September XXth, 2022

Ayanna Thomas, PhD

Editor-in-Chief

*Memory & Cognition*

Dear Dr. Thomas,

We have submitted a revised version of MC-ORIG-22-093 “Is Discriminability a Requirement for Reactivity? Comparing the Effects of Mixed vs. Pure List Presentations on Judgment of Learning Reactivity” for your consideration. We appreciate the thorough examination and the feedback provided by our reviewers and are particularly encouraged that our manuscript was viewed by reviewers as “well-written” and “methodologically sound” and that the present research is “valuable for scientific advancement” as it compares “the two most prominent theories of word-pair reactivity.” In our responses, we list each reviewer’s comments and cite page numbers when referencing specific changes. To facilitate review, all primary modifications to the manuscript have been made using blue-colored font. We look forward to your response and hope that this revised version of the manuscript is now suitable for publication in *Memory & Cognition*.

Sincerely,

Nicholas P. Maxwell, PhD.

School of Psychology

The University of Southern Mississippi

Email: nicholas.maxwell@msutexas.edu

Phone: (601) 266-5411

Fax: (601) 266-5580

Cc:

Mark J. Huff, PhD

The University of Southern Mississippi

**Action Editor (Ayanna Thomas):**

The reviewers agree that this is a well written manuscript presenting expertly conducted experiments. However, the primary concern is contribution. Both reviewers 1 and 3 raise this as an issue and I would like to give you the opportunity to revise the manuscript and address this concern. Additional minor points also require your attention, and I will direct you to the reviews included in this email for more detail.

**Response:**

**Reviewer: 1**

Across 3 experiments, the authors examined JOL reactivity: the finding that making JOLs during study enhances memory on a final test compared to not only studying and not making any judgments. The general paradigm involved participants studying word pairs, making judgments after studying each pair or not, and then taking a cued recall test. JOL reactivity was observed for forward related word pairs (credit-card; Ex. 1), backward related word pairs (card-credit; Ex. 2), and symmetrical pairs (king-queen; Ex. 3). No reactivity was observed for unrelated pairs (Ex. 1-3). These patterns of results emerged regardless of whether pair type was manipulated within or between subjects, i.e., whether participants studied mixed lists (e.g., both forward and unrelated pairs in Ex. 1; both backward and unrelated pairs in Ex. 2; both symmetrical and unrelated pairs in Ex. 3) or pure lists (only forward, backward symmetrical, or unrelated words). Judging the frequency with which the words in a pair co-occur in everyday language while studying produced similar effects on final test performance as making JOLs for each pair.  
  
The paper is well written, and the reported studies are experimentally sound. Overall, the topic of JOL reactivity is a major focus of metacognition research and of interest to readers of Memory & Cognition. A growing body of research has found that making JOLs can affect subsequent memory. Various theories—including the changed-goal and cue-strengthening hypotheses cited in the Introduction—have been proposed to account for how JOLs affect memory. My read of the literature is that prior research largely supports the cue-strengthening hypothesis, but more direct evidence is needed. Prior research has tested the cue-strengthening hypothesis rather indirectly by manipulating the methods used in the study phase or test phase (e.g., pair types: related vs. unrelated; list composition: mixed vs. pure lists; test type: cued vs. free recall) and examined the effect on JOL reactivity. The present studies took a similar approach by manipulating both pair relatedness and list composition. In this way, the three reported experiments seemingly contain conditions that are all replications of prior research but have combined them into a single experiment and/or paper. As a result, I believe that these results represent a modest advancement to the literature. Nevertheless, I believe the present paper is valuable for scientific advancement because it provides more evidence and more consistent evidence about the boundary conditions of JOL reactivity. Indeed, as far as I know, only one other study has examined JOL reactivity with symmetric pairs (Maxwell & Huff, in press) only two studies have examined JOL reactivity with backward pairs (Maxwell & Huff, in press; Mitchum et al., 2016). However, I believe that only mixed lists were used in those two papers. Similarly, only two prior studies have manipulated list composition and they found opposite results (Janes et al., 2018; Tauber & Witherby, 2019). I believe that with some additional writing and possibly additional analyses, these well-conducted experiments could be a solid contribution to the Memory & Cognition.  
  
