Reactivity from Judgments of Learning are not due to Judgments of Learning:

Evidence for a Strategic Relational Encoding Account

Nicholas P. Maxwell & Mark J. Huff

The University of Southern Mississippi

Author Note

Correspondence concerning this article should be addressed to Nicholas P. Maxwell, 118 College Dr, Hattiesburg, MS, 39406. E-mail: [nicholas.maxwell@usm.edu](mailto:nicholas.maxwell@usm.edu). *R* code used for data screening and analyses as well as all applicable stimuli and data files have been made available on our OSF page (https://osf.io/8yvn3/).

Abstract

Previous research has shown that judgments of learning (JOLs) often produce a reactive effect on learning of cue-target pairs in which target recall differs between participants who provide JOLs at study versus those who do not. Positive reactivity, or the memory improvement found when JOLs are provided, is typically observed on related pairs, whereas negative or no reactivity are typically found on unrelated pairs. In three experiments, the present study further examines JOL reactivity effects by comparing JOL and no-JOL groups to other groups who engaged in relational encoding tasks. Experiment 1 replicated positive JOL reactivity effects with related pairs with an extension to symmetrically related pairs. In Experiment 2, a similar positive reactivity pattern was found using a relational encoding task when compared to a standard JOL task. Next, Experiment 3 found that providing judgments of associative memory (JAM)—a task that does not require memory predictions—yielded equivalent reactivity patterns to providing JOLs. Finally, Experiment 4 extended this pattern using frequency judgments. Collectively, our results suggest that previous JOL reactivity patterns are not due to memory forecasting processes found with JOLs, but instead reflect relational encoding that is strategically directed towards related, but not unrelated, pairs.

Word count: 198

*Keywords:* Metamemory; Judgments of Learning; Relational Encoding; Associative Direction; Cued-Recall

Reactivity from Judgments of Learning are not due to Judgments of Learning:

Evidence for a Strategic Relational Encoding Account

An individual’s to ability to accurately monitor the progress of their own learning is a critical component for successful retention. Effective monitoring allows individuals to adjust their study strategies to maximize memory performance (Nelson & Narens, 1990) and provides insights on how best to allocate memorial resources to optimize learning (Soderstrom, Clark, Halamish, & Bjork, 2015; see also Bjork, 1999 for a review). Empirically, information about learning processes can be obtained through metacognitive judgments. Though these tasks have received significant attention from memory researchers (see Bjork, 2016; Metcalfe, 2000, for a historical overview of metacognitive judgments within the field of metamemory), relatively little research has been conducted to assess whether the simple act of providing these judgments at study can affect memory performance and, if so, the memory processes that may contribute to this effect.

A common type of judgment used to assess online metamemorial processes is the judgment of learning (JOL) task. In a standard JOL task, participants are presented with a cue-target study pair (traditionally paired associates) and are asked to respond with the likelihood that they would respond with the correct target at test if provided with only the cue. While these judgments can be made using a variety of scales (e.g., Likert scales or binary “yes”-“no” responses; Hanczakowski, Zawadzka, Pasek, & Higham, 2013), JOLs are typically elicited using a continuous 0 to 100 scale that represents the percent likelihood that the cue-target pair would be successfully recalled at test (e.g., 100% = definitely would remember; 0% = definitely would not remember). The use of a 100-point scale allows for a comparison between predicted recall (via JOLs) and the overall proportion of items later recalled at test.

Recently, several studies have examined whether providing JOLs at study is *reactive*. JOL reactivity refers to changes in memory performance due to providing JOLs at encoding. A simple way to assess whether JOLs produce a reactive effect on learning is to compare recall performance for participants who complete a JOL task at study to those who do not (Soderstrom et al., 2015; Janes, Rivers, & Dunlosky, 2018). Reactive effects could produce a memory benefit (a pattern termed *positive reactivity*) or a memory cost (a pattern termed *negative reactivity*) relative to a control task that does not provide JOLs. Thus, evaluating reactivity simply involves the inclusion of a no-JOL control group, a comparison that is generally absent in JOL studies as researchers have been more interested in condition-specific effects on JOLs themselves (and not overall memory performance), or have made the assumption that the act of providing JOLs at study has no impact later memory. However, given that no-JOL control groups are often absent, this assumption cannot be confirmed.

The lack of no-JOL controls across studies is surprising given early evidence for the reactive effects of JOLs on memory were documented in Arbuckle and Cuddy’s (1969) experiments. In one experiment, metacognitive judgments were elicited using a 1-5 Likert scale, and importantly, participants provided metamemory judgements either during both study and test phases, or only at test. Judgments at study were framed as a JOL (i.e., the predicted ability to correctly recall the target item at test if prompted by the cue), while judgments made at retrieval were elicited as a confidence rating (i.e., confidence that the memory response was correct). This design allowed for a comparison between groups in which metacognitive judgments were provided at both study and test versus a group that only made judgements at test (i.e., a no-JOL control). A positive reactivity pattern emerged in which correct recall was greater for participants who provided judgements at encoding. However, it is important to note that although Arbuckle and Cuddy reported that JOLs can boost recall, participants in both the JOL and no-JOL groups provided confidence ratings at test, making it unclear whether confidence ratings were a requisite for positive reactivity.

More recently, Soderstrom et al. (2015) had participants study a list of cue-target pairs which contained both related and unrelated pairs. After studying each pair, one group of participants were instructed to provide JOLs, while a no-JOL control group studied each pair in isolation. Participants were then tested on their recall of the target word when presented with the cue without additional metacognitive judgments made at retrieval (cf. Arbuckle & Cuddy, 1969). Overall, target recall was greater for participants who provided JOLs initially versus those who did not; however, this positive reactivity pattern was restricted to related pairs. For unrelated pairs, target recall did not differ between the JOL and no-JOL groups. A similar pattern was reported by Janes et al. (2018), who also showed that initial JOLs produced positive reactivity for targets from related but not unrelated pairs. Furthermore, Witherby and Tauber (2017) found evidence for positive reactivity on related pairs after a 48-hour retention interval, providing evidence for positive reactivity after a delay.

In contrast to the positive reactivity for JOLs associated with related pairs as reported by Soderstrom et al. (2015) and Janes et al. (2018), Mitchum, Kelly, and Fox (2016) reported a divergent pattern of reactivity. In their study, participants who provided JOLs at study showed no difference in later recall relative to a no-JOL group on related pairs and produced a negative reactivity pattern relative to the no-JOL group for unrelated pairs. Mitchum et al. interpreted this discrepancy as arising from methodological differences between their study and Soderstrom et al., such as differences in experimenter-paced study and the inclusion of a generation task in their second experiment. However, in a subsequent experiment that used experimenter paced study, Mitchum et al. again reported no evidence for positive reactivity on related pairs and negative reactivity on unrelated pairs. Taken together, these studies demonstrate that providing JOLs at study can induce reactivity on target learning, but the direction of the resulting reactivity is mixed, with positive or no reactivity reported when pairs are related and negative or no reactivity reported with unrelated pairs.

**Mechanisms of JOL Reactivity**

Three mechanisms have been proposed to account for JOL reactivity (see Mitchum et al., 2016). First, the *positive reactivity hypothesis* states that given monitoring is essential for determining the effectiveness of the learning process (e.g., Nelson & Narens, 1990), retention will benefit from any additional monitoring that occurs as a byproduct of providing JOLs at encoding. Because JOLs are provided for all pairs at study, a global memory improvement should occur across study materials relative to a non-JOL control. Next, the *dual-task* *hypothesis* suggests that generating JOLs at encoding will produce negative reactivity across study materials versus a no-JOL control, as providing JOLs is resource demanding and may interfere with the learning of word pairs (Hertzog, Dunlosky, Powell-Moman & Kidder, 2002). Finally, the *changed-goal hypothesis* proposes that JOL reactivity occurs due to online changes in participant study goals that arise during encoding. According to this hypothesis, participants set an initial goal of memory mastery and strategically allocate more encoding time and/or effort towards studying items perceived as challenging to remember relative to those perceived as easy. However, certain conditions may induce a change of study goal in which easier items are prioritized. For example, Metcalfe & Kornell (2003) presented participants with English-Spanish vocabulary pairs and found that when study time was limited, participants prioritized learning of pairs perceived as “easy” due to a shared root word (i.e., cognate pairs such as *park* - *parque*) versus more difficult pairs that did not contain the same root word (i.e., non-cognate pairs such as *dog – perro*).When providing JOLs (specifically those utilizing a 0-100 rating scale), it becomes clear to participants that not all items will be equally recalled. Thus, participants may use their perceptions of item difficulty when completing a JOL task to shift their study goals towards mastering easier items.

Within the context of JOL reactivity effects on word pairs, the changed-goal hypothesis assumes that study lists will provide participants with at least two distinct pair types and that participants will be able to discern them. This hypothesis predicts that providing JOLs will induce positive reactivity for pairs that are perceived as easy to remember, but negative reactivity for pairs that are perceived as difficult to remember. This is because when individuals detect differences in difficulty between pair types, they will prioritize encoding of the easier to remember related pairs at a cost of encoding more difficult unrelated pairs. Thus, for related and unrelated pairs, the changed-goal hypothesis predicts a divergent memory pattern when comparing JOL to non-JOL group due to participant perceptions of pair difficulty.

Although JOL reactivity patterns based on pair association have been mixed (e.g., Janes et al., 2018; Mitchum et al., 2016; Soderstrom et al., 2015), a meta-analysis conducted by Double, Birney, and Walker (2018) which included 17 published and non-published experiments comparing JOL to non-JOL groups provided no support for the positive reactivity and dual-task hypotheses and showed only partial support for the changed-goal hypothesis. Specifically, providing JOLs yielded a positivity effect for related target recall, but showed no reactive effect on recall of unrelated targets relative to no-JOL controls. In terms of the changed-goal hypothesis, it therefore appears 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.

**Associative Direction and JOL Accuracy**

The associative direction between related word pairs has been shown to directly influence both how well individuals recall items at test and the accuracy of JOLs made at study. Koriat and Bjork (2005; see too Koriat & Bjork, 2006) demonstrated that across three experiments, JOLs for pairs associated in the forward direction (e.g., credit-card) were accurate at predicting later recall of the target item. When forward association strength between pairs was weak (e.g., article-newspaper), JOLs were less predictive of later recall relative to when the forward association between pairs was strong (e.g., lost-found). For weak forward pairs, JOLs ratings were similar to those given to strong associates, but recall was reduced as weakly related cues were less effective in aiding retrieval of the target. Thus, the *calibration* between JOLs and recall was moderated by the strength of the forward cue-target association.

In addition to forward associates, Koriat and Bjork (2005; Experiment 2) also evaluated the correspondence between JOLs and target recall for pairs associated in the backward direction (e.g., card-credit). Like weak associates, backward associates received high JOL ratings, but recall for the target word was considerably lower relative to forward pairs. Dubbed the *illusion of competence,* this overestimationpattern has been extended to other pair types. Castel et al. (2007) showed that the illusion of competence extended to identical pairs in which the cue is perfectly predictive of the target (e.g., lost-lost). More recently, Maxwell and Huff (in press) showed that the illusion of competence holds for backward associates after controlling for lexical and semantic properties of the cue and target (e.g., word length, concreteness, etc.) and extends to symmetrical associates (e.g., off-on). Thus, the direction of association more than the associative strength, contributes to the illusion of competence.

