**Action Editor Comments**

Reviewer 1’s point 4 is where I want to start.  Reviewer 1’s point is that you compare the making of JOLs to other ways of increasing relational processing.  By doing so, you show that because other factors that increase relational processing improve memory, then that must be why JOLs increase memory. Reviewer 1 argues that this reasoning is circular and writes, “I concede that the findings are suggestive, and consistent, but the strongest test of the idea would be to try to refute the account.”  The last sentence of Reviewer 1’s review offers a number of variables that would effectively serve as means to seek to “disconfirm your hypothesis.”

Reviewer 3 also makes a very similar point to start off that review (Point 1).  Reviewer 3 is also concerned whether the different tasks first produce the same change in processes and what might be different among them. Reviewer 3 writes, “Potentially, JAMs and explicit relational encoding instructions trigger more relational encoding than JOLs (see also p. 23), but JOLs compensate for this with memory forecasting, resulting in similar reactivity effects. The experiments nicely show that memory forecasting is not necessary for reactivity effects per se since JAMs, frequency judgments and explicit relational encoding all led to such effects. Yet, this does not necessarily imply that JOLs' reactivity effects are not due (in part) to memory forecasting. The authors should elaborate more on this point and tone down their conclusion.”  Reviewer 3, in effect, clearly states Reviewer 1’s concerns about the nature of the tasks and how they might affect performance and how different variables might lead to similar effects.  This makes Reviewer 1’s point about doing an experiment with the goal of “disconfirming the hypothesis” important.

Please strongly consider one of Reviewer 1’s suggestions for such an experiment or include one of your own. If you choose not to, I will need a very clear explanation of why such a study is not necessary. I really like the idea of adding a study in which you compare a JOL condition to a low-level-of-processing condition. It might also be interesting to encourage participants to use item-specific encoding (for each item in a paired associate) and see if you that produces negative reactivity with highly related pairs (i.e., Mulligan and Peterson, 2015).

***Response:***Thank you for pointing out this issue. The finding that JOLs produce memory benefits that are consistent with several different types of relational tasks provides reliable evidence that JOLs are likely to encourage relational encoding. Despite this pattern, we agree that stronger evidence for this pattern would be a comparison to a non-associative task, which is in part why we had already included this task in our initial submission (Experiment 4). Indeed, we compared a low-LOP vowel counting task (which was a specific task suggested by Reviewer 1) to a JOL group, a relational encoding group, and a no-JOL group. Providing support for our relational encoding argument, the JOL and relational groups were equivalent except on unrelated pairs (which was due to explicit relational encoding being non-strategic as we argue), and importantly, the vowel group showed negative reactivity on forward and symmetrical pairs in which the cue word is most predictive of the target. Thus, patterns consistent with JOL reactivity are only found under conditions that are likely to foster relational encoding, shallow tasks that emphasize item-specific characteristics of the words do not show reactivity and even show a memory cost for two of our related pair types.

Although our initial submission had already contained this comparison, we were not very clear in our writing about communicating it’s importance. We have now reframed Experiment 4 (pgs. 28-29) to emphasize the need for a disconfirming comparison and how it supports our processing argument.

1. I also think it is important to address both Reviewer 1 and Reviewer 2’s general concerns about the use of the word “strategy.” Please address the various points that Reviewer 1 makes about how this term is used in your paper, and the specific suggestions Reviewer 2 has for you for discussing the role of strategy in JOLs.

***Response:*** We have revised our manuscript based on the reviewer comments regarding strategy and have provided responses to their reviewer comments below.

1. All three reviewers are also concerned about being careful in what you generalize your results to.  As Reviewer 1 points out, in this study, you can only generalize your results to paired-associates in list format.  Reactivity in metacognition has fortunately now looked at a number of different metacognitive judgments using different stimuli in different contexts.

***Response:***Agreed. We have made revisions accordingly. We thank you for your suggestions and synthesis of the reviewer comments.

