January XX, 2023

Signy Sheldon, PhD

Action Editor

*Psychological Research*

Dear Dr. Sheldon,

My co-authors and I have submitted a revised version of PRPF-D-21-00324 “Item-Specific and Relational Encoding are Effective at Reducing the Illusion of Competence” for your consideration. We are encouraged our reviewers viewed our manuscript as “clearly written and well-organized” and that our experiment was viewed as “well designed.” Below, we list each reviewer’s comments and provide our responses. In addition to these responses, our primary revision includes a new experiment which replicates our first experiment using a “think aloud” procedure. Our reviewers questioned the validity of the item-specific and relational encoding tasks and requested a replication. The “think aloud” procedure confirmed that participants were using the directed encoding task and, critically, replicated the findings of our initial experiment.

We note here that our new experiment was collected in-person, like our initial experiment, and was completed in 2022 as COVID-19 research restrictions were slowly lifted at our institution. As a result, data collection was slower than anticipated which delayed our revision. Despite this challenge, we persisted with the replication/extension which shores up the concerns regarding methodology.

To facilitate review, all primary modifications to the manuscript have been made using blue font, and we include page numbers when referencing specific changes to the manuscript. We look forward to your response and hope that our revised manuscript is now suitable for publication in *Psychological Research*.

Sincerely,

Mark J. Huff, PhD

School of Psychology

The University of Southern Mississippi

Email: mark.huff@usm.edu

Phone: (601) 266-5411

Fax: (601) 266-5580

Cc:

Nicholas P. Maxwell, PhD

The University of Southern Mississippi

Emily Cates

The University of Southern Mississippi

**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 if you can provide these additional experiments, and address the other concerns raised by the reviewers.

***Response:*** We appreciate you and our reviewers pointing out this issue regarding participants’ adherence to the item-specific/relational instructions. Given these concerns, our revised manuscript now includes a second experiment (introduced on pg. xx) in which we combined the item-specific/relational encoding instructions from our original experiment (now Experiment 1) with the “think-aloud” procedure suggested by Reviewer 1 (see Reviewer 1, comment 2). In doing so, this allowed us to confirm that participants were correctly and consistently implementing their respective encoding tasks across all pair types at study while also providing an opportunity to replicate our originally reported patterns (See Reviewer 2, comment 2).

In addition to our inclusion of a second experiment designed to address each reviewer’s primary concerns surrounding the item-specific/relational encoding manipulations and replications, our revised manuscript additionally addresses Reviewer 2’s concerns regarding our definitions of calibration and resolution (comment 4) and have updated our analyses based on their suggestions (e.g., comment 7, comment 9).

**Reviewer 1**

**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.

***Response:*** Thank you for bringing these concerns to our attention. Yes, we agree that JOLs that are made based on the mnemonic cues that are available when JOLs are made. Because encoding processes greatly strengthen the quality of these cues, it is reasonable to examine the effects of encoding tasks on JOLs. The item-specific and relational framework is a classic approach for differentiating between two types of “deep” encoding processes that result in fundamentally different type of mnemonic cues. Therefore, our study not only examines how overall improved quality of cues may affect JOLs, but also whether focusing of differentiating vs. shared features might affect JOL ratings. As you note, our encoding-based approach differs from previous manipulations which have sought to improve JOL accuracy by lowering JOLs to more closely match recall (e.g., Koriat & Bjork 2006a, 2006b). Instead, we reasoned that if participants are better able to gauge the availability, they may be better able to adjust their JOL ratings in accord with future recall. We have updated the Introduction to convey our position more clearly (pg. xx) while also referencing the studies you mention. Additionally, please also see our response to Reviewer 2, comment 1.

**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.

***Response:*** As noted on pg. 15 of our initial submission, participants first completed a 10-trial practice set in which they applied their respective encoding strategies. Prior to beginning the full experiment, participants were required to inform the experimenter which strategy they were using. Additionally, following completion of each block, participants were again required to tell the experimenter which study strategy they had been instructed to use. We have clarified this point in the Experiment 1 Procedure section (pg. xx).

