[TITLE HERE]

Nicholas P. Maxwell & Mark J. Huff

The University of Southern Mississippi

**Author Note**

Correspondence concerning this article should be addressed to Nicholas P. Maxwell, School of Psychology, The University of Southern Mississippi, 118 College Dr. #5025, Hattiesburg, MS 39406, United States. All study materials, data files, and *R* code used for analyses have been made available via OSF (https://osf.io/mfbnz/).

Abstract

[WORDS WILL GO HERE….]

Word count: XXX

*Keywords:* XXX; XXX; XXX; XXX

[TITLE HERE]

Metacognitive judgments are commonly used to obtain information about the learning process. These judgments can take many forms, depending on the nature of the research and the specific metacognitive processes of interest (e.g., confidence judgments at retrieval; Huff, Meade, & Hutchison, 2011, feeling-of-knowing; Metcalfe, 2000, etc.). Commonly, the judgment of learning task (JOL) is used to assess online metamemory processes at encoding. In a standard JOL task, participants study cue-target pairs (e.g., cat-dog) and are tasked with predicting the likelihood that they would successfully recall the target (e.g., dog) if shown only the cue (e.g., cat) at test. Although JOLs can be elicited using a variety of scales (see Hanczakowski, Zawadzka, Pasek, & Higham, 2013, for a review), they are commonly framed as a 0-100 percentage representing the probability that the target item would be correctly recalled at test (e.g., 100% = definitely will remember, 0% = definitely would not remember). By having participants elicit JOLs in this manner, researchers can easily assess predicted recall (measured via JOLs) and the proportion of target items correctly recalled at test.

While research on JOLs has generally assumed that having participants make these judgments at encoding does not affect learning, a growing body of research suggests that opposite, positing instead 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 generally not attend to (Ericsson & Simon, 1993). Within the context of JOLs, reactivity refers to any changes in memory performance that result from participants providing JOLs at encoding. The simplest way to assess memory changes due to JOLs is by comparing recall performance for participants who complete a JOL task at encoding to a separate group of particpants who complete a control task in which pairs are silently read rather than judged (e.g., Janes, Rivers, & Dunlosky, 2018; Soderstrom, Clark, Halamish, & Bjork, 2015). Reactivity effects can potentially manifest in two ways depending on whether the JOL task produces a memory boost or cost compared 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. Though testing for reactivity simply requires including a no-JOL control group, studies investigating JOLs have generally omitted this comparison group. Instead, researchers have either assumed that these judgments have no influence on recall or have been more interested in examining factors that influence the magnitude and accuracy of JOLs (e.g., associative direction; Koriat & Bjork, 2005; Maxwell & Huff, 2021; font-size Rhodes & Castel, 2008; etc.).

That previous JOL studies have commonly not control for JOL reactivity effects is surprising, especially considering that early studies have documented their reactive effects on memory. For example, Arbuckle and Cuddy (1969) compared recall for participants who made both JOLs at study and confidence judgments at test to a second group who engaged in silent reading at study and only made confidence judgments at test. This allowed for a comparison of recall rates between a group of participants who made JOLs at encoding versus those who engaged in silent reading. Overall, this study yielded a positive reactivity pattern such that making JOLs improved subsequent recall. However, it should be noted that all participants made confidence judgments at retrieval, regardless of whether they were assigned to the JOL or no-JOL group.