The illusion of competence serves as an example of how the directional correspondence between related pairs can affect the predictive capacity of JOLs on later recall. Regarding JOL reactivity, most studies use forward associate pairs in which the cue is highly predictive of the target. In a notable exception, Mitchum et al. (2016, Experiment 1), compared target recall using forward associates, backward associates, and unrelated pairs that were presented within the same study list. Study latencies were also measured. As reported above, no reactivity was found for either backward or related pairs. Yet, despite this null pattern, the authors concluded that the changed-goal hypothesis was partially supported as JOL participants spent less time studying unrelated pairs relative to participants who did not make JOL ratings. This pattern suggests that participants were prioritizing memorization of related pairs by allocating additional study time.

Although Mitchum et al. (2016) showed reactivity results inconsistent with other JOL reactivity studies (e.g., Janes et al., 2018; Soderstrom et al., 2015), it is also worth pointing out another inconsistency in their data—no illusion of competence pattern emerged for backward pairs (cf. Castel et al., 2007; Koriat & Bjork, 2005; Maxwell & Huff, in press). Though Mitchum et al. reported reduced recall rates for backward than forward pairs across JOL and non-JOL groups, these differences were much smaller than those typically reported. This discrepancy may have resulted from how association was measured across these studies. Koriat and Bjork (2005) for instance used Hebrew word pairs derived from a set of Hebrew free association norms, while Mitchum et al. used English word pairs derived from the University of South Florida Free Association Norms (USF norms; Nelson, McEvoy, & Schreiber, 2004) as well as a relatedness score calculated with Latent Semantic Analysis (LSA; Landauer & Dumais, 1997). Maxwell and Huff (in press) similarly utilized the USF norms as in Mitchum et al. and used pairs that were identical in associative strength (.37 in both studies). However, a robust illusion of competence pattern was found.

A second possibility for this discrepancy is that while association between pair types was assessed and manipulated, neither Koriat and Bjork (2005) nor Mitchum et al. (2016) controlled for lexical and semantic item characteristics in cues and targets that may have covaried across pair types. Characteristics such as word length, frequency, and concreteness have each been shown to affect later recall (Balota & Neely, 1980; Madan, Glaholt, & Caplan, 2010; Criss, Aue, & Smith, 2011) and could be confounded with associative direction in these studies. Thus, given discrepancies in recall that occur as a result of pair direction (i.e., the illusion of competence), it remains unclear whether pair direction could moderate JOL reactivity (i.e., reactivity greater for forward vs. backward pairs).

The goal of the present study was therefore to examine pair associations as a means of testing potential mechanisms that contribute to JOL reactivity. First, Experiment 1 was designed to provide a replication of JOL reactivity patterns reported by Janes et al. (2018) and Soderstrom et al. (2015) to further test the reliability of positive reactivity for related pairs and no reactivity for unrelated pairs while controlling for lexical and semantic characteristics of cues and targets. Additionally, we compared reactivity effects on four different pair types including three types of related pairs (forward, backward, and symmetrical) and unrelated pairs. Next, given that previous research has shown JOL reactivity to be contingent upon pair association, Experiment 2 tested a novel strategic relational encoding account of reactivity. Briefly, the strategic relational encoding account posits that when participants are exposed to both related and unrelated pairs, they strategically emphasize processing of relational characteristics of related (but not unrelated) pairs, leading to their greater recall. To test this account, we compared target recall in JOL and no-JOL groups relative a relational encoding group where participants were explicitly instructed to relate study pairs together. We further compared these groups relative to a shallow levels-of-processing task as another control group. Experiments 3 and 4 then provided stronger tests of the strategic relational encoding account by comparing recall for JOL and no-JOL groups to the judgment of associative memory task (JAM; Experiment 3) and a frequency-rating task (Experiment 4), both of which were designed to mimic the relational processing participants engage in when making JOLs while removing the predictive component associated with JOLs. Thus, Experiments 3 and 4 also evaluated whether JOL reactivity effects are due to the memorial forecasting that is required when providing a JOL, or due to rating cue-target pairs within the same context which encouraged relational encoding. To preview, across experiments, we found reliable positive JOL reactivity for all three related pair types, consistent with the general pattern in the literature (cf. Double et al., 2018), and found that the benefit to related pairs is equivalent to related pairs that are studied using an explicit relational encoding task, a finding consistent with a strategical relational encoding account. We then show that both JAMs and frequency judgments also elicit similar patterns of reactivity as JOLs. Collectively, our experiments reveal that reactivity patterns are not unique to JOLs and reflect strategic use of relational encoding directed towards related pairs.

**Experiment 1: JOL Reactivity on Related and Unrelated Pairs**

In Experiment 1, we first sought to replicate and extend previous JOL reactivity patterns by comparing target recall following study of related and unrelated pairs. The changed-goal hypothesis predicts that JOL reactivity should produce a benefit to related pairs and a cost to unrelated pairs as participants shift their study goals to prioritize the easier related pairs over unrelated pairs. However, given that prior studies only find partial support for the changed-goal hypothesis with positive reactivity for related pairs and no effect on unrelated pairs (Double et al., 2018; Janes et al., 2018; Soderstrom et al., 2015), we expected that our experiment would yield this same pattern. Thus, Experiment 1 served as an additional test of positive JOL reactivity for related pairs.

An additional goal of our experiment was to evaluate positive reactivity effects on different types of related pairs. To this end, we directly compared forward and backward pairs, but we also included symmetrical pairs—a related pair type that has not been tested in reactivity experiments. We expected that positive reactivity would be found across all three related pair types despite differences in recall rates that have been shown across these pair types (Maxwell & Huff, in press). Importantly, we also controlled for potential lexical and semantic item effects that were not equated for across pair types in previous studies (e.g., Soderstrom et al., 2015; Janes et al., 2018). All related and unrelated pairs were matched on word frequency, concreteness, and length and related pairs were further matched on associative strength. Thus, Experiment 1 provides a more precise test of JOL reactivity patterns while controlling for important lexical and semantic item effects.

**Method**

**Participants**

Seventy-eight participants were recruited online through Prolific (www.prolific.co) and were compensated at a rate of $8.00/hour. Participants were randomly assigned to either the JOL or no-JOL group (39 per group). A sensitivity analysis conducted with G\*Power 3 (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that this sample size provided adequate power (0.80) to detect medium-sized main effects/interactions (Cohen’s *d* = 0.50) or larger. All participants were native English speakers with normal or corrected-to-normal vision.

**Materials**

Stimuli pairs were taken from Maxwell and Huff (in press). These pairs consisted of 180-word pairs generated from the University of Florida Free Association Norms (Nelson et al., 2004). Pairs were split into four types consisting of 40 forward pairs (e.g., credit-card), 40 backward pairs (e.g., card-credit), 40 symmetrical pairs in which forward and backward strength were equivalent (e.g., ball-bounce), and 40 unrelated pairs (e.g., artery-bronze). Additionally, 20 non-tested buffer pairs were generated to control for primacy and recency effects. Item pairs were distributed across two study lists of 90 items which were used in two separate study/test blocks. Thus, each list contained 20 items of each of the four pair types and 10 buffer items. All pairs are available at https://osf.io/8yvn3/.

Study lists were created such that the 80 tested pairs were always proceeded and followed by five buffer pairs to reduce primacy and recency effects. Additionally, lists were constructed such that pair types were equated on frequency (SUBTLEX; Brysbaert & New, 2009), word length, and concreteness (from the English Lexicon Project; Balota et al., 2007) and related pair types were further equated associative strength (e.g., FAS and BAS values derived from the Nelson et al. (2004) free association norms; See Tables A1-A2 in the Appendix for associative strength and lexical properties for each pair type). Finally, counterbalanced versions of each study list were created that flipped the order of words with each of the four pair types (i.e., king-queen becomes queen-king). While the order was switched across all pair types, this was especially impactful for forward and backward pair types given forward pairs were transformed to backward pairs, making these pair types perfect controls. The cued-recall test was generated from all 80 cue items (excluding buffers) by replacing the target item with a question mark (i.e., credit - ?). Test items were presented in a newly randomized order for each participant.

**Procedure**

Data collection was conducted online using *Collector*, an open-source program for presenting web-based psychological experiments (Garcia & Kornell, 2015). In both the JOL and No-JOL groups, participants were instructed that they would view a series of cue-target word pairs and that their memory for the target item would be tested. Participants in the JOL group received further instruction to rate the likelihood that they would be able to remember the target word if shown only the cue at test. Judgments were elicited using a scale of 0-100, in which 0 indicated that they would be completely unable to recall the item at test, while a rating of 100 represented full certainty in their ability to correctly recall the target. After receiving instructions, participants began the first study list. Study was self-paced, with both groups pressing the Enter key to advance to the next pair. Additionally, participants in the JOL group were asked to type a JOL rating before advancing to the next study pair. JOL ratings were provided concurrently with study such that ratings were typed while the pair was displayed.

Following presentation of the first study list, participants completed a two-minute filler task in which they were asked to list the 50 U.S. states in alphabetical order. This was immediately followed by a cued-recall test that presented participants with the cue word from each of the previously studied items. Participants were asked to type the correct target item. If participants could not retrieve the correct item, the Enter key could be pressed to advance to the next pair. Following the first cued-recall test, participants began the second block, which followed the format of the first block. Participants were fully debriefed following the completion of the second cued-recall test. Each experimental session lasted approximately 30 minutes.

**Results**

A *p* < .05 significance level was used for all analyses. Partial eta-squared (*ηp*2)and Cohen’s *d* effect sizes are reported for all significant analyses of variance (ANOVAs) and *t*-tests. For all *t*-tests, we report standard test statistics, but note that all comparisons hold when using a Bonferroni correction. Additionally, for all non-significant main effects and post-hoc comparisons, we report a Bayesian estimate of the strength of the evidence supporting the null hypothesis (Masson, 2011; Wagenmakers, 2007). This analysis compares two models, one in which a significant effect is assumed, and one that assumes a null effect. From this analysis, a probability estimate is generated, a *p*-value termed *p*BIC (Bayesian Information Criterion), which estimates the probability that the null hypothesis is retained. This estimate is sensitive to the sample size, providing increased confidence in null effects reported. Finally, while not imperative to the research question at hand, encoding durations for all experimental groups as a function of pair type and group are reported in our Supplemental Materials with data made available on our OSF page (https://osf.io/8yvn3/).

Figure 1 (top panel) plots mean JOL ratings and cued-recall rates for each pair type for participants in the JOL study group, while the bottom panel compares recall rates for participants who made JOLs at study versus those who silently read pairs at study. A liberal scoring criterion was adopted for recall such that misspellings and grammatical errors (i.e., changes in tense) were counted as correct. For completeness, all comparisons between of JOL ratings and correct recall proportions for each pair type are displayed in Appendix Table A3, and all comparisons between correct recall proportions for JOL and no-JOL groups are reported in Table A4. All analyses have been collapsed across block order1.

