Spring 5-12-2022

The Influence of Word Pair Associative Direction on Judgment of Learning Reactivity

Nicholas Maxwell
University of Southern Mississippi

THE INFLUENCE OF WORD PAIR ASSOCIATIVE DIRECTION ON JUDGMENT OF LEARNING REACTIVITY

by

Nicholas P. Maxwell

A Dissertation
Submitted to the Graduate School,
the College of Education and Human Sciences
and the School of Psychology
at The University of Southern Mississippi
in Partial Fulfillment of the Requirements
for the Degree of Doctor of Philosophy

Approved by:

Dr. Mark J. Huff, Committee Chair
Dr. Lin Agler
Dr. Alen Hajnal
Dr. Lucas Keefer
Dr. Hans Stadthagen

May 2022

Published by the Graduate School

THE UNIVERSITY OF
SOUTHERN
MISSISSIPPI®

ABSTRACT

Judgments of learning (JOLs) are commonly used by researchers to assess whether individuals can accurately predict later memory performance. While the JOL literature has generally operated under the assumption that providing judgments at study does not affect the learning process, recent studies have shown a reactivity effect in which memory differs between participants who do and do not make JOLs at study. The effects of providing JOLs on memory have been mixed: Some studies report memory improvements (i.e., positive reactivity), while others report memory costs (i.e., negative reactivity). Additionally, little work has evaluated the effects of associative direction (i.e., credit-card vs. card-credit) and list structure (i.e., mixed vs pure lists) on JOL reactivity. Across four experiments, JOLs produced a reactive effect on learning which was consistently moderated by pair relatedness. Related pairs repeatedly showed positive reactivity, while no reactivity was observed for unrelated pairs. Importantly, this pattern extended to a novel frequency judgment task, suggesting that reactivity is not unique to JOLs and instead reflects relational encoding rather than metacognitive processes. Findings from Experiments 2-4 showed that this pattern emerged regardless of whether pair types were presented in mixed lists or pure lists, indicating that exposure to different pair types is not a requisite for reactivity to occur.

ACKNOWLEDGMENTS

First and foremost, I would like to thank my faculty mentor, Dr. Mark Huff. His support and guidance have been critical to my successes as a student and as a researcher. I would also like to thank Dr. Erin Buchanan for setting me up for success in my doctoral studies. Not only did Erin get me interested in studying memory and language, she also taught me how to code (which saved me so much time completing this project!). Finally, I would like to thank former and current students in the MACA Lab who I have had the privilege of working with over the past four years. Specifically, Matthew Gretz, Kendal Smith, and Jacob Namias.

DEDICATION

To my wife, Juliah. Your constant love and support over the past four years kept me on track and helped make this possible.

TABLE OF CONTENTS

LIST OF TABLES

LIST OF ILLUSTRATIONS

CHAPTER I - INTRODUCTION

The ability for individuals to accurately monitor the progress of their own learning is critical for successful retention. Effective monitoring allows individuals to adjust their study strategies to maximize item retention at test (Nelson & Narens, 1990), while also providing insight on how best to allocate resources to optimize future learning (Soderstrom, Clark, Halamish, & Bjork, 2015; see also Bjork, 1999 for a review). Empirically, information about the learning processes can be obtained through metacognitive judgments (i.e., having individuals make judgments about their memorial abilities). While metacognitive measures have received significant attention from memory researchers (see Metcalfe, 2000; Rhodes, 2016, for historical overviews of metacognitive judgments), comparatively little work has been conducted assessing whether the act of providing judgments at study affects memory performance and, if so, what factors potentially moderate this effect. The goal of the present study was to explore these factors by (1) replicating previous findings which have shown metacognitive judgments can influence later memory (i.e., a reactivity effect), (2) testing whether reactivity is sensitive to the associative direction of the study pairs, and (3) testing the effects of study list composition (i.e., pure vs. mixed lists) on reactivity.

While metacognitive judgments can be elicited using a variety of formats, they are generally categorized as *prospective* or *retrospective* judgments depending upon the time in the memory process in which judgments are elicited (see Schraw, 2009, for an overview of judgment types and their applications). Prospective judgments (i.e., those made at encoding) can take on several forms, including judgments of learning (JOLs; i.e., having individuals rate the likelihood that they could correctly retrieve a target item if shown only the cue) and feeling of knowing judgments (FOKs; i.e., individuals report the likelihood that they will later recall an item they cannot currently remember; Metcalfe, Schwartz, and Joaquim, 1993). Retrospective judgments, on the other hand, are provided at retrieval and include confidence ratings (i.e., confidence that a retrieved item was previously studied; Huff, Meade, & Hutchison, 2011) and ease of learning judgments made at retrieval (i.e., difficulty in retrieving a memory item; Schraw, 2009). Prospective judgments therefore provide researchers with an online estimate of encoding effectiveness at study, while retrospective judgments attempt to gauge online metacognitive monitoring at test.

Although prospective and retrospective judgments are critical for determining how individuals perceive the effectiveness of their own encoding and retrieval processes, the present study focuses exclusively on prospective judgments, and specifically, those made using JOLs. Within this task, participants are presented with a set of study items (typically a cue-target pair such as *mouse* - *cheese*) and are asked to estimate the likelihood that the target (e.g., *cheese)* would be later recalled on a future test if only the cue (e.g., *mouse*) is provided. JOLs can be elicited in several ways including binary JOLs (e.g., ‘yes’ – ‘no’ responses; Hanczakowski, Zawadzka, Pasek, & Higham, 2013; Arbuckle & Cuddy, 1969, Experiment 1), Likert scale responses (Arbuckle & Cuddy, 1969, Experiment 2), and scaled JOLs, which are made using a continuous 0 to 100 scale that represents the percent likelihood that the target item would be successfully recalled at test (e.g., 100% = definitely would remember; 0% = definitely would not remember). Of the various collection methods, scaled JOLs are used most frequently as they provide an easy comparison between predicted recall (via JOLs) and the subsequent proportion of items correctly recalled at test (i.e., predicted vs actual recall performance; see Higham, Zawadzka, & Hanczakowski, 2016, for a discussion of common judgment scales used for metacognitive judgments).

Judgment of Learning Reactivity

Although researchers commonly use JOLs as a metric of metamemory, until recently, few studies have explicitly examined the effects providing JOLs on subsequent cued-recall. Most JOL studies investigate various factors that affect the accuracy of these judgments rather than their effects on memory more generally (e.g., the illusion of competence, Koriat & Bjork, 2005; the delayed JOL effect, Nelson & Dunlosky, 1991; etc.) or operate under the assumption that having participants make these judgments at encoding does not affect learning. A growing body of research, however, suggests that that JOLs are *reactive* on learning. A measure is said to be reactive whenever it draws attention to any cues or information that individuals would otherwise not attend to (Ericsson & Simon, 1993). Within the context of JOLs, reactivity refers to memorial changes that result from participants providing JOLs at encoding. Thus, the easiest way to test for JOL reactivity effects is to simply compare memory performance between participants who complete a JOL task at encoding to a separate group of control participants who do not provide judgments and instead read pairs silently (e.g., Janes, Rivers, & Dunlosky, 2018; Soderstrom, Clark, Halamish, & Bjork, 2015). Reactivity effects can manifest in two ways, depending on whether JOLs produce a benefit or cost to memory relative to the control group. Accordingly, *positive reactivity* refers to increases in memory performance as a function of making JOLs at encoding, while *negative reactivity* refers to any memory costs that may occur.

While testing for reactivity simply requires including a no-JOL control group, studies investigating JOLs often omitted this comparison. However, the lack of no-JOL controls across these studies is surprising given early evidence for reactive effects of JOLs were documented in Arbuckle and Cuddy's (1969) seminal study. In their second experiment, metacognitive judgments were elicited using a 1-5 Likert scale, and critically, participants provided judgements either during both the study and test phases or only during the test phase. Ratings made at study were framed as a JOL (i.e., subjects indicated their ability to correctly recall pairs at test), while judgments made at retrieval were framed as a confidence rating (i.e., confidence that the response provided is 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. Importantly, a positive reactivity pattern emerged in which correct recall was greater for participants who made judgements at encoding. Though the authors did not provide an in-depth discussion of these findings, they noted that making predictions did not produce a negative reactivity pattern and therefore did not interfere with recall. Of course, it is important to note that while Arbuckle and Cuddy reported that JOLs can boost recall, both the JOL and non-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) compared JOL and no-JOL groups by having participants study a list of cue-target word pairs in which half consisted of related pairs, while the other half was unrelated. Participants were then tested on their recall of the target word when presented with the cue without making additional metacognitive judgments (cf. Arbuckle & Cuddy, 1969). Overall, target recall was found to be 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, but only when study pairs were related.

In contrast to the positive JOL reactivity for related pairs in Soderstrom et al. (2015) and Janes et al. (2018), Mitchum et al. (2016) reported a divergent pattern of reactivity. In their study, participants who provided JOLs at encoding showed no difference in later recall relative no-JOL group on related pairs and produced a negative reactivity pattern relative to the no-JOL group for unrelated pairs, though it is likely that this pattern emerged due to methodological differences between their study and the one conducted by Soderstrom et al. Taken together, these studies demonstrate that providing JOLs when studying cue target pairs can induce reactivity on target learning, but the direction of reactivity has been mixed with positive or even no reactivity reported for related pairs and negative or no reactivity reported for unrelated pairs.

Mechanisms of JOL Reactivity

Several mechanisms have been proposed to account for JOL reactivity (see Mitchum et al., 2016 and Soderstrom et al., 2015). The *positive reactivity hypothesis* states that because monitoring is essential for determining the effectiveness of the learning process (e.g., Nelson & Narens, 1990), retention will benefit from additional or more effective processing that occurs as a byproduct of providing JOLs at encoding, as this additional monitoring of their study causes participants to engage more deeply with the material relative to silent reading. Because JOLs are provided for all pairs at encoding, this hypothesis predicts a global memory improvement for all items relative to a no-JOL control group. Alternatively, the *dual-task hypothesis* predicts the opposite pattern such 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).

Next, 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, when beginning a study task, participants initially set a 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 being easy. However, certain conditions may induce a change of study goal in which easier items are prioritized. For example, Metcalfe and 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, *park* - *parque*) versus more difficult pairs that did not contain the same root word (i.e., non-cognate pairs, *dog – perro*). When providing JOLs, it becomes clear to participants that not all items will be recalled equally, and participants may use these perceptions of item difficulty to shift their study goals towards mastering easier items within a study list.

Within the context of JOL reactivity on word pairs, the changed-goal hypothesis assumes that study lists will provide participants with at least two discernable pair types. This hypothesis predicts that making JOLs will induce positive reactivity for pairs perceived as easy to remember, but negative reactivity for pairs perceived as difficult to remember. This is because when individuals detect differences in difficulty between pair types, they 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 JOLs to a no-JOL group due to participant perceptions of pair difficulty.

Finally, Soderstrom et al. (2015) introduced a *cue-strengthening account,* which is based upon Koriat's (1997) cue-utilization theory. This account posits that JOLs call attention to certain intrinsic cues about study pairs (e.g., perceived difficulty, pair relatedness, etc.) and that reactivity will occur when those cues are made available at test. Within the context of cued-recall of word pairs, the act of making JOLs at encoding likely reinforces relatedness cues that are used when participants make JOLs. By further strengthening these cues, the JOL task functions akin to a generation task (e.g., Slamecka & Graf, 1978), boosting recall for pairs that receive JOLs at study. According to this account, JOL reactivity should occur whenever relatedness cues are made easily discernable (as in the case of related pairs), while no reactivity would be expected when relatedness cues are weak or nonexistent (e.g., unrelated pairs). Recent work by Myers, Rhodes, and Hausman (2020) supports this account, as they found positive reactivity on related pairs when participants completed cued-recall and recognition tests in which cues were available at test, but these patterns did not extend to free-recall testing in which relatedness cues were absent.

