Judgments of Learning Improve Memory for Word Lists via Enhanced Item-Specific Encoding: Evidence from Categorized, Uncategorized, and DRM Lists

Nicholas P. Maxwell

Midwestern State University

Word Count: 8509

**Author Note**

Correspondence concerning this article should be addressed to Nicholas P. Maxwell, Department of Psychology, Midwestern State University, 3410 Taft Blvd, Wichita Falls, TX, United States. Study materials, data files, and *R* code used for all analyses have been made available via OSF: https://osf.io/t453a/.

Abstract

Judgments of Learning (JOLs) have been repeatedly shown to be reactive on memory. However, the specific processes underlying JOL reactivity differ based on the type of stimuli participants study and the method by which memory is assessed. Recently, item-specific encoding has been proposed as a mechanism explaining JOL reactivity on word list learning. To test this account, participants studied categorized and uncategorized word lists (Experiments 1A/1B) or DRM lists (Experiment 2) while providing item-level JOLs, global JOLs, or silently reading each word. Across experiments, item-level JOLs improved correct memory for all list types but only when recognition testing was used (Experiments 1B and 2). Separately, global JOLs improved free-recall of categorized but not uncategorized lists (Experiment 1A) and were non-reactive on correct recognition (Experiments 1B and 2). Finally, Experiment 2 found that global but not item-level JOLs increased the DRM false memory illusion. Taken together, when JOLs are elicited individually, they improve memory via item-specific processes. However, when JOLs emphasize list-wise relations (e.g., global JOLs), reactivity may instead reflect a relational encoding process.

Word Count: 173

*Keywords*: Judgments of Learning; Reactivity; Categorized Lists; DRM Lists; Recognition

Judgments of Learning Improve Memory for Word Lists via Enhanced Item-Specific Encoding: Evidence from Categorized, Uncategorized, and DRM Lists

When learning new information, it is critical to assess whether the knowledge being acquired has been sufficiently encoded. Metamemory, or the processes by which individuals monitor their learning, is crucial as it helps individuals identify which items need additional study while simultaneously letting them modify their actions (see T. O. Nelson & Narens, 1990). To investigate questions surrounding metamemory, researchers commonly have participants make Judgments of Learning (JOLs), which involve studying items (commonly cue-target word pairs) and predicting the likelihood of correctly remembering them on a later test (see Rhodes, 2016; Schwartz & Metcalfe, 2017, for reviews). Thus, JOLs provide researchers with a simple measure for assessing how various factors affect the study process, including perceptual features (e.g., font-size; Rhodes & Castel, 2008; Chang & Brainerd, 2022) and the extent to which studied items are related (e.g., Koriat & Bjork, 2005; Castel, McCabe, & Rhodes, 2007).

Historically, JOLs have been viewed as neutral measures with little influence on memory, particularly when participants are instructed to elicit them concurrently with or immediately following study (though see Spellman & Bjork, 1992, who argued that delayed JOLs were likely to impart a memorial benefit). However, recent studies have repeatedly demonstrated that JOLs are *reactive* on learning (e.g., Janes, Rivers, & Dunlosky, 2018; Maxwell & Huff, 2022; Mitchum, Kelley, & Fox, 2016; Soderstrom, Clark, Halamish, & Bjork, 2015; see Double, Birney, & Walker, 2018, for review). As such, making JOLs appears to modify participants’ memory for studied items, likely by making certain features of the stimuli more salient at encoding (Ericsson & Simon, 1993). Thus, providing JOLs at encoding could potentially produce memory benefits (i.e., *positive reactivity*) or costs (i.e., *negative reactivity*), which would be evident when comparing memory for participants making JOLs at encoding to a separate group completing a no-JOL control task (e.g., silent reading).

Several theories have been proposed to explain JOL reactivity. For example, the changed-goal hypothesis (Mitchum et al., 2016) suggests that the requirement to provide JOLs at encoding alerts participants to salient differences in item difficulty which subsequently causes them to shift from an initial study goal of mastering all list items to only mastering those which they perceive as being easy to learn. As such, this hypothesis predicts positive reactivity on easy items (e.g., related study items) and negative reactivity on more difficult items (e.g., unrelated items). Alternatively, Soderstrom et al. (2015) proposed a cue-strengthening account which posits that JOLs strengthen intrinsic cues which participants use to inform the magnitude of their judgments (see Koriat, 1997) and, importantly, that these cues will benefit memory so long as the method of testing is sensitive to the specific cues which have been strengthened. Because relatedness cues are highly salient, the cue-strengthening account therefore predicts positive reactivity on related study materials (e.g., paired-associates). However, based on this account, unrelated stimuli may still show positive reactivity so long as the method of testing is sensitive to other cues which may also be strengthened (e.g., familiarity cues; see Maxwell & Huff, 2024).