More recently, [SODERSTROM]

[MITCHUM ET AL. HERE]

[JANES, DOUBLE ET AL., MAXWELL HUFF]

**Theories of JOL Reactivity**

Various theories have been proposed to explain JOL reactivity. [MITCHUM ET AL 2016 – CHANGED-GOAL]

[SODERSTROM AND COLLEAGUES CUE STRENGTHENING]

[general pattern of reactivity]

Finally, Maxwell and Huff (under review) showed that reactivity effects not limited to JOLs and extend to other encoding tasks that promote the selective use of relational encoding at study. Both judgments of associative memory and frequency of co-occurrence judgments showed reactivity patterns that closely matched those reported for JOLs. [EXPAND – TALK EXPLICIT RELATIONAL TASK AND WHAT IT ALL MEANS]

**Mixed Lists vs Pure Lists**

With few exceptions, studies investigating JOL reactivity have generally done so using a mixed list design, such that participants are always presented with a list containing both related and unrelated pairs at encoding. According to the change-goal hypothesis, participants’ ability to discriminate between related and unrelated pairs is the driving force behind reactivity effects. Therefore, this hypothesis predicts that reactivity would only occur when using a mixed-list design, as this comparison process cannot occur in a pure list. [CUE-STRENGHENING HERE] Finally, the positive JOL and dual-task hypotheses each predict that reactivity effects would occur globally across pair types, regardless of whether the study list contains a comparison group. Thus, an easy way to test between these hypotheses is via the inclusion of pure list groups in which participants only study one pair type. Finally, because Soderstrom et al.’s (2015) cue-strengthening account assumes that reactivity effects operate strategically, our inclusion of pure lists in each experiment provided a test of the strategy use component of this account.

While studies investigating reactivity effects have generally used mixed list designs, we note that both Janes et al. (2018) and Tauber and Witherby (2019) each included pure group comparisons. First, Janes et al. (2018) compared JOL reactivity effects for mixed vs pure list designs by having participant either study mixed lists of forward associates and unrelated pairs, pure lists of forward associates, or pure lists of unrelated pairs. Overall, the authors found that positive reactivity effects found for forward associates when using mixed lists failed to emerge when using a pure list presentation, 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 associates presented using a pure list. However, because all of their study lists were pure lists of related pairs, 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, the present study was designed to provide additional tests of list type effects on reactivity by comparing recall for a group of participants who studied mixed lists wherein related and unrelated pair types were presented in the same study list to participants who studied pure lists of only related or unrelated word pairs. In doing so, [ASSOCIATIVE DIRECTION STUFF] First, Experiment 1 [FORWARD], [DETAIL 2 AND 3 HERE] Finally, because Maxwell and Huff (under review) showed that reactivity effects extend to other judgment tasks, each experiment included a frequency judgment comparison group. This allowed us to test whether reactivity effects for frequency judgments would replicate for mixed groups and whether these judgments would continue to show reactivity patterns that matched JOLs for pure groups.

**Experiment 1: Forward versus Unrelated Pairs**

The goals of Experiment 1 were twofold. First, we sought to replicate positive reactivity findings for related pairs presented in mixed lists as initially reported by Soderstrom et al. (2015). We then tested whether this pattern would extend to pure lists by comparing participants who studied pure lists of forward associates to those who studied pure lists of unrelated pairs. Finally, across all list types, we include a group of participants who make frequency judgments at encoding. Like JOLs, frequency judgments… [EXPAND]. However, [NO PREDICTIVE COMPONENT]. [WHY WE INCLUDE AND WHAT’S EXPECTED]

By comparing between mixed and pure lists, Experiment 1 provided tested both the changed-goal hypothesis and the strategic nature of the cue-strengthening account. Because the changed-goal hypothesis states that reactivity results from changes in study goals that are triggered by participants discerning between related and unrelated pairs at encoding, this account predicts that reactivity would only occur for mixed lists. This is because pure lists lack the easy/difficult comparison that is required to trigger the change in study goal. Thus, the changed-goal hypothesis predicts a null effect of reactivity for pure lists, regardless of relatedness.

