**Action Editor**

First, please attend to reviewer 2's general comment about strategy use (this point is also echoed by Reviewer 1 though not as forcefully). Second, two of the reviewers are still concerned about the use of "silent reading" as a control condition to compare with JOLs.  Please address that as well. I also still agree with Reviewer 3's point ("To point 1") about toning down your conclusions based on the experiments because of the possibility of another factor causing the overlap between the different methods. I also think you should be a bit more clear that there may be multiple causes of reactivity and the differences between the studies you review may reflect that differential strength of different mechanisms that influence reactivity. Personally, I loved the introduction - the review of the mechanisms that may cause reactivity effects was clear and thorough.  But I also see the point Reviewer 3 makes about reducing the introduction and focusing a shorter paper on the results of the current study.  So, in your response, please indicate why you are either keeping the extensive review or if you decided to shorten the paper and focus it on the empirical data.

***Response:*** Thank you for your feedback. We have revised our manuscript based on the reviewer comments below while paying special attention to tone down our conclusions and note the possible existence of other potential causes of reactivity.

Regarding the introduction, we are glad you see the value of our comprehensive discussion of the literature. Based on your positive comments regarding its length and Reviewer 2’s comments regarding the importance of discussing the literature, we have elected to keep it intact. Although we can appreciate the importance of brevity, we feel our introduction provides important framing that is necessary to place our experiments within the context of the broader literature on JOL reactivity effects (please see our response to Reviewer 2).

**Reviewer 1:**

**Comment 1:** The authors have done an excellent job in responding to the previous round of editorial suggestions. While I remain a little uncertain of the use of the term "strategy", and the value of the theory, the authors have articulated their position much more clearly, and it is up to others to follow up this research if they wish.

***Response:*** Thank you. We are glad that our revisions allowed us to convey our position on strategy use more clearly.  
  
**Comment 2:** I have only one minor query that needs addressing. The authors report (on p.38) that for JOLs and JAMs "judgment values were highly correlated across tasks, rs > .94".  How were these correlations calculated? Were they the cross-item correlations per experiment? i.e. the JOL vs JAM given to each experimental item averaged across participants?  If so, then these correlation values may be inflated because the items were pre-selected to he high (related) and low (unrelated), rather than a continuous set of values across the range.  This would be like looking at the correlation between age and height in  samples of children preselected to be younger (aged 5-6) and older (aged 9-10). There would be a high correlation, but this would be due to the between-group differences rather than because within each age sample the correlation would be equally high.

***Response:*** Yes, the correlations initially reported on page 38 denoted cross-item correlations. Based on your suggestion, we have re-analyzed this data separately for related and unrelated pairs. Overall, correlations between each judgment type remained strong (related pairs, *r*s ≥ .65, *p*s < .001) or moderate (unrelated pairs, *r*s ≥ .41, *p*s < .001), and are still consistent with the pattern reported in our General Discussion regarding the similarities between the task types. We have updated the correlation section in the General Discussion (page 38) accordingly and now include scatterplots of these relationships in the Appendix (pages 67-68).

**Comment 3:** It would be more informative to show the plot of the correlation to see the extent to which JOLs and JAMs are associated for each of the item sets.

***Response:*** Agreed. Please our response to Comment #2.

**Reviewer 2:**  
  
**Comment 1:** In paragraph 1, the authors make it seem like there is little work on reactivity which is not clearly not the case based on all the work cited in the intro.

***Response:*** We disagree with this characterization of our introduction. In our first paragraph, we state that relative to studies investigating JOLs, “comparatively few studies have examined whether the act of providing metamemory judgments at study influences subsequent memory performance.” We simply mean that compared to the number of studies investigating various factors related to JOLs (i.e., manipulations that affect their accuracy), far fewer studies have focuses specifically on the effect of making these judgments on memory relative to a no-judgment control. Further, we make it clear in our introduction that the field has shown a recent interest in reactivity. On page 3, we write, “Recently, several studies have examined whether JOLs are reactive on learning.” Additionally, in our discussion of Experiment 1, reference this trend again, noting on page 18, “While the literature on JOL reactivity has recently experienced an increased focus…”

**Comment 2:** In the authors' response regarding the control group in reactivity studies, they state "a no-JOL control task like silent reading". I'm not sure exactly what they mean here. What I was trying to convey in my original review is that reactivity studies often compare a JOL group to a no-JOL group that controls for the time it takes to make a JOL such that participants making JOLs do not get more total study time (they could potentially be rehearsing words while making the JOL).

