**Action Editor**

Thank you very much for submitting your work to Psychological Research. I have obtained two reviews from leading experts in your field of research. Both reviewers note strengths in your work, noting that it is a nice extension from prior work and the reported experiment is well designed. However, there are also sustainable issues with theoretical clarity and the methodology implemented. You can see the theoretical concerns raised by the reviewers, which could be addressed with re-framing your paper. However, as noted by the reviewers, and I agree, many of the methodological issues require follow-up experimentation to ensure the manipulation of the instructions was induced and to address concerns that findings and conclusions might not replicate in future research.  Thus, I can’t accept your paper in its current form but would re-evaluate a revised version f you can provide these additional experiments, and address the other concerns raised by the reviewers.

**Reviewer 1**

In the current study, the researchers investigated the effects of item-specific and item-relational processing on JOLs and actual memory performance for different types of word associations. The effects of item-specific and item-relational information on actual memory performance have been widely researched for a long time. Similarly, the types of word pair associations (e.g., related, identical, unrelated, vs) on actual and predicted memory have been widely discussed in literature over the last decade. However, how item-specific and relational information can be used concerning metacognitive measures such as JOLs is becoming more popular and are being primarily discussed in terms of reactivity effects. The researchers of the current study have taken a different spin on this topic and specifically investigated whether associations between words in different directions and strength can change calibration and resolution for these variables. Mostly, they are building upon their previous paper (Maxwell & Huff, 2021), examining the underlying mechanisms for their findings. The paper is clearly written and organized, specifically in terms of results. I also appreciate the researchers for attempting to gain deeper insight into previous findings. I believe that it is a good start; however, I have a few concerns in this paper, which keep me from recommending the publication of this manuscript in its current form. These concerns mainly relate to the theoretical motivation and the methodology of the study.

**Comment 1:** For the theoretical motivation of the study, I am not sure why the authors decided to use item-specific and relational encoding strategies to "change" JOLs. JOLs are usually considered inferential and based on intrinsic, extrinsic, and mnemonic cues (for a review, please see Koriat, 1997; Rhodes, 2016). These cues are helpful in making memory predictions to the extent that the cues are indicative of actual memory performance. If participants happen to focus on cues that will not help actual memory performance, they may not be able to discriminate between what will be remembered and forgotten, which will decrease resolution. The intrinsic cues are those cues that are inherent to words or word pairs, such as word frequency, concreteness, or the associative strength of the word pairs ( or whether they are backward forward or symmetrical). The study conditions for the list (e.g., how participants are processing the list) constitute the extrinsic cues that may affect JOLs. In the current study, the authors contend that using these extrinsic cues may potentially change the use of intrinsic cues in terms of both (1) calibrating one's memory performance (2) discriminating between what will be remembered and forgotten. The authors do not really make a real case as to why this should be the case, and the authors' predictions concerning encoding strategies are not precise. Their previous paper showed that different types of associations might allow for better calibration and resolution for JOLs, but the current paper falls short of making specific predictions as to why this might be the case.  This is one of my primary concerns for this paper because this specific concern does not allow me to understand how the results should be interpreted, why the calibration gets better, or the resolution gets worse. When I read the paper, I do not walk away with a principled understanding of why encoding strategies modify resolution and calibration for different types of word associations.

**Comment 2:** My second major concern is methodological: In the current study, the authors stated that they indued item-specific and item-relational processing strategies by giving participants instructions, but there is no manipulation check for this variable. Obviously, the instructions that they provide change the results, producing similar results for item-specific and item-relational encoding and different from pure reading instruction, but the experiment requires a manipulation check. This manipulation check can be a "think-aloud" procedure, or it can basically ask at the end of each trial whether the participant was able to use the strategy given to them. When the experimenters instruct participants to use a particular strategy, this does not mean that participants can and will encode the words using that strategy. More controls for these encoding strategies are essential to infer that the results are due to the strategy, not only thinking about the word pairs more deeply.

**Comment 3:** This brings me to another issue about the methodology. When the authors gave examples of item-specific vs. relational instructions, they used the cat-turtle word-pair as an example, which is a lovely example to process similarities or differences across word pairs, because both items are members of the animal category, allowing similarities or differences to emerge easily and naturally. However, when we have an example such as "credit-card" or "card-credit" (which was another example from the manuscript), I am not sure that this is conducive to the encoding strategy that they are mentioning. Again, having a "think-aloud" procedure or at least confirmation from the participants that they used a given strategy may take care of these concerns, but the study may have too many alternative interpretations in its current form. The actual memory performance may be simply better in the strategy conditions because participants were unsuccessful in using the assigned strategy, which may also lead to deeper processing than reading and create more traces for later retrieval of information. Either way, the current study will need at least one or two follow-up experiments to check for the efficiency of the manipulations.