Overall, findings from studies investigating JOL reactivity with cue-target word pairs have largely supported the cue-strengthening account (e.g., Chang & Brainerd, 2023; Maxwell & Huff, 2023) and, moreover, suggest that JOLs specifically encourage relational encoding when elicited for this stimuli type (e.g., Halamish & Undorf, 2023; Maxwell & Huff, 2024; Rivers, Dunlosky, Janes, Witherby, & Tauber, 2023). Importantly, these studies have revealed that positive JOL reactivity is generally moderated by the presence of pre-existing cue-target relations. Specifically, when pairs are semantically related (e.g., dog – paw), JOLs generally improve memory for the target item; however, this memory benefit does not extend to unrelated pairs (e.g., dog – cup) and these pairs may even show negative reactivity (see Undorf, Schäfer, & Halamish, 2024). Additionally, participants often indicate on post-study questionnaires that perceived relatedness was the primary factor which they considered when assigning their JOLs (Rivers et al., 2023). Taken together, these findings are consistent with a cue-strengthening account and, moreover, suggest that JOLs specifically encourage participants to reflect upon relational properties of stimuli at encoding.

**JOL Reactivity and Item-Specific Encoding**

Although previous JOL reactivity studies have often emphasized the link between cue-target relations and reactivity, there is growing evidence that JOLs also encourage *item-specific encoding* of stimuli, particularly when participants study words that are presented individually in lists rather than cue-target pairs (e.g., Senkova & Otani, 2021; Zhao et al., 2023). Per the item-specific/relational framework (Einstein & Hunt, 1980; Hunt & Einstein, 1981), encoding tasks differ regarding their likelihood of encouraging participants to process unique properties of stimuli that differentiate between individual items (i.e., item-specific encoding) or properties which are shared between items (i.e., relational encoding). While item-specific and relational encoding tasks emphasize different aspects of studied materials, both have been shown to improve memory performance relative to silent reading, with item specific information being particularly beneficial for recognition memory and relational and item-specific information facilitating free-recall. Thus, JOL reactivity could potentially reflect enhanced item-specific encoding, relational encoding, or a combination of both processing types, though whether JOL reactivity ultimately reflects item-specific or relational processes may be contingent upon the method by which memory is tested.

Importantly, whether JOLs are more likely to encourage item-specific or relational encoding may also depend upon the type of stimuli which participants study. Because previous JOL reactivity studies have extensively used cue-target word pairs, which easily lend themselves to relational encoding (i.e., related cue-target pairs have highly salient relatedness cues), theories of JOL reactivity have often emphasized the role of relational encoding as an underlying of reactivity (e.g., Soderstrom et al.’s, 2015 cue-strengthening account). However, JOLs may also encourage item-specific encoding, particularly when they are provided for words presented individually in a list rather than cue-target pairs. Like cue-target pairs, word lists may similarly contain pre-existing relations (i.e., words within a list can be categorized around a specific topic or theme). However, how individuals process listwise relations when making JOLs differs from how cue-target pair relations are processed (see Koriat, 1997). Furthermore, compared to cue-target pairs, word lists are by nature more likely to encourage item-specific encoding, as words are often presented in isolation in which list-wise relations are less salient relative to pairs. As a result, making JOLs on individual words (i.e., item-level JOLs) may be more likely to encourage item-specific rather than relational encoding compared to when they are elicited on cue-target pairs.

Recent studies have explored the potential for item-level JOLs to encourage item-specific encoding of individual words. For example, Senkova and Otani (2021) reported that making JOLs individually for each word in a study list improved free-recall of categorized word lists and, furthermore, demonstrated that the memorial benefits of these JOLs approximated benefits observed when participants used other encoding tasks which similarly emphasized item-specific processing (e.g., pleasantness ratings in Experiment 1 and an imagery task in Experiment 2). Separately, Zhao et al. (2023) found that item-level JOLs were non-reactive on free-recall but instead improved recognition when participants studied uncategorized lists (which relies extensively upon item-specific processing). Importantly, Zhao et al. also demonstrated that item-level JOLs impaired temporal order memory for list items (which was assessed via an order reconstruction task that is inherently relational).

To explain their findings, Zhao et al. (2023) proposed an item-order account of JOL reactivity (see also McDaniel & Bugg, 2008). Based on this account, word lists contain item-specific information regarding the individual words and relational information regarding potential list-wise relations. Importantly, this account proposes a dissociation between item-specific and relational information, such that tasks emphasizing one aspect impair the other. As such, the item-order account makes diverging predictions regarding reactivity based on test format. Specifically, the item-order account predicts that item-level JOLs would produce positive reactivity on word lists, but only when memory is assessed via recognition testing. However, because free-recall also relies upon relational cues (McDaniel & Bugg, 2008; Rawson & Zamary, 2019), the item-order account predicts no memorial benefits when participants are tested using this method, as item-level JOLs appear to inhibit relational cues. Item-level JOLs would therefore not be expected to be reactive on free-recall or could even produce negative reactivity within this context. As such, the item-order account predicts that item-level JOL reactivity is moderated by test type, with recognition but not free-recall showing a memorial benefit.

