Is Discriminability a Requirement for JOL Reactivity? Comparing the Effects of Mixed vs. Pure List Presentations

Nicholas P. Maxwell & Mark J. Huff

The University of Southern Mississippi

.

Word Count: XXX

**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/3fztn/).

Abstract

Previous research has shown that judgments of learning (JOLs) are reactive when used to assess learning of cue-target word pairs. This reactivity often produces memory improvements (i.e., positive reactivity) but only for related word pairs. For unrelated pairs, no reactivity is generally reported. Researchers have primarily investigated reactivity using study lists which contain at least two distinct pair types (i.e., related vs. unrelated pairs). Using these mixed lists, reactivity may occur because participants use pair information to inform their study goals (i.e., prioritizing related pairs at the expense of unrelated pairs). The present study tested whether this process was a requisite for reactivity. First, Experiment 1 replicated previous work showing that in mixed lists, JOLs produce positive reactivity on related pairs but are not reactive on unrelated pairs. Importantly, Experiment 1 also showed that these patterns extended to pure lists, which only present participants with one pair type. Next, Experiments 2 and 3 extended these patterns to backward and symmetrical paired associates. Finally, across experiments, reactivity patterns reported for JOLs extended to frequency of co-occurrence judgments, regardless of list. Taken together, our findings provide further support for a cue-strengthening account of JOL reactivity rather than a goal-changing account.

Word count: 198

*Keywords:* Judgments of Learning; Reactivity; Frequency Judgments; Within vs. Between Designs

Is Discriminability a Requirement for JOL Reactivity? Comparing the Effects of Mixed vs. Pure List Presentations

Judgments of learning (JOLs) are used to measure online metamemory processes at encoding. While JOLs can be elicited for a variety of stimuli (e.g., text passages, Townsend & Heit, 2011; sentences, Luna, Albuquerque, & Martín-Luengo, 2019; etc.), they commonly present participants with cue-target pairs (e.g., word pairs like cat-dog), and participants are instructed to estimate their likelihood of correctly recalling the target (e.g., dog) at test if shown only the cue (e.g., cat). While JOLs are frequently used to investigate metacognitive processes, a growing body of research suggests that these judgments are reactive on learning, particularly when they are used to assess learning of cue-target pairs. (e.g., Janes, Rivers, & Dunlosky, 2018; Maxwell & Huff, in press; Soderstrom, Clark, Halamish, & Bjork, 2015). Reactivity occurs whenever a measure causes participants to attend to information they might otherwise ignore, leading to changes in performance (Ericsson & Simon, 1993). Regarding JOLs, reactivity results in memory changes, which can manifest as either memory improvements (i.e., *positive reactivity*) or memory costs (i.e., *negative reactivity*). Testing for these memory changes is relatively simple and involves comparing recall for participants who make JOLs at encoding to a no-JOL control group (e.g., silent reading). However, this comparison group has commonly been omitted from previous JOL studies, as researchers have often been more interested in factors affecting the accuracy of JOLs (e.g., associative direction; Koriat & Bjork, 2005; Maxwell & Huff, 2021; immediate vs. delayed JOLs; Dunlosky & Nelson, 1994; Nelson & Dunlosky, 1991; etc.) rather than the effects of these judgments on memory.

While JOL researchers have shown a recent interest in reactivity, evidence for JOL reactivity is not new and can be traced back to early research conducted by Arbuckle and Cuddy’s (1969). Arbuckle and Cuddy compared recall between two groups of participants: Those who made JOLs at study and confidence judgments at test and those who silently read each item at study and made confidence judgments at test. This design allowed for a comparison of recall rates between participants making JOLs at encoding to a group of participants who engaged in silent reading. Overall, the authors showed evidence that JOLs produce positive reactivity, as participants who made JOLs at encoding showed improved recall compared to those who silently read pairs. However, because all participants also made confidence judgments at retrieval, it was unclear whether providing judgments at test was also a requirement for reactivity to occur.