**Reviewer 2**

This paper reports a single experiment that examined the effects of manipulating encoding strategies (silent reading vs. item-specific encoding vs. relational encoding) on the correspondence between the level of people's predictions of their future memory (judgments of learning, JOLs) and people's cued recall performance. Study materials were word pairs that differed in associative strength and direction (unrelated vs. symmetrically associated vs. forward associated vs. backward associated). Results showed that JOLs were largely unaffected by encoding strategy. In contrast, encoding strategy had marked effects on recall performance. Because of this, the usual pattern of overconfident JOLs for unrelated and backward pairs was reduced or eliminated in the item-specific and relational encoding conditions.   
  
Metacognitive illusions such as the illusion of competence addressed in this submission are of interest to many researchers in the fields of memory and metamemory. The reported experiment is well designed and the obtained results are clear. At the same time, I see several fairly major theoretical issues and also have concerns regarding the reported analyses and conclusions. I therefore cannot recommend this manuscript for publication, at least not in its current form.   
  
**Comment 1:** Usually, research on alleviating metamemory illusions focuses on whether people can be guided to make more accurate metamemory judgments. Prominent examples include Yan, Bjork, and Bjork's (2016) study on the interleaving illusion and, more relevant to the present submission, Koriat and Bjork's (2006a, 2006b) studies on mending the illusion of competence. The current study sets out to examine whether the illusion of competence can be reduced by item-specific encoding and/or relational encoding. While the results indeed demonstrate that deep encoding strategies alleviate the illusion of competence, this improvement in calibration is almost entirely due to changes in memory performance. So, I am wondering: What do we learn from this research? Does it tell us that changing people's memory is the most promising way to improve JOL accuracy?   
  
**Comment 2:** As a related point, I am concerned that the current findings and conclusions might not replicate in future research. The calibration of JOLs is often found to differ across experiments, even if they are very similar. For instance, in Koriat and Bjork's (2005) study, JOLs for forward-associated pairs were underconfident in Experiment 1 but well-calibrated in Experiments 2 and 3. It therefore would be essential to report some replication of the current findings in an additional experiment. I'd recommend that a potential new experiment aims to achieve considerably lower or higher memory performance so that the authors can test whether the changes in calibration they observed in the current experiment are specific to the obtained level of memory performance.  
  
**Comment 3:** The authors emphasize purported effects of word pair type and encoding condition on calibration plots. To me, all calibration plots seem to reveal very similar hard-easy-effects (= JOLs overestimate memory performance for hard items and underestimate memory performance for easy items) and differences between word types and conditions appear to be limited to the overall level of recall. Consistent with my interpretation, the reported a 3 (encoding group) x 4 (pair type) x 11 (JOL bin: 0, 1, …, 9, 10) ANOVA does not reveal significant interactions between encoding group and JOL bin. While my interpretation might be wrong, the critical point is that the reported analysis evaluating potential differences between calibration plots fails to provide conclusive evidence for the authors' interpretation (and against mine). I'd recommend that the authors replace the ANOVA including JOL bin with a multilevel regression analysis. In a regression model, they could easily test whether recall performance increases with JOLs (rather than only whether it differs in some way across JOL bins), whether the strength of a potential increase differs across word types and conditions, and it would also be possible to test for nonlinear relations between JOLs and recall (e.g., curvilinear relations). In addition, a regression analysis would be more powerful than the reported ANOVA and would circumvent any issues with violations to the assumption of sphericity (which appear likely but are not addressed in the paper). In summary, the authors' conclusions concerning calibration plots require more justification.  
  