Consistent with Zhao et al.’s (2023) item order account, Kaya and Mulligan (in press) recently reported that JOL reactivity on free-recall also exhibits design effects that are in line with an item-specific/relational account of reactivity. Specifically, the authors found that JOL reactivity was greater when participants studied items in a mixed list context (i.e., participants made JOLs for some but not all studied items) rather than in separate pure lists (i.e., JOLs were elicited for all items in one list but no items in the other). To explain this finding, Kaya and Mulligan proposed that JOLs impair relational encoding of all list items, including adjacent no-JOL items, while simultaneously encouraging item-specific encoding only on the individual items which received JOLs. Thus, unlike reactivity observed with cue-target word pairs, which is thought to reflect an inherently relational process (see Halamish & Undorf, 2023; Maxwell & Huff, 2024), there is growing evidence that item-level JOL reactivity observed on word lists reflects item-specific encoding.

**The Present Study**

As noted above, previous research has demonstrated that while item-level JOLs are generally reactive on word lists, these effects are moderated by test type, with reactivity occurring when memory is assessed via recognition but not free-recall testing. Therefore, the present study assessed JOL reactivity across two testing contexts: Free-recall (Experiment 1A) and recognition testing (Experiments 1B and 2). Additionally, because few studies have directly compared reactivity between item-level and global JOLs (see Chang & Brainerd, 2024, for a recent exception), each experiment also tested for changes in memory for participants who provided global JOLs following each study list. Unlike item-level JOLs which are thought to encourage item-specific encoding, global JOLs would be more likely to encourage processing of intra-list relations, particularly when all list items are centered around a specific category. As such, each experiment compared memory between participants who provided item-level JOLs, global JOLs, and a no-JOL control task.

First, Experiment 1A sought to replicate findings that item-level JOLs are non-reactive on free-recall (Chang & Brainerd, 2024; Zhao et al., 2023), while Experiment 1B was designed to replicate findings showing that item-level JOLs are reactive on uncategorized lists when recognition testing is used (e.g., Zhao et al., 2023) and, additionally, test whether this effect would extend to categorized lists. In doing so, Experiments 1A/1B provided separate tests of the item-order account, which makes diverging predictions regarding item-level JOL reactivity when item memory is assessed via free-recall and recognition testing.

Second, Experiment 2 used Deese-Roediger-McDermott lists (DRM; Deese; 1959; Roediger & McDermott, 1995) to test the item-order account’s central claim that item-level JOLs specifically encourage item-specific but not relational encoding of words. Like categorized lists, DRM lists center around a specific category exemplar. However, all items in DRM lists are also strongly related to a non-studied critical lure (e.g., *bed*, *rest*, *dream*, *nap*, and *pillow* are each related to the non-presented word *sleep*). At test, false memory for the critical lure is generally high, regardless of whether memory is assessed via free-recall or recognition (see Gallo, 2006). However, previous studies have found that item-specific and relational encoding tasks differentially affect the DRM illusion on recognition memory, with item-specific, but not relational tasks reducing this illusion compared to silent reading (e.g., McCabe, Presmantes, Robertson, & Smith, 2004; Huff & Bodner, 2013; Huff & Bodner, 2019). Thus, Experiment 2 provided an additional test of the item-order account while also assessing the contributions of item-specific and relational processing to JOL reactivity.

**Experiment 1A: Categorized versus Uncategorized Lists and Free-Recall Testing**

Experiment 1A tested the item-order account of JOL reactivity while also testing whether global JOLs would be reactive on free-recall of categorized and uncategorized word lists. As such, free-recall was compared between three groups: Participants who provided concurrent JOLs individually for each item (item-level JOLs), participants who made a single, list-wise JOL immediately following each list (global JOLs), and a no-JOL control group in which participants silently read each item without providing JOLs. Based on the item-order account, item-level JOLs were not expected to improve free-recall of categorized lists, given that this JOL type should inhibit the relational processes that facilitate free-recall (see Table 1 for a summary of anticipated JOL reactivity patterns based on the item-order account).

For global JOLs, the item-order account makes no specific claims regarding whether this JOL type would facilitate or inhibit free-recall relative to the no-JOL group. However, based on the item-specific/relational framework (Einstein & Hunt, 1980; Hunt & Einstein, 1981), tasks which emphasize intra-list relations are likely to improve free-recall. Therefore, any positive reactivity on global JOLs would be taken as evidence suggesting that global JOLs encourage relational encoding, particularly if this effect is moderated by list relatedness (i.e., positive reactivity on categorized lists but no or negative reactivity on uncategorized lists). As such, Experiment 1A provided an additional test of the item-order account while also testing for potential global JOL reactivity within this context.

