July 15th, 2024

Dr. Yoonhee Jang, PhD

Associate Editor

*Memory & Cognition*

Dear Dr. Jang:

I have submitted my revision of MC-ORIG-024-064 “Investigating the Effects of Item-Specific and Relational Encoding on Judgment of Learning Reactivity for Categorized, Uncategorized, and DRM Lists” for your consideration. I appreciate the thorough examination provided by yourself and the reviewers. I am pleased that Reviewer 1 found the results “interesting and informative”, and that Reviewer 2 only suggested minor edits to the manuscript. Below, I list my responses to each reviewer’s comments in addition to your own and cite page numbers when referring to specific changes. To facilitate the review process, all primary modifications to the manuscript were made using blue-colored font. I look forward to your response and hope that the revised manuscript is now suitable for publication in *Memory & Cognition*.

Sincerely,

Nicholas P. Maxwell, PhD

Department of Psychology

Midwestern State University

nicholas.maxwell@msutexas.edu

Ph: 940.397.4340

**Action Editor: Dr. Yoonhee Jang**  
  
**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:*** The item-order account makes no specific claims regarding the presence or absence of negative JOL reactivity. Instead, this account merely posits that item-level JOLs should only improve memory when memory is assessed via recognition testing but not free-recall. This point is now noted on pg. 7. Additionally, pg. 5 of the Introduction now notes the potential for item-level JOLs to sometimes produce negative reactivity on unrelated study materials.

Regarding the negative reactivity observed in Experiment 1A: Item-level JOLs produced negative reactivity on memory relative to global JOLs but not the no-JOL group. There was a mistake in this section which incorrectly attributed this comparison to the item-level JOL/no-JOL comparison. This has been corrected in the revision (pg. 16; please see my response to Reviewer 1, Comment 8). While free-recall in the item-level JOL group as numerically lower than the no-JOL group (.48 vs. 55), this comparison failed to reach significance, *t*(74) = 1.67, *SEM* = .04, *p* = .10.

**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. Zheng et al.’s (2024) paper appears to have been published in *Metacognition and Learning* shortly after I submitted my initial manuscript for review. I have now added a brief discussion of Zheng et al.’s findings to the General Discussion and discuss how they relate to the recognition findings in Experiment 1B (pg. 28).

Yonelinas’s (2002) paper was initially cited as a literature review supporting the claim that recognition and recall tests are likely to emphasize different cues. I have removed this citation from the Introduction and have reworked the paragraph on pg. 28 of the General Discussion for clarity. Please also see my response to Reviewer 1, Comment 8.

**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:*** This is an interesting point. Experiment 1A utilized multiple study/test block-cycles as I was concerned that performance on a 48-item free-recall test would be near floor, especially since this study was conducted online rather than in a lab setting. Separately, the recognition experiments (Experiments 1B and 2) used a single block design given concerns that performance in a multi-block design would be near ceiling (please see my response to Reviewer 1, Comment 5).

In the revised manuscript, I have taken care to note the differences between in presentation sequence between Experiments 1A/1B. This is now noted as a potential limitation on pgs. 28-29 of the General Discussion. Additionally, I have revised the manuscript to carefully note that while the reactivity patterns observed in Experiments 1A and 1B suggest a dissociation between test format and JOL type, future work will be needed to fully explore this account. To that end, page 29 now notes potential areas for future research which were initially suggested by Reviewer 2 (please see my response to Reviewer 2, Comment 7).

**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:*** Please see my response to Reviewer 2, Comment 3 regarding the use of self-paced versus experimenter paced encoding.

**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 counterbalanced sets of study lists were generated. This was done to control for potential item effects on memory and, additionally, so that studied and control items in Experiments 1B and 2 could be counterbalanced across participants. As such, the primary purpose of generating two separate sets of lists was to provide further confidence that any potential memory changes between groups resulted from the JOL manipulations rather than from the materials that were studied. The Method sections of each experiment have been updated to more accurately reflect this (Experiment 1A, pgs. 11-12; Experiment 1B, pg. 15; Experiment 2, pg. 23; see my response to Reviewer 2, Comment 4).