Regarding the cue-strengthening account, our use of pure lists allowed us to test whether reactivity effects result from the use of strategic processes at encoding. While Soderstrom et al. (2015) made no claims regarding the strategic nature of JOL reactivity when proposing this account, Metamemorial processes are generally assumed to operate strategically (e.g., Nelson & Narens, 1990). Findings from Maxwell and Huff (under review) are consistent with this notion, as judgments of associative memory (JAM; Maki, 2007) and frequency of co-occurrence judgments mimicked JOL reactivity patterns. Thus, reactivity will emerge whenever the encoding task calls attention relatedness cues, and participants use these cues to inform monitoring. However, any strategic processes based on relatedness should only occur within the context of a mixed list (i.e., participants cannot selectively prioritize related pairs over unrelated pairs when lists only contain related pairs). If participants are using relatedness cues to strategically adjust monitoring, reactivity should only be observed using mixed lists. Alternatively, if reactivity is contingent solely on the presence of relatedness, pure lists should mimic the relatedness pattern observed for mixed lists (i.e., positive reactivity when lists contain related pairs, no reactivity for lists of unrelated pairs).

**Methods**

**Participants**

228 participants were recruited from The University of Southern Mississippi’s undergraduate psychology research pool (SONA) and completed Experiment 1 online in exchange for course credit. Additionally, 78 participants were recruited through Prolific Academic (www.prolific.co) and were compensated at a rate of $3.90/half hour.1 An a priori power analysis conducted via *G\*Power 3.1* (Faul, Erdfelder, Lang, & Buchner, 2007) determined that this sample size would be sufficient to detect medium main effects and interactions (*d* = 0.50) for both the mixed list and pure list analyses. Participants were randomly assigned to one of nine groups based on list composition (mixed list, pure forward pairs, pure unrelated pairs) and encoding tasks (JOLs, frequency judgments, or silent reading; see Table 1 for each group’s final *n* following data screening). All participants were native English speakers and reported normal or corrected vision.

**Materials**

200 word pairs were generated using the University of South Florida Free Association Norms (USF norms; Nelson, McEvoy, & Schreiber, 2004). Pairs were divided into six study lists: Two mixed lists, two pure lists of forward associates, and two pure lists of unrelated pairs. Mixed list and pure forward associate lists were 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). Lexical and associative overlap measures for all stimuli are reported in Tables A1-A2.

Study pairs within all lists were randomized with the constraint that five buffer pairs were always presented at both 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, we detail the process used to create the mixed and pure lists.

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

***Pure Lists*** We then created four pure lists (two for each pair type). Starting with the related pair lists, each list contained 40 forward associates (consisting of the 40 pairs presented in the mixed list as well as 40 pairs not assigned to a mixed lit) that were evenly split into two lists of 40 pairs each. The remaining 20 forward associates served as primacy and recency buffers (10 per list). The second set of pure lists consisted of unrelated pairs and contained the 40 unrelated pairs used in the mixed list and 40 unrelated pairs not assigned to a pure list. As with the related lists, the remaining 20 unrelated pairs were used as buffer items. Thus, each pure list regardless of type contained of 40 study pairs and 10 buffer items.

**Procedure**

Experiment 1 was conducted online using *Collector*, an open-source program for presenting web-based psychological experiments (Garcia & Kornell, 2015). Across all encoding groups, participants were informed that they would be studying a set of cue-target word pairs and following study, their memory for the target item would be tested. Participants assigned to the JOL group received additional directions to rate their likelihood of correctly recalling the target word at test if prompted by only the cue. Participants in the JOL group were instructed to make judgments on a 0 to 100 scale corresponding to their percent likelihood of correct retrieval at test (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). For participants in the frequency judgment group, additional instructions were provided to rate the likelihood that the two words would appear together in everyday language. These judgments were elicited using the same 0-100 scale used by the JOL group. Finally, participants in the no-JOL control group were instructed to silently read each pair before moving to the next one.