***Response:*** By “a no-JOL control task like silent reading,” we simply mean comparing participants in the JOL group to a group of participants who intentionally encoding study pairs without any specification of an encoding task to use (thus they are “silently reading” the pairs). This is a standard procedure used to assess JOL reactivity effects (e.g., Soderstrom et al., 2015; Janes et al., 2018). Further, though many researchers have controlled for encoding durations by equating study time between the JOL and no-JOL groups, we note that Janes et al. (2018) specifically investigated the effects of study-pacing (e.g., self-paced vs. experimenter paced) on reactivity. Reactivity only occurred when study was experimenter paced (which was matched between the JOL and no-JOL groups), however, we reliably found reactivity with participant-paced study suggesting that experimenter-paced vs. participant-paced study may not matter.

**Comment 3:** On page 5 the authors added that JOLs encourage participants to process the info more deeply than silent reading. I don't agree with that statement. JOLs do not encourage deeper levels of processing, even if that may sometimes occur. For example, while some JOLs may lead some learners to engage deeper levels of processing, other focus on physical characteristics of words like font size (shallow processing; e.g., Rhodes & Castel, 2008).

***Response:*** We agree that JOLs do not encourage participants to engage in deep encoding in all conditions, rather they apply it strategically as you suggest (“even if it may sometimes occur”). We have no evidence and are unaware of any evidence that indicates that JOLs would yield an individual difference in whether participants engage in deep process. Rather, we argue that deep encoding is applied at the item level.   
  
**Comment 4:** The introduction is still exceedingly thorough, perhaps to a fault. There are 11 pages of text before the first experiment and it reads more like a review paper than an introduction to empirical work, with the nuances and details of prior experiments discussed in great detail. I defer to the authors on the length of the paper, but my favorite papers are more parsimonious.

***Response:*** While we understand concerns about our manuscript’s length, we believe that our extensive review of the literature provides important context which allows this study to be properly framed within the existing literature on JOL reactivity. As a result, we have elected to keep it as is.

**Comment 5:** I still don't get why the authors summarize the results prior to the first experiment. It makes their discussion of their hypotheses in each subsequent experiment feel pointless. I know in my last review I said I defer to the authors here, but this doesn't seem like the best way to present their findings.

***Response:*** Given the complexity of the analyses, we provide this brief summary to help readers comprehend the results. We wanted to be clear regarding our hypotheses particularly regarding strategy use and believe this approach best supports the reader.  
  
**Comment 6:** The authors continue to claim that related pairs are prioritized at encoding (e.g., page 18 line 17) but I'm only seeing indirect evidence for this claim. More direct evidence could include spending more time studying those pairs

***Response:*** We previously addressed this comment by reporting response latencies for each encoding task in the Supplemental Analyses. However, as we also had noted in the General Discussion on page 40, caution should be used when interpreting RTs, given that data collection was conducted online and because judgments in the JOL, JAM, and Frequency tasks were made concurrent with study. Therefore, we are unable to separate encoding latencies from the time taken to provide a JOL rating. We have updated this section to convey this point more clearly.

Furthermore, we point out that encoding durations can be difficult to interpret. For instance, 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.

**Comment 7:** Relational processing was never properly defined. As such, I wonder if the memory benefit from "relational processing" of related words is not "relational processing" per se, but the more effective encoding strategies used when the words are related compared to when they are not related. For example, how are participants doing the JAM task not engaging in relational processing when the words are unrelated? They still have to think about the relationship between them, even though there is none.

***Response:*** On page 18, we define relational processing as encoding that occurs whenever “participants emphasize shared features or characteristics of a study set (Einstein & Hunt, 1980; Hunt & Einstein, 1981).” This definition was included in our initial submission.