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 and non-JOL groups provided no support for the positive reactivity and dual-task hypotheses, showed only partial support for the changed-goal hypothesis, and fully supported a cue-strengthening account. Specifically, providing JOLs yielded a positivity effect for related pair target recall, but showed no reactivity on cued-recall of unrelated targets relative to no-JOL controls. 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 unrelated pairs.

Associative Direction and JOL Accuracy

While relatedness has been shown to affect JOL reactivity, both the strength and direction of cue-target pair associations have been shown to influence the accuracy of JOLs. For example, across three experiments, Koriat and Bjork (2005; see too Koriat & Bjork, 2006) showed that for forward pairs (e.g., credit-card), JOLs were generally accurate at predicting later recall of the target item. However, for weak forward pairs (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 were similar to those given to strong pairs, but recall was reduced as weakly related cues were less effective in aiding target retrieval. Thus, calibration between JOLs and recall was moderated by the strength of the forward cue-target pairs.

In addition to forward pairs, 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 forward pairs, backward pairs received high JOL ratings, however, recall for the target word was considerably lower relative to forward pairs. Dubbed the *illusion of competence*, this overestimation pattern has been extended to other pair types. For example, Castel et al. (2007) showed that the illusion of competence extended to identical pairs in which the cue and target match (e.g., lost-lost). More recently, Maxwell and Huff (2021) showed that the illusion of competence holds for backward pairs after controlling for lexical and semantic properties of the cue and target (e.g., word length, concreteness, etc.) and extended this pattern to symmetrical pairs (e.g., off-on). Thus, the direction of association more so 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 ability of JOLs to predict later recall. Regarding JOL reactivity, the related pairs used in most studies have been in the forward direction 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 pairs, backward pairs, 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 forward 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, which suggested that related pairs were being prioritized with additional study time.

Although Mitchum et al. (2016) showed reactivity results that were inconsistent with findings from other JOL reactivity studies (e.g., Janes et al., 2018; Soderstrom et al., 2015), it is worth noting an additional inconsistency in their data—no illusion of competence pattern emerged for backward pairs (cf. Castel et al., 2007; Koriat & Bjork, 2005; Maxwell & Huff, 2021). 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, as participants had high percentages of correct recall on both backward and unrelated pairs. 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 (2021) similarly utilized the USF norms, as in Mitchum et al., and used pairs that were identical in associative strength (0.37 in all experiments); however, a robust illusion of competence pattern was found.

A further possibility for this discrepancy is that while the 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 of 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; Criss, Aue, & Smith, 2011; Madan, Glaholt, & Caplan, 2010) and could be confounded with associative direction in these studies. Thus, given discrepancies in recall that occur due to pair direction (i.e., the illusion of competence), it remains unclear whether pair direction could moderate JOL reactivity.

The Present Study

Given the effects of associative direction on cued-recall, the present study sought to examine the association between cue-target word pairs as a means of testing potential mechanisms that contribute to JOL reactivity. This was tested in Experiment 1, which provided a replication of JOL reactivity patterns initially reported by Soderstrom et al. (2015) while controlling for lexical and semantic characteristics of cues and targets. Specifically, this experiment compared reactivity effects across four different pair types, including three types of related pairs (forward, backward, and symmetrical) and unrelated pairs. To date, no study outside of Mitchum et al. (2016) has investigated the influence of pair direction on JOL reactivity, and furthermore, no study has directly investigated reactivity effects using symmetrical paired associates.

Next, Experiments 2-4 tested the effects of list composition on JOL reactivity. With few exceptions (e.g., Janes et al., 2018; Tauber & Witherby, 2019), JOL reactivity studies have primarily presented study pairs using *mixed list* presentations in which study lists contain at least two unique pair types (e.g., a list containing a mixture of forward and unrelated pairs). However, in a *pure list*, participants study only one type of word pair (e.g., only forward paired associates) rather than multiple types of pairs that can be readily discriminated between. Thus, pure lists lack the “easy-difficult” comparison that is central to Mitchum et al.’s (2016) changed-goal hypothesis. By comparing reactivity effects between mixed and pure list presentations, Experiments 2-4 were designed to provide a direct test of the changed-goal hypothesis as well as further tests of the cue-strengthening account. Additionally, these experiments provided individual tests of each type of paired associate (forward, backward, or symmetrical) compared to unrelated pairs.

Finally, each experiment included an additional group of participants who completed a frequency judgment task at encoding in lieu of making a JOL. By encouraging participants to process the cue and target together, this task was designed to mimic the processing used by the JOL task, but without including a memory prediction. Therefore, the frequency judgment task was included to assess whether memory forecasting via JOLs is a requisite for reactivity to occur or if reactive effects can be induced via other, non-metacognitive judgment tasks that encourage participants to engage in relational encoding.

CHAPTER II - EXPERIMENT 1

The primary goal of Experiment 1 was to replicate previous JOL reactivity patterns initially reported by Soderstrom et al. (2015) and Janes et al. (2018) while also testing whether reactivity effects observed for forward pairs would extend to backward and symmetrical pairs. As such, participants studied three types of related pairs (forward, backward, and symmetrical) and a set of unrelated pairs. Importantly, Experiment 1 controlled for potential item effects that were not equated for across pair types in previous studies investigating reactivity (e.g., Soderstrom et al., 2015; Janes et al., 2018). Specifically, lexical and semantic properties such as word frequency, concreteness, and word length were matched across pair types. Related pairs were further matched on associative strength. Given that associative strength has been shown to affect cued-recall performance (e.g., Nelson et al., 2004), it was critical to ensure that related pairs were matched across associative strength measures (e.g., Forward Associative Strength; FAS; for forward pairs, Backward Associative Strength; BAS; for backward pairs, FAS/ BAS for symmetrical pairs).

Overall, it was expected that any observed reactivity would follow patterns previously reported by Double et al. (2018) and support the cue-strengthening account. Specifically, positive reactivity was expected to occur for forward pairs, and no reactivity was anticipated for unrelated pairs. Furthermore, any positive reactivity observed for forward pairs was expected to extend to backward and symmetrical pairs.

Finally, because pair relatedness is often used as a cue to inform JOL ratings when participants study cue-target pairs (Koriat, 1997), it may be the case that reactivity occurs due to JOLs encouraging participants to engage in relational encoding at study, rather than as a byproduct of participants generating a memory prediction. Because JOL reactivity patterns appear to be contingent on pair relatedness (i.e., positive reactivity is generally observed only when pairs are related), other tasks which encourage the use of relational encoding would be expected to produce similar reactivity patterns as JOLs. Further, based on the cue-strengthening account, reactivity would be expected to occur anytime the encoding task calls attention to cues that are used to inform retrieval. Because frequency judgments also encourage participants to relate cue-target pairs together, it was expected that any reactivity observed for JOLs would extend to frequency judgments.

Method

Participants

A total of 118 participants were recruited from The University of Southern Mississippi's online undergraduate psychology research pool (SONA) and received course credit in exchange for completing the study. Participants were randomly assigned to either the JOL group (n = 40), the no-JOL group (n = 39), or the frequency judgment group (n = 39). This sample was based on an a priori power analysis conducted with *G*Power 3.1* (Faul, Erdfelder, Lang, & Buchner, 2007), which indicated that 114 participants would be needed to detect small effects and interactions (d = 0.25). Participants were all native English speakers and reported normal or corrected-to-normal vision.

Materials

To create the stimuli, 180 cue-target word pairs were taken from the University of South Florida Free Association Norms (Nelson et al., 2004). These pairs consisted of 40 forward pairs (e.g., *bounce-ball*), 40 backward pairs (e.g., *ball-bounce*), 40 symmetrical pairs (i.e., pairs equivalent in forward and backward strength; e.g., *on-off*), 40 unrelated pairs (e.g., *building-cat*), and 20 unrelated non-tested buffers to control for primacy and recency effects. Pairs were equally distributed across two study lists, each of which contained 20 symmetrical, forward, backward, and unrelated pairs and 10 buffers pairs. Participants were presented with lists in two separate study-test blocks— the order of which was counterbalanced across participants. Study lists were organized such that five non-tested buffer pairs were always presented at the beginning and end of each list, with the remaining pairs randomized anew for each participant. Thus, each study block contained 90 pairs (80 tested and 10 buffer pairs).

Within each block, pair types were equated on forward and backward associative strength (FAS and BAS) using the Nelson et al. (2004) free-association norms and lexical and semantic properties including word length, SUBTLEX frequency (Brysbaert & New, 2009), and concreteness values derived from the English Lexicon Project (Balota et al., 2007). Associative strength and semantic/lexical properties for each pair type are reported in the Appendix (Tables A1-A2). Furthermore, all study blocks were matched on these properties such that mean associative overlap and lexical/semantic properties were equivalent between pair types and across study lists. For all pair types, counterbalanced versions of the study lists were used that switched the order of the word pairs (i.e., forest-tree vs. tree-forest). This allowed for greater control of item differences, particularly on forward and backward pairs, as the same items were used in both directions across counterbalances. Pair order was similarly flipped and counterbalanced across unrelated and symmetrical pairs.

The cued-recall test in each block contained all 80 cues from the original study items (minus buffers) presented in a newly randomized order for each participant. At test, participants viewed the cue item which was presented next to a question mark (e.g., cat - ?).

Procedure

Data collection occurred online using *Collector*, an open-source program for presenting web-based psychological experiments (Garcia & Kornell, 2015). Across 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 following study. 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. Participants in the frequency judgment group were instructed to rate the likelihood that the two words would appear together in everyday language and made these ratings using the same 0-100 scale used by the JOL group. Finally, participants in the no-JOL group were instructed to read each pair silently before continuing to the next pair. Study was self-paced, with participants in all groups pressing the Enter key to advance to the next pair.

After receiving their respective encoding instructions, participants began the first study list. Additionally, participants in both the JOL and frequency groups were asked to type their respective rating before advancing to the next study pair. Thus, both JOL and frequency ratings were provided concurrently with study such that these ratings were typed while the cue-target pair was displayed on the computer screen. 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, with the target replaced with a question mark (e.g., dog - ?). 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 same format as 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 (η_p^2) and Cohen's *d* effect sizes are reported for all significant analyses of variance (ANOVAs) and *t*-tests, respectively. Standard test statistics are reported for all *t*-tests; however, all comparisons hold when applying a Bonferroni correction. Additionally, for all non-significant main effects and post-hoc comparisons, a Bayesian estimate of the strength of the evidence supporting the null hypothesis is reported (Masson, 2011; Wagenmakers, 2007). In this analysis, two models are compared. In the first, a significant effect is assumed, while the second model assumes a null effect. From this analysis, a probability estimate is generated, a *p*-value termed p_{BIC} (Bayesian Information Criterion), which is an estimate of the probability that the null hypothesis is retained. This estimate is sensitive to the sample size, providing increased confidence in null effects reported.

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 encoded pairs via silent reading or provided frequency judgments. 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 JOL ratings and correct recall proportions for each pair type are displayed in Table A3, and all comparisons between correct recall proportions for the JOL, frequency judgment, and no-JOL groups are reported in Table A4. The following analyses first test for an illusion of competence pattern in the JOL group, given that this pattern has not been reported consistently in JOL reactivity studies (cf. Mitchum et al., 2016). The second set of analyses then tests for reactivity patterns as a function of associative pair direction by comparing the JOL, frequency judgment, and no-JOL groups across each of the four pair types.

