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 are glad that the manuscript was viewed as “well-written” and “methodologically sound” and, further, are particularly encouraged that this set of studies was viewed as being “valuable for scientific advancement.” In our responses below, we list each reviewer’s comments and cite page numbers when referencing specific changes to the manuscript. 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 our 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:** We have the revised our manuscript in accordance with each reviewer’s suggestions. Specifically, we have updated the introduction (pg. xx) to highlight important methodological discrepancies between previous JOL reactivity studies which also used backward associates (see our response to Reviewer 1, comment 1), clarified our interpretation of backward pair reactivity (Reviewer 1, comment 4), clarified the discrepancies between previous comparisons of mixed/pure list reactivity (Reviewer 3, comment 1), and have addressed the importance of assessing reactivity effects for other types of related word pairs (backward and symmetrical associates) in isolation from standard forward associates (pg. xx; Reviewer 3, comment 1). In doing so, we believe that the novel contributions of our manuscript have now been strengthened.

Additionally, while we understand the concern with contribution, we note that both Reviewers 1 and 3 each mentioned the importance of the replication aspect of our study, with Reviewer 1 going so far as to state that “the present paper is valuable for scientific advancement because it provides more evidence and more consistent evidence about the boundary conditions of JOL reactivity.” Further, a close reading of each reviewer’s comments suggests that only Reviewer 3 raised concerns about the contribution of our manuscript. Even so, Reviewer 3 was quick to note the value in our replication of existing reactivity patterns and extension to new list types (pure backward and pure symmetrical), given that relatively little work has explored JOL reactivity effects on pure lists. In our revised manuscript, we now highlight this replication aspect in the General Discussion (pg. xx), in addition to the previously mentioned changes.

**Reviewer: 1**

**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:*** Thank you for your comment. Both Maxwell and Huff (2022) and Mitchum et al. (2016) used mixed lists designs for all experiments. Regarding JOL reactivity for backward pairs, Maxwell and Huff showed that 1) backward pairs produced positive reactivity patterns mirroring other related pair types (forward and symmetrical associates) and 2) these patterns replicate when participants make non-metacognitive Judgments of Associative Memory (JAMS; i.e., likelihood that the cue word would be given as a common response to the target) and frequency judgments (i.e., likelihood cue and target would co-occur in natural language). These patterns were taken as evidence in favor of Soderstrom et al.’s (2015) cue-strengthening account. Mitchum et al., however, found no JOL reactivity on backward pairs, though we also note they similarly showed no reactivity on forward associates. Given the lack of difference between forward and backward associates, Mitchum et al. elected to only investigate forward associates in their subsequent experiments.

We have updated this section accordingly and now discuss methodological discrepancies between both studies while also clarifying how our present finding of backward associate reactivity in mixed lists replicates previous findings from Maxwell and Huff (pg. xx).

**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:*** We have rephrased this section accordingly. On page xx, we now note that JOL reactivity patterns may result from cue-strengthening but leave open the possibility that both judgment types may operate using different mechanisms. The corresponding sections for Experiments 2 (page xx) and 3 (page xx) have similarly been updated.

Additionally, we now state on page xx that “the memory pattern found using JOLs is similar to those found using other, non-metacognitive judgment tasks that similarly emphasize cue-target relations.”

**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 forward associate comparison group.”

***Response:*** We have clarified our predictions regarding backward pairs in Experiment 2. Specifically, we have updated the Experiment 1 discussion (page xx) to note that for backward pairs, relatedness cues are less likely to be available at retrieval (as opposed to these cues being completely absent). Additionally, we have clarified our predictions in the Experiment 2 introduction (page xx) to be consistent with the Experiment 1 discussion.

**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:*** Any reactivity for backward associates likely reflects the JOL task encouraging participants to engage in relational processing at study. For example, Maxwell and Huff (2022) proposed that positive reactivity on related pairs was likely due to the cue-strengthening aspect of the JOL task implicitly encouraging participants to engage in relational encoding at study. Thus, related pair types consistently receive a memorial benefit when JOLs are made at encoding, regardless of their associative direction. Unrelated pairs, however, do not receive this benefit, as these pairs do contain obvious relatedness cues. We have updated the Experiment 2 Discussion on page xx to be consistent with this account.

**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:*** When completing JOL tasks, participants use many cues about the conditions of their learning (extrinsic cues) and the inherent properties of the stimuli (intrinsic cues) to inform their judgments (see Koriat, 1997). Relatedness cues refer to a specific type of intrinsic cue that is unique to cue-target word pairs. For example, participants inherently know that the pair *cat-dog* is more semantically related than the pair *cat-door*, and as a result, will assign the former a pair higher JOL rating than the latter. Soderstrom et al. (2015) suggested that related pairs show positive reactivity because these cues become strengthened when participants are required to make JOLs at encoding. Unrelated pairs, however, lack these cues. Thus, the requirement to make JOLs only benefits related word pairs.