**Method**

**Participants**

One hundred twenty-nine participants were recruited from Prolific (www.prolific.co) and were compensated at a rate of $4.00 per 20-minute session. This sample size was informed by an a priori power analysis conducted with *G\*Power 3.1* (Faul, Erdfelder, Buchner, & Lang, 2009), which suggested that 111 participants would be needed to detect small-to-medium main effects/interactions (*α* = .05, 1 – *β* = .80, *d* = 0.30). However, data collection was extended as funding permitted to account for increased variability from online testing. Data were omitted from seven participants who failed to complete the filler task (which suggested failure to adhere to task instructions) and 9 participants with extreme recall rates (above 95% which suggested cheating or below 5% which suggested failure to attend to the study lists). As such, a total of 113 participants were included in the final dataset (item-level JOLs, *n* = 36; global JOLs, *n* = 38; no-JOL control group, *n* = 39). All participants were native English speakers and were required to have obtained at least a high school degree or equivalent.

**Materials**

Ninety-six English words served as stimuli and were arranged into eight 12-word lists. Of these lists, four were categorized, such that each word was an exemplar of a specific category. Categorized lists were taken from Van Overschelde, Rawson, and Dunlosky’s (2004) categorical word norms. The four remaining lists were uncategorized based on these norms and, as such, all words within uncategorized lists were semantically. Uncategorized lists were created by randomly selecting words from unused categories in the Van Overschelde et al. norms with the constraint that all words within each uncategorized list were taken from separate categories. The eight lists were then arranged into two counterbalanced sets of four lists. Each counterbalance contained two categorized and two uncategorized lists. Importantly, categorized and uncategorized word lists were matched between counterbalances on SUBTLEX frequency (Brysbaert & New, 2009), concreteness (D. L. Nelson, McEvoy, & Schrieber, 2004), and length. A series of 2 (Counterbalance Version: A vs. B) × 2 (List Type: Categorized vs. Uncategorized) between subjects ANOVAs confirmed that lists did not differ on these lexical properties; *F*s ≤ 2.71, *p*s ≥ .10 (see Appendix Table A1 for mean lexical/semantic properties for each counterbalance). The full stimuli set used in Experiments 1A/1B has been made available at: https://osf.io/9qxne.

**Procedure**

Experiment 1A was administered online using Collector, a platform for conducting online, browser-based cognitive psychology experiments (Garcia & Kornell, 2015). Following informed consent, participants were randomly assigned to one of the two list counterbalances and one of the three encoding groups (item-level JOL, global JOL, or no-JOL control task). Participants in all groups were told that they would be viewing a series of words which would be presented one at a time on their computer screen and that their memory for each word would be tested. The two JOL groups then received further instructions regarding their respective judgments. Participants in the item-level JOL group were asked to provide a probability rating reflecting their ability to remember each word on a later memory test. Item-level JOLs were elicited concurrently with study and were provided while each word was displayed on the screen. Separately, participants in the global JOL group were informed that following the last item in each list, they would be asked to provide a single JOL representing their ability to remember the previously studied items on a later test. Global JOLs were framed as the percentage of list items which would be correctly remembered at test. Thus, for both groups, JOLs were framed as the percent likelihood of correctly remembering items and were elicited via a continuous 0-100 scale (i.e., 0 = definitely will not remember, 100 = definitely will remember). To mitigate potential anchoring effects, participants in both groups were instructed to be as accurate as possible when providing their JOLs and were encouraged to use the full range of the response scale. Participants in the no-JOL group did not receive additional judgment instructions and were instead simply instructed to silently read each pair.

After receiving their respective encoding instructions, participants began studying the first list. For all groups, encoding was self-paced, and participants pressed the ENTER key to advance to the next word. Item-level JOL participants provided their JOLs concurrently with each word, and global JOL participants were prompted to enter their list-wise JOL immediately following the last list item. Following the first study phase, participants completed a short filler task in which they were presented with a random consonant and were instructed to list as many words starting with the letter as they could generate in a 30 second span (e.g., list all words starting with the letter “M”). After the time limit had been reached, participants completed a free-recall task which prompted them to type as many previously studied words as they could recall. For all encoding groups, free-recall was self-paced, with participants clicking a button at the bottom of the screen to advance. After completing the free-recall task, participants began the second study list. This process then repeated until participants had cycled through all four study/recall blocks. List presentation order was randomized for all participants, and all words were additionally randomized within lists. The full experiment took approximately 20 minutes to complete.