First, a 4 (Pair Type: Forward vs. Backward vs. Symmetrical vs. Unrelated) × 2 (Measure: JOL vs. Recall) repeated measures ANOVA was conducted which assessed 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, $F(3, 117) = 293.33$, $MSE = 151.31$, $\eta_p^2 = .88$, in which 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, ts ≥ 2.38, ds ≥ 0.18. JOL ratings were only marginally greater than cued-recall rates (52.25 vs. 47.11), $F(1, 39) = 3.56$, $MSE = 590.62$, $p = .07$, $\eta_p^2 = .08$, $p_{\text{BIC}} = .53$, however a significant interaction confirmed the presence of an illusion of competence pattern, $F(3, 117) = 57.32$, MSE =

68.40, η_p^2 = .59. For backward pairs, JOLs greatly exceeded subsequent cued-recall rates (59.69 vs. 35.44), *t*(39) = 6.79, *SEM* = 3.69, *d* = 1.27. However, for unrelated pairs, the illusion of competence did not occur, as JOLs and recall were equivalent (16.77 vs. 17.53), *t* < 1, p_{BIC} = .86, and this equivalence was also found on symmetrical pairs, (68.54 vs. 62.91), *t*(39) = 1.69, *SEM* = 3.44, *p* = .10, p_{BIC} = .61. Finally, an underestimation pattern was found for forward pairs in which JOLs were lower than subsequent recall (64.03 vs 72.57), *t*(39) = 2.90, *SEM* = 3.04, *d* = 0.52.

Next, a 4 (Pair Type: Forward vs. Backward vs. Symmetrical vs. Unrelated) × 3 (Study Group: JOL vs. Frequency vs. No-JOL) mixed measures ANOVA was used to test for reactivity patterns in the JOL and frequency groups. An effect of Pair Type was found, *F*(3, 348) = 590.71, *MSE* = 99.13, η_p^2 = 0.84, indicating that collapsed across encoding groups, 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, η_p^2 = .12, 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, this analysis yielded a significant interaction, *F*(6, 348) = 12.34, *MSE* = 1205.07, η_p^2 = .17. Follow-up *t*-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, *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 ts ≥ 4.23, ds ≥ 0.96, except for the comparison between the JOL and frequency judgment groups, $t < 1$, $p_{\text{BIC}} = .85$. For backward pairs, correct recall in the JOL (35.44) and frequency judgment (31.23) groups were greater than the no-JOL group (23.01). All comparisons differed significantly, ts ≥ 1.96, ps < .05, except for the JOL and frequency judgment group, which did not differ, $t < 1$, $p_{\text{BIC}} = .90$. Finally, for unrelated pairs, recall rates were equivalent across the JOL (17.53), frequency judgment (13.34), and no-JOL (14.94) groups, ts ≤ 1.02, ps ≥ .31, $p_{\text{BIC}} \geq .88$. 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.

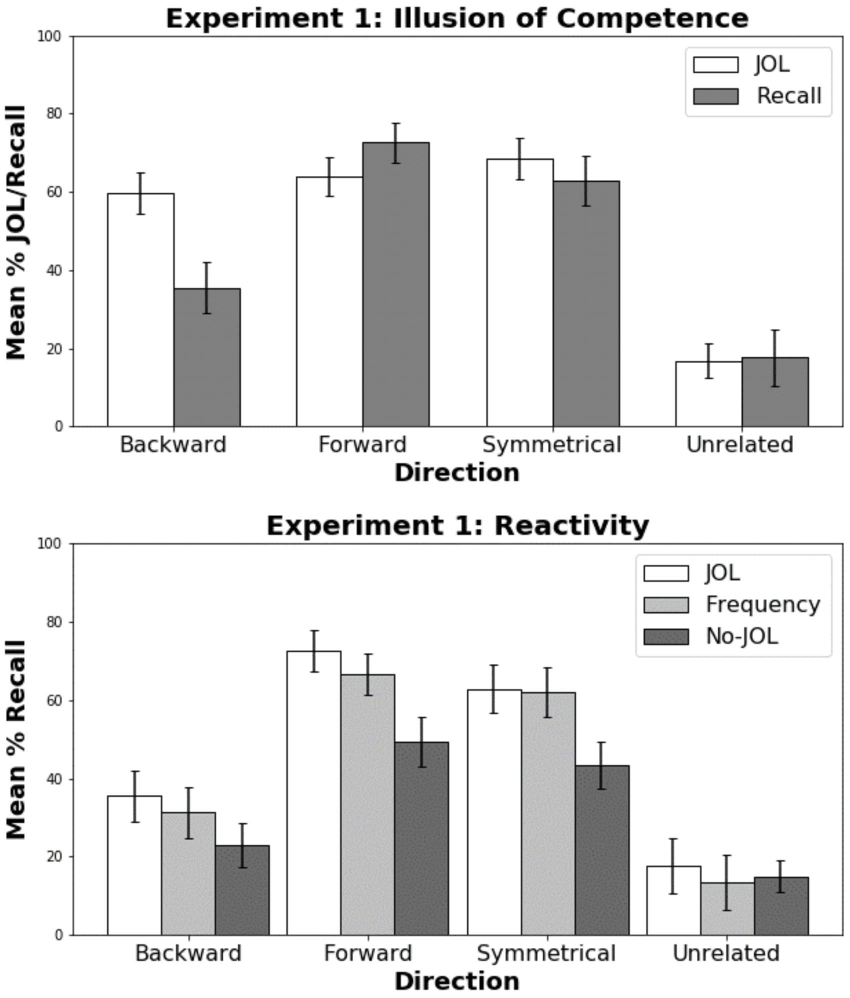

Figure 1. Experiment 1 Results

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

Discussion

The results from Experiment 1 are quite clear. First, consistent with prior JOL studies (e.g., Koriat & Bjork, 2005; Maxwell & Huff, 2021), the illusion of competence replicated for backward pairs in the JOL group. For this pair type, JOLs exceeded later recall rates, and this pattern was particularly robust, given that the cue word at test was a poor predictor of the target. The presence of the illusion of competence for this pair type indicates that JOLs were poorly calibrated to later recall. In contrast, JOLs for forward pairs, in which the cue was a strong predictor of the target at test, were better calibrated to later recall and underpredicted later recall. This pattern, however, did not extend to symmetrical and unrelated pairs, as recall of these pair types was well calibrated with JOLs. Regarding JOL reactivity, providing JOLs at study greatly increased correct recall of 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.

Second, the finding that JOL reactivity effects on related pairs generalize to different types of directional paired 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 (e.g., forward pairs). The positive reactivity patterns across related pairs and the lack of reactivity observed for unrelated pairs is therefore consistent with JOL reactivity patterns reported in other studies (e.g., Double et al., 2018; Janes et al., 2018; Soderstrom et al., 2015), and is consistent with a cue-strengthening account (Soderstrom et al., 2015) rather than the changed-goal hypothesis (Mitchum et al., 2016).

Finally, of particular interest from Experiment 1, is the finding that frequency judgments followed a similar reactivity pattern as JOLs, an observation which yields several important findings regarding reactivity effects in recall of cue-target pairs. First, the similarity between reactivity patterns for JOLs and frequency judgments suggests that the type of task employed at encoding may not be a critical factor in determining whether a reactivity pattern emerges. Instead, the qualitative processing given to the cue and target by the rating task may be more impactful. Second, 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 additional, non-JOL rating tasks. Finally, the finding that reactivity does not operate globally across all pair types suggests that reactivity processes are applied strategically, with an emphasis placed on related pairs over unrelated pairs. This point is discussed in greater depth in the General Discussion.

Mixed vs. Pure List Designs

With few exceptions, studies investigating JOL reactivity have done so using mixed-list designs in which participants study lists containing both related and unrelated pairs. A mixed-list design is central to the changed-goal hypothesis, as it states that participants' ability to discriminate between different pair types is the primary factor behind reactivity effects. Thus, this hypothesis predicts that reactivity would only occur when a mixed-list design is used, as this "easy-difficult" comparison cannot occur in a pure list in which there is only one pair type. Regarding the cue-strengthening account, however, reactivity would be expected to occur whenever the encoding task emphasizes cues used at retrieval, regardless of whether pairs are presented using mixed or pure lists. Therefore, the use of pure lists in Experiments 2-4 provided a method to test these competing accounts.

Although studies investigating reactivity effects have generally used mixed-list designs, both Janes et al. (2018) and Tauber and Witherby (2019) each included pure-group comparisons. First, Janes et al.'s (2018) Experiment 2 compared JOL reactivity effects for mixed- vs pure-list designs by having participants study (1) mixed lists of forward paired associates and unrelated pairs, (2) pure lists of forward pairs, or (3) pure lists of unrelated pairs. Overall, the authors found that positive reactivity patterns normally observed on forward pairs with mixed lists failed to emerge when a pure list was used, suggesting that reactivity effects were contingent on participants being able to discriminate between different pair types. Conversely, Tauber and Witherby (2019) showed a reactivity effect for forward pairs presented using a pure list. However, because Tauber and Witherby only used pure related lists, it remains unclear how these observed reactivity effects compare to a mixed list (i.e., whether reactivity effects would be greater when using a mixed list relative to a pure list) or whether this effect would also extend to a pure list of unrelated pairs.

Given these discrepancies, Experiments 2-4 were designed to provide further tests of list type on reactivity by comparing recall for a group of participants who studied mixed lists to separate groups of participants who studied either pure lists of only related or unrelated word pairs. In doing so, these experiments provided stronger tests of reactivity effects for each of the three related pairs used in Experiment 1 (forward, backward, and symmetrical) by presenting them alongside unrelated pairs (mixed lists) or in isolation (pure lists). First, Experiment 2 attempted a direct replication of Janes et al.'s second experiment by comparing reactivity effects for forward and unrelated pairs across mixed and pure lists. Experiments 3 and 4 then expanded upon Experiment 2 by comparing unrelated pairs to backward and symmetrical pairs, respectively. Experiments 2-4, therefore, provided three separate tests of list effects on reactivity.

Finally, because Experiment 1 showed that reactivity effects extend to other, non-metacognitive judgment tasks, each experiment included a frequency judgment comparison group. This additional comparison was included to (1) test whether the reactivity effects for frequency judgments in Experiment 1 would replicate for mixed groups and (2) test whether these judgments would continue mirror JOL reactivity pairs when made within a pure-list context.

CHAPTER III - EXPERIMENT 2

The goals of Experiment 2 were twofold. First, Experiment 2 sought to replicate positive reactivity findings for forward pairs presented in mixed lists as initially reported by Soderstrom et al. (2015). Next, Experiment 2 tested whether this pattern would extend to pure lists by comparing participants who studied pure lists of forward pairs to those who studied pure lists of unrelated pairs. Finally, consistent with Experiment 1, all list types included a group of participants who made frequency judgments at encoding. For participants completing the frequency judgment task, it was expected that any reactivity observed for JOLs would be mirrored by this task.

By comparing reactivity effects between mixed and pure lists, Experiment 2 provided an additional test of the changed-goal hypothesis. Because the changed-goal hypothesis states that reactivity occurs due to changes in participants' study goals that are triggered when they discern between easy and difficult pairs at encoding, this account predicts that reactivity should only occur for mixed lists, given that pair relatedness is commonly used as a marker of difficulty. Therefore, the changed-goal hypothesis predicts a null effect of reactivity for pure lists, regardless of whether pure lists contain related or unrelated pairs. The cue-strengthening account, however, makes no claims regarding easy/difficult comparisons. Instead, this account predicts that positive reactivity will occur on related pairs provided the encoding task emphasizes relatedness cues. Findings from Experiment 1 are consistent with this notion, as frequency judgments (which call attention to pair relatedness) mimicked JOL reactivity patterns and similarly induced positive reactivity on related pairs. Thus, if pure lists display the same reactivity patterns previously reported for mixed lists (i.e., positive reactivity for related pairs, no reactivity for unrelated pairs), this would indicate further evidence for a cue-strengthening account.

Method

Participants

A total of 347 participants were recruited to take complete Experiment 2. Participants were recruited from two sources: Undergraduate students from The University of Southern Mississippi's undergraduate psychology research pool who completed the study in exchange for course credit (*n* = 260) and individuals who were recruited through Prolific Academic (www.prolific.co) who were compensated at a rate of $3.90/half hour (*n* = 87). Of these 347 participants, 111 participants were randomly assigned to the mixed list group, which used a 3 × 2 mixed design in which pair relatedness was manipulated within subjects. The remaining 236 participants were randomly assigned to either the pure related or unrelated list groups, which employed a 3 × 2 between-subject design. For both groups, sample sizes were based on a set of a priori power analyses conducted with *G*Power 3.1*, which indicated that at least 42 participants would be needed to detect a medium effect with mixed lists (*d* = 0.50) and 158 participants would be needed to detect the same effect when analyzing pure lists. However, groups were oversampled due to an anticipated increase in participant performance variability via online data collection.