Reanalyzing each experiments’ free-recall/recognition data with Counterbalance Version included as an additional between-subjects factor yielded no main-effects or interactions with counterbalance version in Experiment 1A (*F*s ≤ 1.97, *p*s ≥ .14) and no main effect in Experiment 1B (*F*(1, 107) < 1, *MSE* = .03, *p* = .59. The List Type (Categorized vs. Uncategorized) × Counterbalance Version (A vs. B) interaction was significant (*F*(1, 107) = 8.40, *MSE* = .01, *η*p2 = .07) but, importantly, this interaction was not qualified by a significant Encoding Group × List Type × List Version interaction (*F*(1, 107) < 1, *MSE* = .01, *p* = .07). Thus, in Experiment 1B, it is unlikely that the counterbalance version differentially effected recognition within each encoding group. Finally, no main effects or interactions with Counterbalance Version emerged in Experiment 2 (*F*s ≤ 1.86, *p*s ≥ .18). Taken together, it is unlikely that use of counterbalanced study lists differentially influenced free-recall or correct recognition.

Finally, Table A1 (now on pg. 44) has been updated to display mean/sd values for concreteness, length, and frequency split by list counterbalance and list type (categorized vs. uncategorized) rather than collapsed across counterbalances as presented in the initial manuscript. Additionally, three 2 (Counterbalance: A vs. B) × 2 (List Type: Categorized vs. Uncategorized) between-subjects ANOVAs yielded no main effects or interactions with counterbalance for frequency (*F*s < 1, *p*s ≥ .62), concreteness (*F*s ≤ 1.52, *p*s ≥ .20), or length (*F*s ≤ 2.71, *p*s ≥ .10). These findings are now reported in the Experiment 1A materials (pg. 12). When considered alongside the above analyses showing that memory did not differ as a function of counterbalance version, there is considerable evidence that the counterbalance version that participants were randomly assigned to did not influence their overall performance on the study task or memory test.

**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:*** Only one type of distractor item was reported as, by nature, items must be studied within a categorized/uncategorized context to belong to either group (i.e., it would not be known to particpants at test whether distractor items were categorized or uncategorized, even though these items were indeed taken from non-studied categorized/uncategorized lists. This is particularly likely as all 96 test items were presented in a randomized order, making it difficult for participants to draw connections between non-presented items). Because of this, false alarms rates were collapsed across all item types. Similarly, *d′* and *c* were computed across lists, rather than being computed separately for each list type. I have added a footnote on pg. 18 clarifying this point. Additionally, this approach where false alarms, *d′* and *c* are collapsed across relatedness categories is consistent with prior studies which have assessed JOL reactivity on recognition of related and unrelated study materials (e.g., Myers, Rhodes, & Hausman, 2020).

**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:*** Yes, you are correct. I initially separated false alarms for list item controls and critical lure controls given that this has often been done in other DRM studies (e.g., Huff, Maxwell, & Mitchell, 2022 and Huff, Bodner, & Gretz, 2020, for examples). The term ‘control’ was also taken from this literature. Consistent with your thoughts, *d′* and *c* were calculated using all false alarms, regardless of control type. This has now been clarified in the Table A6 note (pg. 49).

**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 word generation filler task was modeled directly after Huff, Maxwell, and Mitchell (2022) 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:*** While the item-order account makes differing predictions based on the type of processing which JOLs may encourage (i.e., relational processing vs. item-specific processing), this account makes no predictions regarding whether reactivity would be moderated by pre-existing relations between study materials (i.e., categorized vs. uncategorized lists). This is because in Zhao et al.’s (2023) study, participants only studied uncategorized word lists. As such, a secondary goal of Experiment 1B was to test whether the positive reactivity which Zhao et al. observed on uncategorized lists would extend to categorized lists.