First, we conducted a 4 (Pair Type: Forward vs. Backward vs. Symmetrical vs. Unrelated) × 2 (Measure: JOL vs. Recall) repeated measures ANOVA to assess whether the illusion of competence first reported by Koriat and Bjork (2005) replicated for participants in the JOL group. A main effect of Pair Type was found, *F*(3, 114) = 421.81, *MSE* = 99.94, *ηp*2 = .67, in which JOLs/recall rates were highest for forward pairs (65.10), followed by symmetrical pairs (61.32), backward pairs (43.40), and unrelated pairs (14.14). Post-hoc *t*-tests showed that JOLs/recall rates significantly differed across all comparisons, *t*s ≥ 4.42, *d*s ≥ 0.32. Next, a significant effect of measure was observed, *F*(1, 38) = 10.02, *MSE* = 521.91, *ηp*2 = .08, in which JOL ratings (50.07) exceeded later recall rates (41.90). Importantly, a significant interaction between Pair Type and Measure, *F*(3, 114) = 68.55, *MSE* = 49.40, *ηp*2 = .14, confirmed the presence of an illusion of competence pattern. Follow-up *t-*tests indicated a robust illusion of competence for backward pairs whereby JOLs greatly exceeded later recall accuracy (55.18 vs. 31.62), *t*(38) = 7.59, *SEM* = 3.21, *d* = 1.56. Additionally, the illusion of competence extended to unrelated pairs (19.43 vs. 8.85), *t*(38) = 3.97, *SEM* = 2.75, *d* = 0.87, and symmetrical pairs (64.83 vs. 57.78), *t*(38) = 2.32, *SEM* = 3.14, *d* = 0.47, though the difference between judgments and recall smaller than backward pairs. Finally, for forward pairs, this pattern reversed—JOL ratings were significantly lower than recall rates (60.87 vs. 69.34), *t*(38) = 2.93, *SEM* = 2.98, *d* = 0.57, indicating that participants underestimated their performance for this pair type and performed better than predicted at test.

Next, we tested JOL reactivity patterns by comparing the pair types across study groups using a 4 (Pair Type: Forward vs. Backward vs. Symmetrical vs. Unrelated) × 2 (Study Group: JOL vs. No-JOL) mixed ANOVA. A main effect of Pair Type was found, *F*(3, 228) = 512.24, *MSE* = 75.53, *ηp*2 = .64, indicating that across study groups, correct recall was greatest for forward pairs (58.69), followed by symmetrical pairs (46.89), backward pairs (23.88), and unrelated pairs (9.26). Post-hoc *t*-tests indicated that all comparisons differed significantly, *t*s ≥ 7.79, *d*s ≥ 1.27. An effect of Study Group was also found, *F*(1, 76) = 26.01, *MSE* = 623.74, *ηp*2 = .20, in which correct recall in the JOL group (41.89) exceeded the no-JOL group (27.47), indicating that providing JOLs at study produced a reactive effect across pair types. Importantly however, a significant interaction was found, *F*(3, 228) = 28.71, *MSE* = 75.53, *ηp*2 = .09, and post-hoc tests indicated that positive reactivity was confined to related pairs. Correct recall in the JOL group exceeded that of the no-JOL group for forward pairs (69.29 vs. 48.07), symmetrical pairs (57.76 vs. 36.03), and backward pairs (31.67 vs. 16.09), *t*s ≥ 4.90, *d*s ≥ 1.11. However, for unrelated pairs (8.85 vs. 9.68), no reactivity was found, *t* < 1, *p*BIC = .88. Thus, JOLs only appear to benefit cued-recall performance when item pairs are related.

**Discussion**

The results from Experiment 1 are quite clear. Providing JOLs at study greatly increased correct recall for targets for forward, backward, and symmetrical related pairs relative to a no-JOL control. For unrelated pairs however, providing JOLs had no effect on later recall compared to the no-JOL group. The finding that JOL reactivity effects on related pairs generalize to different types of directional associates that are matched on several lexical and semantic characteristics indicates that JOL reactivity effects occur for related pairs more broadly and are not specific to one associative direction. The JOL reactivity pattern is therefore consistent with most reactivity studies (Double et al., 2018; Janes et al., 2018; Soderstrom et al., 2015).

Also consistent with prior literature were the illusion of competence patterns found for backward, symmetrical, and unrelated pairs in the JOL group. For these pair types, JOLs exceeded later recall rates, and this pattern was particularly robust for backward pairs given the cue word that was provided at test was a poor predictor of the target. The illusion of competence indicates that JOLs were poorly calibrated to later recall. In contrast, JOLs for forward pairs, in which the cue was a better predictor of the target at test, were better calibrated to later recall and even underpredicted later recall. These patterns are generally consistent with previous studies (e.g., Koriat & Bjork, 2005; Castel et al., 2007), though forward JOLs are not always found to underpredict later recall (however, see Koriat & Bjork, 2006, Experiment 1).

Regarding JOL reactivity, the finding that positive reactivity effects are found for related pairs, but negative reactivity is not found for unrelated pairs is inconsistent with the changed-goal account. Related pairs, regardless of their associative direction, are clearly prioritized at encoding, leading to their elevated recall. Given this pattern, we instead suggest that participants are engaging in strategic relational encoding in which shared characteristics between related pairs are processed more deeply when providing JOLs, leading to their enhanced retrieval. However, this relational processing is moderated by pair relatedness, such that participants do not extend this relational encoding to unrelated pairs, possibly because processing shared characteristics between unrelated items would be relatively difficult. Thus, participants strategically prioritize related pairs by focusing on their shared characteristics. However, the lack of relational encoding directed towards unrelated pairs when JOLs are provided does not mean that participants reduce their encoding relative to unrelated pairs in the no-JOL group. This is because recall of unrelated pairs does not show a cost in the JOL group relative to the no-JOL group. Instead, unrelated pairs are encoded and recalled similarly across both groups.

Our interpretation of relational encoding on related word pairs is based on the Item-Specific/Relational framework (Einstein & Hunt, 1980; Hunt & Einstein, 1981). This framework proposes that encoding processes can emphasize the processing of unique item-specific characteristics of items in a memory set or shared relational characteristics between items, both of which can aid retention. Item-specific and relational encoding types are not mutually exclusive, but different tasks and study materials can increase the likelihood that one type of encoding is prioritized over another. For instance, a single study task (e.g., generation) can induce either item-specific or relational encoding depending upon qualitative differences in how the task is completed (Huff & Bodner, 2013). Importantly, it has been shown that study materials themselves can marshal encoding processes. In particular, related (vs. unrelated) study materials have been shown to promote relational encoding at study (Huff & Bodner, 2014; Hunt & Seta, 1984). Collectively then, relational encoding has been shown to benefit retention, especially compared to a shallow or neutral-control task (e.g., reading silently; Huff & Bodner, 2013; 2014; 2019), and this encoding type can be activated based on obvious relations between study items. Therefore, although providing JOLs is not inherently a relational task, it may be more likely to promote relational encoding when participants detect related pairs relative to silent reading.

We note that while similar, our strategic relational encoding account differs from the cue-strengthening account proposed by Soderstrom and colleagues (Soderstrom et al., 2015; Myers, Rhodes, & Hausman, 2020). Soderstrom et al. (2015) proposed that when individuals are tasked with making JOLs at encoding, the act of making the JOL strengthens the cues used forming the judgment (i.e., cue utilization theory; Koriat, 1997) and that the JOL task functions akin to a generation task (e.g., Slamecka & Graf, 1978), further strengthening the relational cues between the cue and the target. However, pair relatedness is not a requirement for the generation effect to occur, as the generation effect has been observed using single words rather than paired associates (Mulligan, Smith, & Buchin, 2018). Furthermore, the generation effect has been shown to extend to unrelated item pairs (McCurdy, Viechtbauer, Sklenar, Frankenstein, & Leshickar, 2020). Thus, if JOLs are operating in a manner similar to generation, positive reactivity should be observed for unrelated pairs. Instead, we suggest that JOL reactivity is primarily driven by implicit relational encoding that occurs as a byproduct of additional monitoring at encoding due to making JOLs. However, because this operates selectively, reactivity is only observed for related study pairs.

We test the strategic relational encoding account in Experiment 2 by comparing JOL and no-JOL groups relative to a relational encoding group who are instructed to explicitly use a relational encoding task for all study pairs. Although it is difficult to determine the exact processes involved in JOL reactivity, if recall for related pairs is equivalent between the JOL and relational encoding groups, this would provide evidence that relational processing contributes to JOL reactivity patterns.

**Experiment 2: JOLs versus Relational Encoding**

In Experiment 2, we tested whether positive reactivity found for related pairs following JOLs is equivalent to the memorial benefits typically observed when participants engage in relational encoding (Huff & Bodner, 2014). We therefore included a relational encoding group in which participants were instructed to explicitly think about how each of the word pairs are similar, which was used to encode all word pairs. We note that these explicit relational instructions differ from the strategic relational encoding account which posits that relational encoding that accompany JOLs is only applied to related but not unrelated pairs. Having participants in the relational group apply this task to all pairs (rather than selectively to related pairs as predicted by strategic relational encoding) was used because explicit relational encoding instructions have been shown to spillover to other encoding tasks when encoding is manipulated within-subjects (Huff, Bodner, & Gretz, in press). Given these carryover issues, it was reasonable to have participants utilize relational encoding for all pair types, rather than a subset which could potentially bleed over to unrelated pairs. In addition to the relational encoding group, we also included a shallow levels-of-processing group (i.e., vowel-counting task) to serve as an additional control. This group allowed us to ensure that positive reactivity effects were not merely due to having participants complete a task at study and that relational encoding was producing deep processing, leading to an overall memory benefit.

Consistent with Experiment 1, we expected a positive reactivity pattern for the JOL versus no-JOL group. Additionally, we expected that relational encoding would also induce a recall benefit that would mimic the positive reactivity pattern in the JOL group on related pairs, consistent with the strategic relational encoding account. However, we also expected that the recall of unrelated pairs would be greater in the relational encoding group relative to the JOL group. This is because the explicit relational task forces participants to utilize relational encoding regardless of pair type, which will likely benefit memory for unrelated pairs. Finally, we expected that the shallow group would produce lower levels of recall, possibly even lower than the no-JOL group since shallow processing is ineffective for promoting long-term memory.

**Methods**

**Participants and Stimuli**

A total of 167 participants were recruited for Experiment 2. Participants were recruited from two sources. First, we recruited 84 undergraduate psychology students recruited from University of Southern Mississippi who completed the study online for partial course credit. The remaining 83 participants were recruited online via Prolific and were compensated at a rate of $8.00/hour2. Participants were randomly assigned to the JOL group (*n* = 39), the no-JOL group (*n* = 40), the relational-encoding group (*n* = 45), and the shallow group (*n* = 43). All participants were native English speakers with normal or corrected-to-normal vision.

**Materials and Procedure**

The same materials from Experiment 1 were again used in Experiment 2. The same general procedure was similarly used with the exception of two additional encoding tasks. Participants in the relational-encoding group were instructed to think about how the two concepts were related to one another. The pair *cat-turtle­* was provided as an example, and participants in this group were instructed to consider overlapping features shared between the two concepts while studying the pairs (i.e., both are animals, have four legs, and can be kept as pets, etc.). In the vowel-counting group, participants were instructed to report the number of vowels in both the cue and target items. Both the relational encoding and shallow groups did not provide JOL ratings at study as in the no-JOL group and were instructed to apply their encoding strategy to all study pairs. After viewing each pair and studying it using their respective encoding strategy, participants pressed the enter key to move to the next pair. Participants in the JOL and no-JOL groups followed the identical procedure to Experiment 1 including completion of the 2-minute filler task and cued-recall test.