**Reviewer #1:**

**1: Framing**

Overall, I think the authors could be more constrained or cautious about the conclusions they draw about "JOLs", and instead frame their work as being about item-based JOLs for word-pairs.   
In paragraph 1, the authors make the case for the importance of studying judgements of learning because they allow "individuals to adjust their study strategies" (p.2, emphasis added for reasons discussed in point 2). I agree, but when I think of strategic control of learning, in particular I think of global judgements of learning for text materials (or topics for an exam), with the opportunity to select topics for revision or restudy, not JOLS for cue-target pairs presented once only.

By paragraph 2, the framing narrows entirely to JOLs for word-pairs, presented in list format. I acknowledge that this is a common method for testing (item-based) JOLs, but it is not the only way in which JOLs can be elicited, or how they may influence learning. This isn't acknowledged anywhere in the manuscript, except in the general discussion where the discrepant findings of Senkova and Otani (2021) are discussed. But even here, this difference is relegated to a methodological point, rather than acknowledging the more general point that this paper isn't about JOLs per se, but is specifically about JOLs for word-pairs presented in list format.

***Response:*** Good point. We’ve updated the Abstract, Introduction (second paragraph page 2), and General Discussion to reflect our use of item-based JOLs which, as you mention, are among the most commonly used materials in the JOL literature.

**2: Strategy**

I struggled to understand the use of the term "strategic" throughout the manuscript, and in my description of the studies above, I avoided using the term, but I believe I captured the essence of the findings.

It was never clear to me what "strategic" meant with regards to the experimental data, or what participants might be doing when they make JOLs. This happens throughout the manuscript, but I will provide some examples here:

a)    The Introduction makes two statements about the nature of a strategy:  On page 7, the authors state that "individuals prioritize encoding of related pairs when making JOL ratings, but this priority is not accompanied by a concomitant cost to the encoding of unrelated pairs". Later, in the overview of the experiments on pages 10-11,  the authors state that reactivity only emerges for related pairs, because metacognition is strategic, "it is assumed that this pattern occurs because participants selectively emphasize processing of related (but not unrelated) pairs at encoding, leading to their greater recall".   
  
These are the clearest statements of what the authors mean by strategy, but there is no subsequent operationalisation of the concepts of prioritization or selective emphasis in the subsequent experiments. I assume that the authors mean something like "devoting more time, or encoding resources", but they don't expand on these concepts, or report any analysis of how people encode. Instead, all conclusions are drawn from the recall data, which are the data that the theory is supposed to be explaining.   
  
Note - there is reference to RT analysis in the supplemental materials, but I couldn't navigate my way around that.

***Response:*** Yes, you are absolutely correct. By strategic, we are suggesting that participants are choosing to process certain pairs for different durations or more or less effectively than other pairs. It is important to make this distinction clear at the beginning, and we therefore provide a more explicit definition on pg. 11. The use of strategic encoding in our study is based solely on recall performance. Although we collected encoding RTs for completeness (which are available in both our supplement and our OSF page), the RT data are difficult to interpret because task responses were made concurrently with study. For instance, in the JOL task, participants are both encoding pairs and providing JOLs, so we do not have a measure how of how long participants are only encoding a pair or only providing a JOL. What’s more, encoding durations are not informative regarding encoding effectiveness or efficiency. Classically, the generation effect still emerges when compared to control conditions with longer encoding durations (Slamecka & Graf, 1978) and the production effect similarly holds when encoding durations are equated to a “neutral” control task (Icht et al., 2014). Thus, RTs may be an unreliable measure of encoding effectiveness and provide no information about qualitative processes within the task.

b)    Experiment 1 is introduced as a test of the changed goal hypothesis and the cue-strengthening hypothesis. There is no use of the term strategy or strategic in this introduction, but if I had to select which of these two accounts involves more "strategy" I would choose "changed goal". The results then favoured the cue-strengthening hypothesis, and the authors wrote: "This notion is complimentary  to previous research on JOL reactivity conducted by Soderstrom et al. (2015) who proposed that JOLs were reactive because they strengthened cues used at retrieval (e.g. pair relatedness). Though they made no explicit claims regarding the strategic nature of any JOL inducing relational encoding, previous work on metacognition (e.g. Nelson & Narens, 1990) has already proposed that metacognitive processes operate in a strategic manner. Therefore, our findings in Experiment 1 provide further support for Soderstrom et al.'s (2015) account while simultaneously providing additional evidence for strategy use regarding reactivity". (p 17-18, emphases added).