**Comment 4:** I see some issues with the definition of and differentiation between calibration and resolution. First, these two aspects of JOL accuracy should be introduced much earlier than on Page 9 ff. so that readers understand right from the start that the illusion of competence impairs the calibration of JOLs rather than their resolution. Second, please note that testing and practice do not only affect JOL resolution but also affect JOL calibration. The current phrasing on P. 10 suggests that effects of these factors are limited to resolution. Third, the statement that "JOL calibration can be easily assessed by plotting mean JOL ratings against mean recall proportions, so long as JOLs and recall are measured using the same scale." (P. 9) seems to imply that calibration plots are the primary method used to examine calibration. I don't think that this is correct. In my view, most studies evaluating calibration have focused on bias (= the signed difference between mean JOLs and mean memory performance). What is more, effects seen in calibration plots are often hard to examine with inferential statistics, a problem that also plagues this work (see above). Finally, the current writing sounds a bit as if the current authors were the first to use calibration plots in combination with JOLs (P. 9-10: "These calibration plots allow researchers to assess whether JOLs are over or underconfident [see Maxwell & Huff, 2021]"). It should be mentioned that calibration plots have been previously used in the JOL literature, for instance by Koriat, Sheffer, and Ma'ayan (2002).  
  
**Comment 5:** The definition of a posteriori relatedness as "any perceived relatedness between pairs that becomes more apparent to participants when words are presented together" (P. 5) and, more specifically, categorizing "strong associates in which the pair order has been flipped (i.e., backward pairs such as card-credit, baby-stork, etc.)" (P. 5) as a posteriori pairs differs from the definition of a posteriori relatedness introduced by Koriat and Bjork (2005). Koriat and Bjork reserve the term a posteriori pairs for pairs with very low forward and backward associations. I'd recommend sticking to the established definition by Koriat and Bjork. Also, the current authors seem to make use of Koriat and Bjork's definition rather than their own one when stating "The illusion of competence pattern found with a posteriori and backward pairs has similarly been reported by Castel et al. (2007)" (P. 5).  
  
  
**Comment 6:** Separate 3 (encoding group) x 4 (pair type) ANOVAs for (1) JOLs and (2) recall should be reported prior to the 2 (measure: recall vs. JOL) x 3 (encoding group) x 4 (pair type) ANOVA. This would provide a better test whether differences in calibration across groups and pair types were due to changes in JOLs, in recall, or both.   
  
**Comment 7:** It is important to note that the reported 2 (measure: recall vs. JOL) x 3 (encoding group) x 4 (pair type) ANOVA evaluates calibration, because significant effects of or involving measure point to differences in the level of recall and JOL. It should therefore be reported under the heading "Calibration" rather than prior to this heading.   
  
**Comment 8:** The reported conclusion that there an illusion of competence is not always justified. For instance, the two-way interaction between measure and pair type (P. 17, Lines 17-19) in the 2 (measure: recall vs. JOL) x 3 (encoding group) x 4 (pair type) ANOVA is interpreted as evidence for an illusion of competence. However, this conclusion requires qualification given that there also was a significant three-way-interaction.  
  
**Comment 9:** Each reported gamma correlation needs to be tested against zero in order to evaluate whether resolution for the respective pair type and encoding condition is reliable. Also, any conclusion about effects of encoding group or pair type on resolution requires statistical tests. Given that no inferential statistics are reported in the submitted manuscript, there is no justification for conclusions such as "both item specific and relational encoding resulted in reduced resolution compared to silent reading" or "for unrelated pairs, resolution was increased for participants who completed item-specific and relational encoding tasks" (P. 21).  
  
**Comment 10:** In my view, it would be important that the authors provide better justification for their decision to aggregate data across blocks in the paper itself rather than in the Supplemental Materials. This would require mentioning whether there were any significant interactions involving block.  
  
  
**Minor Issues**  
  
- P. 6: Please note that there is evidence against the notion that identical pairs are more fluent than related or unrelated pairs (Mueller, Dunlosky, & Tauber, 2016)  
  
- P. 10: Given that most later studies (e.g., Mueller, Dunlosky, Tauber, & Rhodes 2014; Undorf & Zimdahl, 2019) found that the font size effect on JOLs demonstrated by (Rhodes & Castel, 2008) is NOT based on perceptual fluency, the following sentence requires revision: "Thus, factors that have been shown to directly influence the magnitude of JOLs such as associative direction (Koriat & Bjork, 2005; Maxwell & Huff, 2021) and perceptual fluency (Rhodes & Castel, 2008) would be expected to produce changes in calibration"  
  
- P. 19: "for items given an initial judgment" - I'd recommend rephrasing this sentence, because readers might wonder whether the initial judgment is different from the JOL or whether several JOLs were obtained.  
  
- P. 40, Table A3: Understanding the heading and the column is difficult.