**Reviewer #1:**

This manuscript reports 4 experiments exploring the reactive effect of JOLs on subsequent cued recall of word-pairs. Experiment 1 (and all subsequent studies) replicated previous work showing that making JOLs improved subsequent cued recall for related pairs, but not unrelated pairs. Subsequent experiments showed that similar effects are observed if participants judged the associative strength of the cue-target pairs, judged their co-occurrence or if a relational processing task was induced at learning (relative to study alone). The authors conclude that there is no memory benefit from the memory forecasting aspect of JOLs, but instead the benefit arises from participants adapting more relational encoding of the cue-target pairs, relative to studying alone.

The authors are to be commended for their generally clear writing, and for having run a coherent set of experiments aimed at elucidating the basis for the reactive benefit of making judgements of learning. However, I am not yet convinced by the empirical contribution made by this set of studies, for the reasons I articulate below.   
  
**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.

**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.   
  
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 Soderstom et al (2015) saw no need to invoke strategic processes, what has changed?

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.   
  
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?

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.   
  
**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.   
  
**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?   
  
  
**Reviewer #2: Summary:**  
  
The authors examined reactivity for related (forward, backward, and "symmetrical") and unrelated word pairs. In Experiment 1, making JOLs led to better recall for related pairs but not unrelated pairs. In Experiment 2, judging the relatedness of the pairs led to similar reactivity as JOLs. Similarly, in Experiment 3, judging the frequency of co-occurrence resulted in similar reactivity as JOLs. In Experiment 4, participants either related pairs together, counted the vowels, made JOLs, or did not make JOLs during the encoding phase. JOLs and relational processing instructions led to reactivity for related pairs with relational processing instructions also leading to reactivity for unrelated pairs (relative to not making JOLs or counting vowels).   
  
**Evaluation:**  
  
The manuscript provides an excellent review of past research, is clearly written, and the results are analyzed thoroughly and appropriately. In addition, each experiment nicely sets up and builds off the previous one. The manuscript is somewhat long (some redundancy could be eliminated throughout and the General Discussion could be more concise).   
  
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.   
  
(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.   
  
(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.   
  
(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.   
  
(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.   
  
(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?  
  
(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.   
  
(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.   
  
(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.   
  
(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.   
  
(11). Minor comments:  
  
In paragraph 1, the authors should consider also citing Rhodes (2016), which is a nice review of work on JOLs.

We’ve added this citation. We appreciate the additional literature.  
  
On page 3 paragraph 1, it may be worth mentioning that reactivity studies usually control for study time (when fixed) for participants who do not make JOLs.   
  
On page 4 paragraph 2, can the authors explain what they mean by the negative reactivity pattern? In this paragraph, it feels like the authors are using negative reactivity and no reactivity as synonyms.

Added!  
  
In the last paragraph on page 7, the authors mention calibration as a measure of JOL accuracy. I think the authors should either define calibration (and probably resolution too) or just say accuracy here.   
  
At the bottom of page 10 the authors mention how "metacognitive processes are thought to operate strategically" but I think there is a lot of relevant work on this that could have been cited, such as agenda-based regulation (e.g., Ariel, 2013; Ariel & Dunlosky, 2013; Ariel et al., 2009). The authors mostly rely on the (seminal) Nelson and Narens (1990) paper, but there has been a lot of work over the last 30 years on strategic metacognitive processes.   
  
It seems odd to preview the results at the end of the Introduction, though perhaps that is the authors' prerogative. Maybe that could be saved for the discussion, especially given that the results are discussed in both the individual experiment discussions and the General Discussion.   
  
On page 18 the authors may want to avoid the sweeping claim that no other work has examined reactivity with judgments other than JOLs. There have been many different types of judgments used in memory tasks with these judgments compared to JOLs (e.g., judgments of importance; see Murphy & Castel, 2021). The authors may also want to note that reactivity does not exclusively involve more or less recall; reactivity can also refer to a change in what is remembered.   
  
When the authors introduce the shallow levels of processing group in Experiment 4, it would be useful if they would discuss how relational processing facilitates deeper levels of processing—associated with better memory—whereas shallow processing is associated with comparatively poorer memory. A relevant discussion occurs on page 36, but it would benefit the reader if moved to the Experiment 4 introduction.   
  
On page 28 line 28, the authors suggest that the vowel-counting task is not an explicit encoding task. Were those participants still aware of the memory test? Especially with 2 lists, it is difficult to see how how vowel counting should be considered incidental encoding. In several places where it is relevant, the authors should clarify the degree to which participants were told there would be a later test and of what kind.    
  
The authors' focus on whether learners might default to low JOLs for unrelated pairs seems unnecessary, so perhaps this section can be removed (along with the informal discussion of standard deviations and latencies).   
  
References  
  
Ariel, R. (2013). Learning what to learn: The effects of task experience on strategy shifts in the allocation of study time. Journal of Experimental Psychology: Learning, Memory, and Cognition, 39, 1697-1711.   
Ariel, R., & Dunlosky, J. (2013). When do learners shift from habitual to agenda-based processes when selecting items for study? Memory & Cognition, 41, 416-428.   
Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: when agendas override item-based monitoring. Journal of Experimental Psychology: General, 138, 432-447.   
Murphy, D. H., Agadzhanyan, K., Whatley, M. C., & Castel, A. D. (2021). Metacognition and fluid intelligence in value-directed remembering. Metacognition and Learning.   
Rhodes, M. G. (2016). Judgments of learning. In J. Dunlosky and S. K. Tauber (Eds.), The Oxford handbook of metamemory (pp. 65-80). New York: Oxford University Press.   
  
  
**Reviewer #3:**

In the manuscript "Reactivity from judgments of learning is not due to memory forecasting: Evidence from associative memory and frequency judgments", the authors report four experiments that investigate reactive effects of judgments of learning (JOLs) on cued-recall of word pairs. Across experiments, providing JOLs at study vs. mere encoding of pairs increased recall of related but not for unrelated word pairs. Similar positive reactivity effects for related pairs also emerged when participants performed other tasks at study that drew attention to word relatedness but did not require memory forecasting. Similar increases in recall were also obtained with an explicit relational encoding task. The authors conclude that (1) JOLs lead to positive reactivity by drawing attention to word relatedness, whereby relational encoding is strategically used for related pairs but not for unrelated pairs; and that (2) memory forecasting in JOLs is not required for this positive  
reactivity to occur.

The reported research is quite interesting. The manuscript is very well-written, clear, and mostly concise. The experiments are well-conducted, and the authors provided a balanced interpretation and discussions of their results. Nevertheless, clarifying and elaborating on the following points will improve an already good manuscript in a revision.

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.

We’ve updated the descriptions of the JAM and frequency judgment tasks on page 10 to emphasize the similarities between these tasks.

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?

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.

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.

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?

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.

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.

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

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?

The table notes for Tables A3 and A4 now contain a brief explanation for how to compute the interval.

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

The respective table here is Table A4 (the reactivity comparisons). We’ve updated the language to make this more apparent.