**Comment 1:** P. 21 line 20s-38: What did Maxwell & Huff (in press) and Mitchum et al., (2016) find with regards to reactivity and backwards pairs? Were those studies done with mixed list or pure list designs? Similarly, what did Maxwell & Huff (in press) find with regards to reactivity and symmetric pairs? Was the study done with mixed or pure list designs? Explaining the methods and findings of relevant prior research would help highlight the novel contribution of the present research and clarify whether the results from the present study align with previous work.

***Response:*** [words here]  
  
**Comment 2:** P. 16 line 15: I’d recommend tempering the language “suggesting that JOL reactivity results from strengthening relational cues rather than via metamemorial or predictive processes.” Making JOLs and providing frequency judgments could have produced the same memorial effect via different mechanisms. It doesn’t seem that Experiment 1 was specifically design to test, experimentally, whether JOLs and frequency judgments affect memory through similar mechanisms. The same comment applies to Experiments 2 and 3 as well. Relatedly, I’m not sure it makes sense to say that the word extends makes sense in the claim that “JOL reactivity extends to other non-metacognitive judgment tasks” (e.g., p. 29) since one task involves JOLs and one doesn’t. I would recommend rephrasing this and similar claims to something like “non-metacognitive judgment tasks produce a similar memorial benefit as JOLs.”

***Response:*** [words here]

**Comment 3:** The predictions for Experiment 2 could be clarified. The discussion of Experiment 1 notes that “Thus, while backward pairs are thematically related relatedness cues are not available at retrieval. As a result, it is unclear whether cue-strengthening can occur with backward pairs, given that the target item is a less obvious response to the cue.” (p. 16 line 42). But then, the beginning of Experiment 2 predicts the that “because relatedness cues for backwards pairs is not readily available at retrieval (i.e., the target is a less common response to the cue), any reactivity effects for backward pairs should be reduced compared [to] forward pairs [in] Experiment 1.” (p. 17, line 38). I’d recommend clarifying whether the prediction was reduced reactivity for backward pairs or if you had no a priori prediction. In either case, Experiment 2 does not compare forward pairs to backward pairs and, therefore, cannot provide additional direct evidence for the cue-strengthening hypothesis. Along these lines, I noticed that the manuscript mentions that “no reactivity study using related pairs has only targeted backward pairs without a

***Response:*** [words here]

**Comment 4:** P. 21 line 6: Relatedly, how do the results of Experiment 2 “provide additional support for the cue-strengthening account” since reactivity was observed for backwards pairs? My understanding of the predictions for Experiment 2 was that although JOLs cause participants to consider cues such as item relatedness, those cues wouldn’t be (as) helpful on a cued recall test for backwards pairs.

***Response:*** [words here]

**Comment 5:** What exactly are the relatedness cues that are posited as being activated by making JOLs and thus strengthened by relatedness processing? How can these cues explain why backward pairs showed reactivity, too?

***Response:*** [words here]

**Comment 6:** P. 28 line 5: What does strategically mean in the claim that “relational encoding is applied strategically”? Does this mean selectively for related pairs only? Does strategically imply a that JOLs cause conscious, volitional shift in encoding strategy? If so, wouldn’t the hypothesis be that relational processing would be less likely to occur as a result of pure lists for JOLs? Couldn’t it be that relational encoding is applied to all pairs but because there is no strong connection between unrelated pairs, such relational encoding offers little memorial benefit?

***Response:*** [words here]

**Comment 7:** Ex. 1: Based on the Introduction, the evidence for the changed-goal hypothesis would be reactivity for forward pairs in the mixed list but not pure list condition. In contrast, the cue-strengthening hypothesis would predict reactivity for forward pairs in both the mixed and pure list conditions. Wouldn’t the most appropriate analysis be to examine only the related pairs and conduct a 2 (JOL vs. read only) x 2 (mixed vs. pure) ANOVA? My understanding is that the goal-change hypothesis predicts an interaction, but the cue-strengthening hypothesis does not. As the results are currently reported, the reader must infer a lack of an interaction because the overall pattern of results is the same for related pairs in the mixed and pure conditions. However, this comparison has not been statistically evaluated. The same comment applies to Experiment 2 and 3 analyses.

