Memory & Cognition (2023) 51:1198–1213
https://doi.org/10.3758/s13421-022-01381-4

Is discriminability a requirement for reactivity? Comparing the effects of mixed vs. pure list presentations on judgment of learning reactivity

Nicholas P. Maxwell[1,2] · **Mark J. Huff**[1]

Accepted: 10 December 2022 / Published online: 28 December 2022

Abstract

Providing judgments of learning (JOLs) at study tends to produce reactive effects on recall of cue–target word pairs. This reactivity generally produces memory improvements (i.e., positive reactivity) but only for related word pairs. For unrelated pairs, reactivity is typically not observed. Researchers have primarily investigated reactivity using study lists that contain at least two distinct pair types (i.e., related vs. unrelated pairs). Using these mixed lists, reactivity may occur because participants use distinguishing pair characteristics to inform their study goals (i.e., prioritizing related vs. unrelated pairs). The present study examined whether detection of separate pair types within mixed lists is a requisite for reactivity to occur. Experiment 1 replicated previous work showing that in mixed lists, JOLs produced positive reactivity on related pairs but are nonreactive on unrelated pairs. Importantly, Experiment 1 also found that these patterns extended to pure lists, in which only one pair type is presented. Experiments 2 and 3 then extended these patterns to backward and symmetrical paired associates. Finally, across experiments, reactivity patterns reported for JOLs extended to frequency of co-occurrence judgments across pair and list types. Our findings that reactivity patterns consistently emerge using pure lists supports a cue-strengthening account of reactivity.

Keywords Judgments of learning · Reactivity · Frequency judgments · Within vs. between designs

Judgments of learning (JOLs) are used to assess the metamemory processes participants engage in at encoding. While JOLs can be elicited for various types of study materials (e.g., text passages, Townsend & Heit, 2011; sentences, Luna et al., 2019), participants commonly study cue–target word pairs (e.g., *cat–dog*) and estimate their likelihood of correctly recalling the target (e.g., *dog*) at test if just shown the cue (e.g., *cat*). While JOLs are used to gauge metacognitive processes (see Rhodes, 2016, for review), a growing body of research suggests that these judgments are *reactive* towards learning (e.g., Janes et al., 2018; Maxwell & Huff, 2022; Soderstrom et al., 2015). Reactivity occurs whenever a task encourages participants to attend to information they might otherwise ignore, leading to changes in performance (Ericsson & Simon, 1993). Regarding JOLs, reactivity may produce memory benefits (i.e., *positive reactivity*) or memory costs (i.e., *negative reactivity*). Testing for these memory changes is simple and merely requires comparing recall for participants who make JOLs at encoding to a no-JOL control group (e.g., silent reading). However, this comparison group is often absent, particularly in studies in which JOLs are made immediately following study, as researchers have often been more interested in factors affecting JOL accuracy (e.g., associative direction; Koriat & Bjork, 2005; Maxwell & Huff, 2021; multiple study trials; Koriat et al., 2002; Meeter & Nelson, 2003) than the direct effects of these judgments on memory.

Although JOL studies commonly omit no-JOL group comparisons, interest in the effects of these judgments on memory is not new. Research suggests that JOLs made following a delay can produce memory benefits (e.g., Akdoğan et al., 2016; Spellman & Bjork, 1992; see Rhodes & Tauber, 2011). However, researchers have only recently begun to explore whether concurrent and immediate JOLs (i.e., those elicited at or immediately following encoding) are similarly reactive. This is surprising, given reactive effects of immediate JOLs were reported over 50

✉ Nicholas P. Maxwell
nicholas.maxwell@msutexas.edu

[1] School of Psychology, The University of Southern Mississippi, 118 College Dr. #5025, Hattiesburg, MS 39406, USA

[2] Present Address: Midwestern State University, Wichita Falls, TX, USA

years ago by Arbuckle and Cuddy (1969). In their early work, 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 provided confidence judgments at test. This design allowed for a comparison of recall rates between participants making JOLs at encoding to participants who only engaged in silent reading. Overall, JOLs produced positive reactivity; however, because all participants made confidence judgments at retrieval, it was unclear whether providing these judgments at test was also a requirement for reactivity to occur.

More recently, Soderstrom et al. (2015) tested for reactivity by comparing recall between participants who made JOLs at encoding to a silent reading control group. Across groups, participants studied cue–target word pairs, in which half were related (e.g., *mouse–cheese*) and the other half were unrelated (e.g., *mouse–bread*). Following the JOL/study phase, participants completed a cued-recall test, which did not require participants to make additional metacognitive judgments (cf. Arbuckle & Cuddy, 1969). Overall, Soderstrom et al. reported a positive reactivity pattern in which cued-recall performance was greater for participants who made JOLs than the silent-reading control. However, this pattern was moderated by pair relatedness as only related pairs showed positive reactivity. Recall of unrelated pairs did not differ between encoding groups. Subsequent studies by Janes et al. (2018) and Maxwell and Huff (2022) replicated this pattern using immediate and concurrent JOLs, respectively, with both studies similarly showing that JOLs produce positive reactivity selectively on related pairs.

Although recent studies show that immediate JOLs produce positive reactivity on related pairs but have no effect on unrelated related pairs (e.g., Janes et al., 2018; Maxwell & Huff, 2022; Soderstrom et al., 2015), Mitchum et al. (2016) reported a different pattern. Specifically, they found no reactivity for related pairs, and negative reactivity for unrelated pairs. To date, it is unclear why this pattern emerged as similar methodologies were used relative to other studies (e.g., Maxwell & Huff, 2022; Soderstrom et al., 2015). However, a meta-analysis by Double et al. (2018) reported positive reactivity for related pairs and no reactivity for unrelated pairs across the 17 experiments they analyzed.

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). The changed-goal hypothesis proposes that reactivity occurs because participants shift study goals as they progress through a study list. According to this account, participants initially approach study tasks with a broad goal of mastering all list items. However, when instructed to make JOLs at study, participants realize that not all pairs will be remembered equally well, particularly when lists contain a mix of pairs perceived as easy and difficult to remember (i.e., related vs. unrelated pairs). As a result, participants use perceptions of item difficulty to adjust their study strategies, prioritizing the encoding of pairs perceived as easy at the expense of more difficult pairs. Thus, the changed-goal 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). Because this account depends on a comparison process, it 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 (2015) cue-strengthening account proposes that making JOLs directs participants' attention towards intrinsic cues about each study pair, which participants use to inform their JOLs (e.g., pair relatedness; see Koriat, 1997). According to this account, reactivity occurs anytime the cues that are emphasized by JOLs at study are available at test (e.g., cued-recall testing). As a result, positive reactivity should occur on related pairs, but no reactivity for unrelated pairs, given this pair type's lack of relatedness cues which are used to inform JOLs. Furthermore, the cue-strengthening account makes no predictions regarding list composition, as reactivity in this account does not require an easy/difficult comparison, just the availability of intrinsic cues that could direct attentional processes at encoding.