Within each list group, participants were further assigned to one of three groups based on encoding task (JOLs, frequency judgments, or silent reading). This resulted in a total of nine groups in (see Table 1 for each group's final *n* following data screening). All participants were native English speakers who reported normal or corrected vision.

Materials

To create the stimuli used in Experiment 2, 200 word pairs were generated from the University of South Florida Free Association Norms (USF norms; Nelson, McEvoy, & Schreiber, 2004). These pairs were then divided into six study lists: Two mixed lists, two pure lists of forward pairs, and two pure lists of unrelated pairs. Mixed list and pure list forward pairs were each matched on mean levels of forward associative strength (FAS) and backward associative strength (BAS). Additionally, all lists were matched on word length, SUBTLEX frequency values (Brysbaert & New, 2009), and concreteness values derived from the English Lexicon Project (Balota et al., 2007). Associative overlap measures and lexical characteristics for all stimuli are reported in Tables A1 and A5, respectively.

Following the design of Experiment 1, study pairs across lists were randomized with the constraint that five non-tested buffer pairs were always presented at the beginning and end of each study list. All participants were presented with two study lists of the same type (i.e., participants in the pure unrelated condition would only receive the two pure unrelated study lists), which were organized into two study-test blocks. Block presentation order was counterbalanced across participants. Below, the process used to create the mixed and pure lists is described in further detail.

Mixed Lists. To generate the mixed lists, 40 forward pairs (e.g., chisel-hammer) and 40 unrelated word pairs (e.g., justice-maroon) were randomly selected from the initial pool of 200 pairs. An additional 20 pairs (10 forward pairs and 10 unrelated pairs) were then selected to serve as non-tested buffer items to control for primacy and recency effects. Pairs were divided into two study lists, each consisting of 20 forward and 20 unrelated study pairs as well as 10 buffer items (five related and five unrelated). As a result, each mixed list contained a total of 50 pairs.

Pure Lists. Next, four pure lists were generated (two for each pair type). Starting with the related pure lists, each list contained 40 forward pairs, with list one consisting of the 40 pairs presented in the mixed list, and the other containing 40 forward pairs not assigned to a mixed list. The remaining 20 forward pairs served as primacy and recency buffers (10 per list). The second set of pure lists contained unrelated pairs and followed the same process used to create the related pure lists. Specifically, the first pure unrelated list used the 40 unrelated pairs presented in the mixed lists, while the second one contained 40 unrelated pairs not assigned to a mixed list. Like the related lists, the remaining 20 unrelated pairs were used as buffer items. Thus, each pure list regardless of pair type contained of 40 study pairs and 10 buffer items.

Procedure

Experiment 2 was conducted using the same equipment as Experiment 1 and followed the same general procedure, with the primary difference being the use of only forward and unrelated pairs (rather than all four pair types) and the inclusion of pure-list groups. Participants were randomly assigned to either the mixed- or pure-list groups and were then further randomly assigned to complete either the JOL, frequency judgment, or silent reading encoding tasks. In the mixed groups, participants completed two study-test blocks containing both forward and unrelated pairs, and, depending on the encoding group to which they were assigned, provided JOLs, frequency judgments, or engaged in silent reading. In contrast, participants assigned to the pure groups completed two study-test blocks that contained only forward or unrelated pairs. All encoding instructions and test instructions were identical to those used Experiment 1, including the filler task that was completed in between the study and test blocks.

Results

Figure 2 displays findings from Experiment 2. The top panel plots mean recall rates for participants who made JOLs, frequency judgments, or engaged in silent reading of mixed-list pairs The bottom panel displays mean recall rates for pure-list participants. For completeness, all comparisons between forward and unrelated pairs are provided in Table A6. Responses from 39 participants were excluded for one of the following reasons: (1) Low recall rates (e.g., correct recall rates < 5%) which suggested that participants did not correctly follow study instructions, or (2) recall rates of 100% across all blocks/pair types (which suggested participants were cheating during online testing). Additionally, data were omitted for one pure group participant due to a coding error. As a result, 307 participants were included in the following analyses (105 in the mixed-list analyses; 202 in the pure-list analyses). Final group *n*s are displayed in Table 1.

Mixed Lists

First, a 2 (Pair Type: Forward vs. Unrelated) × 3 (Study Group: JOL vs. Frequency vs. No-JOL) mixed ANOVA was used to test for reactivity effects for pairs presented via mixed lists. First, a main effect of Pair Type was found, $F(1, 102) = 1309.60$, $MSE = 99.84$, $\eta_p^2 = .93$, such that collapsed across encoding tasks, mean recall was higher for forward pairs (71.74) relative to unrelated pairs (21.69). However, the effect of Study Group was only marginally reliable, $F(2, 102) = 2.64$, $MSE = 485.32$, $p = .08$, $p_{BIC} = .88$. Importantly, a significant interaction between Pair Type and Study Group was found, $F(2, 102) = 12.41$, $MSE = 99.84$, $\eta_p^2 = .20$. Post-hoc *t*-tests indicated that for forward pairs, correct recall in both the JOL (75.59) and frequency judgment (76.68) groups exceeded that of the no-JOL group (62.98). All comparisons differed, *t*s $\geq$ 3.30, *d*s $\geq$ 0.78, except for the difference in recall between the JOL and frequency judgment groups, *t* < 1, *SEM* = 3.57, *p* = .74, p_{BIC} = .89. However, for unrelated pairs, recall rates did not statistically differ between the JOL (18.14) and frequency judgment groups (25.27) and the no-JOL (21.86) group, *t*s < 1, *p*s $\geq$ .38, p_{BIC}s $\geq$ .85, though the comparison between the JOL and frequency judgment groups was marginal, *t*(68) = 1.91, *SEM* = 3.78, *p* = .06, *d* = 0.45, p_{BIC} = .58. Thus, when pairs were presented using mixed lists, JOL ratings and frequency judgments produced statistically equivalent reactivity patterns for related pairs but produced no reactivity on unrelated pairs.

Pure Lists

A 2 (Pair Type: Forward vs Unrelated) × 3 (Study Group: JOL vs Frequency vs No-JOL) between-subject ANOVA tested whether reactivity patterns observed for mixed lists would hold when pairs were presented in a pure-list context. Overall, this analysis yielded a significant effect of Pair Type, *F*(1, 196) = 468.13, *MSE* = 262.08, η_p^2 = .70. Collapsed across encoding tasks, mean recall was higher for forward pairs (71.74) versus unrelated pairs (21.69). Next, a significant effect of Study Group emerged, *F*(2, 196) = 3.52, *MSE* = 262.08, η_p^2 = .03, such that collapsed across pair type, mean recall was highest when participants made frequency judgments (50.69), followed by the JOL (51.40) and No-JOL groups (46.65). Post-hoc testing, however, revealed no significant differences in recall between encoding groups, *t*s < 1, *p*s $\geq$ .36, p_{BIC}s $\geq$ .88.

Critically, a significant interaction emerged, *F*(2, 196) = 7.37, *MSE* = 262.08, η_p^2 = .07. Follow-up testing revealed that for forward pairs, correct recall was greater in the

JOL (83.19) and frequency judgment (77.78) groups relative to the no-JOL group (65.88). All comparisons differed significantly, *t*s ≥ 2.62, *d*s ≥ 0.70, except for the difference between the JOL and frequency judgment groups, *t*(60) = 1.36, *SEM* = 4.05, *p* = .18, p_{BIC} = .76. For unrelated pairs, correct recall did not differ between the between the JOL (23.25), frequency judgment (28.01), or the No-JOL (27.45) groups, *t*s ≤ 1.42, *p*s ≥ .16, p_{BIC} ≥ .76. Therefore, pure lists demonstrated similar reactivity patterns as mixed lists.

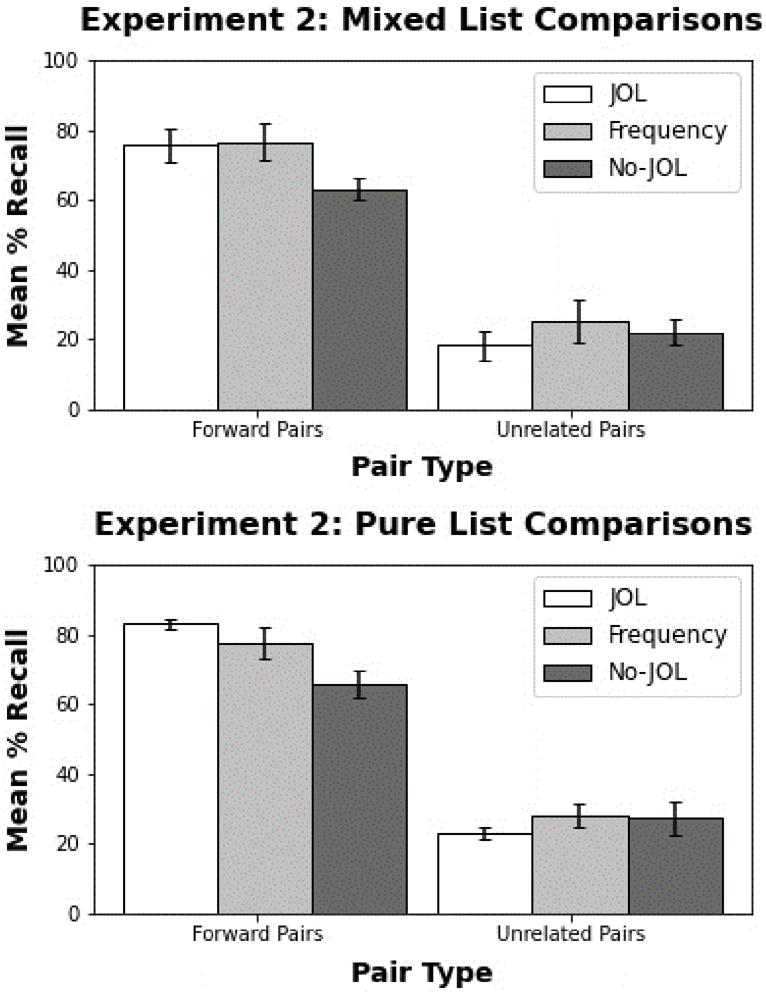

Figure 2. Experiment 2 Results.

Mean percent recall for participants in Experiment 2 who completed the JOL, frequency judgment, or No-JOL silent reading tasks for mixed lists (top panel) or pure lists (bottom panel). Error bars represent 95% confidence intervals.

Table 1 *Final Sample Sizes for all Comparison Groups in Experiments 2-4.*

Experiment	Encoding Task	Mixed	Pure Forward	Pure Backward	Pure Symmetrical	Pure Unrelated
Exp. 2	JOL	36	31	--	--	35
	Frequency	34	31	--	--	37
	No-JOL	35	34	--	--	34
Exp. 3	JOL	40	--	41	--	35
	Frequency	43	--	42	--	37
	No-JOL	37	--	37	--	34
Exp. 4	JOL	35	--	--	32	35
	Frequency	36	--	--	36	37
	No-JOL	35	--	--	35	34

Note: Cells reflect final *n*s for each group following data screening. The five left-most columns denote list type. The pure unrelated group in Experiment 2 was used as the pure unrelated comparison in Experiments 3 and 4.

Discussion

The primary goal of Experiment 2 was to test the effect of list type on reactivity. In doing so, Experiment 2 assessed reactivity effects for a group of participants who studied a mixed list of forward and unrelated pairs and tested whether these effects would extend to pairs presented in a pure-list context in which only one pair type was studied. Starting with participants in the mixed-list group, the predicted pattern of reactivity emerged. Compared to the control group, making JOLs increased correct recall of forward pairs—a positive reactivity pattern—but produced no recall benefit for unrelated pairs. This finding directly replicates previous work on JOL reactivity (e.g., Janes et al., 2018; Soderstrom et al. 2015) while also replicating JOL reactivity patterns observed in Experiment 1. Finally, reactivity patterns observed for JOLs again extended to frequency judgments, further suggesting that JOL reactivity is contingent on relational encoding rather than metamemorial or predictive processes.