To avoid potential confusion, the sentence which originally appeared on pg. 12 has been omitted. Instead, the Experiment 1B hypotheses section (pgs. 13-14) now notes that JOL reactivity patterns tend to be larger and in the positive direction when recognition testing is used rather than discussing specific differences in list relatedness.  
  
**Comment 10:** Several comments regarding the presentation of statistical analyses in the results section.

***Response:*** *p* values and effect sizes have been added to all significant and non-significant results, respectively.  
  
  
**Comment 11:** 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 to the Experiment 1B results section on pg. 18.

**Comment 12:** Minor comments regarding subheadings, references, and tables.

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

**Comment 13:** Several comments regarding the Appendix Tables  
  
***Response:*** Appendix Table A1 (pg. 44) has been updated to show the characteristics for each counterbalance, rather than collapsing across them as in the initial submission. In addition, I now report inferential statistics for Table A1 (these statistics are reported on pg. 12 of the Experiment 1A methods; please see my response to Comment 5). Next, Experiment 1B results have been moved to a new table (Table A3, pg. 46), which now displays Hits, False Alarms, *d′,* and *c.* Appendix Table A6 (pg. 49; Table A3 in the initial submission) now reports mean hit rates, *d′*, *c*. Additionally, based on Reviewer 2’s feedback, Table A4 has been created on pg. 47 and displays mean JOL ratings for both JOL groups as functions of list type/experiment. Finally, Table A5 (pg. 48) has been created to display lexical properties for DRM lists, including mean BAS between list items and critical lures. Like Table A1, values in Table A5 are also split by list counterbalance.

**Reviewer: 1**  
  
**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. 4-5). Additionally, I have also added more discussion of how findings reported in the present study align with the cue-strengthening account (pg. 28; please see my response to Comment 2).

**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. For example, the finding that global JOLs reactivity patterns in Experiment 1A were moderated by pre-existing stimuli relations (i.e., positive reactivity on categorized lists but no reactivity on uncategorized lists) mirrors patterns that are reported on cue-target word pairs. As such, I know discuss on pg. 28 how the diverging reactivity patterns reported in Experiments 1A/1B related back to the cue-strengthening account.

**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:*** 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, though recent findings from Zheng et al. (2024) suggest that item-level JOLs likely enhance familiarity (see my response to the Action Editor, Comment 2). To avoid confusion, I no longer include the item-order account in this paragraph. Instead, this paragraph (which is now on pg. 28) notes that these patterns are consistent with prior studies and may also reflect differences in cues emphasized by the different testing methods (i.e., cue-strengthening).

**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 have now added a table (Table 1; pg. 41) which provides the predictions for each experiment based on the item-order account.

**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).

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:*** This is an interesting point. Experiment 1A used a series of study/test blocks given concerns about floor effects due to participants completing a lengthy free-recall test (i.e., studying all 48 items). Separately, Experiment 1B used a single study/test block due to concerns about ceiling effects on recognition as each test would only contain 24 items if following the design used in Experiment 1A. (i.e., 12 studied and 12 non-studied). However, as you note, this approach does make comparing between Experiments 1A and 1B difficult. As such, I have made a point of tempering the language surrounding the dissociation between JOL tasks and test type observed between Experiments 1A and 1B. Additionally, I now discuss the difference in study-test cycles as a potential limitation in the General Discussion (pg. 28-29) while also noting that the general patterns in Experiment 1A are still consistent with findings from Zhao et al. (2023), who used a presentation sequence more similar to the one used Senkova and Otani (2021).

**Comment 6:** 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. 11 of the initial submission (now on pg. 13), JOLs were self-paced for all participants. In addition to JOLs, the filler tasks were used as an attention check (e.g., data from seven participants in the Experiment 1A no-JOL group were omitted due to not completing the filler tasks). Finally, data was screened for cheating (mean recall/recognition > 95% across all list types) or failure to complete the memory tests (mean recall/recognition < 5%). These procedures are described on pg. 11.