Prior research generally supports a cue-strengthening account over a changed-goal account. For example, Myers et al. (2020) found that positive reactivity on related pairs depended upon the availability of cues at test, as positive reactivity emerged on cued recall and recognition but not free recall in which cues are absent at retrieval. Additionally, Maxwell and Huff (2022) reported that positive reactivity on related pairs was not limited to JOLs and extended to other, non-metacognitive judgment tasks that similarly emphasize relatedness cues, including judgments of associative memory (JAMs; Maki, 2007; Valentine & Buchanan, 2013) and frequency of co-occurrence judgments. Thus, reactivity occurs whenever the judgment task emphasizes the processing of cues that are subsequently available at retrieval, and not on an adjustment of study goals.

Mixed-list versus pure-list designs

When investigating reactivity mechanisms, direct comparisons between mixed-list and pure-list designs are informative. For instance, a mixed-list design is central to the changed-goal hypothesis, as shifting study goals requires the perception of both easy and difficult pair

types. Separately, according to the cue-strengthening account, reactivity could occur whenever the encoding task emphasizes cues used at retrieval, regardless of whether pairs are presented within a mixed or pure list context.

Although studies investigating reactivity effects generally use mixed-list designs, Janes et al. (2018) and Tauber and Witherby (2019) included pure-group comparisons. First, Janes et al.'s (2018) Experiment 2 compared JOL reactivity effects for mixed- versus pure-list designs by having participants study (1) mixed lists of forward associates and unrelated pairs, (2) pure lists of forward pairs, or (3) pure lists of unrelated pairs. Overall, the authors replicated previously reported positive reactivity patterns mixed-list related pairs, but this same pattern did not occur on pure lists, suggesting that reactivity effects were contingent on participants being able to discriminate between different pair types. Tauber and Witherby (2019), however, reported positive reactivity on forward pairs presented via a pure list. However, Tauber and Witherby were unable to directly compare the changed-goal and cue-strengthening accounts, as a mixed-list comparison was not included. Thus, it is unclear how these observed reactivity effects would compare with a mixed list (i.e., whether reactivity effects would be greater when using a mixed list vs. a pure list) or whether this effect would also extend to a pure list of unrelated pairs.

Given these discrepancies, and the absence of consistent comparison groups within the literature, the present study sought to provide a direct test of list-composition effects on reactivity. Specifically, our study compared cued-recall in mixed lists containing related and unrelated pairs to a separate group of participants who studied either pure lists of only related or unrelated word pairs. 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. Additionally, because Maxwell and Huff (2022) showed that reactivity effects extend to other, non-metacognitive judgment tasks, each experiment included an additional frequency-judgment group in which participants rated the likelihood that paired items would appear together in everyday language rather than making a JOL. 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 patterns when elicited within a pure-list context. Thus, the present study provides three separate tests of list effects on JOL and frequency judgment reactivity while also isolating these effects for three types of related word pairs, including backward and symmetrical pairs which have not been included in previous reactivity studies.

Experiment 1: Forward versus unrelated pairs

Experiment 1 had three main goals. First, we sought to replicate positive reactivity on related pairs presented via mixed lists as initially reported by Soderstrom et al. (2015). Second, we 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 included a group of participants who provided frequency judgments at encoding. Like JOLs, frequency judgments implicitly encourage the processing of intrinsic features of cue–target pairs, including pair relations. However, frequency judgments do not require participants to forecast subsequent memory and therefore are less likely to encourage metacognitive processes. Based on findings by Maxwell and Huff (2022), we expected that frequency judgments would produce reactivity patterns mirroring JOLs.

By comparing reactivity between mixed and pure lists, Experiment 1 directly tested the changed-goal hypothesis while also testing the cue-strengthening account. In doing so, Experiment 1 sought to replicate Janes et al.'s (2018) Experiment 2, while also assessing if JOL reactivity for related pairs only occurs for mixed but not pure lists. We also assessed whether JOL reactivity patterns would extend to a frequency judgment tasks in each list type. Because shifting goals requires discerning between related and unrelated pairs, the changed-goal hypothesis predicts that reactivity would only occur for pairs presented in mixed lists and a null effect of reactivity for pure-list pairs, regardless of relatedness. However, because the cue-strengthening account makes no claims regarding comparison processes, this account simply predicts positive reactivity would occur on related pairs, provided the encoding task emphasizes relatedness cues that are accessed at retrieval. Thus, the cue-strengthening account predicts a reactivity effect, regardless of whether participants study mixed or pure lists. If pure lists produce the same reactivity patterns previously found in mixed lists (i.e., positive reactivity for related pairs, no reactivity for unrelated pairs), this would provide further evidence for a cue-strengthening account over a goal-changing account.

Methods

Participants

A total of 347 online participants were recruited to complete Experiment 1. Participants were recruited from two sources: Undergraduate students from The University of Southern Mississippi's psychology research pool, who completed the study in exchange for course credit (n = 260), and individuals who were recruited through Prolific (www.proli

fic.co), who were compensated at a rate of $3.90/half hour ($n$ = 87). Of these 347 participants, 111 were randomly assigned to the mixed-list group, which used a 3 × 2 mixed design that manipulated pair relatedness 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 et al., 2007), which indicated that at least 42 participants would be needed to detect medium effects/interactions (d = 0.50) with mixed lists, while 158 participants would be necessary for the same effect size with pure lists. However, groups were oversampled due to an anticipated increase in participant performance variability from 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/control). This resulted in a total of nine groups (see Table 1 for each group's final n following data screening). All participants were native English speakers. Responses from 39 participants were excluded for one of the following reasons: (1) Low recall rates (e.g., correct recall <5%), which suggested that participants did not correctly follow study instructions, or (2) recall rates of 100% across all blocks/pair types, which suggested 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 for analysis (105 in the mixed-list analyses; 202 in the pure-list analyses).