After receiving encoding instructions, participants were presented with the first study list. Study was self-paced for each group, and participants pressed the Enter key to move to the next pair. For participants in the JOL and frequency judgment groups, judgments were made concurrently at study, such that participants typed their respective rating while the pair was displayed on the computer screen. Participants in these groups advanced to the next study pair after providing their rating. Following the conclusion 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. After completion of the filler task, participants completed a cued-recall test in which the cue word from each of the previously studied items was presented paired with a question mark (e.g., *cheese* ­- ?). Participants were instructed to type the target item that was originally paired alongside the cue. Participants were asked to press the Enter key to advance to the next pair if they struggled to recall the correct target. After completion of the first cued-recall test, participants immediately began the second study/test block, which used the format as the first. Following the completion of the second cued-recall test, participants were fully debriefed. Experimental sessions took approximately 30 minutes to complete.

Participants were randomly assigned to either the mixed or pure list groups and were further randomly assigned to complete either the JOL task, frequency judgment task, or to read pairs silently at encoding. In the mixed groups, participants studied two blocks of items that contained both forward and backward pairs and, depending on the encoding group to which they were assigned, either 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.

**Results**

Figure 1 displays findings from Experiment 1. The top panel plots mean recall rates for participants who made JOLs, frequency judgments, or engaged in silent reading of mixed list pairs, while the bottom panel displays mean recall for participants who were presented with the pure lists. For completeness, all comparisons are provided in Table A3. For all analyses, significance was set at the *p* < .05 level. We report partial eta-squared (*ηp*2)and Cohen’s *d* effect sizes for all significant analyses of variance (ANOVAs) and *t*-tests. Additionally, for all non-significant comparisons, we conducted further analyses assessing the strength of evidence supporting the null hypothesis via Bayesian estimates (Masson, 2011; Wagenmakers, 2007). This analysis compares two models. The first assumes a significant effect, while the second assumes a null effect. This analysis generates a probability estimate (termed *p*BIC; Bayesian Information Criterion), which represents the probability of the null hypothesis being retained. *P*BICs are advantageous as the analysis is sensitive to the sample size, providing increased confidence in null effects reported. This analysis is therefore supplementary and used to confirm null effects detected with standard null-hypothesis-significance testing.

Responses from 25 participants were excluded for one of the following reasons: Low recall rates (e.g., correct recall rates < 5%) which suggested that participants did not correctly follow study instructions or for recall rates of 100% across all blocks/pair types (which suggested participants were cheating during online testing). Additionally, data were omitted for one participant in the pure related JOL group due to a coding error. As a result, 278 participants were included in the following analyses (93 in the mixed list analyses and 185 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 measures ANOVA was used to test for reactivity effects for pairs presented via mixed lists. A main effect of Pair Type was detected, *F*(1, 90) = 1186.80, *MSE* = 94.96, *ηp*2 = .93, such that collapsed across encoding tasks, mean recall was higher for forward pairs (72.12) relative to unrelated pairs (22.52). However, the effect of Study Group was non-significant, *F*(2, 90) = 1.74, *MSE* = 94.96, *p* = .18, *pBIC* = .94. Importantly, this analysis yielded a significant interaction between Pair Type and Study Group, *F*(2, 90) = 11.75, *MSE* = 94.96, *ηp*2 = .21. Post-hoc *t*-tests indicated that for forward pairs, correct recall in both the JOL (76.62) and frequency judgment (75.82) groups exceeded that of the no-JOL group (63.48). All comparisons differed, *t*s ≥ 2.78, *d*s ≥ 0.70, except for the difference in recall between the JOL and frequency judgment groups, *t*(61) < 1, *SEM* = 3.71, *p* *=* .85, *p*bic = .89. However, for unrelated pairs, recall rates did not statistically differ between the JOL (19.37), frequency judgment (25.73), and no-JOL (22.80) groups, *t*s ≤ 1.57, *p*s ≥ .12, *p*bic ≥ .70. 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**