**Results**

The top panel of Figure 2 displays mean JOL ratings and cued-recall rates for each pair type for participants in the JOL group, while the bottom panel displays mean cued-recall rates for each of the four encoding strategies as function of pair type. To test for the illusion of competence, we first conducted a 4 (Pair Type: Forward vs. Backward vs. Symmetrical vs. Unrelated) × 2 (Measure: JOL vs. Recall) repeated measures ANOVA, assessing only participants who completed JOL encoding task. Consistent with Experiment 1, a main effect of Pair Type was found, *F*(3, 114) = 363.39, *MSE* = 112.72, *ηp*2 = .68, in which JOLs/recall rates were highest for forward pairs (65.68), followed by symmetrical pairs (63.15), backward pairs (44.43), and unrelated pairs (16.06). All comparisons differed significantly, *t*s ≥ 2.48, *d*s ≥ 0.22. A significant effect of Measure was also found, *F*(1, 38) = 50.54, *MSE* = 464.04, *ηp*2 = .29, such that JOL ratings (56.03) exceeded cued-recall rates (38.69). Finally, a significant interaction between Pair Type and Measure was found, indicating the presence of an illusion of competence, *F*(3, 114) = 56.41, *MSE* = 61.67, *ηp*2 = .15. Post-hoc tests indicated that an illusion of competence occurred for backward pairs such that JOLs greatly exceeded later recall rates (62.18 vs. 26.67), *t*(38) = 12.02, *SEM* = 3.05, *d* = 2.63. This pattern also occurred on unrelated pairs (22.30 vs. 9.81), *t*(38) = 4.07, *SEM* = 3.16, *d* = 0.96, and symmetrical pairs, (71.89 vs. 54.42), *t*(38) = 6.49, *SEM* = 2.79, *d* = 1.18. The illusion of competence, however, was not found on forward pairs, (67.63 vs. 63.74), but unlike Experiment 1, JOLs were equivalent to recall rates *t*(38) = 1.38, *SEM* = 2.91, *p* = .17, *p*BIC = .71.

We then examined reactivity patterns across encoding tasks, using a 4 (Pair Type: Forward vs. Backward vs. Symmetrical vs. Unrelated) × 4 (Study Group: JOL vs. No-JOL vs. Relational Encoding vs. Shallow Encoding) mixed ANOVA. An effect of Pair Type, *F*(3, 489) = 691.11, *MSE* = 78.13, *ηp*2 = .41, indicated that correct recall was highest for forward pairs (52.17), followed by symmetrical pairs (42.95), backward pairs (22.28), and lowest for unrelated pairs (13.73), which all differed statistically from each other, *t*s ≥ 10.72, *d*s ≥ 0.44. A main effect of Study Group was also found, *F*(1, 163) = 10.56, *MSE* = 1166.90, *ηp*2 = .14, in which correct recall was highest in the relational encoding group (41.06), followed by the JOL group (38.61), the no-JOL group (28.11), and shallow group (23.18). Post-hoc *t*-tests indicated that cued-recall rates in the JOL and relational encoding groups differed significantly from the no-JOL and shallow groups tasks (*t*s ≥ 4.14, *d*s ≥ 0.93), but did not differ between each other, *t* < 1, *p*BIC = .88. Additionally, there was no difference between the no-JOL and shallow groups, *t*(69) = 1.48, *SEM* = 3.39, *p* = .14, *p*BIC = .76.

The effects of Pair Type and Study Group were qualified by a significant interaction, *F*(9, 489) = 13.29, *MSE* = 78.13, *ηp*2 = .04. Beginning with forward pairs, correct recall was highest in the JOL group (63.78), followed by the relational group (58.17), the no-JOL control group (48.06), and the shallow group (39.19). All comparisons differed significantly (*t*s ≥ 2.13, *d*s ≥ 0.47), with the exception of the JOL and relational group, *t*(75) = 1.37, *SEM* = 4.18, *p* = .18, *p*BIC = .79, which were equivalent. This same pattern was also found with symmetrical pairs: Correct recall was highest in the JOL group (54.17), followed by the relational group (50.06), the no-JOL group (38.13) and shallow group (29.83). All comparisons differed significantly, *t*s ≥ 2.06, *d*s ≥ 0.45, again with the exception of the JOL and relational groups which were equivalent, *t* < 1, *p*BIC = .79. For backward pairs, correct recall was highest in the relational group (30.89), followed by the JOL group (26.60), the no-JOL group (17.23), and the shallow group (14.13). Follow up *t*-tests showed that recall rates in the JOL and relational groups differed from both the no-JOL and shallow groups (*t*s ≥ 3.24, *d*s ≥ 0.77). Recall did not differ between the JOL and relational group (26.60 vs. 30.89), or between no-JOL and shallow groups (17.13 vs. 14.13), *t*s < 1, *p*s ≥ .33, *p*BICs ≥ .85. Finally, for unrelated item pairs, recall rates were highest for the relational group (25.11) relative to the JOL task (9.87), the no-JOL group (9.13), and the shallow group (9.59, *t*s ≥ 3.73, *d*s ≥ 0.74). All other comparisons were non-significant, (*t*s < 1, *p*s ≥ .73, *p*BICs ≥ .90).

**Discussion**

Experiment 2 produced three notable outcomes. First, the illusion of competence pattern observed in Experiment 1 was again in evidence. Second, the JOL reactivity pattern found in Experiment 1 was found such that providing JOLs at study facilitated recall for related targets, but not for unrelated targets. Third, and most importantly, the JOL reactivity pattern found in related pairs was mimicked by the relational encoding group that was instructed to explicitly relate pairs together at encoding. This similarity suggests that JOL participants are engaging in deep relational encoding of related pairs despite not receiving instructions to do so. Positive reactivity was similarly found when comparing the JOL and relational groups to the shallow-encoding group, indicating that reactivity effects hold relative to a shallow control task. Recall did differ however between the JOL and relational group for unrelated pairs. This pattern is likely due to relational participants employing their encoding task across all pair types, rather than selectively limiting it to only related pairs as is likely occurring in the JOL group.

While similarities between relational encoding and the JOL group for related pair recall are consistent with a strategic relational encoding account, a stronger test of this account would be to contrast the JOL task with a similar relational-type task that can be strategically applied at study. The relational encoding task given to participants in Experiment 2 was not strategic because participants did not selectively apply relational encoding to related versus unrelated pairs. Instead, participants were explicitly instructed to use relational encoding when studying all pair types. In Experiment 3, we evaluate the strategic use of relational encoding by comparing the JOLs to the JAM task (Maki, 2007), which similarly requires participants to process the cue and the target items together to provide an estimate. If participants display similar reactivity patterns when employing the JAM task compared to the standard JOL task, this suggests that participants are strategically deploying relational encoding to facilitate encoding of related over unrelated pairs.

**Experiment 3: JOLs versus Judgments of Associative Memory**

The goal of Experiment 3 was to provide a further test of the strategic relational encoding account by comparing JOL reactivity effects to the JAM task. In the JAM paradigm, participants are presented with a cue-target pair and are asked to respond with the percent likelihood that an individual would respond to cue with the presented target (Garskof, & Forrester, 1966; Nelson, Dyrdal, & Goodmon, 2005; see Maki, 2007 for a review). These ratings are typically framed as the number of individuals out of 100 who would respond to the cue item with the presented target. In doing so, the JAM task is heavily dependent upon associative cues, as it gauges perceived association. Thus, like JOLs, JAMs should encourage relational processing, especially when the cue and target are strongly related.

By encouraging participants to process both the cue and target together, this task was designed to mimic the processing used by the JOL task. We elected to use JAMs due to this task’s similarity to the JOL task, as both require participants to think about related aspects of the study pairs (either conceptually or their use together) and assign a judgment value. Further, both tasks be elicited using the same scale, allowing for easy comparison. If participants are using relational encoding strategically on related word pairs, they would be able to use this encoding on both the JOL and JAM tasks. Of course, a key difference between the two tasks is that JOLs require participants to provide a recall forecast at encoding, whereas JAMs do not. Thus, an important question regarding JOL reactivity effects is whether memory predictions are necessary to produce a memory improvement. According to the strategic relational encoding account, only the use of relational encoding given to pairs at study will benefit memory, not necessarily whether a memory prediction is made. Thus, we expected memory forecasting via JOLs would not be necessary to produce reactivity effects.

**Methods**

**Participants**

70 participants were recruited from the University of Southern Mississippi’s undergraduate research pool and completed the study for partial course credit. Additionally, 28 participants were recruited from Prolific and completed the study at a rate of $8.00/hour, leading to a total of 98 participants who completed Experiment 33. Participants were randomly assigned to either the JOL group (*n* = 33), the no-JOL group (*n* = 32), or the JAM group (*n* = 33). All participants were native English speakers who reported normal or corrected-to-normal vision.

**Materials and Procedure**

Experiment 4 used the same materials as the previous experiments and followed the same general proceduredescribed in Experiment 1 with the following exception. In addition to the standard JOL and control groups, participants were also randomly assigned to a JAM task group in which they were asked to rate the likelihood in which the second item would be given as a response to the first. Like JOLs, JAM ratings were elicited using a continuous 0-100 scale. JAM instructions were modeled after the associative judgment task used by Maxwell & Buchanan (2020; exact instructions available at [OSF LINK]). Specifically, JAMs were framed as the number of individuals out of 100 who would respond with the target word if shown only the cue (i.e., as is typical in a free association task). As with the JOL task, JAMs were elicited concurrently with study, and study was self-paced across all groups. Thus, only the focal point of the two judgments differed.

**Results**

Figure 3 (top panel) plots mean JOL ratings and cued-recall rates for each pair type for participants completing the JOL task. The bottom panel displays mean recall as function of encoding group and pair type. Using a 4 (Pair Type: Forward vs. Backward vs. Symmetrical vs. Unrelated) × 2 (Measure: JOL vs. Recall) repeated measures ANOVA, we tested for the illusion of competence in the JOL group. Consistent with our predictions, this analysis yielded a significant effect of Pair Type, *F*(3, 96) = 269.87, *MSE* = 127.66, *ηp*2 = .89 that closely followed the patterns reported across the previous experiments. Specifically, mean JOLs/recall rates were highest for forward pairs (69.02), followed by symmetrical pairs (65.36), backward pairs (47.76), and were lowest for unrelated items (18.61). Comparisons differed statistically across each pair type, *t*s ≥ 3.04, *d*s ≥ 0.29. Next, collapsing across Pair Type, mean JOL ratings were significantly greater than mean correct cued-recall (55.16 vs. 45.36), *F*(1, 32) = 10.32, *MSE* = 693.79, *ηp*2 = .24. Finally, a significant interaction between Pair Type and Measure confirmed that the illusion of competence replicated, *F*(3, 96) = 38.71, *MSE* = 64.82, *ηp*2 = .55. Starting with backward pairs, post-hoc analyses revealed that JOLs greatly exceeded subsequent later recall (60.15 vs. 35.61), *t*(32) = 6.92, *SEM* = 3.78, *d* = 1.54. Next, the illusion of competence extended to unrelated pairs, (23.94 vs. 13.41), *t*(32) = 2.77, *SEM* = 3.71, *d* = 0.59, and symmetrical pairs, (70.14 vs. 60.68), *t*(32) = 2.89, *SEM* = 4.15, *d* = 0.61. Finally, for forward pairs, JOLs and recall did not significantly differ (66.25 vs. 71.74), *t*(32) = 1.44, *SEM* = 3.58, *p* = .16, *pBIC* = .67.