***Response:*** [words here]

**Comment 8:** Experiments 2 and 3: Were the data collected at the same time as Experiment 1 such that participants were randomly assigned? If not, then I would recommend taking the unrelated pure bars out from the Experiment 2 and 3 graphs (pp. 37-38). I would also hesitate to use the Experiment 1 data in the Experiment 2 and 3 analyses of the pure list conditions. As noted above, the key comparisons seem to be the mixed vs. pure lists for related pairs. Therefore, I would also recommend examining the results of the pure list conditions via a one-way ANOVA: JOL vs. frequency vs. no JOL for just the backward pairs (Experiment 2; p. 20 lines 3-36) and just the symmetrical pairs (Experiment 3; p. 24 line 24 – p. 25 line 10).

***Response:*** [words here]

**Comment 9:** P. 10 line 31: What was the medium effect of interest in the mixed-lists design power analysis? The between variable? The within variable? The interaction?

***Response***: For both

**Comment 10:** P. 13: Was study time self-paced or experimenter paced? Relatedly, based on my read of the manuscript, the goal-changed hypothesis suggests that participants would need to be aware of the different types of items and “prioritiz[e] easy pairs at the expense of more difficult ones” (p. 16, line 31.) If this is the case, would studying need to be self-paced for reactivity to emerge?

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

**Comment 11:**  P. 13: Please report the pBIC for statistically significant effects as well so that readers can interpret the strength of the evidence for the null and strength of the evidence for the alternative for those significant analyses. After all, the p value is not an indicator of the strength of the evidence. It may help to briefly provide information on how to interpret pBIC as well.

***Response:*** The *p*BIC statistic estimates the strength of evidence in favor of the null hypothesis. We included this additional statistic as a supplement to traditional null hypothesis testing whenever null effects were detected. Because *p*BICs are assessing evidence in favor of the null hypothesis, they are essentially meaningless when calculated for significant effects.

[INTERPRETATION PARAGRAPH]

**Comment 12.** P. 14 and subsequent results: Please report all inferential statistics (ts and ds and pBICs) rather than using >. This will be helpful for future meta-analyses. These could be reported in an appendix or supplemental materials, though.

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

**Comment 13.**  Ex. 1-3: Did you analyze the experiments with source (university vs. Prolific) to verify the pattern of results was not affected by how participants were recruited?

***Response:*** For all experiments, each ANOVA model initially included an additional between-subjects factor assessing the effects of participant recruitment platform on cued-recall performance. Starting with Experiment 1, no main effects of Recruitment Platform were detected, regardless of whether participants studied mixed or pure lists, *F*s < 1, *p*BICs ≥ .87. Additionally, no interactions with platform were detected, including the three-way interactions between Recruitment Platform, Pair Direction, and Encoding Task, *F*s < 1, *p*BICs ≥ .98.

In Experiment 2, cued-recall of mixed-list items was marginally higher for Prolific participants versus university students (39.74 vs. 32.77, respectively; *F*(1, 114) = 3.43 *MSE* = 587.36, *p* = .07, *p*BIC = .66. Importantly, however, no interactions with platform were detected, *F*s < 1, *p*BICs ≥ .83. For recall of pure list items, no differences were detected between Prolific and university participants (32.87 vs. 35.16; *F* < 1, *p*BIC = .93), and again, no interactions with Recruitment Platform were detected, *F*s < 1, *p*BICs ≥ .91.

Finally, in Experiment 3 cued-recall of mixed-list items showed no difference between Prolific and university participants (45.21 vs. 43.92, *F* < 1, *p*BIC = .91), and no interactions occurred with Recruitment Platform, *F*s ≤ 1.41, *p*BICs ≥ .96. For pure lists, cued-recall again did not statistically differ between Prolific participants and university students (39.63 vs.50.51; *F*(1, 197) = 1.35, *MSE* = 242.43, *p*BIC = .88). However, a significant Recruitment Platform × Encoding group interaction was detected, *F*(2, 197) = 4.11, *MSE* = 242.43, *ηp*2 = .04. Post-hoc testing revealed that this interaction was driven by recall differences between participants completing the JOL task, as recall performance in the JOL group was greater for university students (57.56) relative to Prolific participants (28.96; *t*(65) = 4.22, *SEM* = 6.95, *d* = 1.12). Upon further inspection, however, it was revealed that the majority of the Prolific JOL participants were randomly assigned to study pure unrelated lists (79%). Thus, the decreased performance of Prolific participants likely reflects the more difficult nature of the word pairs they studied, rather than any qualitative differences in participants based on recruitment source. Additionally, comparisons the other encoding groups revealed no differences in cued-recall performance between Prolific participants and university students, (*t*s < 1, *p*s ≥ .77, *p*BICs ≥ .89). Finally, no other interactions with Recruitment Platform were detected, *F*s < 1, *p*BICs ≥ .91). Taken together, it is likely that recruitment source had no effect on our reactivity findings.