Next, a 2 (Pair Type: Forward vs Unrelated) × 3 (Study Group: JOL vs Frequency vs No-JOL) between subjects ANOVA tested whether these reactivity patterns would replicate when participants studied pairs presented with pure lists. Overall, this analysis yielded a main effect of Pair Type, *F*(1, 180) = 568.22, *MSE* = 225.02, *ηp*2 = .76. Accordingly, collapsed across encoding tasks, mean recall was higher for forward pairs (76.43) versus unrelated pairs (23.87). Additionally, a significant effect of Study Group emerged, *F*(2, 180) = 6.92, *MSE* = 225.02, *ηp*2 = .07, such that mean recall was highest when participants made frequency judgments (53.79), followed by particpants in the JOL (52.54) and No-JOL groups (43.41). Post-hoc testing, however, indicated that this effect was largely driven by differences between participants in the Frequency and the No-JOL groups, which were marginally significant, *t*(122) = 1.97, *SEM* = 5.25, *p* = .05, *pBIC* = .62. All other comparisons were non-significant, *t*s ≤ 1.64, *p*s ≥ .10, *p*bic ≥ .74.

Critically, a significant interaction emerged between Pair Type and Study Group *F*(2, 180) = 7.63, *MSE* = 225.02, *ηp*2 = .08. Follow-up testing revealed that for forward pairs, correct recall in the JOL (83.19) and frequency judgment (80.65) groups was higher relative to the no-JOL group (65.08). All comparisons differed significantly, *t*s ≥ 3.65, *d*s ≥ 0.93, except for the difference between the JOL and frequency judgment groups, *t*(60) < 1, *SEM* = 3.42, *p* *=* .45, *p*bic = .85. For unrelated pairs, correct recall did not differ as function of encoding group, as recall rates were statistically equivalent between the JOL (20.88), frequency judgment (26.94), and No-JOL (23.71) groups, *t*s ≤ 1.61, *p*s ≥ .11, *p*bic ≥ .68. Thus, pure lists demonstrated the same reactivity patterns as mixed lists.

**Discussion**

The goal of Experiment 1 was to test the effects of list type on reactivity. We compared reactivity effects between participants who studied a mix of forward and unrelated pairs and those who studied pure lists of only forward associates or only unelated pairs. First, for participants in the mixed list group, the predicted pattern of reactivity emerged. Compared to the control group, making JOLs resulted in increased correct recall of forward associates—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).Furthermore, this reactivity pattern extended to frequency judgments, replicating previous findings by Maxwell and Huff (under review) and providing further evidence that JOL reactivity is contingent on relational encoding rather than metamemorial or predictive processes.

Next, these reactivity patterns replicated using pure lists. JOLs and frequency judgments again boosted recall of related pairs, while having no effect on recall of unrelated pairs. Thus, pure lists showed identical reactivity patterns to mixed lists, regardless of whether participants made JOLs or frequency judgments at encoding. This similarity in reactivity patterns between mixed and pure lists yields an important finding regarding reactivity effects on cued-recall. Specifically, the observation that reactivity for pure lists mimics that of mixed lists suggests that reactivity does not occur as the result of participants prioritizing easy pairs at the expense of more difficult ones (e.g., the changed-goal hypothesis). Instead, reactivity effects appear to be driven almost exclusively by pair relatedness. This account is in line with the cue-strengthening account (Soderstrom et al., 2015), which suggests that reactivity effects will only emerge if the JOL task strengthens cues used to inform the JOL (i.e., pair relatedness) and those same cues used to inform the JOL must then be available at test (i.e., a cued-recall test in which the desired target can be triggered by the cue). However, because these processes are assumed to operate strategically (i.e., at encoding, participants adjust their study strategies based on salient cues inherent to the stimuli such as relatedness; Nelson & Narens, 1990), the emergence of reactivity for pure list related pairs suggests instead that this effect is driven primarily by enhanced processing due to relatedness (i.e., relational encoding), rather than any adjustments in study strategy that occur due to differences in pair relatedness.