I simply do not understand this claim: what additional evidence is provided for strategic involvement in the reactivity beyond the final recall measure? If the only evidence is the final recall, then the argument is entirely circular, and the need for "strategy" is not made. Given that Soderstrom et al (2015) saw no need to invoke strategic processes, what has changed?

***Response:***The evidence here is regarding the type of strategy that is being implemented, not whether a strategy is being used or not. According to the changed goals account, participants strategically emphasize processing on related pairs, and strategically de-emphasize processing on unrelated pairs, leading to a memory cost on unrelated pairs. Given that we and most other researchers do not find a cost, this suggests that participants are engaging in a cue-strengthening process in which related pairs are favored, but this is not negatively affecting encoding of unrelated pairs.

We agree that relying upon recall alone makes the argument circular as we cannot use latencies or other online measures (an important point we now make in our GD, pg. 40), but strategic implementation of memorial processes is a fundamental component of Nelson and Naren’s framework. Soderstrom et al. similarly interpreted their findings using this framework including strategy use but were less clear about how strategies are applied (e.g., relational vs. item-specific) and under what contexts (related vs. unrelated pairs and what type of related pairs). We provide some additional clarity regarding qualitative strategy but do not find the circularity issue to be problematic as this is a criticism for some of the most well-known memory phenomena (e.g., deep LOP effects).

Most importantly, however (and as you note in Comment e), we have non-circular evidence for strategy use in Experiment 4 in our explicit relational group, who showed a memory benefit on unrelated words when participants were required to use a relational strategy for all pair types. As we indicate, if participants were not using a strategy when providing JOLs, then they would use relational encoding for all pair types including unrelated pairs. However, this was not the case.

c)      In the discussion of Experiment 1, the authors state "As demonstrated in Experiment 1, related pairs, regardless of their associative direction, are prioritized at encoding, and thus receive a recall boost. Given this pattern, it is likely that participants are strategically processing related pairs over unrelated pairs, leading to only a memory benefit for related pairs"(p.17, emphasis added). But the only evidence for prioritization at encoding is the recall boost. Thus, this is an entirely circular argument.

***Response:*** See our response for Comment b.  
  
d)      A similar logic applies in Experiment 2, where the authors argue that "the finding that reactivity does not operate globally across all pair types…further suggests that reactivity processes are applied strategically, with an emphasis on related over unrelated pairs".  Again, I don't understand this claim, which again seems to rest on the outcome: a difference is found therefore it must be strategic in origin. Why is this, particularly when the same finding is found with the JAM task?

***Response:***See our response for Comment b. The hallmark finding of Experiment 2 is that the JAM task yields the same pattern as JOLs, which indicates these patterns can occur without memory forecasting. Therefore, the same relational encoding processes that are applied strategically with JOLs may also be applied strategically when providing JAMs.

e)      In the General Discussion the authors state "Thus, only related pairs are processed using a relational encoding strategy, as participants modify their study strategy whenever they encounter this pair type". What is the evidence that (i) this is strategic and (ii) there is modification of such a strategy? If the answer is the recall data, then this becomes circular.   
  
In defence of the authors, there is an interesting finding in Experiment 4 which compares JOLs to an explicit relational judgement applied to all items. Here, there is greater memory benefit to unrelated items from an explicit relational strategy, compared to JOLs. This is interesting, though it is rather buried in the complexity of the 4 x 4 ANOVA, and it merits replication. However, there are two potential explanations: the one favoured by the authors is that JOLs are the same as relational encoding, but participants choose not to employ relational processing for unrelated pairs, for some unspecified strategic reason. This is what I think they mean by a "strategic" account. But given that JOLs require a judgement of the relationship between cue and target, I struggle to see how people can avoid making this judgement prior to knowing the status of the pair. Moreover, it is unclear why they would adjust their strategy in this fashion. A second potential reason is that JOLs and relational processing are non-identical, and this difference is non-strategic in nature. This view is not considered.