More recently, Soderstrom et al. (2015) tested for reactivity effects by comparing recall for between participants who made JOLs at encoding and those in a silent reading control group. Across groups, participants studied cue-target word pairs, half of which were related (e.g., mouse-cheese) while the other half were unrelated (e.g., mouse-bread). Following the JOL/study phase, participants completed a cued-recall test, which, importantly, did not require participants to make additional metacognitive judgments (e.g., confidence judgments; Arbuckle & Cuddy, 1969). Overall, Soderstrom et al. showed a positive reactivity pattern in which cued-recall performance was greater for participants who made JOLs initially relative to those in the control group. However, this effect was moderated by pair relatedness, such that only related pairs benefited from the requirement to make JOLs. When pairs were unrelated, no differences in recall were detected between the two groups. Subsequent studies by Janes et al. (2018) and Maxwell and Huff (in press) replicated this pattern, with both showing that JOLs produce positive reactivity on cued-target pairs, but only when word pairs are related.

However, although several studies have shown that JOLs produce positive reactivity on related pairs (e.g., Janes et al., 2018; Maxwell & Huff, in press; Soderstrom et al., 2015), Mitchum, Kelley, & Fox (2016) demonstrated the opposite pattern of reactivity, such that recall did not differ between participants in the JOL and control groups when for related pairs. Additionally, JOLs produced a negative reactivity pattern on unrelated pairs, such that participants making JOLs showed a memory cost for this pair type. However, it is likely that this pattern emerged due to methodological differences between their study and the one conducted by Soderstrom et al. (2015), as to date no other reactivity study has reported negative reactivity on unrelated pairs (see Double, Birney, & Walker, 2018). Thus, previous research has repeatedly shown that JOLs are reactive when applied to cue-target pairs, though the effect is moderated by pair relatedness.

**Theories of JOL Reactivity**

While several theories to explain JOL reactivity have been proposed, the two most prominent accounts are the *changed-goal hypothesis* (Mitchum et al., 2016) and the *cue-strengthening account* (Soderstrom et al., 2015). First, Mitchum et al.’s changed-goal hypothesis proposes that reactivity occurs because participants change study goals as they progress through a study list. According to this account, participants initially approach a study task with a broad goal of mastering the entire list. However, when instructed to make JOLs at study, participants realize that not all pairs will be remembered at the same rate, particularly when study lists contain a mix of easy and difficult pair types (i.e., related vs. unrelated pairs). As a result, participants use their perceptions of item difficulty to alter their study strategies, prioritizing pairs perceived as easy at the expense of more difficult pairs. Thus, the changed-goals hypothesis predicts positive reactivity for pairs perceived as easy to learn (e.g., related pairs) and negative reactivity for pairs perceived as difficult (e.g., unrelated pairs). However, because this account is dependent on a comparison process, the changed-goal hypothesis assumes that study lists will contain at least two discernable pair types (i.e., related vs. unrelated pairs). Reactivity would not be expected to occur when lists contain only one pair type (e.g., only related or unrelated pairs).

Alternatively, Soderstrom et al.’s cue-strengthening account proposes that the act of making JOLs directs participants’ attention towards specific intrinsic cues about each study pair that participants use to inform their JOLs (e.g., pair relatedness; see Koriat, 1997). According to this account, reactivity occurs anytime these cues emphasized by JOLs at encoding are made available at test (e.g., cued-recall testing). As a result, this account predicts positive reactivity on related pairs, but no reactivity for unrelated pairs, given this pair type’s lack of relatedness cues. Furthermore, the cue strengthening account makes no predictions regarding list composition, as reactivity in this account does not require an easy/difficult comparison.

Recent studies by Myers, Rhodes, and Hausman (2020) and Maxwell and Huff (in press) lend support to the cue-strengthening account. First, Myers et al. (2020) demonstrated that positive reactivity only emerged on related pairs when participants completed cued-recall or recognition testing. For free-recall testing in which relatedness cues are absent, JOLs produced no reactivity on related pairs. Second, Maxwell and Huff showed that positive reactivity on related pairs was not limited to JOLs and extended to other judgment tasks that similarly emphasized relatedness cues at encoding, including judgments of associative memory (JAMs; Maki, 2007; Valentine & Buchanan, 2013) and frequency of co-occurrence judgments. Taken together, these findings suggest that reactivity occurs due to the encoding task emphasizing cues used at retrieval, rather than inducing a change in participant study goal.