Since previous work on JOL reactivity has primarily focused on the differences between paired-associates presented in the forward direction (e.g., credit-card) and unrelated study pairs (e.g., credit-leaf; see Soderstom et al., 2015), Experiment 2 examined recall performance for backward paired-associates (e.g., card-credit) and unrelated pairs. Like forward pairs, participants rate their ability to remember backward pairs highly (i.e., they are perceived as being related and thus easy to recall and thus backward pairs receive high JOLs at study; see Maxwell & Huff, 2021). However, at test, participants perform poorly on these pairs relative to forward associates, since target items in backward pairs are low probability responses (e.g., the illusion of competence; Koriat & Bjork, 2005). Thus, backward pairs provide a situation in which the cue-target word pair is related, but related cues used at encoding are not made available at test. Our use of backward associates allowed us to test the second claim from Soderstrom et al.’s (2015) cue-strengthening account, which states that the cues used to inform the JOL must be available at test for reactivity to occur. Finally, consistent with Experiment 1, we included a frequency judgment group to serve as an additional point of comparison.

**Experiment 2: Backward vs Unrelated Pairs**

The previous experiment showed that positive reactivity effects for related pairs occur for related pairs presented in pure lists. However, because Experiment 1 used only forward associates, it is unclear whether associative direction moderates this effect. Specifically, based-on the cue-strengthening account, positive reactivity for related pairs would only be expected to occur when the cue is predictive of the target as in forward pairs, but not when the cue is a poor predictor of the target, such as in backward pairs. Alternatively, the changed-goals hypothesis predicts that backward pairs would still produce a reactive effect, as the perceived relatedness between the cue and target should still produce a change in study. Because backward associates still contain relatedness cues at encoding that are not available for unrelated pairs, we expected some reactivity to occur for this pair type. However, given that these cues are not made readily available at test, reactivity effects should be smaller for backward associates compared to forward associates. For unrelated pairs, we again expected that reactivity would not occur. Given that negative reactivity predicted by the changed-goal hypothesis has consistently failed to emerge. Finally, frequency judgments should again display reactivity patterns that mimic those found for JOLs.

**Methods**

**Participants**

Experiment 2 followed the same design used in Experiment 1. A total of XX participants were recruited from the University of Southern Mississippi psychology research pool and completed the study online in exchange for partial course credit. Because backward pairs generally show smaller recall effects relative to forward pairs (e.g., Koriat & Bjork, 2005; Maxwell & Huff, 2021), we increased participant recruitment relative to Experiment 1. Accordingly, an a priori power analysis conducted with *G\*Power 3.1* indicated that this provided a sufficient sample size for detecting small effects and interactions, *d* = 0.25. Participants were randomly assigned to one of six groups (the mixed JOL group, the mixed frequency group, the mixed no-JOL group, the pure backward JOL group, the pure backward frequency group, and the pure backward no-JOL group; *n*s are reported in Table 1). For comparison purposes, the three pure unrelated groups (JOL, frequency judgment, and no-JOL) from Experiment 1 served as pure list comparison groups for unrelated pairs.

**Materials**

Experiment 2 used the same study lists as the previous experiment, with the following modifications. First, while the same unrelated word pairs that were used in the previous experiment were retained, the forward associates were replaced with backward associates (e.g., butter-peanut). This pair type is characterized by a strong associative relationship when reading the pair from the target to the cue, but not from the cue and the target. Backward associates were generated by [FLIPPING FORWARD PAIRS IN EX 1] Each mixed study contained 10 buffer items, 20 backward associates, and 20 unrelated study pairs. Additionally, two pure lists each containing only backward associates were created, providing a baseline for backward pair recall in the absence of unrelated study pairs. Each pure list contained 50 pairs (10 buffer items and 40 backward associate study pairs), with each pure list consisting of the 20 backward pairs presented in the mixed groups and an additional 20 backward pairs not presented in the mixed lists.

**Procedure**

The procedure used in Experiment 2 was identical to that of Experiment 1 with the exception that all forward pairs were replaced with backward pairs.