***Response:*** We do not argue that JOL reactivity (and JAM/Frequency judgment reactivity) is non-strategic simply because we have no evidence that JOLs are non-strategic. The finding that reactivity effects emerge selectively for related by not unrelated pairs indicates a strategic process and our comparison to a non-strategic relational encoding provides additional evidence that participants are utilizing relational encoding processes more effectively for related pairs compared to unrelated pairs.  
  
**3: Strategy (part 2)**  
The authors justify their claim that JOLs might involve a strategic component by repeatedly citing Nelson and Narens' (1990) framework paper. But this paper is about the strategic nature of metacognition in the widest sense: it isn't evidence that there must therefore be a strategic component to judgements of item-based JOLs for related vs unrelated word pairs. That is, it conflates points 1 and 2 above. I entirely agree with the original claim that metacognition may involve strategic allocation of resources: I need convincing that the differential pattern of reactivity for related and unrelated pairs requires a strategic explanation.

***Response:***A strategic explanation is needed because related and unrelated pairs show different recall patterns in reactivity studies. Our finding that a non-strategic relational group produces a consistent memory benefit across pair types relative to JOLs which only benefit related pairs is evident that metacognitive processes are likely being implemented strategically  
  
**4: Confirmation bias**  
Experiment 1 essentially replicates what is known (with a minor extension), and sets up the following experiments. I found the experiments generally well-conceived, and interesting, but ultimately not-convincing, because they all sought to confirm the authors' hypothesis. That is, they compared JOLs with other means of boosting relational processing and concluded they because they both boost memory, they must be the same thing.  I concede that the findings are suggestive, and consistent, but the strongest test of the idea would be to try to refute the account. Effectively, the authors are basing their claim on a confirmation of the antecedent: The prediction is if A then B. B is then observed, but this does not imply that A caused it. A may have been caused by many things.

There is also an additional difficulty because the authors' proposal is that (a) that JOLs induce relational processing, but (b) this is strategically applied. But the strategic element makes the theory hard to test.  Both Experiments 2 and 3 show equivalent reactivity for JOLs and associative tasks (JAMs and Frequency judgements), supporting the first part of the theory (JOLs are associative). But in Experiment 4, judgements of the relationship produce a different pattern of reactivity for unrelated items. This is not taken as evidence against the associative nature of JOLs, but for the claim that it is strategic. To me, this is close to an untestable idea, unless there is independent evidence of a change in strategy.   
One way to test their idea is to actively seek to disconfirm it: for example, what would the authors predict with regards tasks that encourage participants to process the material in a non-associative fashion: e.g. to contrast JOLs with tasks in which participants judge which of the pair is the has the most vowels, is the largest, is easiest to imagine (and so forth). What conclusion would be drawn if these tasks also show the same reactivity pattern?

***Response:*** We appreciate your discussion here regarding confirmation, which is an important point for researchers to keep in mind. We refer you to our response to the Action Editor’s Comment 1, which addresses this issue specifically. In short, our initial submission already included a vowel task as you suggest here and we found that indeed, there was no improvement in correct recall in any of the pair types relative to the JOL group. The similarity in recall between JOLs and relational-type tasks but not a shallow item-specific task supports our argument regarding use of strategic relational encoding.  
  
**Reviewer #2: Summary:**  
  
Remaining Issues that Need to Be Addressed     
  
(1). The authors need to clarify their claims about strategy use. For example, on page 11 line 31, the authors suggest that participants are strategically processing related pairs. What considerations back up this claim? Are learners prioritizing those pairs or are they more automatically encoded? Is it strategic processing or just that related pairs engage deeper levels of processing? Said differently, is relational processing is a strategic process or an automatic process? This issue is addressed to some degree in Experiment 4 but the evidence that JOLs lead learners to engage in strategic processing of related pairs is not entirely. Rather, it seems like the relational processing is occurring simply a byproduct the required judgments.

***Response:*** Please see our response to Reviewer 1’s second point, in which we clarify our claims regarding strategy use. Regarding automatic encoding, if enhanced relational processing occurred as a byproduct making JOLs, one would expect to find relational processing globally acrosspair types rather it being limited to only relational pairs as is consistently observed. Moreover, just because something occurs as a byproduct does not negate strategy use. Related pairs may indeed more easily facilitate relational encoding relative to unrelated pairs, but participants are still selectively engaging in relational encoding as a function of pair type rather than applying relational encoding across all pair types.  
  