Regarding unrelated pairs, although encoding tasks like JAM certainly have a relational undertone, the JAM task never explicitly directs participants to relate pairs together. Given that unrelated pairs show a memory improvement when encoded using a relational strategy (as in Experiment 4), the lack of reactivity on unrelated pairs for JOLs, JAMs, and Frequency judgments suggests that participants are not applying a relational encoding strategy when making these judgments on unrelated pairs.

**Comment 8:** Page 27 line 56: I don't see evidence that tests the claim that related pairs are receiving more processing than unrelated pairs.

***Response:*** We based this claim on recall rates. Relative to unrelated pairs, related consistently showed a memory improvement. Even backward pairs, which are difficult for participants to recall (see Koriat & Bjork, 2005; Maxwell & Huff, 2021), showed a memory improvement after being judged. Unrelated pairs, however, consistently showed no recall benefit in any of the three judgment groups.  
  
**Comment 10:** I'm still not sure the authors provide evidence to support their claims about strategically employing relational processing. For example, the authors assume that participants making JOLs are "choosing to use relational encoding on different subsets of pair types" (page 11) or that "relational encoding is applied selectively" when making JOLs (also page 39: "participants modify their study strategies based on pair type") but I'm not seeing evidence for that claim, only inferences.

***Response:*** The consistent finding that judgment participants are increasing their recall of related pairs and not unrelated pairs is good evidence that participants are prioritizing the relational characteristics of related pairs. Our comparison to a non-strategic explicit relational group, which was equal to the JOL group and also boosted unrelated recall is good evidence for the application of strategy. Yes, we acknowledge that we infer these processes, but what cognitive or memory process does not involve an inference process? Indeed, given cognitive processes are latent, researchers have no other option but to make inferences to make processing claims. Many strategic memory processes such as the distinctiveness heuristic, recall-to-reject, phantom recollection, and many others are based on inferences based on memory performance. Even measures used to infer memory strategies such as signal detection and the drift diffusion model are based on inferences. We therefore view criticisms of inferences as weak, given these are omnipresent in memory experiments who examine cognitive processes.

**Comment 11:** Rather than selectively applying a relational processing strategy when studying related pairs as a technique to enhance memory, it could be more of an incidental benefit of using those cues to inform their judgments. For example, if one was selectively applying relational processing, one would first have to evaluate the relatedness of the pair (similar to the relational encoding group in Experiment 4) before employing different encoding strategies based on the relatedness. Since the recall patterns  
between participants making JOLs and the relational encoding group diverge, this seems like evidence that the positive reactivity for related words is more incidental than strategic.

***Response:*** If reactivity simply reflected an incidental benefit due to cues informing judgments, it would be expected to occur for all pairs, regardless of relatedness. The observation that reactivity consistently is moderated by pair relatedness suggests that these processes are being applied selectively. Thus, the way individuals process JOLs involves relational encoding, because the task consistently benefits related pairs.

**Comment 12:** My understanding is that the novel findings are that 1) memorial forecasting is not required to observe reactivity and 2) that JOLs engage relational processing that benefits memory for related word pairs. These results provide a small advancement of the literature but the broader implications should be made more clear. Why should someone not studying reactivity care? Even a few sentences would be beneficial.

***Response:*** We appreciate this suggestion. Until recently, the assumption within the literature has been that having participants make metacognitive judgments does not influence memorial processes. Given recent work showing reactivity effects, however, this is not the case. We have updated the Conclusion (pages 42-43) addressing this.  
  
**Comment 13:** I think it should be made explicitly clear what two variables the new correlations on page 38 included.

***Response:*** The correlations reported on page 38 are between mean judgment values (JOL, JAM, and Frequency judgment ratings) at the item-level. We have updated this section to convey this more clearly.

Thank you for serving as a reviewer. We have appreciated your discussion of the points discussed in our manuscript.  
  
**Reviewer 3**

**Comment 1:** It is a valid argument that it is more parsimonious to assume that the similar reactivity patterns across JOL, JAM, and frequency judgment tasks result from similar underlying processes rather than different processes. And indeed, the almost perfect correlations across these judgments are compelling in this regard. Nevertheless, without a more fine-grained analysis of the underlying strategies, e.g., by online measures, as the authors suggest in the General Discussion and/or by experimentally discouraging relational encoding, the conclusion that reactivity effects of JOLs are due to relational encoding alone is too strong.