To test for reactivity effects, we next conducted a 4 (Pair Type: Forward vs. Backward vs. Symmetrical vs. Unrelated) × 3 (Study Group: JOL vs. JAM vs. No-JOL) mixed measures ANOVA. Overall, this analysis yielded a significant effect of Pair Type, *F*(3, 285) = 616.18, *MSE* = 81.46, *ηp*2 = .60, such that collapsed across study groups, correct recall was highest for forward pairs (64.92), followed by symmetrical pairs (56.22), backward pairs (33.16), and lowest for unrelated pairs (14.82). These differences were found to be significant across all comparisons, *t*s ≥ 8.08, *d*s ≥ 0.45. Next, an effect Study Group was detected, *F*(2, 95) = 3.90, *MSE* = 827.92, *ηp*2 = .06, in which collapsed across pair types, correct recall was highest when participants made JOLs (45.36) and JAMs (44.85) at encoding relative to participants in the no-JOL control group (36.46). All comparisons differed statistically, *t*s ≥ 2.28, *d*s ≥ 0.57, with the exception of the comparison between the JOL and JAM groups, *t*(64) < 1, *SEM* = 3.57, *p* = .88, *pBIC* = .88.

Importantly, a significant interaction between Pair Type and Study Group emerged, *F*(6, 285) = 9.82, *MSE* = 81.46, *ηp*2 = .04. Follow-up *t*-tests revealed that for forward pairs, correct recall in both the JOL (71.74) and JAM (67.58) groups exceeded that of the no-JOL group (55.16). Comparisons across all statistically differed, *t*s ≥ 2.93, *d*s ≥ 0.65, with the exception of the comparison between the JOL and JAM groups, which were equivalent, *t*(64) < 1, *SEM* = 4.47, *p* *=* .35, *p*bic = .84. Next, a similar pattern was observed for symmetrical pairs. Correct recall was greater for the JOL (60.68) and JAM (61.29) groups compared to the no-JOL group (46.41). Again, all comparisons statistically differed *t*s ≥ 3.22, *d*s ≥ 0.80, except for the comparison between the JOL and JAM groups, *t*(64) < 1, *SEM* = 4.54, *p* *=* .89, *p*bic = .87. For backward pairs correct recall in the JOL (35.61) and JAM (36.36) groups was again greater relative to the no-JOL group (27.34). Correct recall in the JAM and the No-JOL groups differed significantly, *t*(63) = 2.11, *SEM* = 4.35, *d* = 0.52, while the comparison between JOLs and the No-JOL group was marginal, *t*(63) = 1.93, *SEM* = 4.37, *p* *=* .06, *p*bic = .56, *d* = 0.48. Recall did not differ between the JOL and JAM groups, *t*(64) < 1, *SEM* = 4.21, *p* *=* .86, *p*bic = .88. Finally, for unrelated pairs, recall rates were equivalent across the JOL (13.41), JAM (14.68), and no-JOL (16.95) groups, *t*s ≤ 1.23, *p*s ≥ .22, *p*bic ≥ .79. As such, both the JOL and JAM tasks resulted in statistically equivalent reactivity patterns on correct recall for related pairs while producing no reactivity on unrelated pairs.

**Discussion**

Experiment 3 provided an additional test of the strategic relational encoding account of reactivity by comparing the standard JOL task to a JAM. In both tasks, participants processed the cue-target relations prior to providing a judgment using the same 0-100 scale. Although the judgment type differs (recall forecasting vs. relatedness estimates), the reactivity patterns observed for related and unrelated pairs did not differ as a function of judgment type, indicating that similar processing occurred between the two task types. Compared to the no-JOL control group, both the JOL and JAM groups showed increased correct recall of targets across forward, backward, and symmetrical pairs—a positive reactivity pattern, but produced no effect on recall of unrelated targets.

The similarity in recall rates between the JOL and JAM encoding groups yields several important findings regarding reactivity effects in recall of cue-target pairs. First, providing a memory prediction does not appear to be a requisite for positive reactivity on related pairs given the equivalence between the JOL and frequency groups. This finding is important in reference to other studies that have reported JOL reactivity patterns (e.g., Soderstrom et al., 2015; Mitchum et al., 2016) which have only compared JOL and no JOL groups and have not measured recall differences relative to non-JOL encoding tasks. Second, the similar reactivity patterns in the JOL and JAM tasks indicates that type of task employed at encoding may not be a critical factor as to whether or not a reactivity pattern emerges. Instead, the qualitative processing given to the cue and target by the task may be more impactful. Finally, the finding that reactivity does not operate globally across all pair types suggests that reactivity processes are applied strategically, with an emphasis on related over unrelated pairs.

While the JAM task does explicitly instruct participants to relate item pairs together at encoding, pair relatedness is still a focal point of this task, as participants are essentially tasked with estimating the associative strength between items. A more complete assessment of the strategic relational encoding account would be to test whether reactive effects replicate using a deep encoding task that places less emphasis on relational processing. As such, Experiment 4 introduces a frequency judgment task in which particpants are tasked with rating the likelihood that two words would be used together. Like JAMs, frequency judgments are sensitive to the context in which words are used. However, because the frequency judgment task places less emphasis on pair relatedness relative to JAMs, this task would be less likely to promote relational processing. Thus, the use of frequency judgments in Experiment 4 allowed us to test whether the reactivity patterns predicted by the strategic relational encoding account replicate when using an encoding task that places less emphasis on item relatedness.

**Experiment 4: JOLs vs Frequency Judgments**

The primary goal of Experiment 4 was to provide a more complete test of the strategic relational encoding account by comparing JOL reactivity effects to a frequency judgment task. In the frequency judgment task, participants are asked to estimate the likelihood that the cue and target words would appear together within the English language. We note that while the frequency task is still sensitive to pair relatedness, this sensitivity is diminished relative to the JAM task, as frequency judgments do not require participants to quantify pair relatedness in terms of association strength.

Overall, we expected that any observed reactivity would adhere to the strategic relational encoding account. Specifically, we anticipated that JOL group would again show positive reactivity for related pairs (forward, backward, and symmetrical), and would not differ on unrelated pairs relative to a no-JOL control, based on the previous experiments. Furthermore, consistent with the results of Experiment 3, we also expected that this pattern of reactivity would extend to the frequency judgment group, such that positive reactivity would be observed for related, but not unrelated pairs. Furthermore, we expected these patterns would be equivalent to the JOL group due to relational encoding of related pairs fostered by both tasks.

**Methods**

**Participants**

A total of 118 participants took part in Experiment 4 and were randomly assigned to either the JOL group (*n* = 40), the no-JOL group (*n* = 39), or the frequency judgment group (*n* = 39). All participants were recruited from the University of Southern Mississippi’s undergraduate research pool who completed the study online for partial course credit. All participants were native English speakers and reported normal or corrected-to-normal vision.

**Materials and Procedure**

Experiment 3 used the same set of materials as the previous experiments and followed the same general procedure as Experiment 1 with one exception. In addition to the JOL and no-JOL groups, Experiment 3 included a frequency judgment group in which participants were asked to rate the likelihood in which the cue and target items would appear together in everyday language. The frequency judgment task utilized the same 0-100 rating scale employed by the JOL task in which higher ratings corresponded to more frequent occurrences. As with the JOL task, frequency judgments were made concurrently with study such that participants typed their ratings while the pairs were displayed on the screen. Thus, the only difference between the two tasks was the focus of the judgment.

**Results**

The top panel of Figure 4 compares mean JOL ratings and cued-recall rates for each pair type for participants completing the JOL task at encoding, while the bottom panel reports mean recall rates as function of encoding group and pair type. First, to test for the illusion of competence in the JOL group, a 4 (Pair Type: Forward vs. Backward vs. Symmetrical vs. Unrelated) × 2 (Measure: JOL vs. Recall) repeated measures ANOVA was used. As expected, this analysis revealed a main effect of Pair Type, *F*(3, 117) = 293.33, *MSE* = 151.31, *ηp*2 = .56, following the same pattern reported in the previous two experiments. JOLs/recall rates were highest for forward pairs (68.29), followed by symmetrical pairs (65.73), backward pairs (47.56), and lowest for unrelated items (17.14). All comparisons differed statistically, *t*s ≥ 2.38, *d*s ≥ 0.18. JOL ratings were only marginally greater than cued-recall rates (52.25 vs. 47.11), *F*(1, 39) = 3.56, *MSE* = 590.62, *p* = .07, *ηp*2 = .02, *p*BIC = .53, however a significant interaction confirmed the presence of an illusion of competence pattern, *F*(3, 117) = 57.32, *MSE* = 68.40, *ηp*2 = .10. For backward pairs, post-hoc tests indicated that JOLs greatly exceeded subsequent cued-recall rates (59.68 vs. 35.44), *t*(39) = 6.79, *SEM* = 3.69, *d* = 1.27. For unrelated pairs, the illusion of competence did not occur, as JOLs and recall were equivalent (16.74 vs. 17.53), *t* < 1, *p*BIC = .86. Similarly, JOLs and recall were calibrated on symmetrical pairs, (68.56 vs. 62.91), *t*(39) = 1.69, *SEM* = 3.44, *p* = .10, *p*BIC = .61. Finally, as found in Experiment 1, a pattern of underestimation was detected for forward pairs in which JOLs were generally lower than subsequent recall (72.57 vs. 64.03), *t*(39) = 2.90, *SEM* = 3.04, *d* = 0.52.

Next, we conducted a 4 (Pair Type: Forward vs. Backward vs. Symmetrical vs. Unrelated) × 3 (Study Group: JOL vs. Frequency vs. No-JOL) ANOVA to evaluate reactivity effects. An effect of Pair Type was detected, *F*(3, 348) = 590.71, *MSE* = 99.13, *ηp*2 = 0.50, indicating that correct recall was highest for forward pairs (62.94), followed by symmetrical pairs (56.13), backward pairs (29.97), and lowest for unrelated pairs (15.31). Differences were significant across all comparisons, *t*s ≥ 10.80, *d*s ≥ 0.79. An effect Study Group was also found, *F*(2, 116) = 6.00, *MSE* = 1205.07, *p* = .003, *ηp*2 = .08, indicating that correct recall was highest when participants made JOLs (47.13) and frequency judgments (43.30) relative to the no-JOL control group (32.66). All comparisons were significant, *t*s ≥ 2.97, *d*s ≥ 0.67, except for the JOL and frequency groups, *t* < 1, *p*bic = .86.

Critically, a significant interaction was found, *F*(6, 348) = 12.34, *MSE* = 1205.07, *ηp*2 = .04. Follow-up tests indicated that for forward pairs, correct recall in both the JOL (72.57) and frequency judgment (66.58) groups exceeded that of the no-JOL group (49.42). All comparisons differed, *t*s ≥ 3.91, *d*s ≥ 0.88, except for the JOL and frequency judgment groups, which were equivalent, *t*(76) = 1.50, *SEM* = 4.07, *p* *=* .14, *p*bic = .74. Symmetrical pairs displayed a similar pattern. Recall was greater in the JOL (62.91) and frequency judgement (62.05) groups relative to the no-JOL group (43.27), and again, all comparisons differed *t*s ≥ 4.23, *d*s ≥ 0.96, with the exception of the JOL and frequency judgment groups, which were equivalent, *t* < 1, *p*bic = .85. For backward pairs, correct recall in the JOL (35.44) and frequency judgment (31.28) groups were greater than the no-JOL group (23.01). All comparisons differed significantly, *t*s ≥ 1.96, *p*s < .05, except for the JOL and frequency judgment group, which did not differ, *t* < 1, *p*bic = .90. Finally, for unrelated pairs, recall rates were equivalent across the JOL (17.53), frequency judgment (13.34), and no-JOL (14.94) groups, *t*s ≤ 1.02, *p*s ≥ .31, *p*bic ≥ .88. Thus, both JOL ratings and frequency judgments produced statistically equivalent reactivity patterns on correct recall for related pairs but produced no reactivity on unrelated pairs.