Materials

To generate the stimuli, 200 word pairs were taken from the University of South Florida Free Association Norms (USF norms; D. L. Nelson et al., 2004). These pairs were divided into six study lists: Two mixed lists, two pure lists of forward pairs, and two pure lists of unrelated pairs. Mixed and pure list forward pairs were matched on mean 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 from the English Lexicon Project (Balota et al., 2007). Associative overlap measures and lexical characteristics for all stimuli are reported in the Appendix in Tables 2 and 3, respectively.

Study pairs across lists were randomized with the constraint that five nontested buffer pairs were 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 procedure used to create the mixed and pure lists is described in further detail.

Mixed lists To create 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 selected as nontested buffer items to control for primacy and recency effects. Pairs were divided into two study lists, each consisting of 20 forward pairs, 20 unrelated pairs, and 10 buffer pairs (five related and five unrelated). As a result, each mixed list contained a total of 50 pairs.

Pure lists Four pure lists were generated (two for each pair type). For 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

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

Cells reflect final ns 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

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 contained 40 unrelated pairs not assigned to a mixed list. Like the related lists, the remaining 20 unrelated pairs were used as buffers. Thus, regardless of pair type, each pure list contained of 40 study pairs and 10 buffer pairs. Finally, all pure lists were matched to mixed lists on semantic and lexical characteristics.

Procedure

Data collection occurred online using Collector, an open-source program for presenting psychological experiments (Garcia & Kornell, 2015). Participants were first randomly assigned to either the mixed- or pure-list groups and then further randomly assigned to complete either the JOL, frequency-judgment, or silent-reading tasks. Across groups, participants were informed 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 utilized a 0–100 scale in both groups and were made concurrently with study, such that participants typed their ratings while the pair was displayed on the screen. The only difference between judgment conditions was the framing. For all groups, encoding was self-paced. Participants pressed the ENTER key to advance to the next pair.

After receiving encoding instructions, participants began the first study list. In 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 2-min filler task in which they listed the 50 U.S. states in alphabetical order. This was immediately 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 or to press ENTER 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 Block 1. Participants were debriefed following completion of the second block. The total experiment duration was approximately 30 min.

Results

For all analyses, significance was set at $p < .05$. We report partial eta-squared (η_p^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 separate Bayesian estimation of support for 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). Like p values, p_{BIC} does not specify strength of evidence for the null hypothesis. However, because this probably estimate is sensitive to sample size, it provides increased confidence in reported null effects.

The top panel of Fig. 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

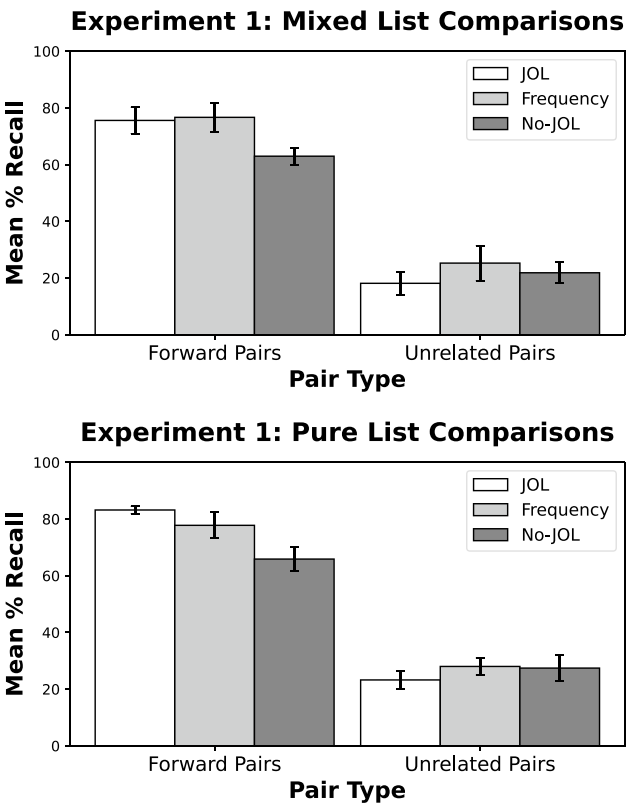

Fig. 1 Mean percentage recall for participants in Experiment 1 who completed the JOL, frequency-judgment, or No-JOL silent reading tasks for mixed lists (top panel) or pure lists (bottom panel). Error bars represent 95% confidence intervals

displays mean recall rates between encoding groups for pure-list participants. For completeness, all comparisons between encoding groups are reported in Appendix Table 4.

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. A main effect of Pair Type was found, $F(1, 102) = 1309.60$, $MSE = 99.84$, $\eta_p^2 = .93$, such that, mean recall was higher for forward (71.74) than unrelated (21.69) pairs. The effect of Study Group was only marginally reliable, $F(2, 102) = 2.64$, $MSE = 485.32$, $p = .08$, $p_{BIC} = .88$, but a significant interaction between Pair Type and Study Group was found, $F(2, 102) = 12.41$, $MSE = 99.84$, $\eta_p^2 = .20$. Post hoc t tests indicated that for forward pairs, correct recall in both the JOL (75.59) and frequency-judgment (76.68) groups exceeded that of the no-JOL group (62.98). All comparisons differed, ts ≥ 3.30, ds ≥ 0.77, except for the difference in recall between the JOL and frequency-judgment groups, $t < 1$, $SEM = 3.57$, $p = .74$, $p_{BIC} = .89$. Importantly, for unrelated pairs, recall rates did not differ between the JOL (18.14), frequency-judgment (25.27) and no-JOL (21.86) group, ts < 1, ps ≥ .38, p_{BIC}s ≥ .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 in mixed lists, JOL ratings and frequency judgments produced equivalent reactivity patterns on related pairs but 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 for pairs in a pure-list context. Overall, this analysis yielded an effect of Pair Type, $F(1, 196) = 468.13$, $MSE = 262.08$, $\eta_p^2 = .70$, in which mean recall was higher for forward (71.74) than unrelated pairs (21.69). A significant effect of Study Group emerged, $F(2, 196) = 3.52$, $MSE = 262.08$, $\eta_p^2 = .03$, such that mean recall was highest in the JOL group (51.40), followed by the frequency-judgment group (50.70) and the no-JOL group (46.65).

Critically, a significant interaction emerged, $F(2, 196) = 7.37$, $MSE = 262.08$, $\eta_p^2 = .07$. Follow-up tests 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, ts ≥ 2.62, ds ≥ 0.65, 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 across the between the JOL (23.25), frequency-judgment (28.01), or the no-JOL (27.45) groups, ts ≤ 1.42, ps ≥ .16, $p_{BIC} ≥ .76$. Therefore, pure lists demonstrated similar reactivity patterns as mixed lists, with reactivity only occurring on related, but not unrelated, lists[1].