The vowel counting task used in Experiment 4 does not provide a sufficiently informative comparison in this regard as it seems to reduce encoding of the single items per se rather than just their relation. Therefore, especially the title and overall conclusion need to be toned down.

***Response:*** This is a fair criticism, and the use of online measures likely constitutes an avenue for future research on mechanisms driving JOL reactivity. Methods such as a “think-out-loud” protocol might be insightful in how participants are applying encoding strategies, provided the protocol itself is benign and does not produce reactivity (though this is debatable given simple JOLs produce reactivity patterns). Regarding vowel-counting, this task was not necessarily designed to *reduce* relational encoding. Instead, this task was selected because it was non-relational in nature, making it more akin to the no-JOL control group.

Given that present study does not include online measures and instead makes its claims based on comparisons to similar encoding tasks, we toned down our conclusions in the General Discussion. We have also amended the title to reflect this change.

**Comment 2:** I agree that adding corrected p-values for all post-hoc comparisons makes the paper less concise, and single p-values are less informative and convincing than a consistent pattern of results and effect sizes across experiments, as was found here. However, I don't quite comprehend how some of the reported, uncorrected t-values indicate significant results, especially after Bonferroni correction (see p. 16). As such, a few results need to be re-checked and if necessary, corrected. I will explain my thoughts:  
  
I assume the post-hoc tests were two-tailed tests, because the p-values reported are based on two-tailed tests, e.g., for the JOL vs. no-JOL comparison for backward pairs in Experiment 2 (p = .06; page 22). In Experiment 2 and 3, 3 tests were conducted for each word pair type: comparing JOL to no-JOL, JOL to JAM/frequency, and JAM/frequency to no-JOL. Bonferroni correction for 3 tests yields a corrected alpha level of .05/3 = .017. First, this makes me question that p = .06 (see above) suggests marginal significance. Second, I don't quite understand how the comparison between the JAM vs. no-JOL group for backward pairs in Experiment 2 was significant on the corrected alpha level, as t(63) = 2.11 results in an uncorrected p-value of p = .039. Similarly in Experiment 3 (p. 26): the smallest t-value of 1.96 (backward pairs) with 77 or 78 df (depending on the groups compared) is not significant on a two-tailed test (p = .054), even before correction. In noticed similar  
unclarities in Experiment 4 (p.32, symmetrical pairs, t = 2.06, uncorrected p = .042 or .043 depending on df).   
  
Of course, the number of tests for which the alpha level needs to be corrected depends on what is considered a family of tests for which the familywise error rate is controlled via Bonferroni correction. In the calculations above, I assumed that the multiple comparisons within each word pair type constitute a family. The corrected alpha level will differ when grouping tests into different families, e.g., by hypothesis, which may arguably be more adequate. The latter would likely result in 4 tests per hypothesis and an even lower corrected alpha-level, e.g., testing JOL vs. no-JOL for the 4 different word pair types (backward, forward, symmetrical, unrelated). Consequently, it should also be clarified for how many tests the alpha-level was corrected and why. The authors may also consider using an alternative correction method, given that Bonferroni correction has been shown to be quite conservative.

***Response:*** We appreciate your attention to detail regarding the post-hoc comparisons. Regarding the marginal *p*-value of .06 in Experiment 2, the value reported here is uncorrected. After applying the Bonferroni correction, this *p*-value increases to .17. Note, however, that on page 16 of our previous revision, we indicated that significant comparisons held after applying this correction. We were careful to make no claims regarding changes to marginal effects.

Second, we’ve corrected the results for the backward pairs comparison in Experiment 3 on page 26 to correctly reflect that the difference in recall between the JOL and no-JOL groups for backward pairs was marginally significant (*p* = .05).

Third, you are correct that the comparison between symmetrical pairs reported in Experiment 4 becomes non-significant once the Bonferroni correction is applied. Thank you for bringing this to our attention. Given this discrepancy (and to avoid any confusion regarding changes to marginal effects), we have elected to remove the sentence on page 16 stating that all significant comparisons held after applying the Bonferroni correction.