**Experiment 1B: Categorized versus Uncategorized Lists and Recognition Testing**

Experiment 1B assessed whether previously reported reactivity patterns observed with recognition testing on uncategorized word lists (e.g., Zhao et al., 2023) would replicate and, additionally, whether this effect would extend to categorized lists. Based on an item-order account, item-level JOLs were expected to produce positive reactivity within this context. Additionally, prior work has found that JOL reactivity effects are generally in the positive direction and are larger for recognition memory versus free-recall (see Zhao et al., 2023). Consistent with this account, Maxwell and Huff (2024) recently demonstrated that item-level JOLs improve recognition of cue-target word pairs, regardless of whether pairs are related or unrelated. Taken together, item-level JOLs were similarly expected to improve recognition memory for both categorized and uncategorized word lists compared to the no-JOL control group.

Finally, like the previous experiment, Experiment 1B also included a separate group of participants who provided global JOLs following the presentation of each study list. However, given that prior studies have not investigated global JOL reactivity using recognition testing, it was unclear the extent to which global JOLs would be reactive on recognition. Because recognition testing is more likely to benefit from item-specific than relational encoding, any potential reactivity from global JOLs was expected to be smaller than item-level JOLs. Thus, Experiment 1B provided an additional test of the item-order account while further exploring the effects of global JOLs on word list learning.

**Method**

**Participants**

A total of 119 undergraduate students were recruited from Midwestern State University and completed Experiment 1B online in exchange for partial course credit. Participants were randomly assigned to one of the three encoding groups described in Experiment 1A. Initial sample sizes for each group were based on Experiment 1A, and the same exclusion criteria were applied. This process removed six participants from the final dataset, leading to a total of 113 participants included in the following analyses (item-level JOLs, *n* = 37; global JOLs, *n* = 40; no-JOL control group, *n* = 36). A sensitivity analysis conducted with *G\*Power 3.1* suggested that the final sample had sufficient power to detect small-to-medium main effects/interactions (*α* = .05, 1 – *β* = .80, *d* = 0.29). All participants were native English speakers.

**Materials and Procedure**

Experiment 1B used the same materials and followed the same general procedure as Experiment 1A, with the following exceptions. First, to account for potential ceiling effects on recognition, participants studied the four lists from each counterbalance back-to-back with no memory tests in between, rather than having their memory tested separately for each list. Thus, following the final item in the first three lists, participants in all groups were informed that they would be presented with a new word list and that they would again be asked to use their respective encoding strategy while studying (item-JOLs, global JOLs, or silent reading). Additionally, like the previous experiment, global JOL participants were instructed to provide their JOLs after the last item in each list.

Second, after the final study list, participants completed a 60 second filler task in which they were asked to list as many words as possible beginning with a randomly selected consonant. Immediately afterwards, participants began the test phase. Participants were informed that they would be viewing a series of words and that some of the words may have been previously studied. They were then instructed to indicate whether each word had been previously studied (“old”) or had not been previously studied (“new”). This test included all 48 items from the previously studied lists, as well as the 48 items from the counterbalanced list-set, which served as non-presented controls. Thus, the old/new recognition test consisted of 96 items, which were presented in a randomized order for each participant. Test performance was self-paced; however, participants were instructed to respond as quickly as possible without compromising their accuracy. All other materials, including the categorized lists and the three sets of encoding instructions (item-JOL, global JOL, and silent reading), were identical to Experiment 1A. Like the previous experiment, list presentation order was randomized, and items within lists were additionally randomized for each participant. Experiment 1B took approximately 20 minutes to complete.

**Results**

For all analyses, significance was set at *p* < .05. Partial eta-squared (*ηp*2)is reported for all main effects and interactions, and Cohen’s *d* effect sizes are reported for post-hoc comparisons. Finally, all non-significant main effects and post-hoc comparisons include a Bayesian strength estimation of the evidence supporting the null hypothesis (see Masson, 2011; Wagenmakers, 2007). This analysis compares two models, which one assume a significant effect and a null effect, respectively, and allows for the computation of a probability estimate (termed *p*BIC; Bayesian Information Criterion). Thus, *p*bics provide a probability estimate of the null hypothesis being retained, such that higher values denote greater strength of evidence in favor of retaining the null hypothesis. Importantly, *p*BICs are sensitive to sample size, providing increased confidence in reported null effects. Finally, free-recall data collected in Experiment 1A was scored in *R* using the *lrd* package (Maxwell, Huff, & Buchanan, 2022), which provides a set of tools scoring lexical data from memory studies while correcting potential spelling errors. Figure 1 depicts changes in free-recall proportions (top panel) and hit-rates/false alarms (bottom panel) as functions of encoding group and list type. For completeness, all recall/recognition comparisons are reported in Appendix Tables A2 and A3, mean JOLs are reported in Appendix Table A4, and self-paced encoding latencies are reported in Table A5.