Importantly, Experiment 2 showed that reactivity effects are not limited to a mixed-list design. Pure lists also showed positive JOL reactivity patterns for related pairs that mirrored mixed lists, and again, this reactivity pattern extended to frequency judgments. Because reactivity extended to pure lists, these effects are not simply the result of a comparison process (i.e., participants prioritizing easy pairs at the expense of more difficult ones as predicted by the changed-goal hypothesis). Instead, reactivity appears driven almost exclusively by pair relatedness, which further supports a cue-strengthening account (Soderstrom et al., 2015). The cue-strengthening account, however, also posits that for reactivity to occur, cues used to inform the JOL (e.g., relatedness) must be made available at test. For backward pairs (e.g., card-credit), the cue and target are related, yet the target item is an uncommon response to the cue. Thus, while backward pairs are thematically related, relatedness cues are not readily available at retrieval. As a result, it is unclear whether cue-strengthening can occur with backward pairs, given that the target item is a less obvious response to the cue.

To test this possibility, Experiment 3 compared mixed- and pure-list reactivity patterns using backward and unrelated pairs. Like forward pairs, participants assign backward pairs high JOL ratings at study (indicating that participants perceive backward pairs as related), but at test, participants struggle to correctly retrieve the target (e.g., the illusion of competence; Koriat & Bjork, 2005). Backward pairs therefore provide a situation in which the cue-target word pair appears strongly related at encoding, but cues used to inform the judgment are not readily available at test. Finally, Experiment 3 similarly included a frequency judgment group, which tested whether JOL reactivity patterns would continue to extend to this encoding task in the absence of forward pairs.

CHAPTER IV - EXPERIMENT 3

The goal of Experiment 3 was to test whether pure-list reactivity effects observed for forward pairs in Experiment 2 would extend to backward pairs. Like the previous experiment, Experiment 3 provided another test of the changed-goal and cue-strengthening accounts of reactivity. Based on the changed-goal hypothesis, positive reactivity would be expected to occur for backward pairs presented in a mixed list, given that this pair type appears easier to encode relative to unrelated pairs. However, no reactivity would be expected for pure lists, regardless of pair type. Regarding the cue-strengthening account, the presence of relatedness cues at encoding should boost recall of backward pairs compared to unrelated pairs, regardless of list type. However, because relatedness cues for backward pairs are not readily available at retrieval (i.e., the target is a less common response to the cue), any reactivity effects for backward pairs should be reduced compared to what was observed for forward pairs an Experiment 2. Finally, frequency judgments should again display reactivity patterns that mimic those found for JOLs, regardless of whether they are made for mixed or pure lists.

Method

Participants

Experiment 3 followed the same design as Experiment 2. A separate set of 253 participants were recruited and completed the experiment online. Of these participants, 204 were undergraduate students from the University of Southern Mississippi who completed the study online in exchange for course credit. The remaining 49 participants were recruited via Prolific Academic and were paid $3.90 per half-hour of participation. Of the 253 participants recruited, 127 were randomly assigned to the mixed-list group, with the remaining 126 participants assigned to the pure related list group. Finally, the 106 participants who were assigned to the pure unrelated group in Experiment 2 served as the pure unrelated comparison group. Thus, the pure-list groups contained a total of 232 participants. For both groups, sample sizes were based on Experiment 2. A sensitivity analysis conducted with *G*Power 3.1* indicated that both the mixed and pure list samples were sufficient for detecting small-medium effects and interactions (*ds* = 0.26 and 0.40, respectively).

Like Experiment 2, participants in each list group were further assigned to randomly complete one of the three encoding tasks (JOLs, frequency judgments, or silent reading). Therefore, the following analyses include a total of nine groups (see Table 1 for final group *n*s following data screening). All participants were native English speakers reporting normal or corrected vision.

Materials and Procedure

Experiment 3 used the same study lists as the previous experiment, with the following modifications. First, while the same unrelated word pairs from Experiment 2 were retained, all forward pairs (e.g., peanut-butter) were replaced with backward pairs (e.g., butter-peanut). In addition to including backward pairs within mixed lists, two pure lists containing only backward pairs were created, which provided a baseline for backward pair recall in the absence of unrelated study pairs. Study lists were identical to Experiment 2 in all other aspects including number of items, the inclusion of buffer pairs, and the study procedure.

Results

Figure 3 (top panel) displays mean recall rates as a function of encoding group for participants who studied mixed lists. The bottom panel compares mean recall for each of the pure list groups. For completeness, comparisons between pair types mixed and pure lists are provided in the Table A8. Data screening followed the same criteria used in Experiment 2, and across groups, responses from 13 participants were omitted. As a result, 120 participants were included in the mixed-list analyses, and 226 participants were included in the pure-list analyses (see Table 1 for final group *n*s).

Mixed Lists

A 2 (Pair Type: Backward vs. Unrelated) × 3 (Study Group: JOL vs. Frequency vs. No-JOL) mixed measures ANOVA was used to test for reactivity effects within mixed lists. This analysis yielded a main effect of Pair Type, $F(1, 117) = 246.79$, $MSE = 87.63$, $\eta_p^2 = .68$. Collapsed across encoding groups, cued-recall was higher for backward pairs (43.90) than unrelated pairs (24.43). The main effect of Encoding Group, however, was non-significant $F(2, 117) = 1.90$, $MSE = 600.55$, $p = .15$, $p_{BIC} = .62$, but the interaction was reliable, $F(2, 117) = 15.83$, $MSE = 87.63$, $\eta_p^2 = .22$. Post-hoc testing confirmed the presence of positive reactivity pattern for backward pairs, as recall was greatest for participants making frequency judgments (48.90), followed by participants in the JOL (46.84) and no-JOL groups (34.85). All comparisons differed significantly (*t*s ≥ 2.72, *d*s ≥ 0.62), except between the JOL and frequency judgment groups, $t < 1$, $p = .66$, $p_{\text{BIC}} = .89$. For unrelated pairs, reactivity was not in evidence as recall rates were statistically equivalent between the frequency (26.75), JOL (20.98), and no-JOL groups

(25.45; *t*s ≤ 1.68, p_{BIC}s ≥ .69). As such, reactivity patterns observed with forward pairs in mixed lists extend to backward pairs.

Pure Lists

Next, a 2 (Pair Type: Backward vs. Unrelated) × 3 (Study Group: JOL vs. Frequency vs. No-JOL) between subjects ANOVA tested whether reactivity occurred for pairs presented within pure lists. Consistent with the previous analyses, a significant effect of pair type emerged, $F(1, 220) = 42.91$, $MSE = 312.67$, $\eta_p^2 = .16$, such that recall of backward pairs (41.95) exceeded recall of unrelated pairs (26.25) when collapsing across encoding groups. However, the effect of Encoding Group was non-significant, $F(2, 220) = 2.08$, $MSE = 312.67$, $p = .13$, $p_{\text{BIC}} = .65$. Finally, the interaction between Pair Type and Encoding Group was right at the conventional level of significance, $F(2, 220) = 2.95$, $MSE = 312.67$, $p = .05$, $p_{\text{BIC}} = .44$, $\eta_p^2 = .03$, and post-hoc comparisons were carried out as originally planned. Starting with backward pairs, correct recall was highest for participants in the frequency judgment group (46.01), followed by participants in the JOL (44.21), and no-JOL groups (34.83). Post-hoc *t*-tests confirmed that all comparisons differed significantly, *t*s ≥ 2.08, *d*s ≥ 0.47, except for the comparison between JOLs and frequency judgments, $t(81) < 1$, $SEM = 4.39$, $p = .67$, $p_{\text{BIC}} = .89$. Recall of unrelated pairs did not differ as a function of encoding group (see Experiment 2). Thus, positive reactivity patterns observed for backward pairs in mixed lists extend to pure lists.

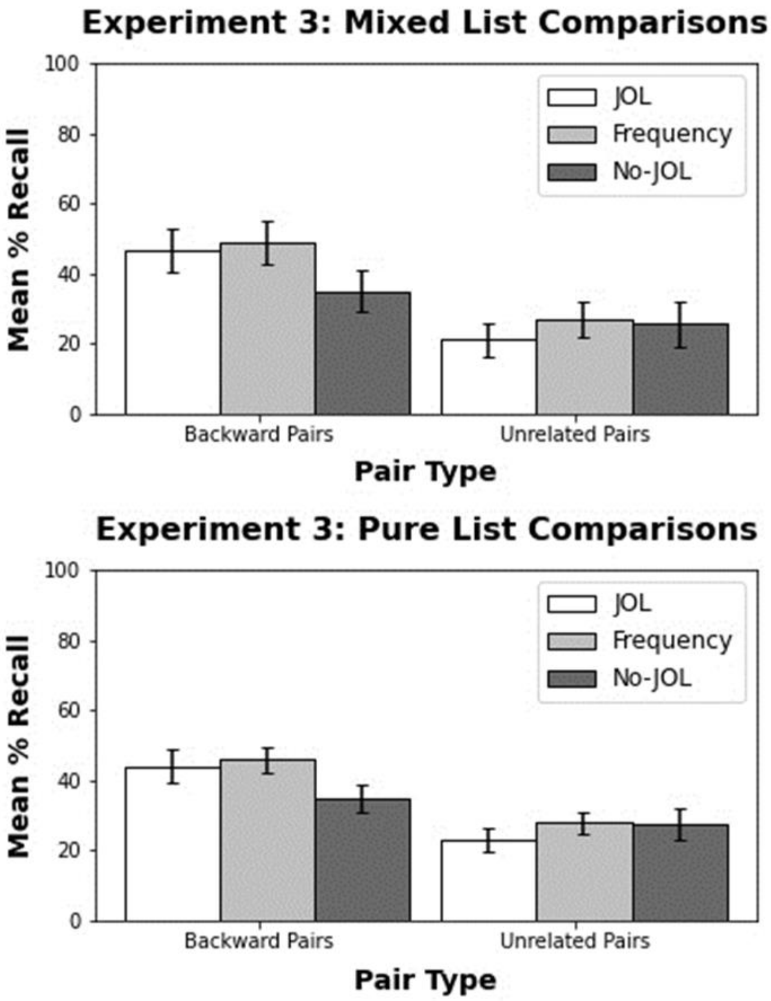

Figure 3. Experiment 3 Results.

Mean percent recall for participants in Experiment 3 who completed the JOL, frequency judgment, or No-JOL silent reading tasks for mixed lists (top panel) or pure lists (bottom panel). Error bars represent 95% confidence intervals.

Discussion

Experiment 3 tested whether reactivity patterns observed for forward pairs in Experiment 2 would also occur with backward pairs in which the target was less predictive of the cue at test. In doing so, this experiment provided an additional test of the cue-strengthening account of reactivity, as backward pairs provide a situation in which cues used to inform the JOL are less likely to be available at test. Furthermore, the inclusion of mixed vs. pure lists allowed for an additional test of the changed-goal hypothesis. Overall, JOLs and frequency judgments each produced reactivity on backward pairs, regardless of list type. For unrelated pairs, however, no reactivity occurred. These findings are consistent with the previous experiments and provide additional support for the cue-strengthening account, as reactivity was again not limited to mixed in lists in which participants could distinguish between related and unrelated pairs.