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. Starting with the mixed-list group, the predicted pattern of reactivity emerged. Relative to the no-JOL group, making JOLs increased correct recall of forward pairs—a positive reactivity pattern—but produced no benefit for unrelated pairs. This finding replicates previous work on JOL reactivity (e.g., Janes et al., 2018; Soderstrom et al., 2015). Finally, reactivity patterns observed for JOLs extended to frequency judgments, replicating findings by Maxwell and Huff (2022). Given the task-type similarities, one possibility is that reactivity occurs on related pairs whenever the judgment task facilitates the processing of relational features between the cue and target. Specifically, JOLs encourage participants to examine cue–target relations to make a memory forecast whereas frequency judgments process cue–target relations based on previously stored semantic knowledge of word frequencies. In both tasks, semantic relations are emphasized, which may contribute to positive reactivity patterns.

In addition to task effects on reactivity, Experiment 1 showed that previously reported reactivity effects are not restricted to mixed-list designs. Pure lists 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). This account, however, also posits that for reactivity to occur, cues used to inform JOLs (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, although backward pairs are thematically related, they are deceptive, as relatedness cues that can aid retrieval are less

[1] Changes in reactivity between mixed and pure lists can also be assessed by analyzing related pairs via a 3 (Study Group: JOL vs. Frequency vs. Read) × 2 (List Type: Mixed vs. Pure) between-subjects ANOVA. Across Experiments, no interactions are detected, Fs < 1; ps ≥ .48, p_{BIC}s ≥ .98. Thus, the overall reactivity pattern for JOLs and frequency judgments does not differ between list types.

likely to be available at test. However, reactivity may still occur for this pair type. Recently, Maxwell and Huff (2022) showed that positive reactivity on forward pairs extends to backward pairs. To explain this finding, they proposed that the presence of intrinsic relatedness cues at encoding may be sufficient to trigger reactivity, as these cues encourage participants to use a relational encoding strategy. Therefore, any reactivity on backward associates may also reflect additional processing via relational encoding in addition to cue strengthening.

To test this possibility, Experiment 2 compared mixed- and pure-list reactivity patterns using backward and unrelated pairs. Like forward pairs, participants typically assign backward pairs high JOLs at study (indicating that participants perceive backward pairs as related), but at test, participants often struggle to correctly retrieve the target (e.g., *the illusion of competence*; Koriat & Bjork, 2005). Backward pairs therefore provide a situation where the cue–target word pair appears strongly related at encoding (via associative relations), but cues used to inform the judgment are weaker 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 versus unrelated pairs

The goal of Experiment 2 was to test whether pure-list reactivity effects on forward pairs in Experiment 1 would extend to backward pairs. Like Experiment 1, Experiment 2 provided another test of the changed-goal and cue-strengthening accounts of reactivity. Based on the changed-goal hypothesis, positive reactivity would only be expected to occur for backward pairs presented in mixed lists, but not pure lists, given backward pairs are ostensibly easier to encode relative to unrelated pairs. 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, as participants are likely to employ a relational processing strategy when encoding this pair type. However, because relatedness cues for backward pairs are less likely to be available at retrieval (i.e., the target is a less common response to the cue), reactivity on backward pairs may be reduced compared to forward pairs in Experiment 1. Finally, frequency judgments should again display reactivity patterns mimicking those found for JOLs in both list types.

Methods

Participants

Experiment 2 used the same design as Experiment 1. A separate 253 participants were recruited and completed the experiment online. Of these participants, 204 were University of Southern Mississippi undergraduates who completed the study online in exchange for course credit. The remaining 49 were recruited via Prolific and received $3.90 per half-hour of participation. Of the 253 participants recruited, 127 were randomly assigned to the mixed-list group, with the remaining 126 participants assigned to the pure-related group. Finally, the 106 participants who were assigned to the pure-unrelated group in Experiment 1 served as the pure-unrelated group in Experiment 2. Thus, the pure-list groups contained a total of 232 participants. For both groups, sample sizes were based on Experiment 1. A sensitivity analysis conducted with G*Power 3.1 indicated that both the mixed- and pure-list samples were sufficient for detecting small-to-medium sized effects and interactions ($ds = 0.26$ and 0.40 for mixed and pure groups, respectively).

Like Experiment 1, participants in each list group were randomly assigned to 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.

Materials and procedure

Experiment 2 used the same study lists as the previous experiment, with the following exception. Specifically, all forward pairs (e.g., *trout–fish*) were replaced with backward pairs (e.g., *fish–trout*). Additionally, two pure lists containing only backward pairs were created, providing 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 Appendix Tables 2 and 5 for stimuli properties).

Results

Figure 2 (top panel) displays mean recall rates as a function of encoding group for mixed-list participants. The bottom panel compares mean recall for pure-list groups. For completeness, comparisons between encoding tasks as functions of relatedness and list-type are reported in Appendix Table 6. Data screening followed the same criteria used in Experiment 1, and across groups, responses from 13 participants were omitted. As a result, 120 participants were included in the mixed-list analyses, and 226 participants in the pure-list analyses (see Table 1 for final group *n*s).

Mixed lists

A 2 (Pair Type) × 3 (Study Group) mixed ANOVA was used to test for reactivity effects within mixed lists. This analysis

yielded an effect of Pair Type, $F(1, 117) = 246.79$, $MSE = 87.63$, $\eta_p^2 = .68$, in which recall was higher for backward (43.90) than unrelated (24.43) pairs. The effect of Encoding Group was nonsignificant $F(2, 117) = 1.90$, $MSE = 600.55$, $p = .15$, $p_{BIC} = .62$, but the interaction was reliable, $F(2, 117) = 15.83$, $MSE = 87.63$, $\eta_p^2 = .22$. Post hoc tests confirmed the presence of positive reactivity for backward pairs, as recall was greatest for the frequency-judgment group (48.90), followed by the JOL (46.84) and no-JOL groups (34.85). All comparisons differed significantly (ts ≥ 2.72, ds ≥ 0.62), except between the JOL and frequency-judgment groups, $t < 1$, $p = .66$, $p_{BIC} = .89$. For unrelated pairs, recall rates were equivalent between the frequency (26.75), JOL (20.98), and no-JOL groups (25.45; ts ≤ 1.68, p_{BIC}s $\geq .69$), indicating no reactivity. Reactivity patterns observed with forward pairs in mixed lists therefore extended to backward pairs.