**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 word pairs. A mixed-list design is central to the changed-goal hypothesis, as this account states that participants’ ability to discriminate between different pair types is the primary factor driving 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 the present study 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, the present study provided 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 1 provided a direct replication of Janes et al.’s (2018) second experiment by comparing reactivity effects for forward and unrelated pairs across mixed and pure lists. Experiments 2 and 3 then expanded upon Experiment 1 by comparing unrelated pairs to backward and symmetrical pairs, respectively. Each experiment, therefore, provided three separate tests of list effects on reactivity.

Finally, because Maxwell and Huff (in press) showed that reactivity effects also extend to other, non-metacognitive judgment tasks, each experiment additionally included a frequency judgment comparison group in which participants rated the likelihood that paired items would appear together in everyday language. This additional comparison was included to (1) test whether the reactivity effects for frequency judgments initially reported by Maxwell and Huff would replicate for mixed groups and (2) test whether these judgments would continue mirror JOL reactivity pairs when made within a pure-list context.

**Experiment 1: Forward vs. 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 implicitly encourage participants to relate items together. However, they do not require participants to make a memory prediction, frequency judgments lack the metacognitive component associated with JOLs. Based on findings by Maxwell and Huff (in press), we expected that frequency judgment would produce a reactivity pattern mirroring JOLs.

By comparing reactivity between mixed and pure lists, Experiment 1 provided a direct test of the changed-goal hypothesis while also providing a further test of the cue-strengthening account. First, because the changed-goal hypothesis states that reactivity results from participants changing study goals as they discern between related and unrelated pairs, reactivity would only be expected to occur when study pairs are presented using mixed lists, as pure lists lack the easy/difficult comparison required to trigger a goal change. The changed-goal hypothesis, therefore, predicts a null effect of reactivity for pairs presented using this list type, regardless of pair relatedness. The cue-strengthening account, however, makes no claims regarding easy/difficult comparisons. Instead, this account predicts positive reactivity for related pairs provided the encoding task emphasizes relatedness cues that are accessed at retrieval, regardless of whether participants study mixed or pure lists. 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 rather than the changed-goal hypothesis.

**Methods**

**Participants**

A total of 347 participants were recruited to complete Experiment 1. 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* (Faul, Erdfelder, Lang, & Buchner, 2007), 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, 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 A2, respectively.

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**

Data was collected online using *Collector*, an open-source program for presenting psychological experiments online (Garcia & Kornell, 2015). 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 all groups, participants were informed that they would see a list of cue-target word pairs and that their memory for the target items in each pair would later be tested. Participants in the JOL and frequency judgment groups were further instructed to make judgments while encoding each study pair. Specifically, participants in the JOL group were instructed to rate the likelihood that they would be able to successfully recall the target item at test if prompted by only the cue. Participants in the frequency judgment group were instructed to rate the likelihood that the cue and target items would appear within the same context in natural language. Judgments in both groups utilized a 0-100 scale and were made concurrently with study, such that participants typed their ratings while the pair was displayed on the screen. Thus, the only difference between judgment conditions was the framing.

After receiving their respective encoding instructions, participants began the first study list. In the mixed list groups, this list contained both forward and unrelated pairs. In contrast, participants assigned to the pure list groups studied lists containing only forward or unrelated pairs. Following completion of the first study list, participants completed a two-minute filler task in which they listed the 50 U.S. states in alphabetical order. This was then followed by a cued-recall test which presented participants with each cue word from the preceding study list in a randomized order. Participants were instructed to type the correct target item from memory and were instructed to press the Enter key if they could not retrieve the correct item. Following completion of the cued-recall test, participants began the second block. This block followed the same format as the first, and participants studied the same list type in block 2 as in block 1. Participants were debriefed following completion of the second block. The total experiment took approximately 30 minutes to complete.