While we understand concerns regarding whether participants were consistently applying these strategies at encoding, previous research (e.g., Huff & Bodner, 2013; 2014) suggests that following a practice set, participants can correctly apply item-specific and relational encoding strategies to study items. Additionally, as you note above, our recall patterns for the item-specific and relational encoding groups suggest that participants were correctly applying these strategies. Based on the item-specific/relational framework (Einstein & Hunt, 1980; Hunt & Einstein, 1981), item-specific and relational encoding strategies are most effective at improving recall when they are used together. Because unrelated pairs naturally encourage participants to engage in item-specific encoding, relational encoding would be expected to be most effective for this pair type. Similarly, related pairs innately encourage relational encoding and thus should benefit most from item-specific encoding. Indeed, these patterns were observed in our initial submission, with relational encoding providing a greater benefit on unrelated pairs vs. item-specific encoding and item-specific encoding providing a greater benefit on related pairs vs. relational encoding.

Given the concerns with both participants’ encoding strategy use and replication, our revised manuscript now includes a second experiment in which participants complete a “think-aloud” procedure while employing their respective encoding tasks with an experimenter present to ensure the correct encoding task was applied (introduced on pg. xx). In doing so, this allowed us to ensure that 1) participants remained on task while applying their encoding strategy and 2) providing a replication. Overall, our original patterns reported in our initial submission replicated in Experiment 2, providing further confidence in the efficacy of these encoding tasks.

**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.

***Response:*** Please see our response to your second comment regarding our inclusion of a “think-aloud” procedure.

Concerning the example instructions, the pair “cat-turtle” was selected as the example provided to participants for the same reasons you mentioned above (namely that it makes for an easy demonstration when explaining how to apply each encoding strategy). However, given the sheer number of stimuli pairs as well as the different types, it is inevitable that some pairs would be easier to encode with a specific strategy relative to others. However, the think-aloud procedure introduced in Experiment 2 ensured that participants were consistently and correctly applying their assigned encoding strategy across all pair types, regardless of the ease with which the strategy could be implemented. Given Experiment 2 produced similar patterns as Experiment 1 (see the Experiment 2 results section on pg. xx), it is likely that the reported changes in memory performance between encoding groups in Experiment 1 were due to participants applying their strategy rather than deep encoding due to strategy failure, a point we now mention in our General Discussion (pg. xx).

We appreciate you taking the time to review our manuscript.

**Reviewer 2**

**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?

***Response:*** As you note, previous research has largely focused on strategies seeking to mend metacognitive illusions by lowering JOLs to match recall (e.g., Koriat & Bjork, 2006a; 2006b). However, given that JOL accuracy reflects both the magnitude of participants’ judgments as well as their later recall performance, focusing solely on manipulating JOLs to be in-line with recall only partially addresses the issue.

A more complete approach would be to use a strategy which modifies participants’ memory *and* theirmetacognitive evaluations. Given that encoding manipulations are commonly used to improve recall because they improve the quality of the cues generated at encoding (see Reviewer 1, Comment 1), we reasoned that that these manipulations provided a logical starting point. Additionally, because these strategies also strengthen cues used to inform JOL ratings, it was unclear whether these manipulations might similarly influence the magnitude of JOLs (see pg. xx of our revised manuscript). Thus, our finding that these manipulations reduce the illusion of competence by improving recall rather than changing JOLs provides a novel contribution, as to date, no study has directly compared the effects of these strategies on JOLs and recall within the same context.

**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.

***Response:*** This is a fair point. However, we note that the read group findings from our initial submission largely replicated patterns previously reported across four experiments by Maxwell and Huff (2021), particularly for backward and unrelated pairs. Similarly, though their focus was on JOL reactivity effects rather than JOL accuracy, Maxwell and Huff (2022) reported a series of analyses in their appendix which further replicated patterns reported in our read group. Importantly, Maxwell and Huff (2022) included each of our four pair types (forward, backward, symmetrical, and unrelated) and replicated this pattern across multiple experiments.

Given the concerns with replication, our revised manuscript now includes a follow-up experiment in which participants employed a think-aloud procedure while completing their respective encoding tasks (see our response to Reviewer 1, comment 2). As reported in the Experiment 2 results section (pg. xx), this additional experiment largely replicated patterns reported in our initial submission. Additionally, these changes do not appear to be limited to the level of memory reported in Experiment 1, as the think-aloud procedure in Experiment 2 consistently improved participants recall relative to Experiment 1. Thus, our inclusion of this replication provides further evidence that item-specific and relational strategies are effective at reducing the illusion of competence, largely through improved calibration due to improved recall.