In addition to providing additional tests of the changed-goal and cue-strengthening accounts of JOL reactivity, Experiment 3 also provided a novel contribution to the reactivity literature by omitting the forward associate comparison group in favor of backward pairs. Studies investigating reactivity have largely focused on comparisons between forward and unrelated pairs (though see Mitchum et al., 2016 who included a backward comparison group), and no study investigating reactivity for related pairs has only assessed reactivity for backward pairs without also including a forward pair comparison group. Given the extensive focus in the literature on using related pairs that are forward pairs, Experiment 4 continued the pattern of isolating each related pair type used in Experiment 1 by testing for reactivity on symmetrical pairs (e.g., king-queen) relative to unrelated pairs. While backward pairs have been used in studies investigating the accuracy of JOLs (e.g., Koriat & Bjork, 2005), to date, little work on JOLs has involved symmetrical pairs (see Maxwell & Huff, 2021). Furthermore, apart from Experiment 1, no study has investigated JOL reactivity effects using symmetrical paired associates. Experiment 4 tested for reactivity effects across mixed and pure lists using symmetrical pairs. In doing so, this experiment provided an additional opportunity to test whether reactivity effects would replicate on pure lists while further testing accounts put forth to explain JOL reactivity.

CHAPTER V - EXPERIMENT 4

Experiment 4 tested whether JOL reactivity would extend to symmetrical pairs (e.g., salt-pepper) when presented in mixed lists with unrelated pairs or when presented in isolation via pure lists. Like backward pairs, symmetrical pairs can be deceptive as they contain strong backward associations. However, these pairs also contain strong forward associations, which should make them easier to learn relative to backward pairs (Maxwell & Huff, 2021). The use of symmetrical pairs in Experiment 4 is important, as it provides a novel pair type with which to test for reactivity effects. Therefore, the use of symmetrical pairs provides a further test of the changed-goal and cue-strengthening accounts while also testing the generality of JOL reactivity effects. Based on the previous experiments, findings were expected to conform to the cue-strengthening account, with positive reactivity occurring for symmetrical pairs and no reactivity for unrelated pairs. Furthermore, this pattern was expected to occur regardless of whether participants studied mixed or pure lists. Finally, frequency judgments were again expected to produce reactivity patterns mirroring JOLs.

Methods

Participants

Two-hundred twenty-seven participants were recruited to complete Experiment 4. Like the previous experiments, participants were either undergraduates recruited from the University of Southern Mississippi's psychology research pool (*n* = 187) who completed the study online in exchange for course credit or individuals recruited through Prolific Academic who completed the study online at a rate of $3.90/half hour (*n* = 40). Of these participants, 113 were randomly assigned to the mixed-list group, with the remainder randomly assigned to the pure symmetrical group (*n* = 114). Like Experiment 3, the 106 participants who studied pure unrelated lists in Experiment 2 again served as the pure unrelated comparison group. Therefore, the pure-list group contained a total of 220 participants. Group sizes were informed by the sample used in Experiment 2, and a sensitivity analysis via *G*Power 3.1* confirmed that the mixed- and pure-list groups were sufficient for detecting small-medium main effects and interactions (*ds* ≥ 0.42).

Like the preceding experiments, participants within both list groups were further assigned to either the JOL, frequency, or no-JOL encoding groups. Nine groups are included in the following analyses (see Table 1 for final group *n*s after data screening).

Materials and Procedure

Experiment 4 used a modified version of the study lists presented in Experiments 2 and 3. While the same unrelated word pairs from the previous experiments were retained, the forward/backward pairs were replaced with symmetrical pairs (e.g., king-queen). Unlike forward and backward pairs which are characterized by an asymmetrical associative relationship (i.e., from cue to target in forward pairs or vice-versa in backward pairs), symmetrical pairs contain relationships in both directions of similar associative strength. All other aspects of the study lists and the study procedure were identical to Experiments 2 and 3.

Results

Figure 4 (top panel) shows recall rates for participants who studied mixed lists as a function of encoding task, while the bottom panel displays mean recall rates for each encoding task across pure list groups. For completeness, all comparisons between related and unrelated pairs are provided in the Appendix (Table A10). Data screening followed the same procedure outlined in Experiment 2, and data from 18 participants were omitted (see Table 1 for final group *n*s).

Mixed Lists

Like the previous experiments, a 2 (Pair Type: Symmetrical vs. Unrelated) × 3 (Study Group: JOL vs. Frequency vs. No-JOL) mixed ANOVA was used to test for reactivity effects in mixed lists. This analysis revealed a significant effect of Pair Type, $F(1, 103) = 825.46$, $MSE = 112.87$, $\eta_p^2 = .89$, as recall of symmetrical pairs (65.09) exceeded recall of unrelated pairs (23.17). The main effect of Encoding Group, however, was non-significant, $F(2, 103) = 1.33$, $MSE = 497.13$, $p = .27$, $p_{\text{BIC}} = .96$. A significant interaction was found, confirming the presence of a reactivity pattern, $F(2, 103) = 12.57$, $MSE = 112.87$, $\eta_p^2 = .20$. For symmetrical pairs, mean recall was highest when participants made frequency judgments at encoding (69.34), followed by JOLs (69.33) and the no-JOL control group (56.51). Follow up *t*-tests confirmed that all comparisons differed significantly (*t*s ≥ 2.78, *d*s ≥ 0.65), except for the comparison between frequency judgments and JOLs, $t < 1$, $SEM = 3.88$, $p = .99$, $p_{\text{BIC}} = .99$. For unrelated pairs, no reactivity was observed. Mean recall did not differ between the JOL (21.24), frequency (23.46), or no-JOL encoding groups (24.80; *t*s < 1, *p*s ≥ .40, p_{BIC}s ≥ .85). Thus, reactivity patterns observed for mixed lists with forward and backward paired associates extend to symmetrical pairs.

Pure Lists

A 2 (Pair Type: Symmetrical vs. Unrelated) × 3 (Study Group: JOL vs. Frequency vs. No-JOL) between subjects ANOVA was then used to test reactivity effects for symmetrical pairs would extend to pure lists. Consistent with the previous experiments, this analysis yielded a significant effect of Pair Type, $F(1, 203) = 407.21$, $MSE = 246.60$, $\eta_p^2 = .67$. Across encoding groups, recall of symmetrical pairs (70.08) was greater than unrelated pairs (26.25). Additionally, significant effect of Encoding Group was detected, $F(2, 203) = 6.84$, $MSE = 246.60$, $\eta_p^2 = .06$, such that recall was highest for participants in the frequency judgment group (52.57), followed by the JOL (47.31) and no-JOL groups (43.39). Post-hoc tests, however, indicated that this effect was driven by differences between the frequency judgment and no-JOL groups, $t(140) = 2.09$, $SEM = 4.44$, $p = .04$, $d = 0.35$. All other comparisons were non-significant, ts ≤ 1.06, ps ≥ .29, p_{BIC}s ≥ .90. Importantly, a significant interaction was again found, $F(2, 203) = 8.12$, $MSE = 246.60$, $\eta_p^2 = .07$. For symmetrical pairs, recall was highest for participants in the frequency judgment group (77.81), followed by the JOL (73.63) and no-JOL groups (58.89). All comparisons differed significantly, ts ≥ 3.80, ds ≥ 0.85, apart from the comparison between the JOL and frequency groups, $t(66) = 1.12$, $SEM = 3.81$, $p = .26$, $p_{\text{BIC}} = .81$. For unrelated pairs, recall did not significantly differ between encoding groups (see Experiment 2). Thus, like the previous experiments, JOLs and frequency judgments again produced a positive reactivity effect on related pairs in a pure list setting.

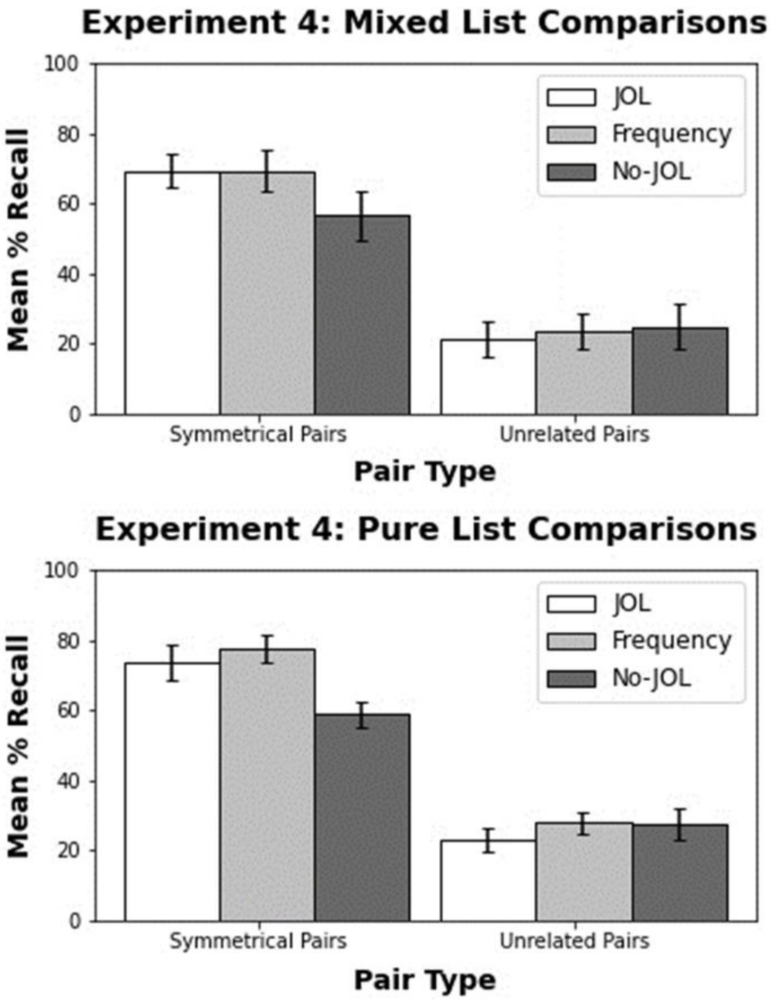

Figure 4. Experiment 4 Results.

Mean percent recall for participants in Experiment 4 who completed the JOL, frequency judgment, or No-JOL silent reading tasks for mixed lists (top panel) or pure lists (bottom panel). Error bars represent 95% confidence intervals.

Discussion

The goal of Experiment 4 was to test whether reactivity effects observed for forward and backward pairs in Experiments 2 and 3 would extend to symmetrical pairs. Overall, both JOLs and frequency judgments produced positive reactivity effects on symmetrical pairs, and as observed in the previous experiments, neither judgment type produced a reactive effect on unrelated pairs. Importantly, reactivity on symmetrical pairs occurred regardless of whether participants studied mixed or pure lists, further suggesting that reactivity is not due to context in which items are studied (i.e., easy/related vs. difficult/unrelated study materials) as posited by the changed-goal hypothesis. Therefore, findings from Experiment 4 are in-line with the previous experiments while providing additional support for the cue-strengthening account.

CHAPTER VI - GENERAL DISCUSSION

The present study provided a further test of JOL reactivity effects on cued-recall while comparing the changed-goal and cue-strengthening accounts which have often been used to explain these patterns. In doing so, this study initially investigated the effects of associative direction on JOL reactivity by including backward and symmetrical paired associates (in addition to standard forward and unrelated pairs). The remaining experiments then tested whether reactivity effects would emerge when related and unrelated pair types were studied in pure lists rather than mixed lists. A secondary goal was to test whether reactivity effects were unique to JOLs. Therefore, in addition to the standard JOL vs. no-JOL comparison that has traditionally been used to explore reactivity, each experiment included an additional group of participants who completed a frequency judgment rating task in lieu of providing JOLs. The inclusion of this group was to evaluate whether a reactivity pattern would also occur when a non-metacognitive judgment task was used.

First, Experiment 1 found positive JOL reactivity on forward pairs that was consistent with previous work by Soderstrom et al. (2015) and Janes et al. (2018), while also extending this pattern to include backward and symmetrical pairs. Importantly, these reactivity patterns occurred using word 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 each of the three related pair types indicated that the associative direction of cue-target pairs does not influence reactivity. Instead, the mere presence of association is likely sufficient to facilitate additional 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 provides further evidence that JOLs may encourage participants to selectively engage in relational encoding of related pair types, which is consistent with findings from Soderstrom et al. (2015), Janes et al. (2018), and Myers et al. (2020).