Pure lists

Next, a 2 (Pair Type: Backward vs. Unrelated) × 3 (Study Group: JOL vs. Frequency vs. No-JOL) between-subject ANOVA tested whether reactivity occurred for pure-list pairs. Consistent with previous analyses, a significant effect of pair type emerged, $F(1, 220) = 42.91$, $MSE = 312.67$, $\eta_p^2 = .16$, such that recall of backward pairs (41.95) exceeded recall of unrelated pairs (26.25). The effect of Encoding Group was non-significant, $F(2, 220) = 2.08$, $MSE = 312.67$, $p = .13$, $p_{BIC} = .65$, but the interaction between Pair Type and Encoding Group was at the conventional level of significance, $F(2, 220) = 2.95$, $MSE = 312.67$, $p = .05$, $p_{BIC} = .44$, $\eta_p^2 = .03$. 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, ts ≥ 2.08, ds ≥ 0.54, except for the comparison between JOLs and frequency judgments, $t < 1$, $SEM = 4.39$, $p = .67$, $p_{BIC} = .89$. Recall of unrelated pairs did not differ as a function of encoding group, ts ≤ 1.42, ps $\geq .16$, $p_{BIC} \geq .76$. Thus, positive reactivity patterns observed for backward pairs in mixed lists extended to pure lists.

Discussion

Experiment 2 tested whether reactivity patterns observed on forward pairs in Experiment 1 would occur using 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 JOLs are less likely to be available at test. Furthermore, the inclusion of both mixed and pure lists allowed for an additional test of the changed-goal hypothesis. Overall,

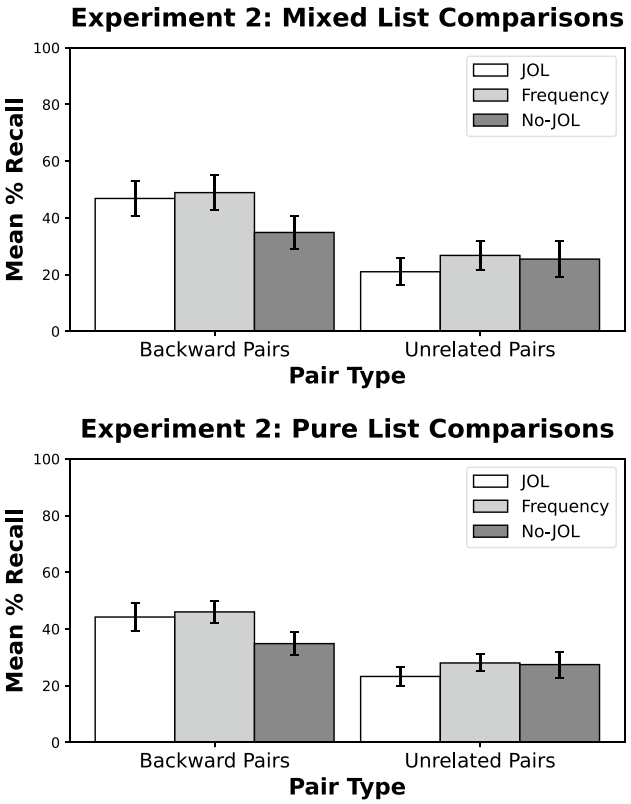

Fig. 2 Mean percentage recall for participants in Experiment 2 who completed the JOL, frequency-judgment, or No-JOL silent reading tasks for mixed lists (top panel) or pure lists (bottom panel). Error bars represent 95% confidence intervals

both JOLs and frequency judgments produced reactivity on backward pairs, regardless of list type. For unrelated pairs, however, no reactivity occurred. These findings are consistent with Experiment 1 and provide additional support for the cue-strengthening account, as reactivity was again not limited to only mixed lists where participants could distinguish between related and unrelated pairs.

In addition to testing the changed-goal and cue-strengthening accounts of reactivity, Experiment 2 provided a novel comparison by replacing forward pairs with backward pairs. In doing so, we compared backward and unrelated pairs in mixed- and pure-list designs and between JOLs and frequency judgments. While previous reactivity studies have traditionally compared between forward and unrelated pairs, we note two exceptions in which backward pairs were presented in mixed lists alongside other related and unrelated pairs. First, Mitchum et al. (2016) showed no differences in reactivity between forward or backward related pairs, as JOLs did not produce a reactive effect on either pair type. However, Maxwell and Huff (2022) showed that positive reactivity patterns on forward pairs extended to backward pairs, and further, these patterns occurred when participants made other judgment types that similarly emphasized pair relatedness (e.g., frequency judgments). Thus,

our findings in Experiment 2 are in line with Maxwell and Huff, while also demonstrating that positive reactivity on backward pairs extends to two novel list types: Mixed lists containing only backward and unrelated pairs, and pure lists of backward pairs.

Given the focus in the literature on forward associative pairs, Experiment 3 further tested for reactivity using a third type of related word pair: Symmetrical associates (e.g., *king–queen*) in which associative strength is balanced in both forward and backward directions. While backward pairs have been used in studies investigating JOL accuracy (e.g., Koriat & Bjork, 2005), few have examined JOLs on symmetrical pairs (cf. Maxwell & Huff, 2021), and only one study (Maxwell & Huff, 2022) has reported mixed-list reactivity patterns on symmetrical pairs. Additionally, no study has assessed reactivity effects using symmetrical associates presented in pure lists. Experiment 3 therefore examined reactivity effects in mixed and pure lists using symmetrical pairs while again assessing whether frequency judgments would continue to mirror JOL reactivity patterns. Thus, Experiment 3 provided an additional test of whether mixed-list reactivity patterns would extend to pure lists while further testing JOL reactivity accounts.

Experiment 3: Symmetrical versus 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 and when presented in isolation via pure lists. Like backward pairs, symmetrical pairs can be deceptive as they contain cues that are less likely to be available at test. 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 3 is important, as it provides a novel pair type with which to test for reactivity effects. Therefore, our use of symmetrical pairs provides a further test of the changed-goal and cue-strengthening accounts while also evaluating the generality of JOL reactivity effects. Based on the previous experiments, findings were expected to conform to a cue-strengthening pattern, 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 or whether participants made frequency judgments or JOLs.

