EXPAND]

**Comment 3:** On p.6, the author stated that the item-order account predicts “positive reactivity on word lists, particularly when the test is sensitive to item-specific and familiarity-based cues.” They also raised a similar point on p.25 that “…because recognition is more sensitive to familiarity cue than relational cue”.  I am confused about why the item-order account would make predictions regarding to familiarity-based cues, as there is no direct mapping between item-specific processing and familiarity (i.e., both item-specific and relational cues can increase familiarity).  Could the author elaborate on that?

***Response:*** This is a valid point. In my initial submission, I was attempting to link findings from the present study showing that item-level JOLs generally produce positive reactivity on all items when memory is assessed via recognition but not recall with findings from other studies which have similarly found this pattern. As you noted, the item-order account makes no specific predictions regarding the role of familiarity cues. I have updated this section accordingly. To avoid confusion, I no longer include the item-order account in this paragraph (which is now on pg. XX). Instead, this paragraph now simply notes that these patterns are consistent with prior studies and may also reflect differences in cues emphasized by the different testing methods.

Additionally, a new paragraph has been inserted on pg. xx which discusses my findings in light of the item-order account before transitioning to the paragraph explaining that these reactivity patterns are also consistent with other studies.

**Comment 4:** On p. 9 and p.12, I see that the predictions for Exp 1A and 1B involve related versus unrelated lists, item versus global jol, and recall versus recognition. For readers’ convenience, I suggest adding a table to explicitly list the multiple predictions involved here.

***Response:*** Thank you for this suggestion. I’ve added a table (Table X; pg. xx) which provides the predictions for each experiment.

**Comment 5:** It caught my attention that the recall test administered in this study is very different from that of Senkova and Otani (2021). Here, the recall test is administered after each list is presented (following a brief filler task), while in the latter two studies, recall test is administered after multiple lists are presented.  In the former type of recall test, it is likely that some words are still lingering in the short-term memory and thus could be directly read out, and participants are likely to reply more on item-specific features for recall.  However, in the latter type of recall test, participants may rely more on interitem relations to reconstruct the items for recall. I think the author should acknowledge this difference when discussing the discrepancy between their results and those of Senkova and Otani (2021).

***Response:*** I have updated the discussion of Senkova and Otani’s (2021) findings on pg. xx accordingly. Specifically, [WORDS HERE]

**Comment 6**: Related to the former point, the difference between the memory test format in Experiment 1A and 1B is not just about recall versus recognition, but also about retention interval (i.e., longer retention interval between study and recognition test than between study and recall test). Is it possible that global JOLs are only reactive in immediate recall test partly because they help bypass the limited capacity of short-term memory?

***Response:*** You raise an interesting point here. [CEILING EFFECTS]

**Comment 7:** I wonder what is the presentation rate for each word and if there is any attention check or cheating check for the no-JOL condition?

**Response:** As noted on pg. XX of the initial submission, JOLs were self-paced for all participants. [ADD A TABLE WITH THE RTS] [FOLLOWING MY PREVIOUS WORK AND MITCHUM ET AL AND JANES ET AL] [DATA WAS SCREENED BASED ON JOLS]

**Comment 8:** For the results, the author reported the pBIC for non-significant effects, please elaborate on how the readers should interpret the pBIC (e.g., what does a pBIC of .9 mean).  Also, the p values are sometimes missing for ANOVAs and post-hoc t tests.

***Response:*** pBIC values provide a *p*-value describing strength of evidence in favor of the null hypothesis being retained. As such, high values should be interpreted as stronger evidence in favor of the null hypothesis. I have updated the paragraph describing pBIC values accordingly (pg. xx).

In my initial manuscript, *p*-values were only reported for non-significant comparisons to avoid cluttering the results section. For all significant comparisons effect-size indices were reported instead. However, for completeness, all *p*-values are now reported where initially omitted.

**Comment 9:** On p.15, the author stated that for uncategorized lists, the difference in free recall between the three JOL conditions was not significant except for between item-JOL and no-JOL conditions (Ms = .48 vs. .55). So, there is a negative reactivity for uncategorized lists? This seems counteractive to me, and I see no discussion about this result in the following text, so I would suggest the authors elaborate on it.

***Response:*** Yes, in Experiment 1A, making item-level JOLs impaired participants’ free-recall compared to the no-JOL condition. As noted on pg. XX of the revised introduction, making JOLs has been shown to sometimes produce negative reactivity when participants provide them for unrelated study items (see a recent paper by Undorf et al., 2024; (https://online.ucpress.edu/collabra/article/10/1/117108/2006810). Because the item-order account makes no predictions regarding negative reactivity and instead simply predicts a lack of positive reactivity for item-level JOLs when free-recall testing is used, this pattern is in-line with this account. As such, in my initial submission, I elected not to focus specifically on negative reactivity and instead focused on the broader patterns.

This finding is now noted in the Experiment 1 discussion (pg. xx) and again noted in the General Discussion (pg. xx). In doing so, I have also the section in the General Discussion now links this pattern back to previous studies which have found that item-level JOLs can sometimes produce negative reactivity on unrelated cue-target word pairs (e.g., Mitchum et al., 2016).