**Comment 14.** P. 21 line 54: Is there evidence that symmetrical pairs are “deceptive”? I’m assuming that deceptive means that they engender overconfident JOLs? Is the absolute accuracy of JOLs important for understanding JOL reactivity?

***Response:*** By deceptive, we mean word pairs in which the relatedness cues that participants used to inform their judgments may be less readily available when the cue item is presented in isolation at retrieval. This is apparent in the case of backward associates (e.g., card – credit), as**… [**WORDS HERE] For symmetrical pairs, [EXPAND]

•       Might it make sense to reorder the experiments as 1, 3, then 2 since the predictions and proposed mechanisms are identical for forward (Ex. 1) and symmetrical (Ex. 3) pairs?

***Response***: While the mechanisms between forward and symmetrical pairs are similar, we elected to present the backward pair findings as Experiment 2, as this pair type is more commonly used in the JOL literature relative to symmetrical pairs.

All other minor spelling and grammatical errors have been addressed. We appreciate your attention to detail.

**Reviewer: 2**

**Comment 1:** The changes goal hypothesis argues that people shift their study time/effort to mastering easy items, while pure lists don’t have the obvious cue of relatedness isn’t is possible that reactivity could be observed here because they are focusing on the easier items (based on some other cue like frequency)?

***Response:*** Mitchum et al.’s (2016) changed-goal hypothesis initially described difficulty of paired-associates in terms of their relatedness (i.e., unrelated pairs are difficult because the cue does not converge upon the target). While other factors can certainly influence pair difficulty (i.e., an individual’s likelihood of correctly recalling an item on a future test), we note that pair relatedness is a strong predictor of future recall (Maxwell & Buchanan, 2020), and furthermore, the JOL task encourages participants to implicitly relate items together at study (see Maxwell & Huff, 2022). Furthermore, semantic information like relatedness likely reflects an automatic process. For example, research on semantic priming has shown that priming effects extend to mediated associates (e.g., lion – stripes), in which paired items are unrelated but are each related to a common third concept (e.g., tiger; see Hutchison, 2003 for a review). Additionally, the DRM false memory illusion has been shown to extend to mediated lists (Huff & Hutchison, 2011), further suggesting the automatic nature of semantic information.

Finally, we note that across our experiments, all list types were matched on several variables that could potentially influence recall, including frequency, length, and concreteness. Thus, unrelated targets were unlikely to differ from related targets in frequency, yet only related targets showed a memory improvement from making judgments at encoding.  
  
**Comment 2:** I wasn’t entirely sure how the sample was determined, the power analysis suggested a significantly smaller sample would have been adequate and Prolific and student samples were combined. Did any of the results differ as a function of sample? What was the logic of topping up the samples with Prolific?

***Response:*** As reported in the Methods section for Experiment 1 (pg. XX), we initially based our samples on a set of a priori power analyses. However, we oversampled due to concerns with participant performance. This study was originally planned to be conducted in-lab, however, due to the Covid-19 pandemic, we were forced to shift our data collection efforts online. We therefore oversampled each group as an additional safeguard against increased participant performance variability that is inherent to conducting research online.

While participants were primarily recruited from the University of Southern Mississippi, we extended our data collection efforts to Prolific, as participant recruitment was dwindling. Our primary goal was to ensure that each cell had at least 35 participants pre-data screening.

Results did not differ as function of recruitment source (please see our response to Review 1’s 13th comment for more details).

**Comment 3:** On pg. 26 its argued that the easy/difficult comparison triggers the change in study goals, but it might be more accurate to say that the metacognitive evaluation produces a change in goals towards mastery, and that re-studying the related word-pairs are the easiest way to achieve that.

***Response:*** We have updated the language on page xx accordingly. We now describe the changed-goal hypothesis as occurring whenever “metacognitive evaluation produces a shift in study goals” rather than being the product of an “easy/difficult” comparison.