Next, Experiments 2-4 tested whether reactivity effects would still occur if pairs were presented via pure lists rather than in mixed lists. In doing so, each of the remaining experiments focused exclusively on one type of related paired associate (forward, backward, or symmetrical) and directly compared it to unrelated pairs using both mixed and pure list contexts. Starting with Experiment 2, JOLs produced a positive reactivity effect on forward pairs, regardless of whether participants studied them within the mixed or pure list setting. For unrelated pairs, however, no reactivity was observed, regardless of list type. This pattern was subsequently extended to backward and symmetrical paired associates in Experiments 3 and 4, respectively. Thus, a key finding from Experiments 2-4 is that reactivity patterns for related pairs emerge in both a mixed list context when presented alongside unrelated pairs and when presented in isolation via a pure list context.

The finding that positive reactivity extends to related pairs in pure lists provides important insights regarding JOL reactivity effects. Regarding the changed-goal hypothesis, Mitchum et al. (2016) proposed that reactivity occurs as a byproduct of participants altering their study goals as a function of pair difficulty (i.e., easy pairs are prioritized at the expense of difficult pairs). However, this account cannot explain reactivity effects in pure lists, given that pure lists lack the comparison needed to trigger a change in study goal. Therefore, pure-list reactivity findings in Experiments 2-4 do not support the changed-goal hypothesis. Regarding Soderstrom et al.'s (2015) cue-strengthening account, the extension of reactivity patterns to pure lists further supports the notion that reactivity is driven by relational encoding that is selectively applied to related pairs. As such, pure list reactivity findings from Experiments 2-4 are in-line with this account.

JOLs are not a Requisite for Reactivity

In addition to testing reactivity effects as a function of associative direction or list type, each experiment also included an additional comparison group in which participants rated the likelihood of words co-occurring together. These groups were included to test whether reactivity effects were unique to JOLs or if they would occur when participants made other, non-metacognitive judgments focusing on pair relatedness. The frequency judgment task was selected because, like JOLs, it allowed for processing of relational characteristics of study pairs without explicitly instructing participants to encode all study pairs using a relational strategy. Moreover, the frequency judgment task utilized the same rating scale as the JOL task. This task therefore resembled JOLs but did not require that participants forecast later recall performance. In doing so, this provided a novel comparison, as to date, studies investigating the reactive effects of JOLs on cue-target word pairs have not compared reactivity to other, non-metacognitive judgment tasks.

Across experiments, frequency judgments produced equivalent positive reactivity on related pairs when compared to JOLs, and critically, no reactivity was found on unrelated pairs. This finding suggests that reactivity is not a byproduct of metacognitive or predictive processes inherent to JOLs, and instead, reactivity likely reflects the use of a relational encoding strategy. Because JOLs call attention to pair relatedness (which is a strong predictor of cued-recall performance; Maxwell & Buchanan, 2020), relatedness cues may become more salient for participants making JOLs at encoding relative to those completing a silent reading task. Based on this account, reactivity would be expected to occur whenever participants complete encoding tasks that encourage the use of relational cues. Results from each experiment support this claim, as frequency judgments consistently produced similar reactivity patterns for related pairs relative to the JOL group.

The similarity of reactivity patterns between JOLs and frequency judgments suggests that both judgments tap into similar underlying relational encoding processes. Based on Koriat's (1997) cue-utilization framework, these encoding tasks 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). As a result, cued-recall performance is enhanced whenever an encoding task draws participants' attention to the relatedness between studied items. However, because this process occurs indirectly (i.e., neither the JOL nor frequency judgment tasks used in this study explicitly instructed participants to relate items together at encoding), only related items receive a memory boost when judged. Thus, reactive effects are not generally observed for unrelated items.

Finally, the finding that reactivity repeatedly occurred only when pairs were related suggests that JOLs and frequency judgments are not merely "deep" encoding tasks. Within the levels of processing framework (Craik & Lockhart, 1972), tasks that encourage deeper processing are those which encourage participants to elaborate on characteristics of items at encoding. While a deep encoding task would be expected to operate globally across all pair types (such as intentional item-specific or relational encoding instructions; e.g., Huff & Bodner, 2014), JOLs selectively affected pairs as a function of relatedness. Thus, it is evident that when making JOLs, participants do not default to the same type of processing for all pair types. While JOLs can facilitate deep encoding and can improve retention relative to silent reading, this additional processing is selectively applied as a function of pair relatedness.

A Case for Strategic Relational Encoding

As reviewed in the Introduction, Soderstrom et al. (2015) proposed that JOLs will induce reactivity whenever 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 that informed JOLs must also 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). Consistent with this account, Myers et al. (2020) extended this pattern to include recognition memory (but not free-recall), providing support for Soderstrom et al.'s first criterion that the JOL task strengthens cue-target associations that are subsequently used at retrieval. The present study provides further support for the cue-strengthening account, as across experiments, JOLs encouraged participants to engage in relational encoding, which was applied selectively to pairs as a function of pair relatedness. Furthermore, the extension of this pattern to pure lists in Experiments 2-4 provides additional evidence that reactivity effects are not context dependent. Therefore, the present study is consistent with previous studies which have indicated that JOL reactivity is found on related pairs and further establishes that the selective use of relational processing contributes to JOL reactivity.

The strategic nature of this relational encoding is consistent with previous work on metamemory and strategy use. For example, in their metamemory framework, Nelson and Narens (1990) posited that participants can adjust their encoding strategies based on cues inherent to the stimuli as participants monitor their study. Moreover, recent work by Undorf and Bröder (2020) suggests that JOLs reflect the strategic integration of a variety cues (e.g., concreteness, valence, etc.) rather than a single mnemonic cue (e.g., encoding fluency; see Koriat, 1997). However, because pair relatedness is a highly salient cue of future recall performance, it is likely that participants use relatedness cues to form the basis of their JOLs. In doing so, they adopt a relational encoding strategy which operates selectively as a function of pair relatedness. As a result, only related pairs are processed using a relational encoding strategy, as participants modify their study strategy based on the type of study pair they encounter. This results in a memory boost for related items that receive additional relational processing at encoding while unrelated pairs show no benefit.

Finally, while strategic relational encoding is evident in mixed lists, the finding that the same reactivity patterns subsequently extended to pure lists should not be interpreted as evidence against a strategy use account. First, reactivity patterns in pure lists mirrored findings from mixed lists, suggesting that for pure lists, only participants studying related lists engaged in relational processing at encoding. Thus, even without an unrelated comparison, only related pairs benefitted from the requirement to make a judgment, as pure unrelated pairs received no memory boost from JOLs relative to the control group. Second, while a strategy use account implies a comparison process, the lack of unrelated pairs within pure related lists simply means that JOLs did not strategically alter participants study strategies within that list type. Because all items in the list were related, participants simply related all pairs together.

Finally, though the present study suggests that JOLs operate selectively on related pairs, this study did not directly assess online changes in participants' study strategies. Instead, recall was compared between participants who completed JOLs and frequency judgments at encoding. Maxwell and Huff (under revision) similarly compared recall between these two tasks while also showing that JOL reactivity patterns extended related pairs which were studied via a non-strategic relational encoding task which participants were instructed to apply globally to all pair types. Unlike JOLs, this relational task also benefited recall of unrelated pairs, providing further evidence that JOLs operate strategically as a function of pair relatedness. Thus, it is likely that the JOL task implicitly encourages participants to relate items together; but only when pairs are related.

Limitations and Future Directions

The present study used cued-recall performance as the primary measure of reactivity, however, these effects may partially represent increased encoding durations for participants who completed judgment tasks at study relative to silent reading. Encoding durations, however, were mixed, with participants in the judgment groups sometimes having higher encoding latencies relative to the control group (e.g., Experiment 1) and other times lower encoding latencies (e.g., Experiment 2; see Tables A11-A12). This variability in encoding durations can likely be attributed to the online nature of the study as well as the concurrent nature of the judgment tasks. Across all experiments, participants made their JOLs/frequency simultaneously with study, rather than immediately following encoding. As a result, encoding durations in the present study represent both the time taken to encode the pair and elicit a judgment, making it difficult to separate encoding duration from the time needed to provide a judgment.

Additionally, while encoding was self-paced in the present study, previous research has used experimenter-paced study to control for potential differences in encoding durations in the JOL group (e.g., Janes et al., 2018; Soderstrom et al., 2015). These studies, however, have repeatedly shown that reactivity effects still emerge even after encoding durations are held constant between JOL and no-JOL groups. Further, Janes et al. (2018) showed that positive reactivity effects on related pairs disappeared when self-paced study was implemented. Finally, it should be noted that while useful for assessing memory, RTs provide only an indirect measure of memory performance, and encoding durations are not always informative regarding encoding effectiveness. Indeed, several studies have found that memory is greater for deep vs. shallow tasks even after controlling for encoding duration (e.g., generation: Slamecka & Graf, 1978; production: Icht, Mama, & Algom, 2014, etc.).

While prior research on JOL reactivity has largely suggested that relatedness cues are a primary factor driving reactivity effects, recent work conducted by Senkova and Otani (2021) proposed that JOL reactivity effects are not due to the use of relational encoding and instead reflect item-specific processing. According to this account, JOLs modify memory by calling attention to the item and modifying its distinctiveness. While Senkova and Otani showed that recall following JOLs was equivalent to recall for lists encoded using item-specific processing tasks (i.e., ratings of pleasantness and imagery), a methodological discrepancy between their study and the present may account for this.

Whereas most studies investigating JOL reactivity have tested for these effects using mixed lists of related and unrelated word pairs (e.g., Janes et al., 2018; Soderstrom et al., 2015), Senkova and Otani instead had their participants study lists of single words. Because participants studied single words as opposed to word pairs, participants could not access relational information from a cue to inform JOL strategy use. Instead, both the JOL and item-specific tasks operated as deep encoding tasks which participants applied universally across all items in the study list (Craik & Lockhart, 1972). Furthermore, Senkova and Otani did not compare JOLs to a relational encoding task, instead electing to only compare JOLs to item-specific encoding. Thus, it remains unclear whether a relational encoding task would produce a similar memory boost as JOLs. It is possible, therefore, that JOLs encourage participants to engage in both item-specific and relational encoding, with participants utilizing whichever information is currently available (i.e., relatedness cues when learning related cue-target word pairs).

Finally, while the present study provides further support that JOL reactivity results from participants selectively engaging in relational strategies at encoding, this study did not directly assess the type of encoding participants engaged in while providing JOLs. Instead, comparisons were made to a similar encoding task (see Huff & Bodner, 2013; Meade, Klein, & Fernandes, 2020, for similar approaches). Additionally, these 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., Hertzog & Dunlosky, 2004; Nelson & Narens, 1990), the present study did not explicitly assess whether participants were altering their 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 strategy used during study as a function of pair type, which could also indicate any encoding changes consistent with a strategy-use account.

Conclusion

Recently, metamemory researchers have become increasingly interested in whether JOLs produce a reactive effect on learning. Several theories have been proposed to explain reactivity effects, including the changed-goal hypothesis (Mitchum et al., 2016) and the cue-strengthening account (Soderstrom et al., 2015). The present study tested these two competing theories by assessing (1) whether reactivity effects would replicate for mixed lists containing four types of study pairs (Experiment 1), (2) how list composition would affect reactivity (Experiments 2-4), and (3) whether reactivity effects were unique to JOLs or if they could extend to other, non-metacognitive judgment tasks (all experiments).

In doing so, this study provided direct comparisons of both accounts of JOL reactivity and constituted the first study in which pure and mixed list contexts were directly compared for multiple types of paired associates. As such, this study was the first to include symmetrical word pairs, a type of paired associate which has received relatively little attention in the JOL literature. Finally, the present study was the first to compare reactivity for JOLs to frequency judgments. Overall, JOL reactivity effects replicated established patterns (positive reactivity for related pairs, no reactivity for unrelated pairs) and extended to pure lists. Importantly, reactivity effects also extended to frequency judgments, suggesting that reactivity is primarily driven by relational encoding, which is selectively applied to related, but not unrelated, study pairs. As such, these findings provide further support for the Soderstrom et al.'s (2015) cue-strengthening account while also providing a greater understanding of the mechanisms driving reactivity effects.