Regarding backward pairs, these intrinsic relatedness cues are still present at encoding; however, the target item does not readily converge on the cue at retrieval (i.e., both words are thematically related, yet the target is not a common response to the cue). Thus, it is likely that cue-strengthening also results in participants engaging in a relational encoding process at study, leading to a memorial benefit for related pairs, regardless of pair direction. We have clarified our backward pair predictions on page xx to align more closely with this account.

**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:*** By strategic, we are suggesting that these judgment tasks direct participants attention towards intrinsic cues about each study pair. Because relatedness is a highly salient cue that is processed automatically (see Hutchison, 2003), we suggest that JOL tasks selectively encourages participants to engage in relational encoding, but only when pairs are related (see our response to Comment 4).

Regarding your final point, it is likely that any relational encoding on related pairs is not occuring on unrelated pairs. For example, Maxwell and Huff (2022) compared reactivity to a direct relational encoding task in which participants were explicitly instructed to relate study pairs together, regardless of relatedness. Overall, JOLs produced the standard reactivity pattern (positive reactivity for related pairs, no reactivity for unrelated pairs), but critically, the relational encoding task produced similar memorial benefits on related pairs while also extending this benefit to unrelated pairs. Thus, if making JOLs resulted in participants applying globally encoding across pair types, a memorial benefit would be expected to occur on both related and unrelated pairs. However, JOLs have routinely been shown to produce a benefit on related, but not unrelated, pairs.

**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:*** We appreciate this suggestion. However, we note that the analyses in our initial submission were directly modeled after Janes et al.’s (2018) Experiment 2, in which the authors separately analyzed reactivity for mixed and pure lists. In running the suggested analyses, we note that all analyses yield significant main effects of Encoding Strategy (*F*s ≥ 9.36; *ηp*2s ≥ .07), and importantly, no significant interactions emerge (*F*s < 1; *p*s ≥ .48, *p*bics ≥ .98). Thus, it is likely that list structure does not differently affect reactivity.

Regarding our inclusion of unrelated pairs in each analysis, the changed-goal hypothesis predicts a negative reactivity pattern on unrelated pairs (in addition to positive reactivity on related pairs). Although the reactivity pattern most reported in the literature has been positive reactivity on related pairs and no reactivity on unrelated pairs (e.g., Janes et al., 2018; Maxwell & Huff, 2022, Soderstrom et al., 2015), we note that Mitchum et al. (2016) reported a divergent pattern in which related pairs showed no reactivity, and JOLs produced *negative* reactivity on unrelated pairs. Given that Mitchum et al. replicated this pattern across several experimental manipulations, we included unrelated pairs in each our analyses, as the potential for negative reactivity on this pair type remains a possibility. Therefore, we believe that our inclusion of unrelated pair types in Experiments 2 and 3 is justified, as it allowed us to provide a further test of whether JOLs would produce negative reactivity on unrelated pairs.

**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:*** Data collection for all experiments was conducted online simultaneously. We have elected retain the pure bars in the Experiment 2 and 3 graphs, as we believe they provide important context for interpreting the reactivity patterns we report.

Running the requested one-way ANOVA models on the pure lists produces similar findings as both the previously suggested set of ANOVAs and those we report in our initial submission. Specifically, an effect of encoding group emerges for related pure lists in all experiments (*F*s ≥ 3.88, *ηp*2s ≥ .06).

**Minor Comments:**

**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***: The medium effect size reported was to detect an interaction between Pair Type and Study Group. We have updated the paragraph on page xx describing the power analysis to more accurately reflect this.

**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:*** As noted in our initial submission, study was self-paced across all experiments. Although goal-changing could potentially still occur when using an experimenter pacing (i.e., participants could potentially “zone-out” or allocate less effort at encoding when more difficult pairs are presented), we reasoned that using self-paced study would allow participants the opportunity to differentially allocate their study time for each pair type (e.g., Mitchum et al., 2016). However, as reported in Tables A8 and A9, no discernable pattern emerged for RTs.

While no clear RT pattern was detected, we note that encoding durations can be difficult to interpret. For example, several well-established memory effects including generation (Slamecka & Graf, 1978) and production (Icht, Mama, & Algom, 2014) have been shown to occur even when encoding durations were equated to a control task. In other words, spending more time encoding an item does not necessarily mean that the item will be better remembered. We also point out the online nature of this study makes interpreting RTs in this situation particularly difficult, as both differences in internet connectivity speed between participants and the lack of an experimenter present during study may have contributed to increased RT variability.

Finally, regarding participants’ perceptions of pair difficulty, it is likely that they are actively aware of the difference between related and unrelated pairs at encoding. For example, JOLs consistently show a relatedness effect across all experiments, such that related pairs receive higher JOLs relative to unrelated pairs. Indeed, relatedness cues are likely one of the strongest indicators of later test performance. Thus, participants are likely to aware of the differences between pair types, and actively use this information to inform their judgments.