***Experiment 1A.*** Free-recall responses were analyzed using a 3 (Encoding Group: Item-Level JOL vs. Global JOL vs. No-JOL) × 2 (List Type: Categorized vs. Uncategorized) mixed-measures ANOVA. Overall, the main effect of Encoding Group was non-significant, *F*(2, 110) = 2.46, *MSE* = .04, *p* = .09, *ηp*2 = .04, *p*bic = .90. However, a significant main effect of list-type indicated that across encoding groups, recall was higher for categorized lists versus uncategorized lists (.72 vs. .54, respectively; *F*(1, 110) = 172.70, *MSE* = .01,*p < .*001, *ηp*2 = .61). Importantly, a significant Encoding Group × Pair Type interaction confirmed the presence of a JOL reactivity pattern, *F*(2, 110) = 3.94, *MSE* = .01, *p* = .02, *ηp*2 = .07. Starting with categorized lists, providing global JOLs increased free-recall relative to the no-JOL group (.76 vs. .68; *t*(75) = 2.38, *SEM* = .03, *p* = .02, *d* = 0.57). However, free-recall for the item-level JOL group (.71) did not statistically differ from the no-JOL and global JOL groups (*t*s ≤ 1.58, *p*s ≥ .12, *d*s ≤ 0.36, *p*BICs ≥ .71). For uncategorized lists, free-recall was highest for participants in the global JOL group (.58), followed by the no-JOL control (.55) and the item-JOL group (.48). All comparisons were non-significant (*t*s ≤ 1.67, *p*s ≥ .10, *d*s ≤ 0.34, *p*BICs ≥ .87) except for the comparison between the item-level JOL and global groups, *t*(73) = 2.16, *SEM* = .05, *p* = .03, *d* = 0.51. Thus, the memorial benefits of global JOLs on categorized lists did not extend to uncategorized lists and, additionally, providing item-level JOLs resulted in a memory cost to this list type relative to global JOLs but not the no-JOL control group.

***Experiment 1B.*** Next, hit rates were analyzed via a 3 (Encoding Group: Item-Level JOL vs. Global JOL vs. No-JOL) × 2 (List Type: Categorized vs. Uncategorized) mixed ANOVA. First, this analysis yielded a significant main effect of Encoding Group, *F*(2, 110) = 16.39, *MSE* = .03, *p* < .001, *ηp*2 = .23, as collapsed across list types, hit rates were highest for participants in the item-level JOL group (.89), followed by the no-JOL (.73) and global JOL groups (.75). Post-hoc testing confirmed that all groups significantly differed (*t*s ≥ 4.61, *ps* < .001, *d*s ≥ 1.07), except for the comparison between the no-JOL and global JOL groups, which was non-significant, *t*(74) < 1, *SEM* = .03, *p* = .45, *d* = 0.15, *p*BIC = .87. Next, a significant main effect of List Type was detected *F*(1, 110) = 67.72, *MSE* = .01, *p* < .001, *ηp*2 = .38, as hit rates were higher for categorized lists versus uncategorized lists (.85 vs. .73, respectively). Finally, this analysis revealed a significant Encoding Group × List Type interaction, *F*(2, 110) = 8.23, *MSE* = .01, *p* < .001, *ηp*2 = .13. Post-hoc testing revealed that hit rates for the item-level JOL group exceeded the no-JOL group for categorized lists (.91 vs. .81) and uncategorized lists (.87 vs. .68; *t*s ≥ 3.03, *p*s≤ .003, *d*s ≥ 0.77). Hit rates for the item-level JOL group similarly exceeded hit rates for the global JOL group for both list types (categorized lists: .91 vs. .83; uncategorized lists: .87 vs. .68; *t*s ≥ 2.65, *p*s≤ .01, *d*s ≥ 0.64). However, no differences in hit rates were detected between the global JOL and no-JOL groups for either list type (categorized lists: .83 vs. .81; uncategorized lists: .68 vs. .65; *t*s ≤ 1, *p*s ≥ .51, *d*s ≤ 0.17, *p*BICs ≥ .87). Thus, the requirement to provide JOLs at encoding benefited hits for both list types, but only when JOLs were elicited individually for each item.

Regarding false alarms (i.e., false recognition of non-presented control items), a significant difference emerged between encoding groups, *F*(2, 110) = 3.91, *MSE* = .02, *p* = .02, *ηp*2 = .07. False alarms were highest for participants in the no-JOL group (.21), followed by the global JOL group (.14) and the item-level JOL group (.11). Post-hoc testing indicated that this effect was driven by differences between the item-level and no-JOL groups, *t*(71) = 2.47, *SEM* = .04, *p* = .02, *d* = 0.61. The difference between the global and no-JOL groups was marginal, *t*(74) = 1.94, *SEM* = .04, *p* = .06, *d* = 0.44, *p*BIC = .78, and no difference in false alarms was detected between the item and global JOL groups, *t*(75) < 1, *SEM* = .03, *p* = .40, *d* = 0.23, *p*BIC = .86.