**Comment 7:** 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. 16).

In the initial manuscript, *p*-values were only reported for non-significant comparisons to help streamline the results sections. For all significant comparisons effect-size indices were reported instead, as these values are arguably more meaningful for significant findings than *p*-values. However, for completeness, all *p*-values are now reported for all significant findings and effect sizes are similarly reported for all non-significant comparisons.

**Comment 8:** 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:*** Thank you for raising this point. Unfortunately, there was a mistake in this section. While the comparison between the .48 and .55 is indeed significant, this comparison refers to the difference between global JOLs (.58) and item-level JOLs (.48). The difference between the no-JOL control group (.55) and item-level JOL group was non-significant, *t*(74) = 1.67, *SEM* = .04, *p* = .10.. However, the sentence which you referenced on pg. 15 of the initial submission incorrectly attributed this negative reactivity to the item-level/no-jol comparison. This has been corrected (now on pg. 17). I appreciate your attention to detail.

Although the comparison between the item-level and no-JOL groups was non-significant, 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). I now note this possibility in the Introduction (pg. 5) and additionally note that the item-order account makes no specific predictions regarding negative reactivity (i.e., this account only states that item-level JOLs should not improve free-recall, see pgs. 7 and 10).

**Comment 9:** 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:*** While findings in the present study suggest that item-level JOLs encourage item-specific encoding of stimuli, these JOLs likely also encourage some degree of relational encoding, albeit to a lesser extent relative to global JOLs. For example, relations between stimuli provide highly salient markers of later memory, and participants likely still use this to inform the magnitude of their JOLs, even for JOLs elicited individually on an item-by-item basis. However, the item-level JOL likely biases participants towards item-specific encoding to a greater extent relative to global JOLs. This point has been clarified on pgs. 29-30 of the General Discussion.

**Comment 10:** Minor spelling and grammatical errors

***Response:*** All minor points have been addressed. I appreciate your attention to detail.  
  
  
**Reviewer: 2**  
  
**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:*** This is a valid concern. Across experiments, mean encoding latencies were significantly greater for participants in the item-level JOL groups (3844.57) compared to the global JOL (2425.88) and no-JOL groups (2509.76), *t*s ≥ 9.37 *p*s ≤ .001.

However, although encoding latencies were greater for item-JOL participants, the use of self-paced encoding is consistent with other studies investigating JOL reactivity. For example, Janes, Rivers, and Dunlosky (2018) compared JOL reactivity effects between self-paced and experimenter paced designs and found that experimenter pacing increased reactivity effect sizes versus self-paced encoding. Separately, Maxwell and Huff (2022; 2023; 2024) have consistently shown that JOL reactivity patterns observed on cued-recall of word pairs in experimenter paced studies extend to a self-paced learning context. As such, there is a precedent in the literature for using self-paced learning to investigate JOL reactivity effects and evidence suggesting that reactivity patterns do not differ with respect to study pacing.

**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 of the materials on pg. 5 (see also my response to the Action Editor, Comment 5 and the revised Experiment 1A materials on pgs. 11-12).

**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 task in Experiment 2 (Table A4, pg. 47). Overall, both JOL types tended to be higher for categorized versus uncategorized list types (though this difference was not significant in Experiment 1B). Although the item-level and global JOL tasks were designed to emphasis item-specific and relational encoding, respectively, intra-list relations the presence (or absence) of intra-list relations was likely still highly salient for both groups, regardless of the specific focus of their JOLs. Because these relations strongly influence the magnitude of JOLs (e.g., Koriat’s, 1997 cue-utilization theory) it is not too surprising that both JOL types were generally greater when words were presented in categorized lists.

**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:*** *p*-values have now been added for all significant main effects, interactions, and post-hoc tests. Please see my response to Reviewer 1, Comment 7.

**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, and I agree that such an analysis would provide additional evidence in favor of an item-order account. However, as you note, these analyses are beyond the scope of this study. I have updated the General Discussion to include analyses of serial position as one potential method for research in this area to explore (see pg. 29). Thank you for this suggestion.

**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.