Methods

Participants

A total of 227 participants were recruited to complete Experiment 3. Like the previous experiments, University of Southern Mississippi undergraduates ($n = 187$) completed the study online in exchange for course credit or were participants recruited through Prolific 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$). The 106 participants who studied pure unrelated lists in Experiment 1 again served as the pure unrelated comparison group. Therefore, pure-list groups contained a total of 220 participants. Group sizes were informed by the sample used in Experiment 1, and a sensitivity analysis via G*Power 3.1 confirmed that the mixed- and pure-list groups were sufficient for detecting small-to-medium main effects and interactions ($ds \geq 0.42$). Like the preceding experiments, participants within both list groups were randomly 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 Appendix Tables 2 and 7 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, comparisons between encoding tasks are provided in the Appendix (Table 8). Data screening followed the same procedure outlined in Experiment 2, and data from 18 participants were omitted (see Table 1 for group *n*s).

Mixed lists

Like the previous experiments, a 2 (Pair Type) × 3 (Study Group) mixed ANOVA was used to test for reactivity effects in mixed lists. An effect of Pair Type was found, $F(1, 103) = 825.46$, $MSE = 112.87$, $\eta_p^2 = .89$, as recall of symmetrical pairs (65.09) exceeded recall of unrelated

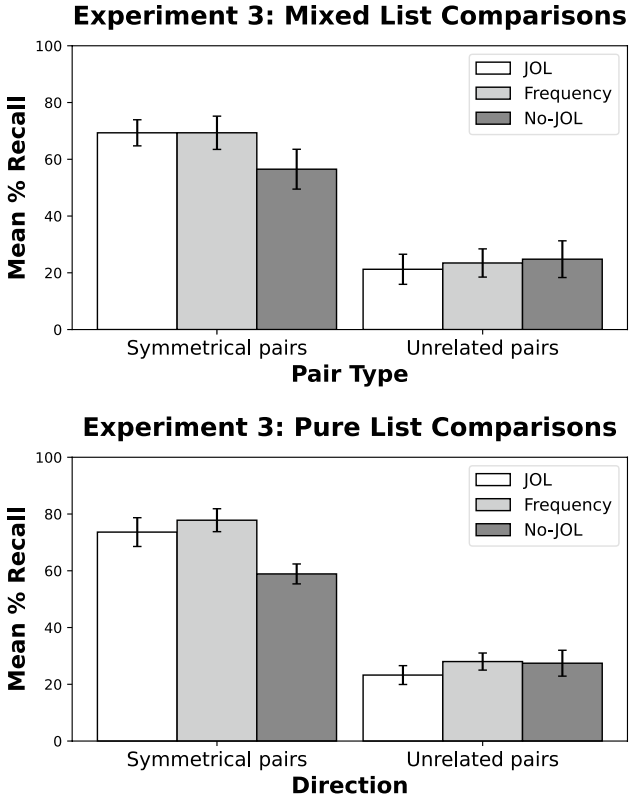

Fig. 3 Mean percentage recall for participants in Experiment 3 who completed the JOL, frequency-judgment, or No-JOL silent reading tasks for mixed lists (top panel) or pure lists (bottom panel). Error bars represent 95% confidence intervals

pairs (23.17). The effect of Encoding Group was non-significant, $F(2, 103) = 1.33$, $MSE = 497.13$, $p = .27$, $p_{BIC} = .96$, but an interaction was found, confirming the presence of a reactivity pattern, $F(2, 103) = 12.57$, $MSE = 112.87$, $\eta_p^2 = .20$. For symmetrical pairs, recall was highest following frequency judgments (69.34), JOLs (69.33), and the no-JOL control (56.51). Follow-up t tests confirmed that all comparisons differed significantly ($ts \geq 2.78$, $ds \geq 0.68$), except 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; $ts < 1$, $ps \geq .40$, $p_{BIC}s \geq .85$). Thus, mixed list reactivity patterns with forward and backward pairs extend to symmetrical pairs.

Pure lists

A 2 (Pair Type: Symmetrical vs. Unrelated) × 3 (Study Group: JOL vs. Frequency vs. No-JOL) between-subject ANOVA was then used to test whether reactivity effects for symmetrical pairs would extend to pure lists. Consistent with the previous experiments, this analysis yielded a significant effect of Pair Type, $F(1, 203) = 407.21$, $MSE = 246.60$, $\eta_p^2 = .67$, in which recall of symmetrical pairs (70.08) was greater than unrelated pairs (26.25). Additionally, significant effect of Encoding Group was detected, $F(2, 203) = 6.84$, $MSE = 246.60$, $\eta_p^2 = .06$, such that recall was highest for participants in the frequency-judgment group (52.57), followed by the JOL (47.31) and no-JOL groups (43.39). Post hoc tests, however, indicated that this effect was driven by difference between the frequency-judgment and no-JOL groups, $t(140) = 2.09$, $SEM = 4.44$, $p = .04$, $d = 0.35$. All other comparisons were non-significant, $ts \leq 1.06$, $ps \geq .29$, $p_{BIC}s \geq .90$. Importantly, a significant interaction was found, $F(2, 203) = 8.12$, $MSE = 246.60$, $\eta_p^2 = .07$. For symmetrical pairs, recall was highest for participants in the frequency-judgment group (77.81), followed by the JOL (73.63) and no-JOL groups (58.89). All comparisons differed significantly, $ts \geq 3.80$, $ds \geq 0.82$, 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 again did not differ between encoding groups, $ts \leq 1.42$, $ps \geq .16$, $p_{BIC} \geq .76$ (see Experiment 1). Thus, like the previous experiments, JOLs and frequency judgments again produced a positive reactivity effect, regardless of list type.

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. Both JOLs and frequency judgments again produced positive reactivity effects on related symmetrical pairs, but neither judgment type was reactive on unrelated pairs. Importantly, reactivity on symmetrical pairs occurred regardless of whether participants studied mixed or pure lists, further suggesting that reactivity is not contingent on the context in which items are studied. Thus, findings from Experiment 3 align with our previous experiments while providing additional support for a cue-strengthening account. Finally, our extension of positive reactivity to symmetrical associates is consistent with Maxwell and Huff (2022) and further suggests that reactivity can occur using pair types that are less likely to cue retrieval of the target word (e.g., backward and symmetrical associates).