Finally, following the designs of Myers et al. (2020) and Maxwell and Huff (2024), signal detection analyses were used to test for differences in discriminability (*d′*) and response criterion (*c*) as functions of encoding group[[1]](#footnote-1). Signal detection indices were computed in *R* via the *psycho* package (Makowski, 2018), which accounts for extreme scores via a log-linear correction based on Hautus’s (1995) guidelines. Overall, mean *d′* differed between encoding groups, *F*(2, 110) = 16.32, *MSE* = 0.75, *p* < .001, *ηp*2 = .23. Mean *d′* was greatest for the item-level JOL group (2.76), followed by the global JOL group (1.98), and the no-JOL control group (1.64). All comparisons differed significantly, *t*s ≥ 3.77, *p*s< .001, *d*s ≥ 0.85, except for the comparison between the global and no-JOL groups which did not reach significance, *t*(74) = 1.83, *SEM* = 0.19, *p* = .07, *d* = 0.42, *p*BIC = .62. Mean *c*, however, did not differ between encoding groups, *F*(1, 110) = 2.51, *MSE* = 0.14, *p* = .09, *ηp*2 = .04, *p*BIC = .46.

**Discussion**

Experiments 1A and 1B provided separate tests of the item-order account of JOL reactivity by testing for item-level JOL reactivity on categorized and uncategorized lists when memory was assessed via free-recall (Experiment 1A) and recognition (Experiment 1B). Additionally, both experiments included a global JOL comparison group in which participants provided a single JOL for each list. As such, each experiment provided a comparison between JOLs task which differentially emphasized item-specific and relational encoding. Starting with Experiment 1A, item-level JOLs did not improve free-recall, regardless of list relatedness, a pattern consistent with previous research by Zhao et al. (2023) and an item-order account of reactivity. Global JOLs, however, improved free-recall of categorized lists relative to the no-JOL control group. Taken together, findings from Experiment 1A provide further evidence for an item order account, which predicts that item-level JOLs should be non-reactive within this context due to this judgment type inhibiting processing of list-wise relations. Separately, the finding global JOLs improved free-recall for items with pre-existing relations suggests that this JOL type may specifically encourage participants to engage in relational encoding, so long as lists contain salient, pre-existing relations.

Next, Experiment 1B found that item-level JOLs improved recognition memory of all items, regardless of whether they were studied in categorized or uncategorized lists. Importantly, this replicates previous findings by Zhao et al. (2023) who similarly reported positive reactivity on uncategorized word lists and additionally extends it to include categorized lists. Global JOLs, however, were non-reactive on recognition, regardless of list type. Finally, signal detection analyses suggested that both types of JOLs generally improved participants’ ability to discern between old and new items compared to the no-JOL group, though increased criterion levels for JOL participants suggest that making judgments may have biased participants’ responses. Considered alongside Experiment 1A, these findings suggest that item-level JOL reactivity is dependent upon test format, with recognition testing, but not free-recall, being particularly sensitive to reactivity from this JOL type.

Overall, the finding that item-level JOLs improved recognition but not free-recall is consistent with an item-order account of reactivity and provides further evidence that JOLs encourage item-specific processing when they are elicited individually for each word in a list. Additionally, the finding that global JOLs only benefitted free-recall of categorized lists aligns with previous studies demonstrating that JOLs improve memory for related but not unrelated cue-target pairs (e.g., Janes et al., 2018; Soderstrom et al., 2015; Maxwell & Huff, 2022). Like JOLs elicited for cue-target word pairs, global JOLs may similarly encourage processing of pre-existing relations between items, benefiting later recall. Taken together, findings from Experiments 1A/1B suggest that both the framing of the JOL task (e.g., item-level vs. global) and the method of testing are both important determinants of JOL reactivity.

While the previous set of experiments suggest that item-level JOLs encourage item-specific encoding, Experiments 1A/1B tested this by comparing reactivity for different JOL types. However, a more complete test would be to compare the effects of item-level and global JOLs using stimuli that are differentially affected by item-specific and relational encoding tasks. As such, Experiment 2 replaced the categorized and uncategorized lists used previously with DRM lists. Like categorized lists, items in DRM lists are associatively related. However, in addition to containing strong list-wise relations, all items are also strongly related to a non-presented critical lure (i.e., the DRM illusion; see Deese, 1959; Roediger & McDermott, 1995). The DRM illusion is highly robust, with previous research showing that false recognition of critical lures can approximate correct recognition rates (see Gallo, 2006). Importantly, previous research has demonstrated that the DRM illusion is differentially affected by item-specific and relational encoding tasks, with item-specific but not relational tasks generally leading to a strong reduction in the DRM illusion when recognition testing is used (McCabe et al., 2004; Huff & Bodner, 2013; see Huff & Bodner, 2019). Thus, Experiment 2 tested whether reactivity patterns observed with recognition testing Experiment 1B would replicate when participants studied DRM lists while testing the effects of both item-level and global JOLs on the DRM illusion.