**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 provides an estimate of the likelihood that the null hypothesis is retained. and address a misunderstanding that you have indicated here regarding the strength of evidence. Like *p*-values,*p*BIC does not specify the strength of evidence for the null. We have updated our explanation of *p*BIC (pg. xx) clarify how this statistic should be interpreted.

Because *p*BIC provides an estimate of the likelihood that the null hypothesis is retained, this statistic provides little information for significant effects. Therefore, we have elected to include this statistic only for non-significant effects.

**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***: We understand your concern regarding metanalyses; however, we have elected to not report these statistics for the sake of concision. However, we note that these statistics are available to other researchers, as all data files and analysis code are hosted on OSF. Our initial submission included multiple links directing interested individuals to our data repository.

**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.

[FOOTNOTES?]

**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 use to inform their judgment values may be less readily available when the cue item is presented in isolation at retrieval. This is readily apparent in the case of backward associates (e.g., card – credit), as*card* does not share semantic features with credit nor is it a common response. JOLs for the pair type commonly overestimate later recall, pattern termed the Illusion of Competence (Koriat & Bjork, 2005). Regarding symmetrical pairs, Maxwell and Huff (2021) showed that this overconfidence pattern similarly extends to this pair type, providing evidence that this pair type similarly leads to overinflated judgments.

While the absolute accuracy of these judgments alone is not necessarily important for interpreting JOL reactivity effects, the cue-strengthening account posits that the relatedness cues used to inform each judgment are critical for reactivity. Thus, because deceptive pairs contain cues that are less likely to be available at test, reactivity may be less likely to occur on these pair types. Alternatively, if reactivity reflects the use of relational encoding (see Maxwell & Huff, 2022), reactivity would likely still occur, the added processing would provide a memorial benefit. We have updated the Experiment 2 Discussion and Experiment 3 introductions (pgs. xx-xx) to more clearly convey our position.

**Comment 15.** 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 Experiment 1 Methods (pg. XX), we initially based our samples on a set of a priori power analyses conducted with G\*Power. This study was originally planned to be conducted in-lab; however, prior to the start of data collection, the Covid-19 pandemic forced us to shift our data collection online. As a result, we 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 with Prolific recruitment was to ensure that each cell had at least 35 participants pre-data screening. However, no differences in cued-recall performance were detected between recruitment sources (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 pg. xx accordingly. We now describe the changed-goal hypothesis as occurring whenever “metacognitive evaluation produces a shift in study goals” rather than it simply being the product of an “easy/difficult” comparison.

Thank you for taking the time to review our manuscript.

**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 similar areas of JOL reactivity (e.g., reactivity for mixed vs. pure lists; Janes et al., 2018, Tauber & Witherby, 2019; effects of associative direction on reactivity in mixed lists, Mitchum et al., 2016, Maxwell & Huff, 2022; whether other, non-metacognitive judgments produce similar reactivity patterns, Maxwell & Huff 2022), the present study marks the first attempt at combining each of these individual lines of research into a single, unified manuscript. As such, we investigate two types of judgments (JOLs vs. Frequency judgments), two list types (mixed vs. pure) and three types of paired associates (forward, backward, and symmetrical).

In addition to the key points you have highlighted above (e.g., reactivity occurs for related pairs in both list types, for both judgment types, and for all related pair types, regardless of associative direction), the present study provides more compelling evidence that pure list reactivity patterns mirror those observed in mixed lists. Previous research has shown mixed results, with some studies (e.g., Tauber and Witherby, 2019) finding reactivity for pure, related lists, while others (e.g., Janes et al., 2018) showed no reactivity on pure lists, regardless of pair relatedness. We also note, that while Tauber and Witherby used showed reactivity on a pure related list, they were primarily interested in whether reactivity effects would be observed in older adults (relative to young adults). Further, Tauber and Witherby’s particpants only studied related word pairs. Thus, an investigation of the changed-goal hypothesis was not possible using their design. We have updated the language on pg. xx to clarify this point. As you note, there is indeed value in replication, especially when comparatively little research has been completed on a topic. Thus, the present adds to the growing body of literature indicating that JOLs are reactive on related pairs, and that this reactivity is not contingent on list composition.

Finally, you are correct that, to our knowledge, Experiments 2 and 3 provide the first reactivity studies in which traditional forward associates are replaced with backward (Experiment 2) or symmetrical paired associates (Experiment 3). This is important, as it provides further evidence that relatedness cues between cue and target (or target and cue in the case of backward pairs) encourage participants engage in relational encoding at study (see Maxwell & Huff, 2022). Thus, these related pairs still show a memory improvement relative to unrelated pairs, in which relatedness cues are absent. We have updated the General Discussion on pg. xx to reflect this novel contribution.

**Minor Comments**

**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, both study and test were self-paced. 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. This has been clarified on pg. xx.

Analyses of the mean RTs reported in Tables 8 and 9 did not yield any patterns that were consistent across studies; however, we included these tables for completeness. However, this may partly be attributed to the online nature of the study (please see our response to Reviewer 1’s 10th comment above).

**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. Thank you for taking the time to review our manuscript.