**Results**

The top panel of Figure 2 displays mean recall rates for participants who studied mixed lists split by encoding task. The bottom panel compares mean recall for each of the pure list groups. For completeness, all comparisons for both mixed and pure lists are provided in Appendix Table AX.

**Mixed Lists**

First, we tested for reactivity in the mixed list group using a 2 (Pair Type: Backward vs Unrelated) × 3 (Study Group: JOL vs Frequency vs No-JOL) mixed measures ANOVA. A [MAIN EFFECT OF PAIR TYPE]. [MAIN EFFECT OF STUDY GROUP] [INTERACTION]

**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. [MAIN EFFECT OF PAIR TYPE]. [MAIN EFFECT OF STUDY GROUP] [INTERACTION]

**Discussion**

[WORDS HERE]

**Experiment 3: Symmetrical vs Unrelated Pairs**

Finally, Experiment 3 tested whether JOL reactivity also occurs for symmetrical pairs (e.g., salt-pepper) and unrelated pairs as an additional test of whether JOL reactivity emerges for item pairs that tend to be judged highly but poorly remembered. 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 (see Maxwell and Huff, 2021, which supports these patterns with symmetrical pairs). The use of symmetrical pairs in Experiment 4 is important, as it provides a novel pair type with which to test for reactivity effects, providing an additional test of the generality the various accounts put forth to explain reactivity effects. As such, the use of symmetrical pairs provides further tests of the changed-goal, positive reactivity, and cue-strengthening accounts while also testing the generality of JOL reactivity effects. Based on both the changed-goal hypothesis and cue-strengthening account, symmetrical pairs presented within a mixed list context should receive a memory boost, while symmetrical pairs presented in pure lists should not. Furthermore, based on the previous experiments, negative reactivity is not expected to occur for unrelated pairs. Finally, frequency judgments should again display the same pattern of reactivity as JOLs.

**Methods**

**Participants**

XX participants were recruited from the University of Southern Mississippi psychology research pool and completed the study online. Participants were randomly assigned to one of six groups (the mixed JOL group, mixed frequency group, the mixed no-JOL group, the pure symmetrical JOL group, pure symmetrical frequency group, and the pure symmetrical no-JOL group; see Table 1 for each group’s *n*’s). Following the design of Experiment 2, each of the pure unrelated groups collected in Experiment 1 served as the pure unrelated comparison groups.

**Materials**

This experiment used a modified version of the study lists presented in Experiments 1 and 2. While the same unrelated word pairs that were used in the previous experiment were retained, the forward/backward associates were replaced with symmetrical associates (e.g., king-queen). Unlike forward and backward associates which are characterized by a strong, asymmetrical associative relationship (i.e., from cue to target), symmetrical pairs have strong forward and backward relationships that occur in both directions. Thus, these pairs have levels of FAS and BAS that are approximately equivalent. As with the previous experiments, pairs were divided amongst mixed and pure study lists. Each mixed study list contained a total of 50 items (10 buffer items, 20 symmetrical associates, and 20 unrelated study pairs). Two pure lists each containing only symmetrical associates were also created, providing a baseline for symmetrical pair recall in the absence of unrelated study pairs. Each pure list consisted of 50 pairs (10 buffer items and 40 symmetrical associate study pairs). Each pure list contained the 20 symmetrical pairs presented in the mixed groups and an additional 20 symmetrical pairs not presented in the mixed lists. Consistent with Experiment 3, unrelated pure group comparisons were made using the pure unrelated group collected in Experiment 2.

**Procedure**

The procedure in Experiment 3 was identical to that used in Experiments 1 and 2 with the exception that all related pairs were symmetrical associates.