(2). The discussion of Koriat's (1997) cue utilization framework is useful, but the term "strategic" implies some sort of top-down implementation of intentional differences in encoding techniques depending on the relationship between two words. However, as evidenced by Experiment 4, it seems like JOLs are just cueing learners to incorporate relatedness as a cue into their metacognitive judgments and by doing so, making JOLs engages deeper levels of processing. It is not entirely convincing that participants were selectively engaging in relational processing for related pairs versus unrelated pairs as an encoding strategy to maximize total recall. This consideration is discussed as a limitation on page 40 but that does not resolve the issue.

***Response:*** Thank you for your thoughts. Yes, by definition, strategy use is a top-down process. In our experiments, even though participants were not told prior to study that they will study different word pairs, pair relations become immediately obvious, and thus participants quickly deploy a strategy to effectively process related pairs. Importantly, participants apply relational encoding selectively as supported by our Experiment 4 groups on unrelated pairs. Regarding whether participants use some other type of processing outside of relational processing, we cannot say based on our data alone. We home in on Hunt & Einstein’s classic framework for separating processing types to remain consistent with the literature but are careful to note that tasks are not process pure and we do not have online measures  
  
(3). In the Introduction, the authors account of the positive reactivity hypothesis does not actually offer much of an explanation as to how JOLs can enhance memory. Rather, they essentially say "monitoring is important for learning and memory benefits from additionally monitoring that occurs due to making JOLs and a memory improvement should be observed as a result." A more substantial discussion of why and how JOLs can enhance memory would be useful.

***Response:*** We’ve expanded our description of the positive JOL hypothesis on page 5 to include a brief discussion of how the positive JOL hypothesis predicts enhanced recall due to JOLs operating as a deep encoding task, which subsequently causes participants to engage more thoroughly with the study materials.

(4). On page 2 the authors refer to "online metamemory processes." What do they mean by online? Also, the authors refer to the JOL "task" which seems odd—that is, JOLs occur during some memory tasks but aren't the "task" itself. Specifically, JOLs can be applied to any kind of memory task, but throughout the manuscript the authors imply that this is the only type of paradigm where JOLs are solicited and can impact memory performance. A broader view of the use of JOLs in memory research would be useful throughout the paper and in the framing of conclusions.

***Response:*** The term “online” corresponds to metamemorial processes that are captured at the time of encoding, not through prospective or retrospective judgments.

We used the word “task” in the sense that the JOL was an activity participants partook in addition to encoding. Indeed, it is not uncommon for researchers to refer to providing judgments of learning as a task and, as our data suggest, providing JOLs can mimic some of the processes found with standard encoding tasks. Thus, we argue that the word task is completely appropriate given this is an activity that participants are engaged in.  
  
(5). The authors may want to also discuss the richness of encoding (where enhancement of memory occurs because people generate ideas about the to-be-remembered information which increases its number of retrieval cues) and dual mechanisms accounts of reactivity (positive and negative reactivity are based on different mechanisms; positive reactivity results as consequence of strengthened cue-target- associations while negative reactivity is due to dual-task interference) in the intro.

***Response:*** By richness of encoding, do you mean “depth of processing”? Under the “Mechanisms of JOL Reactivity” subheading, (page 5), we’ve updated our description of the positive reactivity hypothesis to reflect the notion that JOLs cause participants to engage more deeply with the study material, which in turn leads to reactivity (see our response to point 3).

(6). Experiment 1 materials section: can the authors elaborate on the "buffer pairs" and why they included them to "control for primacy and recency effects"? Would serial position effects differentially influence reactivity for the different pair types?

***Response:*** We included buffer pairs to control for any additional confounds related to presentation order. Given that serial order affects recall, we reasoned that controlling for these effects would allow for a more accurate assessment of factors that contribute to reactivity. We are unsure if serial position effects would emerge for these word pairs and what their relative magnitude would be, but we opted to control for them proactively.  
  