**Discussion**

The primary goal of Experiment 4 was to provide further testing of the strategic relational encoding account of reactivity by assessing whether reactivity patterns observed for JOLs and JAMs replicate when particpants engage in a frequency judgment task at encoding. Consistent with Experiment 3, similar reactivity patterns emerged across both judgment types. Relative to the no-JOL control group, participants completing either the JOL or frequency judgment task at encoding showed increased correct recall for each of the three types of related associates. Again, these tasks produced no reactivity when participants studied unrelated pairs, indicating that reactivity effects operated selectively as a function of pair relatedness. Finally, the results from the frequency judgment task provide further evidence that memory forecasting is not a requirement for reactivity to occur.

**General Discussion**

The primary goals of this study were twofold. First, Experiment 1 sought to replicate previous work showing that JOLs produce a reactive effect on cued-recall of related targets while comparing these reactivity patterns on forward, backward, and symmetrical paired associates—a novel contribution. Second, Experiments 2, 3, and 4 were designed to test a strategic relational encoding account of reactivity. In Experiment 2, we gauged JOL reactivity effects relative to a deep relational encoding strategy, while in Experiment 3, JOL reactivity was compared to the JAM task in which participants made relational, non-metacognitive frequency judgments. Experiment 4 provided an additional test of the strategic relational encoding account by comparing JOL reactivity to a frequency judgment task. Collectively, our results indicate that enhanced relational encoding applied to related but not unrelated pairs primarily contributes to these reactivity benefits and that memory predictions via JOLs are not necessary to produce reactivity effects.

Consistent with Soderstrom et al. (2015) and Janes et al. (2018), Experiment 1 found evidence of positive JOL reactivity on forward pairs, while extending this pattern to include backward and symmetrical pairs. Importantly, these reactivity patterns occurred using pairs that were engineered to control for lexical and semantic item effects, including associative strength that could potentially influence correct recall. The positive reactivity pattern found across related pair types indicates that the associative direction of cue-target pairs does not have an effect on reactivity. Instead, the mere presence of association is likely sufficient to facilitate relational encoding of related pairs. For unrelated pairs however, no reactivity pattern was found as recall was equivalent between the JOL and no-JOL groups. The discrepancy in reactivity for related and unrelated pairs suggests that JOL participants engage in strategic relational encoding of related pair types. As we discuss further below, previous researchers have proposed a similar strategy-based account (e.g., changed-goal hypothesis; Mitchum et al., 2016), but this account does not predict a null reactivity effect on unrelated pairs. Instead, we suggested that positive reactivity of related pairs reflected strategic relational encoding in which participants emphasized the relational characteristics of related pairs at study, facilitating subsequent recall.

To test this account, Experiment 2 compared a relational encoding task in which participants were explicitly instructed to relate all cue-target pairs together at study. Relative to both a no-JOL and a shallow encoding control task, relational encoding produced the same positive reactivity pattern on related pairs as participants who completed the JOL task. Unlike the JOL task, however, the positive reactivity of relational processing was not restricted to related targets as recall of unrelated targets was also greater relative to the no-JOL control group. This latter pattern was unsurprising given participants were instructed to utilize relational encoding for all pair types, precluding any strategic use of relational encoding. Finally, the shallow-vowel counting task did not induce reactivity, suggesting that the qualitative aspects of the encoding task were a driving factor of reactivity relative to merely having participants engage in an additional task at study.

To gauge the strategic component of the relational encoding account, Experiment 3 compared the JOL and no-JOL groups to the JAM task, which required participants to provide relatedness judgments for cue-target pairs. This task was selected as it allowed for processing of the relational characteristics of study pairs without explicit instruction to encode all study pairs using a relational strategy as in Experiment 2. Moreover, the JAM task utilized the same rating scale as the JOL task. The JAM task was therefore closely matched to the JOL task to allow for strategic relational encoding, but it did not require that participants forecast later recall performance. Experiment 3 found equivalent positive reactivity on related pairs when compared to the JOL task and critically, no reactivity was found on unrelated pairs, indicating that reactivity patterns are not exclusive to JOLs and that likely reflect use of strategic relational encoding.

Finally, Experiment 4 compared the JOL and no-JOL groups to a novel frequency judgment task in which participants were required to provide an estimate on the frequency in which the cue-target pair would co-occur in the English language. Like the JAM task used In Experiment 3, the frequency judgment task provided us with a further test of the strategic relational encoding account as it allowed processing of pair relatedness without explicit relational instructions. However, relative to JAMs, frequency judgments are a weaker relational task, as these judgments are not designed to explicitly measure perceived association. Overall, Experiment 4 replicated the previously reported reactivity patterns for JOLs. Furthermore, frequency judgments were found to display positive reactivity on related pairs when compared to the JOL task, and critically, no reactivity was found on unrelated pairs. The extension of this finding to frequency judgments further suggests that reactivity patterns are not exclusive to JOLs and that they likely reflect use of strategic relational encoding.

**Revisiting the Changed-Goal Hypothesis**

As reviewed in the Introduction, one account of JOL reactivity is the changed-goal hypothesis (Mitchum et al., 2016). As per this hypothesis, having participants provide JOLs at study increases participants’ awareness of item difficulty, and as a result, participants will modify their study goals (and therefore, their encoding) to prioritize learning of pairs perceived as easy to remember at the expense of more difficult pairs. Because related pairs are generally viewed by participants as easier to learn, the changed-goal hypothesis posits that providing JOLs will produce a positive reactive effect for forward, backward, and symmetrical pairs and a negative reactive effect for more difficult unrelated pairs. However, previous research has been largely inconsistent with the changed-goals hypothesis (e.g., Double et al., 2018; Janes et al., 2018; Soderstrom et al., 2015), finding only positive JOL reactivity for related pairs, and the absence of negative reactivity for unrelated pairs. Experiments 1-3 of the present study were similarly inconsistent with changed-goal hypothesis, providing further evidence against the viability that participants are altering their study goals in response to related versus unrelated pair types.

Soderstrom et al. (2015) proposed that JOLs will induce reactivity when two criteria are met. First, the JOL task must strengthen cues that inform JOLs (i.e., such as pair relatedness) and, second, the same cues must be available at test (i.e., such as a cued-recall test in which the desired target can be triggered by the presentation of the cue). To test this account, Myers et al. (2020) examined whether the reactive effects of JOLs extended to recognition and free recall tests, as these test types do not present the participants with the cue item at test and are therefore less dependent on the cues activated by the JOL task at encoding. Positive reactivity was found on related pairs, but negative reactivity was not found on unrelated pairs on both cued-recall and recognition tests. This pattern supports Soderstrom et al.’s first criterion that the JOL task must strengthen cue-target associations, but not the second regarding the presence of test cues. Thus, the present study is consistent with other studies indicating that JOL reactivity is contingent upon the presence of relational processing.

Importantly, we propose that reactivity is not unique to JOLs and is largely due to participants strategically employing relational encoding at study based on the related characteristics of the study materials. Because JOLs call attention to pair relatedness (which is a strong predictor of cued-recall performance; e.g., Maxwell & Buchanan, 2020), relatedness cues may become more salient relative to participants in a no-JOL control. Based on this account, reactivity would occur for any task that encourages participants to use a relational strategy at encoding and when there are study items that differ in their relatedness. Results from Experiments 2, 3, and 4 support this claim, as relational encoding (Experiment 2), JAMs judgments (Experiment 3), and frequency judgments (Experiment 4) each produced similar reactivity patterns for related pairs relative to the JOL group. Furthermore, the similarity of reactivity patterns between the JOLs and both JAMs and frequency judgments suggests that each task taps into similar underlying relational encoding processes. Based on Koriat’s (1997) cue-utilization framework, both JOLs and frequency judgments tune participants to specific *intrinsic* cues about the study pairs, providing them with information about inherent properties of the studied material (i.e., pair relatedness). Thus, cued-recall performance is enhanced whenever the encoding task draws attention to the relatedness between studied items, regardless of whether this is done explicitly (e.g., relational study instructions) or implicitly (e.g., JOLs, JAMs, frequency judgments, etc.). However, because this occurred indirectly in Experiments 3 and 4 (as neither the JOL, JAM, or frequency judgment tasks explicitly instructed participants to relate items together at study) only related items receive a memory boost when judged. As such, reactive effects are not generally observed unrelated items unless the task explicitly instructs participants to relate all pairs together.

Our strategic relational encoding account is consistent with previous work on metamemory and strategy use. For example, in their unifying framework of metamemory, Nelson and Narens (1990) posited that participants are able to adjust their encoding strategies based on cues inherent to the stimuli (i.e., engaging in monitoring). Because pair relatedness is a salient cue of future recall performance, it is likely that these relatedness cues trigger participants’ changes in study strategy. Thus, only related pairs are processed using a relational encoding strategy, as participants modify their study strategy whenever they encounter this pair type. This results in a memory boost for related items that receive additional relational processing at encoding while unrelated pairs show no benefit.

Finally, although our proposed account of reactivity is based on strategic relational encoding, the present study does not directly assess whether participants are engaging in relational encoding while providing JOLs. Instead, we rely upon comparisons to similar relational tasks in Experiments 2, 3, and 4 as a means of triangulating encoding processing (see Huff & Bodner, 2013; Meade, Klein, & Fernandes, 2020, for a similar comparison process). Second, our experiments did not include any online measures of strategic encoding at either study or test. While it has been well documented within the metacognitive literature that participants engage in strategic encoding both when acquiring new knowledge and when processing metamemorial information (e.g., Nelson & Narens, 1990; Hertzog & Dunlosky. 2004), our study did not explicitly assess whether participants were altering study strategies as a function of pair type. Rather, strategic changes of encoding strategy were inferred based on differences in cued-recall rates. Future research could utilize more direct measures such as having participants report the type of encoding used during study as a function of pair type which could also indicate any encoding changes which would be consistent with a strategy change.

**Conclusion**

The present study provides further examination of JOL reactivity and its underlying mechanisms. The use of multiple associative pair types provided us with a more precise test of negative reactivity, the changed-goal hypothesis, and allowed us to test whether different associative pair types produce the same reactive benefits as forward associates. Overall, we found that the reactive benefits of JOLs can be extended to both backward and symmetrical pairs. Importantly, our findings from Experiments 3 and 4 indicate that reactive effects associated with JOLs are not exclusive to JOLs and extend to other types of rating tasks that emphasize the relational processing of cue-target pairs. Our experiments demonstrate that reactivity is driven by strategic allocation relational encoding towards related pairs and that memory forecasting from JOLs are not a prerequisite for reactivity.

**Open Practices Statement**

The data for all experiments have been made available at https://osf.io/8yvn3/. None of the experiments were preregistered.

