**Action Editor**

This is what I think is important to do.  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.   [REFERENCE JANES PAPER THAT REACTIVITY DISSAPPEARED IN SELF-PACED, BUT WE’RE SHOWING IT]

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.  [SEE WHAT WE CAN TONE DONE]

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.  [ADD THIS TO THE DISCUSSION]

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 regarding the introduction. While we understand Reviewer X’s concerns about its 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.

**Reviewer 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.  
  
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:*** The correlations originally reported on page xx denoted cross-item correlations that were averaged across participants. 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 pars, *r*s ≥ .41, *p*s < .001). We have updated the correlation section on page xx accordingly.

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:*** We appreciate this suggestion. We now include a set of plots depicting the correlations between JOLs, JAMs, and frequency judgments in the Appendix (pages xx-xx). Based on your previous suggestion, separate plots are constructed assessing the relationship between judgment values for related and unrelated pairs.

**Reviewer 2:**  
  
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:*** This is mischaracterization of our introduction. In our first paragraph, we clearly 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., situations affecting their accuracy), fewer studies have assessed their effects on memory. 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 pages 19-20, “While the literature on JOL reactivity has recently experienced an increased focus…”

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 mean comparing participants in the JOL group to a group of participants who do not make JOLs and instead silently read study pairs. [TALK JANES ET AL PAPER HERE]   
  
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: [SOMETHING HERE]   
  
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:*** Please see our response to the action editor.  
  
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:*** We provide this brief summary to assist the reader with conceptualizing the findings. By providing a bit of foreshadowing, the reader can more easily decompose the results.  
  
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

[REFERENCE SUPPLEMENT, ADDITIONALLY WE ADDRESS THIS IN THE GD “Indeed, several studies have found that memory is greater for deep tasks vs. shallow tasks even after controlling for encoding duration (e.g., generation, Slamecka & Graf, 1978; production, Icht, Mama, & Algom, 2014, etc.).”

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 xx, 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).”

Regarding unrelated pairs, [EXPAND]

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

[related pairs are being better remembered than unrelated pairs, even backward pairs (which are deceptive) are being remembered at a greater rate than unrelated pairs.  
  
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.

[This is cognitive psychology, we have to make inferences. This is a poor criticism because it applies to everything. A hallmark of memory is research is making inferences about processes that cannot be directly measured.]

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. [NO, IF IT WAS INCENDENTAL IT WOULD OCCUR FOR BOTH GROUPS, IF IT’S THE CASE THAT JOLS ARE INCENDENTALLY CAUSING RELATIONAL ENCODING, WE WOULD EXPECT IT TO BE APPLIED TO ALL PAIR TYPES]

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.

[WE ASSUME THAT JOLS OR OTHER TYPES OF METACOGNITIVE JUDGMENTS HAVE NO EFFECT ON METAMEMORY PROCESSES BUT CLEARLY THAT IS NOT THE CASE. – ADD TO THE CONCLUSION. THE WAY SUBJECTS PROCESS JOLS INVOLVES RELATIONAL ENCODING BECAUSE IT TENDS TO FAVOR RELATED PAIRS]  
  
I think it should be made explicitly clear what two variables the new correlations on page 38 included.

***Response:*** The correlations reported on page xx are between participant’s judgment values (JOL, JAM, and Frequency judgment ratings). We have updated this section to convey this more clearly.  
  
  
**Reviewer 3**

The authors have responded to most of my comments with very good revisions which have greatly improved the clarity of the manuscript. Nevertheless, there are two points, that need to be addressed again.  
  
To point 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. [TONE DOWN GD A BIT]

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. [DID THIS REVIEWER SUGGEST VOWEL COUNTING?]  
  
To point 6)  
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). [CHECK THESE, RERUN W/ LSD CORRECTION] [POSSIBLY REMOVE THE SENTENCE]  
  
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.