(7). The authors mention that study time is reported in the Supplemental Materials, but study time should be reported in the manuscript when it is an important consideration. On page 17 line 29, for example, the authors state that related pairs are prioritized at encoding leading to a boost in recall. Does the study time data support this claim? If the JOL group spent longer in the study phase, that would account, at least in part, for positive reactivity. The authors' provide a good discussion of this issue on page 39, but the relevant analyses should be presented in the manuscript.

***Response:*** We felt it best to relegate these analyses to the Supplemental Materials for two reasons. First, while useful for assessing memory, RT data only provides an indirect measure and can be rather noisy. This is compounded by the fact that our data collection occurred online (RTs will be contingent upon internet connections) and, importantly, by our use of concurrent rather than immediate JOLs in which we cannot parse between encoding duration and the time required to provide a JOL. While previous research has shown no difference in JOLs or recall between immediate and concurrent JOLs (see Maxwell & Huff, 2021), we note that the nature of this type of JOL makes it impossible to separate encoding latencies from time spent eliciting a judgment. Second, given the length of the manuscript, we felt it best to keep the RT data in the Supplemental Analyses for concision.  
  
(8). The authors should consider examining relative accuracy (see Rhodes, 2016) of the JOLs (perhaps collapsed across experiments) as an exploratory analysis. Doing so is not totally in line with their research question, but it would be interesting to see if  resolution differed for the different pair types.

***Response:*** While JOL accuracy was not a focal point of this study, we agree that resolution may be of interest to some readers. We have added a table in the Appendix (Table A5) which displays mean Goodman-Kruskal gammas as a measure of resolution. This table is now referenced in the Illusion of Competence analyses in the Appendix (page 65).  
  
(9). Results: there are values in parentheses in the results sections that must be means, but which should be labeled as such (i.e., M = xxx, SD = xxx). It would be nice to see standard deviations too.

***Response:*** We have added a sentence on page 15 regarding how to interpret values in parentheses. In our initial submission, we already included 95% CIs for each comparison (see Table A4). We believe this is a more streamlined presentation of these statistics rather than including them in text.   
  
(10). In the General Discussion, the authors could do a bit more to fit their findings to the different theories of reactivity. Most of the General Discussion summarizes the findings and addresses limitation but a succinct discussion of the theories of reactivity with which the current results are consistent and not consistent would be helpful.

***Response:*** Our findings in the present study primarily support Soderstrom et al.’s (2015) cue strengthening hypothesis. On page 36, we now discuss the extension of positive reactivity patterns to JAMs and Frequency judgments as further support of this account. Additionally, on page 39, we note that “the present study provides further support for the cue-strengthening account…”  
  
(11). Minor comments:  
  
***Response:*** All minor comments have been addressed. We appreciate your attention to detail.

**Reviewer #3:**

1)      The authors assume that the comparison tasks in Experiments 2 and 3 (judgments of associative memory (JAMs); judgments of frequency of co-occurrence) share with JOLs that they draw attention to word relatedness. First, this should be clarified from the start. Specifically, initially the authors explain rather broadly that in all tasks, participants provide ratings of word "pairs within the same context, which could encourage relational encoding" (p. 10) but become more specific over the course of the manuscript. It would help the reader if the authors pinned down the core similarities of these tasks from the start.

***Response:*** We’ve updated the descriptions of the JAM and frequency judgment tasks (pgs. 10-11) to emphasize the similarities between these tasks. We now mention that both tasks encourage participants to engage in relational encoding without overtly instructing participants to relate all items together. Additionally, both tasks are similar in that they remove the metacognitive component associated with JOLs.

Second, if all tasks implicitly emphasize relational encoding by drawing attention to word relatedness, JOLs, JAMs, and frequency judgments may be related to some degree. Do the studies provide some evidence for such relations, e.g., are these judgments correlated?

***Response:*** This is a good point and an interesting suggestion. We ran correlations between JOLs, JAMs, and frequency judgments. If each task is tapping into pair relatedness, we would expect that the measures would be highly correlated with each other. This is the case, as all correlations were significant and had *r*’s ≥ .94. We now report these correlations in the General Discussion and provide a brief discussion (pgs. 37-38).