References

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, *81* (1), 126–131.

Balota, D., A, & Neely, J. H. (1980). Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory, 6* (5), 576-587.

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B, & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39* (3), 445-459.

Brysbaert, M., & New, B. (2009). Moving beyond Kučera and Francis: A critical evaluation of current word frequency norms and the introduction of a new and improved word frequency measure for American English. *Behavior Research Methods, 41*, 977–990.

Bjork, R. A. (1999). Assessing our own competence: Heuristics and illusions. In D. Gopher & A. Koriat (Eds.), *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application* (pp.435–459). Cambridge, MA: MIT Press.

Bjork, R. A. (2016). Prologue: Some metacomments on metamemory. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 1–3). Oxford: Oxford University Press.

Castel, A. D., McCabe, D. P., & Roediger, H. L. (2007). Illusions of competence and overestimation of associative memory for identical items: evidence from judgments of learning. *Psychonomic Bulletin & Review*, *14* (1), 107–111.

Criss, A. H., Aue, W. R., & Smith, L. (2011). The effects of word frequency and context variability in cued recall. *Journal of Memory and Language, 64* (2), 119-132.

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgments of learning. *Memory, 26* (6), 741-750.

Einstein, G. O., & Hunt, R. R. (1980). Levels of processing and organization: Additive effects of individual-item and relational processing. *Journal of Experimental Psychology: Human Learning and Memory, 6* (5), 588-598.

Faul, F., Erdfelder, E., Lang, A. G., & Buchner, A. (2007). G\*Power 3: a flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior* *Research Methods*, *39* (2), 175–191.

Garcia, M. & Kornell, N. (2015). Collector [Computer software]. Retrieved April 3rd, 2020 from https://github.com/gikeymarica/Collector.

Garskof, B. E., & Forrester, W. (1966). The relationship between judged similarity, judged association, and normative association. *Psychonomic Science, 6*, 504.

Hanczakowski, M., Zawadzka, K., Pasek, T., & Higham, P. A. (2013). Calibration of metacognitive judgments: Insights from the underconfidence-with-practice effect. *Journal of Memory and Language, 69*, 429–444.

Hertzog, C., & Dunlosky, J. (2004). Aging, metacognition, and cognitive control. In B. H. Ross (Ed.), *Psychology of Learning and Motivation* (pp. 215−251). San Diego, CA, US: Academic Press.

Hertzog, C., Dunlosky, J., Powell-Moman, A., & Kidder, D. P. (2002). Aging and monitoring associative learning: Is monitoring accuracy spared or impaired*? Psychology and Aging, 17*, 209–225.

Huff, M. J., & Aschenbrenner, A. J. (2018). Item-specific processing reduces false recognition in older and younger adults: Separating encoding and retrieval using signal detection and the diffusion model. *Memory & Cognition, 46*, 1287-1301.

Huff, M. J., & Bodner, G. E. (2013). When does memory monitoring succeed versus fail? Comparing item-specific and relational encoding in the DRM paradigm. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39* (4), 1246-1256.

Huff, M. J., & Bodner, G. E. (2014). All varieties of encoding variability are not created equal: Separating variable processing from variable tasks. *Journal of Memory and Language, 73*, 43-58.

Huff, M. J. & Bodner, G. E. (2019). Item-specific and relational processing both improve recall accuracy in the DRM paradigm. *Quarterly Journal of Experimental Psychology, 72* (6), 1493-1506.

Huff, M. J., Bodner, G. E., & Gretz, M. R. (in press). Distinctive encoding of a subset of DRM lists yields not only benefits, but also costs and spillovers. *Psychological Research.*

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior, 20* (5), 497-514.

Hunt, R. R., & Seta, C. E. (1984). Category size effects in recall: The roles of relational and individual item information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10* (3), 454-464.

Janes, J. L., Rivers, M. L, & Dunlosky, J. (2018). The influence of making judgments of learning on memory performance: Positive, negative, or both? *Psychonomic Bulletin & Review, 25* (6), 2356-2364.

Koriat, A. (1997). Monitoring one’s own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experiment Psychology: General, 126* (4), 349-370.

Koriat, A., & Bjork, R. A. (2005). Illusions of competence in monitoring one’s knowledge during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31* (2), 187–194.

Koriat, A., & Bjork, R. A. (2006). Illusions of competence during study can be remedied by manipulations that enhance learners’ sensitivity to retrieval conditions at test. *Memory & Cognition, 34* (5), 959–972.

Kornell, N., Rhodes, M. G., Castel, A. D., & Tauber, S. K. (2011). The ease-of-processing heuristic and stability bias: Dissociating memory, memory beliefs, and memory judgment. *Psychological Science, 22*, 787-794.

Landauer, T. K., & Dumais, S. T. (1997). A Solution to Plato’s Problem: The latent semantic analysis theory of acquisition, induction, and representation of knowledge. *Psychological Review, 104* (2), 211-240.

Madan, C R., Glaholt, M. G., & Caplan, J. B. (2010). The influence of item properties on association-memory. *Journal of Memory and Language*, *63* (1), 46-63.

Maki, W. S. (2007). Judgments of associative memory. *Cognitive Psychology, 54*(4), 319-353.

Masson, M. E. J. (2011). A tutorial on a practical Bayesian alternative to null-hypothesis significance testing. *Behavior Research Methods, 43*, 679-690.

Maxwell, N. P., & Buchanan, E. M. (2020). Investigating the interaction of direct and indirect relation on memory judgments and retrieval. *Cognitive Processing, 21*, 41-53.

Maxwell, N. P., & Huff, M. J. (in press). The deceptive nature of associative word pairs: Effects of associative direction on judgments of learning. *Psychological Research*, 1-19.

McCurdy, M. P., Viechtbauer, W., Sklenar, A. M., Frankenstein, A. N., & Leshikar, E. D. (2020). Theories of the generation effect and the impact of generation constraint: A meta-analytic review. *Psychonomic Bulletin & Review, 27*, 1139-1165.

Meade, M. E., Klein, M. D, & Fernandes, M. A. (2020). The benefit (and cost) of drawing as an encoding strategy. *Quarterly Journal of Experimental Psychology, 73* (2), 199-210.

Metcalfe, J. (2000). Metamemory: Theory and data. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 197-211). New York, NY, US: Oxford University Press.

Metcalfe, J., & Kornell, N. (2003). The dynamics of learning and allocation of study time to a region of proximal learning. *Journal of Experimental Psychology: General, 132*, 530–542.

Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General, 145* (2), 200-219.

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review, 20*, 378-384.

Mulligan, N. W., Smith, S. A., & Buchin, Z. L. (2019). The generation effect and experimental design. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(8), 1422-1431.

Myers, S. J., Rhodes, M. G., & Hausman, H. E. (2020). Judgments of learning (JOLs) selectively improve memory depending on the type of test. *Memory & Cognition, 48*, 745-758.

Nelson, D. L., Dyrdal, G. M., & Goodmon, L. B. (2005). What is preexisting strength? Predicting free association, similarity ratings, and cued recall probabilities. *Psychonomic Bulletin & Review, 12*, 711-719.

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (2004). The University of South Florida free association, rhyme, and word fragment norms. *Behavior Research Methods,* *Instruments, & Computers*, *36* (3), 402–407.

Nelson, T. O. & Narens, L. (1990). Metamemory: A theoretical framework and new findings. In: *The psychology of learning and motivation*, ed. G. Bower. American Psychologist.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory, 4*(6), 592-604.

Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*, 553–558.

Wagenmakers, E. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review, 14*, 779-804.

Witherby, A. E., & Tauber, S. K. (2017). The influence of judgments of learning on long-term learning and short-term performance. *Journal of Applied Research in Memory and Cognition, 6* (4), 496-503.

Footnotes

1 For completeness, we further analyzed the effect of block order in Experiments 1-3. No interactions with block were found in Experiment 1 or Experiment 3 (*F*s < 2.50, *p*s > .06, *p*BICs > .99), however block did interact with pair type in Experiment 2, *F*(3, 489) = 879.21, *MSE* = 83.64, *ηp*2 = 0.01, though all other interactions with block were not significant (*F*s < 1.63, *p*s > .10, *p*BICs > .99). Post-hoc *t*-tests revealed that correct recall of backward pairs was numerically higher in block 1 (23.66) than block 2 (21.07), however, this comparison failed to reach conventional significance, *t*(165) = 1.77, *SEM* = 1.46, *p* = .08, *p*BIC = .81. All other comparisons were non-significant (*t*s < 1, *p*BICs > .89). Furthermore, the same general pattern of reactivity was detected in Experiment 2 after controlling for block order, indicating order did not contribute to the reactivity patterns reported.

2 Due to the COVID-19 pandemic, data collection was shifted online to Prolific partway through Experiment 2. The forty participants in the no-JOL group were recruited through Prolific. Additionally, 20 participants in the relational group, 19 participants in the shallow group, and 2 participants in the JOL group were recruited via Prolific. For completeness, we note that mean correct recall did not differ between the no-JOL group in Experiment 2 and the undergraduate sample completing the same task in Experiment 3 (28.11 vs. 32.66; *t*(69) = 1.50, *SEM* = 3.08, *p* = .14). Additionally, within Experiment 2, participant responses did not differ as a function recruitment platform participants completing the relational encoding task (44.81 vs. 38.05; *t* < 1, *p* = .33) or the vowel counting task (36.56 vs. 30.47; *t*(43) = 1.07, *SEM* = 5.87, *p* = .29). Thus, recall performance and JOL responses did not appear to differ as a function of participant source.

3 [WORDS HERE]



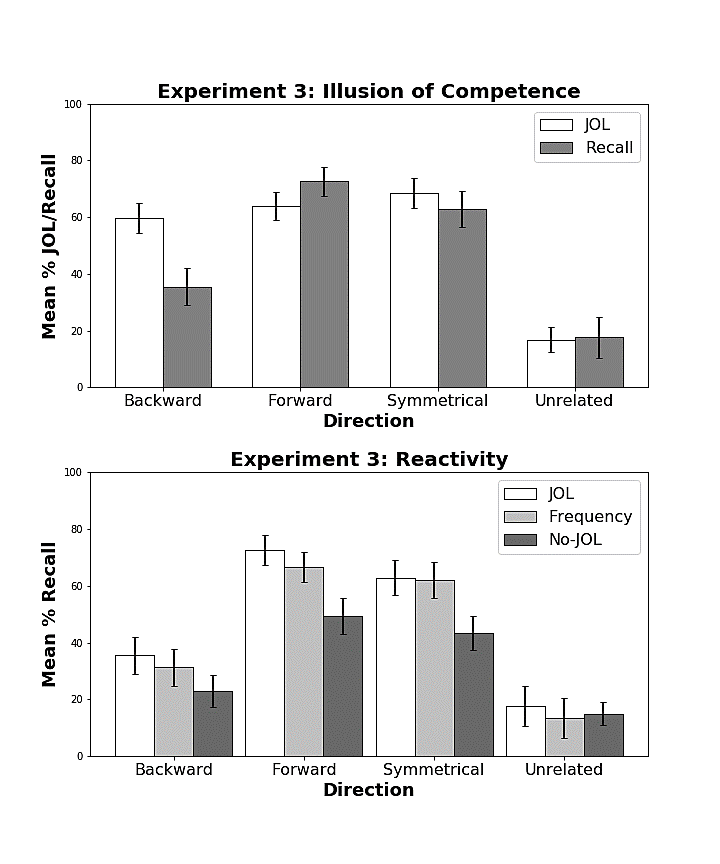
*Figure 1.* Comparison of mean JOL ratings and recall rates in the JOL encoding group (top panel) and recall rates in the JOL and No-JOL groups (bottom panel). Error bars represent 95% confidence intervals.