APPENDIX A

Table A1. *Summary Statistics for Associative Overlap Variables in each Experiment.*

	Pair Type	Variable	*M*	*SD*	*Min.*	*Max.*
Experiment 1	Forward	FAS	.37	.21	.05	.81
		BAS	0	0	0	0
	Backward	FAS	0	0	0	0
		BAS	.37	.21	.05	.81
	Symmetrical	FAS	.19	.13	.01	.46
		BAS	.19	.13	.02	.52
Experiment 2	Pure Forward	FAS	.37	.21	.05	.81
		BAS	0	0	0	0
	Mixed Forward	FAS	.37	.21	.05	.81
		BAS	0	0	0	0
Experiment 3	Pure Backward	FAS	0	0	0	0
		BAS	.37	.21	.05	.81
	Mixed Backward	FAS	0	0	0	0
		BAS	.37	.21	.05	.81
Experiment 4	Pure Symmetrical	FAS	.27	.18	.01	.59
		BAS	.27	.17	.01	.58
	Mixed Symmetrical	FAS	.19	.13	.01	.46
		BAS	.19	.13	.02	.52

Notes. Values are grouped by JOL condition. FAS and BAS values for unrelated pairs are not included as by definition these associations between these items have not been normed. Mean FAS and BAS values are computed by taking the average association strength for each pair.

Table A2. *Summary Statistics for Cue and Target Item Properties in Experiment 1.*

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. Values are grouped by associative direction condition. Forward and backward pairs are grouped by position within cue-target pair. Symmetrical and unrelated pairs are averaged across cues and targets, as they did not differ by position within the pairs. 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 Experiment 1.*

Task	Pair Type	*M*	± *95% CI*	F	B	S
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 Encoding Task as a Function of Pair Type in Experiment 1.*

Encoding Task	Pair Type	*M*	± *95% CI*	F	B	S
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.

Table A5. *Summary Statistics for Cue and Target Item Properties in Experiment 2.*

Pair Type	Position	Variable	*M*	*SD*
Mixed Forward	Cue	Concreteness	5.04	1.15
		Length	5.83	1.89
		Frequency	2.57	0.77
	Target	Concreteness	4.94	1.11
		Length	4.48	1.24
		Frequency	3.72	0.65
Mixed Unrelated	Cue	Concreteness	3.94	3.91
		Length	5.20	1.67
		Frequency	3.79	1.41
	Target	Concreteness	3.92	1.56
		Length	5.22	1.37
		Frequency	3.83	1.30
Pure Forward	Cue	Concreteness	4.81	1.00
		Length	5.85	1.63
		Frequency	2.49	0.65
	Target	Concreteness	4.88	1.07
		Length	4.48	1.38
		Frequency	3.73	0.63
Pure Unrelated	Cue	Concreteness	4.52	1.26
		Length	5.11	1.48
		Frequency	3.05	0.84
	Target	Concreteness	4.64	1.29
		Length	5.08	1.34
		Frequency	3.05	0.81

Notes. Values are grouped by list condition. 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 A6. *Comparisons of Mean Recall Percentages for each Encoding Task as a Function of List and Pair Type in Experiment 2.*

Encoding Task	List Type	Pair Type	*M*	± *95% CI*	U
Mixed	JOL	Forward	75.59	4.63	4.34*
		Unrelated	18.14	3.99	
	Frequency	Forward	76.68	5.11	3.05*
		Unrelated	25.27	6.18	
	No-JOL	Forward	62.98	6.01	2.00*
		Unrelated	21.86	7.50	
Pure	JOL	Forward	83.19	2.56	4.66*
		Unrelated	23.25	3.56	
	Frequency	Forward	77.78	4.60	2.96*
		Unrelated	28.01	3.27	
	No-JOL	Forward	65.88	4.11	2.08*
		Unrelated	27.43	4.66	

Note. The right-most column indicates Cohen's *d* effect sizes for Related-Unrelated comparisons, * = *p* < .05. U = Unrelated pairs.

Table A7. *Summary Statistics for Cue and Target Item Properties in Experiment 3.*

Pair Type	Position	Variable	*M*	*SD*
Mixed Backward	Cue	Concreteness	5.13	1.06
		Length	4.48	1.24
		Frequency	3.72	0.65
	Target	Concreteness	4.82	1.17
		Length	5.83	1.89
		Frequency	2.57	0.77
Mixed Unrelated	Cue	Concreteness	4.73	1.23
		Length	5.20	1.67
		Frequency	3.19	0.93
	Target	Concreteness	4.54	1.33
		Length	5.23	1.37
		Frequency	3.18	0.76
Pure Backward	Cue	Concreteness	5.03	1.13
		Length	4.45	1.27
		Frequency	3.75	0.62
	Target	Concreteness	4.88	1.22
		Length	6.17	1.86
		Frequency	2.48	0.67

Notes. Values are grouped by list condition. 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 A8. *Comparisons of Mean Recall Percentages for each Encoding Task as a Function of List and Pair Type in Experiment 3.*

Encoding Task	List Type	Pair Type	*M*	*95% CI*	U
Mixed	JOL	Backward	46.84	6.07	1.47*
		Unrelated	20.99	4.79	
	Frequency	Backward	48.90	6.20	1.18*
		Unrelated	26.75	4.97	
	No-JOL	Backward	34.85	5.96	0.49*
		Unrelated	25.45	6.47	
Pure	JOL	Backward	44.21	4.96	1.17*
		Unrelated	23.25	3.32	
	Frequency	Backward	46.01	3.76	1.16*
		Unrelated	28.01	3.04	
	No-JOL	Backward	34.83	3.97	0.40
		Unrelated	27.43	4.46	

Note. The right-most column indicates Cohen's *d* effect sizes for Related-Unrelated comparisons, * = *p* < .05. U = Unrelated pairs. Pure unrelated comparison is taken from Experiment 2.

Table A9. *Summary Statistics for Cue and Target Item Properties in Experiment 4.*

Pair Type	Position	Variable	*M*	*SD*
Mixed Symmetrical	Cue	Concreteness	4.70	1.38
		Length	5.21	1.94
		Frequency	3.23	0.67
	Target	Concreteness	4.70	1.38
		Length	5.21	1.94
		Frequency	3.23	0.67
Mixed Unrelated	Cue	Concreteness	4.73	1.23
		Length	5.20	1.67
		Frequency	3.19	0.93
	Target	Concreteness	4.54	1.33
		Length	5.23	1.37
		Frequency	3.18	0.76
Pure Symmetrical	Cue	Concreteness	4.63	1.41
		Length	5.31	1.67
		Frequency	3.24	0.74
	Target	Concreteness	4.68	1.39
		Length	5.16	1.76
		Frequency	3.17	0.71

Notes. Values are grouped by list condition. 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 A10. *Comparisons of Mean Recall Percentages for each Encoding Task as a function of List and Pair Type in Experiment 4.*

Encoding Task	List Type	Pair Type	*M*	± *95% CI*	U
Mixed	JOL	Symmetrical	69.33	4.60	3.21*
		Unrelated	21.24	5.30	
	Frequency	Symmetrical	69.34	5.86	2.76*
		Unrelated	23.46	4.97	
	No-JOL	Symmetrical	56.51	7.02	1.56*
		Unrelated	24.80	6.47	
Pure	JOL	Symmetrical	73.63	4.04	3.18*
		Unrelated	23.25	3.53	
	Frequency	Symmetrical	77.81	3.20	3.59*
		Unrelated	28.01	3.16	
	No-JOL	Symmetrical	58.89	3.51	1.81*
		Unrelated	27.42	4.62	

Note. The right-most column indicates Cohen's *d* effect sizes for Related-Unrelated comparisons, * = *p* < .05. U = Unrelated pairs. Pure unrelated comparison is taken from Experiment 2.

Table A11. *Mean Encoding Latencies as a Function of Pair Type and Encoding Task for Mixed Lists in Experiments 1-4.*

Experiment	Encoding Task	Forward	Backward	Symmetrical	Unrelated
Exp. 1	JOL	6374	7250	6980	7831
	Frequency	7380	6834	6831	8171
	Read	3045	3363	3382	2868
Exp. 2	JOL	4166	--	--	5009
	Frequency	4500	--	--	5992
	Read	6268	--	--	8150
Exp. 3	JOL	--	5527	--	4995
	Frequency	--	5444	--	5179
	Read	--	5396	--	5801
Exp. 4	JOL	--	--	5316	6470
	Frequency	--	--	4322	5310
	Read	--	--	5603	7103

Note: Cells display mean RTs in ms.

Table A12. *Mean Encoding Latencies as a Function of Pair Type and Encoding Task for Pure Lists in Experiments 2-4.*

Experiment	Encoding Task	Forward	Backward	Symmetrical	Unrelated
Exp. 2	JOL	3483	--	--	5197
	Frequency	3616	--	--	6407
	Read	5249	--	--	6376
Exp. 3	JOL	--	6398	--	5197
	Frequency	--	5743	--	6407
	Read	--	6561	--	6376
Exp. 4	JOL	--	--	5026	5197
	Frequency	--	--	4294	6407
	Read	--	--	4739	6376

Note: Cells display mean RTs in ms. Pure unrelated comparison is taken from Experiment 2.

REFERENCES

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: When agendas override item-based monitoring. *Journal of Experimental Psychology: General, 138*, 432–447.

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., Yap, M. J., Hutchsison, 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.

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.

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671-684.

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.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data (Rev. ed.)*. Cambridge, MA: Bradford Books/ MIT Press.

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.

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., & 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., Meade, M. L., & Hutchison, K. A. (2011). Age-related differences in guessing on free and forced recall tests. *Memory, 19* (4), 317-330.

Higham, P. A., Zawadzka, K., & Hanczakowski, M. (2016). Internal mapping and its impact on measures of absolute and relative metacognitive accuracy. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory*. Oxford, UK: Oxford University Press.

Icht, M., Mama, Y., & Algom, D. (2014). The production effect in memory: Multiple species of distinctiveness. *Frontiers in Psychology,* 5, 1–7.

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.

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. (2021). The deceptive nature of associative word pairs: Effects of associative direction on judgments of learning. *Psychological Research*, *85*(4), 1757-1775.

Maxwell, N. P., & Huff, M. J. (under revision). Reactivity from judgments of learning is not due to memory forecasting: Evidence from associative memory judgments and frequency judgments. Revision submitted, *Metacognition and Learning*.

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.

Metcalfe, J., Schwartz, B. L., & Joaquim, S. G. (1993). The cue-familiarity heuristic in metacognition. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 19*, 851–861.

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.

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., McEvoy, C. L., & Dennis, S. (2000). What is free association and what does it measure? *Memory & Cognition, 28* (6), 887–899

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. doi:10.3758/BF03195588

Nelson, T. O. & Dunlosky, J. (1991). When people's judgments of learning are extremely accurate at predicting subsequent recall: The "Delayed-JOL Effect." *Psychological Science, 2*(4), 267 – 270.

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.

Rhodes, M. G. (2016). Judgments of learning. In J. Dunlosky and S. K. Tauber (Eds.), *The Oxford Handbook of Metamemory* (pp. 65-80). New York: Oxford University Press.

Schraw, G. (2009). Measuring Metacognitive Judgements. In D. J. Hacker, J. Dunlosky & A. C. Graesser (Eds.), *Handbook of Metacognition in Education* (pp: 415-429). New York, NY, Routledge.

Senkova, O., & Otani, H. (2021). Making judgments of learning enhances memory by inducing item-specific processing. *Memory & Cognition, 49,* 955-967.

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.

Tauber, S. K., & Witherby, A. E. (2019). Do judgments of learning modify older adults' actual learning? *Psychology and Aging, 34*, 836-847.

Undorf, M., & Bröder, A. (2020). Cue integration in metamemory judgements is strategic. *Quarterly Journal of Experimental Psychology*, *73*(4), 629-642.

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