General discussion

The present study tested the changed-goal and cue strengthening accounts of JOL reactivity by investigating 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, 2022; Soderstrom et al., 2015)

would emerge when pairs were presented in isolation via pure lists. In doing so, each experiment focused exclusively on one type of related pair type (forward, backward, or symmetrical) which were directly compared to unrelated pairs within both mixed- and pure-list contexts. A secondary goal was to further test whether reactivity effects were unique to JOLs. 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 patterns using mixed lists (e.g., Janes et al., 2018; Maxwell & Huff, 2022; Soderstrom et al., 2015), such that JOLs produced positive reactivity on forward pairs but were non-reactive on unrelated pairs. Importantly, this reactivity pattern extended to pure lists, indicating that reactivity is not driven by changes in participant study goals. Additionally, all observed reactivity on JOLs extended to frequency judgments, providing additional evidence that reactivity effects are driven by the encoding task strengthening relatedness cues used at retrieval rather than via a comparative process as posited by the changed-goal hypothesis. This replication of reactivity patterns with mixed lists and extension of previously reported reactivity patterns to pure lists adds to a growing body of literature indicating that JOLs are reactive on forward pairs, while also demonstrating that this reactivity is not contingent on list composition. Experiments 2 and 3 then showed that these positive reactivity patterns for both JOLs and frequency judgments extend to backward and symmetrical pair types, respectively. 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 the experimental design in which pairs are presented.

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 whenever metacognitive evaluations of pair difficulty produce shifts in study goals. However, this account cannot explain reactivity effects in pure lists, given that pure lists lack the easy/difficult comparison necessary to trigger a shift in study goals. Therefore, our pure-list reactivity findings are inconsistent with a changed-goal account. Concerning 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. Pure-list reactivity findings observed in the present study are consistent with this account.

Beyond replicating reactivity patterns observed in Experiment 1, Experiments 2 and 3 provided novel comparisons by extending these findings to backward and symmetrical pairs, respectively. The extension of positive reactivity patterns to each pair type further demonstrates the importance of pair relatedness as a determining factor of reactivity. Furthermore, the extension of this pattern to backward and symmetrical associates—pair types in which relatedness cues are less likely to be available at test—suggests that reactivity on related pairs may also occur in the absence of direct cue–target relations as is found in forward pairs.

Finally, 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 (2022) would similarly extend to pure lists. Like JOLs, frequency judgments direct attention towards relational aspects of study pairs without explicitly instructing participants to employ 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. Additionally, qualitative differences may exist in how JOLs and frequency judgments strengthen relatedness cues, as the former likely encourages processing cue–target relations with the goal of memory prediction, while the latter involves processing relatedness based on co-occurrence frequencies. 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 our comparison of mixed versus pure lists was designed to evaluate the changed-goal and cue strengthening accounts of reactivity, we note that the present study may also provide insight regarding participant strategy use. First, our finding that JOL reactivity extends to frequency judgments replicates previous work by Maxwell and Huff (2022). To explain this observation, Maxwell and Huff proposed that JOLs implicitly encourage participants to relate study pairs together at encoding. However, this relational encoding is applied strategically, such that only related pairs receive a memory benefit. Within this context, our finding that both JOL and frequency judgments are reactive on related pairs presented in pure lists may qualify this strategy use account, indicating that participants may be

able to apply a relational encoding strategy in both mixed and pure list contexts. The pure list pattern is important because it indicates that participants do not need to be exposed to unrelated pairs to instantiate relational encoding. Finally, we note that in addition to testing the cue-strengthening account, Rivers et al. (2021) also assessed strategy use by having participants report the encoding strategies used on each pair following retrieval of each target. Reported strategies did not differ between related and unrelated pairs; however, because strategy use was assessed at retrieval, this measure did not capture online strategy use at encoding. More work is therefore needed to fully understand the role that strategy use plays in JOL reactivity including whether strategies are shifted across items within a study set.

While the present study replicated previous work showing positive reactivity on related pairs, we note that for each experiment, participant study was self-paced. Although other studies investigating reactivity have also made use of self-paced study (e.g., Janes et al., 2018; Maxwell & Huff, 2022; Mitchum et al., 2016), the memory improvements observed for JOLs and frequency judgments could potentially be attributed to participants in the judgment groups encoding pairs for longer durations relative to the silent reading group. However, across experiments and list types, encoding durations were generally longer for participants in the control groups compared to the judgment groups (see Tables 9 and 10 in the Appendix). Thus, the reactivity effects observed in the present study do not appear to be driven by longer encoding durations and instead likely reflect additional processing due to making judgments at encoding.

Conclusion

Researchers have become increasingly interested in the reactive effects of immediate JOLs on cue–target word pairs. The present study tested the changed-goal and cue-strengthening accounts of reactivity by testing between mixed (e.g., related and unrelated pairs) and pure study lists (e.g., only unrelated pairs). Additionally, we assessed whether previously reported reactivity on frequency judgments—a non-metacognitive judgment task that similarly emphasizes cue–target relations—would replicate within this context (Maxwell & Huff, 2022). In doing so, we provided three separate tests of both list-type and encoding-task effects on reactivity while assessing these effects within the same study design. Overall, positive reactivity consistently emerged on related pairs, regardless of pair direction, but no reactivity was observed on unrelated pairs, replicating patterns previously reported on mixed lists (e.g., Janes et al., 2018; Maxwell & Huff, 2022; Soderstrom et al., 2015). Importantly, these patterns persisted, irrespective of judgment type (JOL vs. frequency) or list context (mixed vs. pure). Thus, the present study provides further evidence for a cue-strengthening account of JOL reactivity rather than a goal-changing account.

Appendix

Table 2 Summary statistics for associative overlap variables across each experiment

	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

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 3 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

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 5 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

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 4 Comparisons of mean recall percentages for each encoding task as a function of pair type in Experiment 1

List type	Pair type	Encoding task	*M* (± 95% CI)	JOL			Freq.		
				t(*df*)	*d*	p_{BIC}	*t*(*df*)	*d*	p_{BIC}
Mixed	Forward	JOL	75.59 (4.63)						
		Frequency	76.68 (5.11)	<1 (68)	0.07	.89			
		No-JOL	62.98 (6.01)	3.31 (69)	0.77*	–	3.34 (67)	0.82*	–
	Unrelated	JOL	18.14 (3.99)						
		Frequency	25.27 (6.18)	1.91 (68)	0.45	.58			
		No-JOL	21.86 (7.50)	<1 (69)	0.20	.85	<1 (67)	0.17	.87
Pure	Forward	JOL	83.19 (2.56)						
		Frequency	77.78 (4.60)	<1 (60)	0.35	.76			
		No-JOL	65.88 (4.11)	4.81 (63)	1.21*	–	2.62 (63)	0.65*	–
	Unrelated	JOL	23.25 (3.56)						
		Frequency	28.01 (3.27)	1.42 (70)	0.33	.76			
		No-JOL	27.43 (4.66)	1.00 (67)	0.24	.83	<1 (69)	0.03	.89