**Reviewer: 3**

**Comment 1:** From my read, the key contributions of this manuscript are (a) reactivity is evident in pure and mixed lists, (b) JOLs and frequency judgments produce similar reactivity effects, and (c) reactivity extends to other kinds of related words. All these contributions are nice and fit well with existing JOL theory. My concern is that all these outcomes have also been demonstrated in other recent papers. Regarding point A, Janes et al. (2018, Exp 2) compared a mixed and pure list, Rivers et al. (2021, Exp 2) compared mixed and pure lists using a blocked design (e.g., all unrelated first/ all related first), and Witherby and Tauber (2017) and Tauber and Wither (2019) demonstrated positive reactivity with related pairs in a pure list (granted there was no unrelated pure list comparison). Regarding point B, Maxwell and Huff (2022) showed that frequency judgments elicit reactivity effects as well. Finally, regarding point C, I’m aware of a few papers that have explored whether reactivity extends to other types of related words including symmetrical pairs (Maxwell & Huff, 2022), identical pairs (Halamish & Undorf, in press), and categorized words (Senkova & Otani, 2021). I am not aware of anyone who has looked at backward related pairs, so Experiment 2 of this paper may be the first to do that. That said, I do think it is promising that the outcomes of the present research are consistent and replicate those in all these other studies. I think such replications are valuable especially with how little JOL reactivity research there is.

***Response:*** Although other researchers have investigated other areas of JOL reactivity (e.g., mixed vs. pure lists; Janes et al., 2018, Tauber & Witherby, 2019; pair direction, Mitchum et al., 2016, Maxwell & Huff, 2022), the present study is the first to combine each of these individual pieces into a single, unified manuscript. One goal of this study was to consolidate previous research on JOL reactivity while 1) replicating established reactivity patterns found across most studies (i.e., positive reactivity for related pairs, no reactivity for unrelated pairs) and 2) providing a more direct test of the changed-goal and cue-strengthening accounts of reactivity. [SOMETHING] As you note, these replications are of particular value when investigating topics like JOL reactivity in which relatively little research has been conducted.

My remaining comments are only minor:  
  
**Comment 2:** Timing details are missing from the procedure. How long was each word presented during study? What happened if participants failed to make a judgment in the judgment groups? Did the program advance or did they get another prompt? How often did participants fail to make judgments during study? Was the cued-recall test self-paced? – In the GD I see that the authors note that everything was self-paced. I would make sure to include this in the procedure. I think it would also be worth reporting statistics regarding differences in study/exposure time between groups. If there is, I would not be concerned that the reactivity is due to time differences rather than the judgments because there is an abundance of research using experimenter paced designs to equate exposure time and reactivity effects are still evident. Even so, it is useful to know if differences arose and how big they were. Glancing at the means in Tables 8 and 9, I don’t see any consistent patterns, but it may be worth looking at. This is also something that I believe could be used evaluate the changed-goal hypothesis in a little different way. That is, from that hypothesis, you might expect study time to be greater for easy than hard pairs when participants make JOLs (assuming participants are shifting their goals to prioritize easier items). By contrast, when participants don’t make JOLs they may prioritize difficult over easy items (which is what a discrepancy-reduction model of study time allocation might predict).

***Response***: For all experiments, study and test were self-paced. We have updated the procedure section on page xx to clarify this detail. After making a judgment, participants pressed the ENTER key to move to the next study pair. If participants failed to make judgments, they were excluded from the final analyses. Analyses of the means reported in Tables 8 and 9 did not yield any patterns that were consistent across studies; however, we included these tables for completeness.

[MITCHUM ET AL RESPONSE TIMES – MENTION THAT RESPONSE TIME DATA IS DIFFICULT TO INTERPRET GIVEN THE ONLINE NATURE OF THE STUDY – CITE OUR OTHER REACTIVITY PAPER HERE!]

**Comment 3:** For the significant effects, why don’t the authors report p-values? For example, the p-value is missing on p. 14 for the main effect of pair type and the interaction. I get that the F values reflect the significant effects, but convention is to still report them. Similarly, for the marginal means, SE or SD should also be reported.

***Response:*** We choose to report *p*-values only for non-significant and marginal effects, as for significant effects, the primary interest should be the size of the effect (as indicated by partial eta-squared for ANOVAs or Cohen’s *d* for *t*-tests).  
  
**Comment 4:** On p. 15 it says “recall was highest for frequency judgments (50.69), followed by the JOL (51.40) and …” the authors may want to double check this to fix the order of the groups or the values.

***Response***: We appreciate your attention to detail. This has been corrected.