**Comment 10:** From the bottom of p.26 to the top of p.27, the author explained why item-JOLs did not reduce false memory as predicted by the item-order account.  They reasoned that “the nature of list encourages relational encoding, in addition to item-specific encoding being afforded by JOLs.”  This is not very convincing to me, as previous studies using the same stimuli (i.e., DRM lists) have shown that item-specific encoding operations effectively reduced false memory (the author also cited some studies of that sort at the bottom of p.19). So this result seems to be against the notion that item-level JOLs enhance item-specific processing or at least suggest that item-JOLs does not enhance item-specific processing as effectively as those classic item-specific encoding operations?

***Response:*** [WORDS HERE]  
  
On p.23, “In doing so, each experiment provided additional tests the item-order account by assessing…”, an “of” is missing after “tests.”

As the author used both item-level and global-level JOLs, I suggest using more careful language to make the distinction when discussing JOL reactivity.  For example, on p.9, “Based on Senkova and Otani’s finding, JOLs were expected to benefit free recall of categorized lists”, I think JOLs should only refer to item-specific JOLs, as t Senkova and Otani’s study did not make any prediction about global-level JOLs.

***Response:*** All minor points have been addressed. I appreciate your attention to detail.  
  
  
**Reviewer: 2**  
  
The author assessed the contributions of item-specific and relational processing to JOL reactivity. Participants studied categorized and uncategorized lists of words and either made item-level JOLs, global JOLs, or no JOLs, then took a final free recall or recognition test. Consistent with the item-order account, relative to not making JOLs, item-level JOLs improved recognition performance, whereas global JOLs improved free recall performance for categorized lists and recall of critical lures in DRM lists.  
  
General Assessment and Recommendations  
  
The author builds on prior JOL reactivity research (specifically, research by Senkova & Otani, Zhao and colleagues, Chang & Brainerd) to provide novel and creative tests of the item-order account. I also appreciated that the author included materials, data files, and R code on the Open Science Framework. I have only minor suggestions for improving the manuscript.  
  
**Comment 1**: In the abstract, I would recommend specifying the direction of reactivity observed (not just that “item-level JOLs were reactive on all list types…”)

***Response:*** Thank you for this suggestion. The abstract has been updated accordingly.  
  
**Comment 2:** Please provide the full parameter details of the power/sensitivity analyses (e.g., alpha, power, analysis) so that they can be replicated by another researcher.

***Response:*** The descriptions of the power/sensitivity analyses for all experiments have been updated to include the alpha and power levels.  
  
**Comment 3:** In all three experiments, encoding was self-paced. Were there any differences in encoding/JOL times between the various groups?

***Response:*** [MENTION NEW TABLES AND ANALYSES, REFER BACK TO R1s COMMENT]  
  
**Comment 4:** In Experiment 1B, the author notes that for the old/new recognition test, non-presented controls were developed “from the counterbalanced lists” (p. 13). However, counterbalanced lists have not been previously described. I would recommend including additional details about how these lists were designed (i.e., did the non-presented controls for the studied categorized lists come from the same category?)

***Response:*** The description of the study lists used in Experiment 1B have been updated accordingly. Please see the revised description on pg. XX.  
  
**Comment 5:** I would recommend that the author include (item and global) JOL magnitudes in a Table or at least an Appendix. These values could be used to support claims that global JOLs “encourage participants to reflect on list-wise relations which are beneficial to recall” (p. 9) – for example, if global JOLs were higher for categorized than uncategorized lists, this suggests participants use pre-existing relations in the list as a cue to inform their judgments.

***Response:*** This is an excellent suggestion. I have added a new table to the Appendix which displays mean JOLs as functions of JOL task and list type for Experiments 1A/1B and for JOL type in Experiment 2 (Table AX, pg. xx). [ELABORATE?]  
  
**Comment 6:** I caught multiple analyses reported throughout the manuscript that seem to be missing p values. Please check to ensure all relevant values are reported in each analysis.

***Response:*** These values have been added to the revision. Please see my response to Reviewer 1, comment X.  
  
**Comment 7:** I had a thought about Experiment 1A: I realize that the author cannot measure clustering in free recall because each list was encoded/recalled one at a time. However, I was wondering if there was a similar type of measure to better understand the recall strategy that participants use (and whether this differs by JOL group; cf. Zhao et al.’s order reconstruction task). For example, perhaps participants in the item-JOL group are less likely (than the other two groups) to recall items in the order in which they were studied? I’m not insisting the author conduct this labor-intensive analysis, just curious if there may be additional pieces of evidence for the item-order account within the data.

***Response:*** This is an interesting point. [EXPAND] [GD?]  
  
**Comment 8:** I caught a couple of typos throughout the manuscript. All were minor, and none impaired my understanding of the research. However, I recommend the author thoroughly proofread.

***Response:*** These typos have been corrected. Thank you for taking the time to review this manuscript.