The two right-most column indicate *t* statistic, degrees of freedom, and Cohen's *d* for comparisons between encoding tasks, *$p < .05$. p_{BIC}s are only reported for non-significant comparisons. Freq. = frequency judgment

Table 6 Comparisons of mean recall percentages for each encoding task as a function of list and pair in Experiment 2

List type	Pair type	Encoding task	*M* (± 95% CI)	JOL			Freq.		
				t(*df*)	*d*	p_{BIC}	*t*(*df*)	*d*	p_{BIC}
Mixed	Backward	JOL	46.84 (6.07)						
		Frequency	48.09 (6.20)	<1 (81)	0.06	.89			
		No-JOL	34.85 (5.96)	2.72 (75)	0.62*	–	3.11 (78)	0.67*	–
	Unrelated	JOL	20.99 (4.71)						
		Frequency	26.75 (4.97)	1.68 (81)	0.36	.69			
		No-JOL	25.45 (6.47)	1.11 (75)	0.25	.82	<1 (78)	0.06	.89
Pure	Backward	JOL	44.21 (4.96)						
		Frequency	46.01 (3.76)	<1 (81)	0.13	.89			
		No-JOL	34.83 (3.97)	2.08 (76)	0.54*	–	2.91 (77)	0.66*	–
	Unrelated	JOL	23.25 (3.56)						
		Frequency	28.01 (3.27)	1.42 (70)	0.33	.76			
		No-JOL	27.43 (4.66)	1.00 (67)	0.24	.83	<1 (69)	0.03	.89

The two right-most column indicate *t* statistic, degrees of freedom, and Cohen's *d* for comparisons between encoding tasks, *$p < .05$. p_{BIC}s are only reported for non-significant comparisons. Freq. = frequency judgment; No-JOL = control group

Table 7 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

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 8 Comparisons of mean recall percentages for each encoding task as a function of list and pair type in Experiment 3

List type	Pair type	Encoding task	*M* (± 95% CI)	JOL			Freq.		
				t(*df*)	*d*	p_{BIC}	*t*(*df*)	*d*	p_{BIC}
Mixed	Symmetrical	JOL	69.34 (4.60)						
		Frequency	69.33 (5.86)	<1 (69)	<0.01	.99			
		No-JOL	56.51 (7.02)	3.02 (68)	0.76*	–	2.78 (69)	0.68*	–
	Unrelated	JOL	21.24 (5.30)						
		Frequency	23.46 (4.97)	<1 (69)	0.14	.87			
		No-JOL	24.80 (6.47)	<1 (68)	0.20	.85	<1 (69)	0.08	.89
Pure	Symmetrical	JOL	73.63 (4.04)						
		Frequency	77.81 (3.20)	1.12 (66)	0.26	.81			
		No-JOL	58.89 (3.51)	3.80 (65)	0.82*	–	5.53 (69)	0.96*	–
	Unrelated	JOL	23.25 (3.56)						
		Frequency	28.01 (3.27)	1.42 (70)	0.33	.76			
		No-JOL	27.43 (4.66)	1.00 (67)	0.24	.83	<1 (69)	0.03	.89

The two right-most column indicate *t* statistic, degrees of freedom, and Cohen's *d* for comparisons between encoding tasks, *$p < .05$. p_{BIC}s are only reported for non-significant comparisons. Freq. = frequency judgment; No-JOL = control group

Table 9 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	4,166	–	–	5,009
	Frequency	4,500	–	–	5,992
	Read	6,268	–	–	8,150
Exp. 2	JOL	–	5,527	–	4,995
	Frequency	–	5,444	–	5,179
	Read	–	5,396	–	5,801
Exp. 3	JOL	–	–	5,316	6,470
	Frequency	–	–	4,322	5,310
	Read	–	–	5,603	7,103

Cells display mean RTs in ms

Table 10 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	3,483	–	–	5,197
	Frequency	3,616	–	–	6,407
	Read	5,249	–	–	6,376
Exp. 2	JOL	–	6,398	–	5,197
	Frequency	–	5,743	–	6,407
	Read	–	6,561	–	6,376
Exp. 3	JOL	–	–	5,026	5,197
	Frequency	–	–	4,294	6,407
	Read	–	–	4,739	6,376

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

Funding Study materials, data files, and R code used for analyses have been made available via OSF (https://osf.io/3fztn/). Experiments reported in this study were used to partially satisfy the dissertation requirements of the first author.

References

Akdoğan, E., Izaute, M., Danion, J., Vidailhet, P., & Bacon, E. (2016). Is retrieval the key? Metamemory judgment and testing as learning strategies. *Memory, 24*(10), 1390–1395.

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.

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). 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.

Koriat, A., Sheffer, L., & May'ayan, H. (2002). Comparing objective and subjective learning curves: Judgments of learning exhibit increased underconfidence with practice. *Journal of Experimental Psychology: General, 131*(2), 147–162.

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

Maxwell, N. P., & Huff, M. J. (2022). Reactivity from judgments of learning is not only due to memory forecasting: Evidence from associative memory and frequency judgments. *Metacognition and Learning, 17*, 589–625.

Meeter, M., & Nelson, T. O. (2003). Multiple study trials and judgments of learning. *Acta Psychologica, 113*(2), 123–132.

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

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

Nelson, D. L., McEvoy, C. L., & 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.

Rhodes, M. G. (2016). Judgments of learning. In J. Dunlosky & S. K. Tauber (Eds.), *The Oxford handbook of metamemory* (pp. 65–80). Oxford University Press.

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying judgments of learning on metacognitive accuracy: A meta-analytics review. *Psychological Bulletin, 137*(1), 131–148.

Rivers, M. L., Janes, J. L., & Dunlosky, J. (2021). Investigating memory reactivity with a within-participant manipulation of judgments of learning: Support for the cue-strengthening hypothesis. *Memory, 29*(10), 1342–1353.

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.

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: Judgments of learning may alter what they are intended to assess. *Psychological Science, 3*(5), 315–317.

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.

Open practices statement The data for all experiments have been made available online (https://osf.io/3fztn/). None of the experiments were preregistered.