Third, if different tasks have similar effects, this does not mean that their underlying processes are the same. Potentially, JAMs and explicit relational encoding instructions trigger more relational encoding than JOLs (see also p. 23), but JOLs compensate for this with memory forecasting, resulting in similar reactivity effects. The experiments nicely show that memory forecasting is not necessary for reactivity effects per se since JAMs, frequency judgments and explicit relational encoding all led to such effects. Yet, this does not necessarily imply that JOLs' reactivity effects are not due (in part) to memory forecasting. The authors should elaborate more on this point and tone down their conclusion.

***Response:*** Yes, while we agree with your assessment that producing the same recall pattern does not mean that the same process is being used, it is more likely that similar recall patterns are driven by similar processes than if the recall patterns were dissimilar. We acknowledge that it is possible that memory forecasting might still be in operation, however, we used JAMs and frequency judgments as these tasks very closely approximated JOLs, but without memory forecasts. We have now made this point clearer in our GD on page 37.

2)      In the introduction, the authors describe different accounts of processes that may underlie the reactivity effects of JOLs. Since the current research aims to investigate the mechanisms underlying reactivity effects of JOLs, it should be specified in the introduction (sections) how the experimental manipulations Experiments 2-4 relate to these accounts. The authors should clarify whether these experiments were designed to test any or more of the suggested accounts specifically.

***Response:*** Experiment 1 was designed test these competing accounts by replicating previous research showing a positive reactivity pattern for related pairs and no reactivity for unrelated pairs. We interpreted these findings as further support for the cue-strengthening account. Because the cue-strengthening account posits that reactivity occurs whenever cues used at encoding are available at retrieval, we reasoned that other, non-predictive tasks that emphasize these cues would also show this reactivity pattern. Therefore, Experiments 2-4 were primarily designed to provide further tests of the cue-strengthening account. We have updated the manuscript to clarify this, particularly in the Experiment 2 introduction (pgs. 19-20).

3)      It was not clear to me, why the changed-goal hypothesis would predict positive reactivity for related pairs and negative reactivity for unrelated pairs in the current experiments, without any time constraints. The authors refer to the study by Metcalfe and Kornell (2003) showing that unrelated pairs were prioritized during study when there were no time constraints (p. 5). Would this not predict positive reactivity for unrelated pairs in the present experiments?

***Response:*** As per Mitchum et al. (2016), the changed-goals hypothesis predicts that participants will prioritize encoding of related pairs at the expense of unrelated, resulting in a memory boost when pairs are related but a cost for when pairs are unrelated. While RTs were not our primary measure of encoding (we relied upon cued-recall performance instead given it is a more direct measure), we note that participants spent longer encoding related pairs in Experiments 1 and 2. However, our RT data should be interpreted with caution, considering 1) the online nature this study and 2) that JOLs were made concurrently with study rather than immediately following study. Therefore, the RTs reflect both time spent encoding *and* making a JOL, rather than a separate estimate of each. In the General Discussion, we now include a sentence on RTs being an indirect measure of memory (page 40).

4)      Given that JOL accuracy and illusion of competence played only a very minor role in the current research, the section on this topic in the introduction is overrepresented and should be shortened. The introduction is rather long and would benefit from a stronger focus on the background that is directly related to the following experiments.

***Response:*** We have reduced the length of two paragraphs in the introduction which discuss the illusion of competence (pg. 8).

5)      p. 19 (last paragraph): The authors should specify the pattern of results that would support their hypothesis that memory forecasting is not necessary for positive reactivity to occur.

***Response:*** We have updated this section accordingly (now on page 20, last paragraph)

6)      Exact (corrected) p-values should be reported for all analyses (potentially in the appendix, given their high number).

**Response:** For concision, we chose to report standard *p*-values but note that all comparisons held when using a Bonferroni correction (top of page 16). We feel that effect sizes are more useful when interpreting statistical patterns.

7)      The columns named CI 95% in Tables A3 and A4 do not contain any intervals but single numbers. The authors should ideally provide the intervals or explain how these numbers are to be translated into confidence intervals?

***Response:*** These columns contain the ± 95% CI (i.e., the value one would add or subtract from the mean to compute the 95% confidence interval). We have updated the column names to reflect this.

8)      p. 16 (first paragraph): The respective table seems to be Table A3, not A4.

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