*Figure 2.* Comparison of mean JOL ratings and recall rates in the JOL encoding group (top panel) and recall rates in the JOL, Relational Encoding, Vowel Counting, and No-JOL groups (bottom panel). Error bars represent 95% confidence intervals.

[FIGURE 3 HERE]

*Figure 3.* Comparison of mean JOL ratings and recall rates in the JOL encoding group (top panel) and recall rates in the JOL, JAM, and No-JOL groups (bottom panel). Error bars represent 95% confidence intervals.



*Figure 4.* Comparison of mean JOL ratings and recall rates in the JOL encoding group (top panel) and recall rates in the JOL, Frequency judgment, and No-JOL groups (bottom panel). Error bars represent 95% confidence intervals.

**Appendix**

Table A1

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Condition | Variable | *M* | *SD* | *Min.* | *Max.* |
| Forward | FAS | .37 | .21 | .05 | .81 |
|  | BAS | .00 | .00 | .00 | .00 |
| Backward | FAS | .00 | .00 | .00 | .00 |
|  | BAS | .37 | .21 | .05 | .81 |
| Symmetrical | FAS | .19 | .13 | .01 | .46 |
|  | BAS | .19 | .13 | .02 | .52 |

*Mean Associative Strength Summary Statistics Forward, Backward, and Symmetrical Pairs.*

*Note.* FAS (forward associative strength) and BAS (backward associative strength) values for unrelated pairs as these items share zero associative overlap.

Table A2

*Summary Statistics for Cue and Target Concreteness, Length, and Frequency Item Properties as a Function of Pair Type.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Pair Type | Position | Variable | *M* | *SD* |
| Forward | Cue | Concreteness | 4.97 | 1.22 |
|  |  | Length | 6.20 | 1.86 |
|  |  | Frequency | 3.74 | 0.67 |
|  | Target | Concreteness | 4.96 | 1.14 |
|  |  | Length | 4.46 | 1.27 |
|  |  | Frequency | 2.49 | 0.63 |
| Backward | Cue | Concreteness | 4.96 | 1.14 |
|  |  | Length | 4.46 | 1.27 |
|  |  | Frequency | 2.49 | 0.63 |
|  | Target | Concreteness | 4.97 | 1.22 |
|  |  | Length | 6.20 | 1.86 |
|  |  | Frequency | 3.74 | 0.67 |
| Symmetrical | Cue/Target | Concreteness | 4.70 | 1.38 |
|  |  | Length | 5.21 | 1.94 |
|  |  | Frequency | 3.23 | 0.67 |
| Unrelated | Cue/Target | Concreteness | 4.63 | 128 |
|  |  | Length | 5.21 | 1.52 |
|  |  | Frequency | 2.49 | 0.85 |

*Notes.* Frequency is measured using SUBTLEX word frequency measure (Brysbaert & New, 2009). Concreteness and length were taken from the English Lexicon Project (Balota et al., 2007).

Table A3

*Comparison of Mean JOL Ratings and Correct Recall Percentages across Pair Types for the JOL Group in Experiments 1-4.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Experiment | Task | Pair Type | *M* | *95% CI* | F | B | S |
| Exp. 1 | JOL | Forward | 60.87 | 3.85 |  |  |  |
|  |  | Backward | 55.18 | 4.07 | 0.45\* |  |  |
|  |  | Symmetrical | 64.84 | 3.75 | 0.33\* | 0.77\* |  |
|  |  | Unrelated | 19.43 | 4.76 | 3.00\* | 2.53\* | 3.33\* |
|  | Recall | Forward | 69.34 | 5.39 |  |  |  |
|  |  | Backward | 31.60 | 5.30 | 2.21\* |  |  |
|  |  | Symmetrical | 57.78 | 5.59 | 0.66\* | 1.51\* |  |
|  |  | Unrelated | 8.85 | 2.50 | 4.51\* | 1.72\* | 3.54\* |
| Exp. 2 | JOL | Forward | 67.62 | 3.98 |  |  |  |
|  |  | Backward | 62.18 | 4.24 | 0.39\* |  |  |
|  |  | Symmetrical | 71.92 | 4.21 | 0.31\* | 0.72\* |  |
|  |  | Unrelated | 22.24 | 4.98 | 2.99\* | 3.30\* | 3.98\* |
|  | Recall | Forward | 63.78 | 4.49 |  |  |  |
|  |  | Backward | 26.60 | 4.21 | 2.68\* |  |  |
|  |  | Symmetrical | 54.17 | 5.06 | 0.63\* | 1.85\* |  |
|  |  | Unrelated | 9.87 | 2.85 | 4.50\* | 1.46\* | 3.39\* |
| Exp. 3 | JOL | Forward | 66.25 | 5.68 |  |  |  |
|  |  | Backward | 60.15 | 5.75 | 0.36\* |  |  |
|  |  | Symmetrical | 70.14 | 5.49 | 0.23\* | 0.59\* |  |
|  |  | Unrelated | 23.94 | 8.34 | 1.99\* | 1.70\* | 2.20\* |
|  | Recall | Forward | 71.74 | 5.33 | 2.22\* |  |  |
|  |  | Backward | 35.61 | 5.71 | 0.67\* | 1.46\* |  |
|  |  | Symmetrical | 60.68 | 5.93 | 4.32\* | 1.56\* | 3.25\* |
|  |  | Unrelated | 13.41 | 3.75 |  |  |  |
| Exp. 4 | JOL | Forward | 64.03 | 4.98 |  |  |  |
|  |  | Backward | 59.69 | 5.17 | 0.26\* |  |  |
|  |  | Symmetrical | 68.54 | 5.16 | 0.28\* | 0.53\* |  |
|  |  | Unrelated | 16.77 | 4.42 | 3.11\* | 2.77\* | 3.34\* |
|  | Recall | Forward | 72.57 | 5.20 |  |  |  |
|  |  | Backward | 35.44 | 6.52 | 1.95\* |  |  |
|  |  | Symmetrical | 62.91 | 6.21 | 0.52\* | 1.33\* |  |
|  |  | Unrelated | 17.53 | 7.15 | 3.25\* | 0.80\* | 2.09\* |

*Note.* The three right-most columns indicate Cohen’s *d* effect sizes for post-hoc comparisons, \* = *p* < .05.

Table A4

*Comparisons of Mean Recall Percentages for each Pair Type in Experiments 1-4.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Experiment | Encoding Task | Pair Type | *M* | *95% CI* | F | B | S |
| Exp. 1 | JOL | Forward | 69.34 | 5.39 |  |  |  |
|  |  | Backward | 31.60 | 5.30 | 2.21\* |  |  |
|  |  | Symmetrical | 57.78 | 5.59 | 0.66\* | 1.51\* |  |
|  |  | Unrelated | 8.85 | 2.50 | 4.51\* | 1.72\* | 3.54\* |
|  | No-JOL | Forward | 48.08 | 5.21 |  |  |  |
|  |  | Backward | 16.09 | 3.30 | 2.30\* |  |  |
|  |  | Symmetrical | 36.03 | 4.97 | 0.74\* | 1.48\* |  |
|  |  | Unrelated | 9.68 | 3.16 | 2.80\* | 0.66\* | 1.99\* |
| Exp. 2 | JOL | Forward | 63.78 | 4.49 |  |  |  |
|  |  | Backward | 26.60 | 4.21 | 2.68\* |  |  |
|  |  | Symmetrical | 54.17 | 5.06 | 0.63\* | 1.85\* |  |
|  |  | Unrelated | 9.87 | 2.85 | 4.50\* | 1.46\* | 3.39\* |
|  | Relational | Forward | 58.17 | 6.69 |  |  |  |
|  |  | Backward | 30.89 | 7.56 | 1.12\* |  |  |
|  |  | Symmetrical | 50.06 | 6.73 | 0.35 | 0.78\* |  |
|  |  | Unrelated | 25.11 | 7.49 | 1.36\* | 0.22 | 1.02\* |
|  | Vowel | Forward | 39.19 | 6.72 |  |  |  |
|  |  | Backward | 14.13 | 5.68 | 1.20\* |  |  |
|  |  | Symmetrical | 29.83 | 6.37 | 0.42 | 0.78\* |  |
|  |  | Unrelated | 9.59 | 5.47 | 1.44\* | 0.24 | 1.02\* |
|  | No-JOL | Forward | 48.06 | 4.63 |  |  |  |
|  |  | Backward | 17.13 | 3.45 | 2.34\* |  |  |
|  |  | Symmetrical | 38.13 | 4.65 | 0.66\* | 1.59\* |  |
|  |  | Unrelated | 9.13 | 3.16 | 3.04\* | 0.75\* | 2.26\* |
| Exp. 3 | JOL | Forward | 71.74 | 5.53 |  |  |  |
|  |  | Backward | 35.61 | 5.75 | 2.22\* |  |  |
|  |  | Symmetrical | 60.65 | 5.93 | 0.67\* | 1.46\* |  |
|  |  | Unrelated | 13.41 | 3.75 | 4.32\* | 1.56\* | 3.25\* |
|  | JAM | Forward | 67.58 | 6.74 |  |  |  |
|  |  | Backward | 36.36 | 5.71 | 1.71\* |  |  |
|  |  | Symmetrical | 61.29 | 6.42 | 0.32 | 1.40\* |  |
|  |  | Unrelated | 14.68 | 3.65 | 3.36\* | 1.58\* | 3.08\* |
|  | No-JOL | Forward | 55.16 | 6.28 |  |  |  |
|  |  | Backward | 27.34 | 6.13 | 1.55\* |  |  |
|  |  | Symmetrical | 46.41 | 6.36 | 0.48 | 1.06\* |  |
|  |  | Unrelated | 16.95 | 4.24 | 2.28\* | 0.68\* | 1.89\* |
| Exp. 4 | JOL | Forward | 72.57 | 5.20 |  |  |  |
|  |  | Backward | 35.44 | 6.52 | 1.95\* |  |  |
|  |  | Symmetrical | 62.91 | 6.21 | 0.52\* | 1.33\* |  |
|  |  | Unrelated | 17.53 | 7.15 | 3.25\* | 0.80\* | 2.09\* |
|  | Frequency | Forward | 66.58 | 5.87 |  |  |  |
|  |  | Backward | 31.23 | 6.14 | 1.85\* |  |  |
|  |  | Symmetrical | 62.05 | 6.21 | 0.23 | 1.56\* |  |
|  |  | Unrelated | 13.34 | 4.06 | 3.31\* | 1.08\* | 2.91\* |
|  | No-JOL | Forward | 49.42 | 6.29 |  |  |  |
|  |  | Backward | 23.01 | 5.60 | 1.39\* |  |  |
|  |  | Symmetrical | 43.27 | 6.06 | 0.31 | 1.09\* |  |
|  |  | Unrelated | 14.94 | 4.09 | 2.04\* | 0.52\* | 1.72\* |

*Note.* The three right-most columns indicate Cohen’s *d* effect sizes for post-hoc comparisons, \* = *p* < .05.