**Results**

For all analyses, significance was set at *p* < .05. We report partial eta-squared (*ηp*2)and Cohen’s *d* effect sizes for all significant analyses of variance (ANOVAs) and *t*-tests. Additionally, all non-significant main effects, interactions, and post-hoc comparisons are supplemented by a Bayesian estimate of the strength of evidence in support of the null hypothesis (Masson, 2011; Wagenmakers, 2007). This analysis compares a model assuming a significant effect to a second model assuming a null effect. In doing so, a probability estimate can be generated, representing the likelihood that null hypothesis is retained (i.e., *p*BIC; Bayesian Information Criterion). Because this probably estimate is sensitive to sample size, it provides increased confidence in reported null effects.

The top panel of Figure 1 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 rates between encoding groups for pure-list participants. For completeness, all comparisons between forward and unrelated pairs are provided in Table A3. 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, *ηp*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, *pBIC* = .88. Importantly, a significant interaction between Pair Type and Study Group was found, *F*(2, 102) = 12.41, *MSE* = 99.84, *ηp*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 ≥ 3.30, *d*s ≥ 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 ≥ .38, *p*bics ≥ .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, *ηp*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, *ηp*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 ≥ .36, *p*bics ≥ .88.

Critically, a significant interaction emerged, *F*(2, 196) = 7.37, *MSE* = 262.08, *ηp*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.

**Discussion**

The primary goal of Experiment 1 was to test the effect of list type on reactivity. In doing so, this experiment 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). Finally, reactivity patterns observed for JOLs again extended to frequency judgments, replicating findings by Maxwell and Huff (in press) and further suggesting that JOL reactivity results from strengthening relational cues rather than via a metamemorial or predictive process.

Importantly, Experiment 1 showed that reactivity effects are not limited to mixed-list designs. 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 2 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 2 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.

**Experiment 2: Backward vs. Unrelated Pairs**

The goal of Experiment 2 was to test whether pure-list reactivity effects observed for forward pairs in Experiment 1 would extend to backward pairs. Like the previous experiment, Experiment 2 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 1. 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.

**Methods**

**Participants**

Experiment 2 followed the same design as Experiment 1. 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 1 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 1. 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 for mixed and pure groups, respectively).

Like Experiment 1, 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 2 used the same study lists as the previous experiment, with the following modification. While the same unrelated word pairs from Experiment 1 were retained, all forward pairs (e.g., peanut-butter) were replaced with backward pairs (e.g., butter-peanut). Additionally, 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 1 in all other aspects including number of items, the inclusion of buffer pairs, and the study procedure (see Tables A1 and A4 for stimuli properties).

**Results**

Figure 2 (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 A5. 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, *ηp*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, *pBIC* = .62, but the interaction was reliable, *F*(2, 117) = 15.83, *MSE* = 87.63, *ηp*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*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*bics ≥ .69). As such, reactivity patterns observed with forward pairs in mixed lists extended 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, *ηp*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*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*bic = .44, *ηp*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*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.

**Discussion**

Experiment 2 tested whether reactivity patterns observed for forward pairs in Experiment 1 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 2 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 3 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 Maxwell and Huff (in press), no study has investigated JOL reactivity effects using symmetrical paired associates. As such, Experiment 3 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.

**Experiment 3: Symmetrical vs. Unrelated Pairs**

Experiment 3 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 3. 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 2, the 106 participants who studied pure unrelated lists in Experiment 1 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 1, 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 3 used a modified version of the study lists presented in Experiments 1 and 2. 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 1 and 2 (see Tables A1 and A6 for stimuli properties).

**Results**

Figure 3 (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 A7). 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, *ηp*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*bic = .96. A significant interaction was found, confirming the presence of a reactivity pattern, *F*(2, 103) = 12.57, *MSE* = 112.87, *ηp*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*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*bics ≥ .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, *ηp*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, *ηp*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, *t*s ≤ 1.06, *p*s ≥ .29, *p*bics ≥ .90. Importantly, a significant interaction was again found, *F*(2, 203) = 8.12, *MSE* = 246.60, *ηp*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, *t*s ≥ 3.80, *d*s ≥ 0.85, apart from the comparison between the JOL and frequency groups, *t*(66) = 1.12, *SEM* = 3.81, *p* = .26, *p*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.

**Discussion**

The goal of Experiment 3 was to test whether reactivity effects observed for forward and backward pairs in Experiments 1 and 2 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 3 are in-line with our previous experiments while providing additional support for the cue-strengthening account.