**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.

***Response:*** The primary goal of our calibration plots was to provide qualitative information regarding changes in the JOL level at which the illusion of competence emerged for each pair type. We included the ANOVAs as a method to test whether the general patterns depicted in each plot differed based on pair types and encoding tasks. While we understand your concerns regarding sphericity violations, we have elected to keep these analyses in our revised submission, as the qualitative aspect provided by these plots is their primary importance. Furthermore, our finding that calibration plot patterns replicated in Experiment provides further support regarding these plots.

**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).

***Response:*** We have elected to keep our introduction structured as is, as we believe it is important to first discuss the general pattern of the illusion of competence (i.e., JOLs exceeding recall) and the pair types before shifting our discussion to potential factors that affect it (i.e., changes in calibration vs. resolution).

Regarding your second point, we have added a sentence on pg. xx noting that factors which improve resolution may similarly improve calibration by benefiting retrieval.

Third, you are correct that calibration plots are but one method for assessing JOL calibration. However, given our focus in the present paper on using calibration plots to visualize changes in recall for each JOL bin rather than assessing calibration through changes in bias, we have elected to primarily focus on the use of plots in this section. However, we now note on pg. xx that difference scores assessing bias can also be used as a measure of calibration.

Finally, we have clarified our language regarding calibration plots on pg. xx. In doing so, we now cite additional work using which as used this method to explore the correspondence between JOLs and recall on pg. xx, including the study by Koriat et al. (2002). Thank you for bringing this additional literature to our attention.

**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).

***Response:*** Koriat and Bjork (2005, pg. 188) defined a posteriori relatedness as referring to “the perceived relationship between the cue and the target when both are present, as is the case at the time of study, when JOLs are typically solicited.” Based on this definition, it is logical to conclude that backward associates constitute a specific type of a posteriori pair, as these items appear thematically similar when presented together. However, because they lack strong a priori relatedness, this perceived relatedness is not beneficial at test when the target is presented in the absence of the cue. We have updated the language on pg. xx accordingly to clarify this distinction.

**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.

***Response:*** We agree that separately testing JOLs and recall would provide a better assessment of whether changes in calibration were due to changes in JOLs or recall. Our initial submission included these analyses on pg. 19, and we have edited this section for clarity in our revision (pg. xx).

**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.

***Response:*** Yes, you are correct that this analysis is technically assessing changes in calibration, as it analyzes differences between mean JOLs and Recall for each pair type/group. Our “Calibration” heading in the initial submission was included to specifically report the calibration plots and their corresponding analyses. For clarity, we have updated the section heading on pg. xx to read “Calibration Plots.”  
  
**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.

***Response:*** We have updated our description of the two-way interaction on pg. xx accordingly. We now break down this interaction and report the appropriate means and post-hoc tests. Similarly, we followed this format when reporting our ANOVA results in Experiment 2 (pg. xx).

**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).

***Response:*** We now test each correlation against zero and report these results in Table xx (pg. xx). Additionally, we have updated the Experiment 1 results section on pg. xx to include inferential statistics for our gamma analysis. Specifically, we now report output from a 3 × 4 mixed measure ANOVA as well as corresponding post-hoc *t*-tests when appropriate. Similarly, these additional analyses are reported for the gamma results in Experiment 2 (pg. xx).

**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.

***Response:*** As noted in the supplemental analyses, significant three-way interactions between block, measure, and encoding group and block, measure, and pair type were reported. However, changes in the illusion of competence across block largely reflected changes in recall from block 1 to block 2, with recall sometimes increasing across blocks (e.g., item-specific symmetrical pairs) and other times decreasing (e.g., read-only unrelated pairs). Given the inconsistencies with these patterns, as well as the lack of Block × Encoding Group × Pair Type × Measure interaction, we elected to report these analyses in the supplement for concision within the manuscript.

Additionally, given concerns with potential block effects in Experiment 1 and experiment duration due to the think-aloud procedure, Experiment 2 used a one-block design.

**Minor Issues**

***Response:*** All minor spelling and grammar issues have been corrected. We appreciate your attention to detail.