**Results**

Figure XX (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 across pure list groups. For completeness, all comparisons are provided in Table XX. First, for particpants who studied mixed lists, a 2 (Pair Type: Symmetrical vs Unrelated) × 3 (Study Group: JOL vs Frequency vs No-JOL) mixed measures ANOVA was used to test for reactivity effects. A [MAIN EFFECT OF PAIR TYPE]. [MAIN EFFECT OF STUDY GROUP] [INTERACTION]

Next, a 2 (Pair Type: Symmetrical vs Unrelated) × 3 (Study Group: JOL vs Frequency vs No-JOL) between subjects ANOVA tested whether reactivity occurred for pairs presented within pure lists. [MAIN EFFECT OF PAIR TYPE]. [MAIN EFFECT OF STUDY GROUP] [INTERACTION]

**Discussion**

[WORDS HERE]

**General Discussion**

[WORDS HERE]

**Conclusion**

[WORDS HERE]

**Open Practices Statement**

The data for all experiments have been made available at [LINK]. None of the experiments were preregistered.

References

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

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

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

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.

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., & 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.

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*, 1757-1775.

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.

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

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

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.

Footnotes

1 Due to the COVID-19 pandemic, data collection was shifted online to Prolific partway through Experiment 1. These participants were randomly assigned to one of the 9 (3 encoding tasks × 3 list types), and following data screening, the final dataset included 69 participants who were recruited via Prolific. For mixed lists, 7 Prolific participants completed the JOL task, 8 completed the frequency judgment task, and 3 were assigned to the no-JOL control group. For pure related lists, 8 were included in the JOL group, 7 in the frequency judgment group, and 6 in the no-JOL group. Finally, for unrelated pure lists, 19 Prolific participants completed the JOL task, 7 completed the frequency task, and 4 were included in the no-JOL group. A set of 2 (Pair Type) × 3 (Study Group) × 2 (Recruitment Platform) ANOVAs revealed no main effect of recruitment platform for mixed lists (*F*(1, 87) = 0.58, *MSE* = 531.29, *p* = .45, *pBIC* = .99) or pure lists (*F*(1, 174) = 1.22, *MSE* = 230.20, *p* = .27, *pBIC* = .88). Furthermore, all interactions with recruitment platform across both analyses failed to reach significance, *F*s ≤ 1, *p*s ≥ .68, *pBIC*s≥ .98.

[TABLE 1]

[FIGURE 1]

[FIGURE 2]

[FIGURE 3]

**Appendix**

Table A1

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
|  | Pair Type | Variable | *M* | *SD* | *Min.* | *Max.* |
| Experiment 1 | 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 2 | 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 3 | Pure Symmetrical | FAS |  |  |  |  |
|  |  | BAS |  |  |  |  |
|  | Mixed Symmetrical | FAS | .19 | .13 | .01 | .46 |
|  |  | BAS | .19 | .13 | .02 | .52 |

*Summary Statistics for Associative Overlap Variables across each Experiment.*

*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* |
| 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 A3

*Comparisons of Mean Recall Percentages for each Encoding Task as a function of Pair Type in Experiment 1.*

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| Encoding Task | | List Type | | Pair Type | *M* | *95% CI* | Mixed F | Mixed U | Pure F |
| JOL | Mixed | | Forward | |  |  |  |  |  |
|  | Mixed | | Unrelated | |  |  |  |  |  |
|  | Pure | | Forward | |  |  |  |  |  |
|  | Pure | | Unrelated | |  |  |  |  |  |
| Frequency | Mixed | | Forward | |  |  |  |  |  |
|  | Mixed | | Unrelated | |  |  |  |  |  |
|  | Pure | | Forward | |  |  |  |  |  |
|  | Pure | | Unrelated | |  |  |  |  |  |
| No-JOL | Mixed | | Forward | |  |  |  |  |  |
|  | Mixed | | Unrelated | |  |  |  |  |  |
|  | Pure | | Forward | |  |  |  |  |  |
|  | Pure | | Unrelated | |  |  |  |  |  |

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

Table A4

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 A5]

Table A6

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

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 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 |  |  |
|  |  | Length |  |  |
|  |  | Frequency |  |  |
|  | Target | Concreteness |  |  |
|  |  | Length |  |  |
|  |  | Frequency |  |  |

*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 A7]