**General Discussion**

Overall, the goal of this study was to compare the changed-goal and cue strengthening accounts of JOL reactivity by testing whether reactivity patterns previously reported on mixed lists (i.e., positive reactivity on related pairs, no reactivity on unrelated pairs; Janes et al., 2018; Maxwell & Huff, in press; Soderstrom et al., 2015) would still occur for pairs presented in isolation via pure lists. In doing so, each experiment focused exclusively on one type of related paired associate (forward, backward, or symmetrical) and directly compared it to unrelated pairs within both mixed and pure list contexts. Additionally, a secondary goal was to further test whether reactivity effects were unique to JOLs. Therefore, in addition to the JOL versus no-JOL comparison traditionally used to explore reactivity, each experiment also included a group of participants who completed a frequency judgment task in lieu of providing JOLs. This additional comparison group was included to evaluate whether any observed reactivity patterns would occur when a non-metacognitive judgment task was used.

Overall, Experiment 1 replicated previous JOL reactivity findings for mixed lists (e.g., Janes et al., 2018, Maxwell & Huff, in press, Soderstrom et al., 2015), such that JOLs produced positive reactivity on forward pairs but no reactivity on unrelated pairs. Importantly, this reactivity pattern extended to pure lists, suggesting that reactivity is not driven by changes in participant study goals. Finally, all observed reactivity on JOLs also extended to frequency judgments, providing additional evidence that reactivity effects are driven by the encoding task strengthening relatedness cues rather than a comparative process as posited by the changed-goal hypothesis. Next, Experiments 2 and 3 showed that these positive reactivity patterns for both JOLs and frequency judgments extend to backward and symmetrical paired associates, respectively. Thus, across experiments and list types, negative reactivity for unrelated pairs as reported by Mitchum et al. (2016) consistently failed to occur. Therefore, a key finding from the present study is that JOLs consistently produce positive reactivity on related pairs but no reactivity on unrelated pairs, regardless of whether pairs are presented using mixed or pure list contexts.

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 necessary to trigger a change in study goal. Therefore, our pure-list reactivity findings 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 but not unrelated pairs. As such, pure list reactivity findings observed in the present study are in-line with this account.

In addition to testing for reactivity effects between list types, each experiment included an additional comparison group in which participants rated the likelihood of words co-occurring together. We included these groups to test whether reactivity patterns observed on non-metacognitive judgment in mixed lists reported by Maxwell and Huff (in press) would similarly extend to pure lists. Like JOLs, frequency judgments direct attention towards relational aspects of study pairs without explicitly instructing participants use a relational strategy at encoding. Additionally, this task used the same 0-100 rating scale as JOLs. Thus, the frequency judgment task resembled JOLs but removed the requirement that participants forecast later recall performance. Across experiments, frequency judgments consistently showed reactivity patterns mirroring JOLs, such that frequency that these judgments provided a memory boost to related pairs but no reactivity when pairs were unrelated. Furthermore, like JOLs, frequency judgments were reactive regardless of whether participants studied pairs within mixed or pure lists. Thus, metacognitive processes induced by JOLs do not appear to be a requisite for reactivity to occur.

While the present study was designed to test the changed-goal and cue strengthening accounts of JOL reactivity, we note that this study also provides insight regarding the role of strategy use on reactivity effects. First, the finding that both JOLs and frequency judgments consistently produce positive reactivity on related but not unrelated pairs suggests that these tasks [STRATEGIC IN MIXED BUT MAYBE NOT IN PURE] We note, however, [RIVERS PAPER] [FIND A CRITICISM ABOUT IT] [SUMMARY STATEMENT]

Finally, while the present study used cued-recall performance as the primary measure of reactivity, we note 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 AX-AX). 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.).

[FUTURE DIRECTIONS]

**Conclusion**

[TRANSITION] Both the changed-goal hypothesis (Mitchem et al., 2016) and the cue-strengthening account (Soderstrom et al., 2015) have been proposed as explanations of JOL reactivity. The present study tested these two competing theories by testing how list composition would affect reactivity. [FREQ JUDGMENTS?] Overall, [BRIEF SUMMARY]

**Open Practices Statement**

The data for all experiments have been made available at https://osf.io/3fztn/. 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.

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.

Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur? *Journal of Memory and Language, 33*(4), 545-565.

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.

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.

Luna, K., Albuquerque, P. B., & Martín-Luengo, B. (2019). Cognitive load eliminates the effect of perceptual information on judgments of learning with sentences. *Memory & Cognition, 47*, 106-116.

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., & Huff, M. J. (in press). Reactivity from judgments of learning is not only due to memory forecasting: Evidence from associative memory and frequency judgments. *Metacognition and Learning*, 1-XX.

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.

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.

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., & Dunlosky, J. (1991). When people’s judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The “Delayed-JOL Effect.” *Psychological Science, 2*(4), 267-271.

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.

Tauber, S. K., & Witherby, A. E. (2019). Do judgments of learning modify older adults’ actual learning? *Psychology and Aging, 34*(6), 836-847.

Townsend, C. L., & Heit, E. (2011). Judgments of learning and improvement. *Memory & Cognition, 39*, 204-216.

Valentine, K. D., & Buchanan, E. M. (2013). JAM-boree: An application of observation oriented modeling to judgements of associative memory. *Journal of Cognitive Psychology, 25*(4), 400-422.

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

Table 1

*Final Sample Sizes for all Comparison Groups in each Experiment.*

|  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- |
| Experiment | Encoding Task | Mixed | Pure Forward | Pure Backward | Pure Symmetrical | Pure Unrelated | |
| Exp. 1 | JOL | 36 | 31 | -- | -- | 35 |
|  | Frequency | 34 | 31 | -- | -- | 37 |
|  | No-JOL | 35 | 34 | -- | -- | 34 |
| Exp. 2 | JOL | 40 | -- | 41 | -- | 35 |
|  | Frequency | 43 | -- | 42 | -- | 37 |
|  | No-JOL | 37 | -- | 37 | -- | 34 |
| Exp. 3 | 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 1 was used as the pure unrelated comparison in Experiments 2 and 3.



*Figure 1.* Mean percent recall for participants in Experiment 1 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.



*Figure 2.* 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.



*Figure 3.* 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.

**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 | .27 | .18 | .01 | .59 |
|  |  | BAS | .27 | .17 | .01 | .58 |
|  | 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* | 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 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

*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 | | 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 1.

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 | 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 A7

*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 | | 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 1.

Table A8

Mean Encoding Latencies as a Function of Pair Type and Encoding Task for Mixed Lists in Experiments 1-3.

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Experiment | Encoding Task | Forward | Backward | Symmetrical | Unrelated |
| Exp. 1 | JOL | 4166 | -- | -- | 5009 |
|  | Frequency | 4500 | -- | -- | 5992 |
|  | Read | 6268 | -- | -- | 8150 |
| Exp. 2 | JOL | -- | 5527 | -- | 4995 |
|  | Frequency | -- | 5444 | -- | 5179 |
|  | Read | -- | 5396 | -- | 5801 |
| Exp. 3 | JOL | -- | -- | 5316 | 6470 |
|  | Frequency | -- | -- | 4322 | 5310 |
|  | Read | -- | -- | 5603 | 7103 |

Note: Cells display mean RTs in ms.

Table A9

*Mean Encoding Latencies as Functions of Pair Type and Encoding Tasks for Pure Lists in Experiments 1-3.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Experiment | Encoding Task | Forward | Backward | Symmetrical | Unrelated |
| Exp. 1 | JOL | 3483 | -- | -- | 5197 |
|  | Frequency | 3616 | -- | -- | 6407 |
|  | Read | 5249 | -- | -- | 6376 |
| Exp. 2 | JOL | -- | 6398 | -- | 5197 |
|  | Frequency | -- | 5743 | -- | 6407 |
|  | Read | -- | 6561 | -- | 6376 |
| Exp. 3 | JOL | -- | -- | 5026 | 5197 |
|  | Frequency | -- | -- | 4294 | 6407 |
|  | Read | -- | -- | 4739 | 6376 |

*Note:* Cells display mean RTs in ms. Pure unrelated comparison is taken from Experiment 1.