**Experiment 2: DRM Lists and Recognition Testing**

Experiment 2 provided an additional test of the item-order account by investigating whether item-level and global JOLs would be differentially reactive on correct and false recognition in the DRM paradigm. Based on findings from Experiment 1B, item-level JOLs were expected to improve correct recognition of studied items. Global JOLs, however, were not expected to be reactive on recognition. Regarding the DRM illusion, the item-order account predicts that item-level JOLs would reduce false recognition of critical lures, as per this account, the requirement to provide JOLs individually for each item should inhibit the processing of list-wise relations. This prediction is consistent with previous research showing that tasks which encourage item-specific encoding are generally effective at reducing the DRM illusion (e.g., Burns, Jenkins, & Deans, 2007; Huff & Bodner, 2013; McCabe et al., 2004). Separately, because relational encoding tasks emphasize shared connections between list items, these tasks may additionally increase false recognition of critical lures via spreading activation (e.g., Roediger, Balota, & Watson, 2001; Huff & Hutchison, 2011). Thus, the DRM illusion was expected to be greater for participants making global JOLs versus both the item-level JOL and no-JOL control groups.

**Method**

**Participants**

One hundred and twelve undergraduate students were recruited from Midwestern State University and completed Experiment 2 online for partial course credit. This sample size was based on the previous set of experiments, and participants were again randomly assigned to either item-level, global, or no-JOL encoding groups. Data screening followed the same criteria as previously reported, and data from four participants were excluded. As such, the final dataset contained responses from 108 participants (item-level JOLs, *n* = 36; global JOLs, *n* = 37; no-JOL control group, *n* = 35), and a sensitivity analysis conducted using *G\*Power* 3.1 suggested that the final sample was sufficient to detect a medium effect between encoding groups (*α* = .05, 1 – *β* = .80, *d* = 0.60). Like the previous experiments, all participants were native English speakers.

**Materials and Procedure**

Twenty DRM lists were selected from Roediger, Watson, McDermott, and Gallo (2001) and served as stimuli. These lists contained 12 associates that were additionally related to a non-presented critical lure. Within each list, words were arranged in descending order of backward associative strength (BAS) based on the University of South Florida Free Association Norms (D. L. Nelson et al., 2004). Lists were then divided into two equal sets to serve as counterbalances, which were matched on BAS (see Table A6 for stimuli properties). Next, an 80-item old/new recognition test was generated based on these lists and consisted of 30 previously studied items (taken from positions 2, 8, and 10 from each studied list), 10 critical lures corresponding to each studied list, 30 non-presented items taken from non-studied, counterbalanced lists, and 10 critical lure control items which corresponded to each of the non-studied lists. The general procedure was identical to Experiment 1B such that participants were randomly assigned to one of the two counterbalances and studied all 10 lists back-to-back, with the order of list presentation randomized for each participant. All JOL instructions were identical to the previous experiments, and depending on their encoding group, participants either provided item-level JOLs concurrently with study, made global JOLs following the completion of each list, or read each word silently. Like the previous set of experiments, encoding was self-paced for all groups. Experiment 2 took approximately 30 minutes to complete.

**Results**

**Analyses of Correct and False Recognition**

Figure 2 shows mean hit rates for studied items and false recognition of critical lures for each encoding group (see Appendix Table A7 for all comparisons including false alarms for control items). First, to test for potential JOL reactivity on correct recognition of list items, hits were analyzed via a one-way between-subjects ANOVA. This analysis yielded a significant difference between encoding groups, *F*(2, 105) = 12.44, *MSE* = .01, *p* < .001, *ηp*2 = .19, as hits were highest for participants in the item-level JOL group (.85), followed by the global JOL (.76) and no-JOL control groups (.72) Post-hoc *t*-tests confirmed that all comparisons differed significantly (*t*s ≥ 3.33, *p*s ≤ .001, *d*s ≥ 0.75), except for the comparison between the global JOL and no-JOL groups, which was non-significant, *t*(70) = 1.56, *SEM* = .03, *p* = .12, *d* = 0.32, *p*BIC = .74.

Turning to false recognition of critical lures, a separate one-way ANOVA revealed a significant difference between encoding groups, *F*(2, 105) = 3.89, *MSE* = .06, *p* = .02, *ηp*2 = .07, such that the DRM illusion was highest for participants making global JOLs (.76), followed by the no-JOL control group (.63), and the item-level JOL group (.62). All comparisons differed significantly (*t*s ≥ 2.39, *p*s ≤ .02, *d*s ≥ 0.55), except for the comparison between the item-level JOL and no-JOL control groups, *t*(69) < 1, *SEM* = .06, *p* = .88, *d* = 0.04, *p*BIC = .89.

**Analyses of *d′* and *c***

Following the design of Experiment 1B, signal detection analyses were used to test for changes in *d′* and *c* between encoding groups for both correct and false recognition. Starting with correct recognition, discriminability was highest for item-level JOLs (2.05), followed by global JOLs (1.71), and the no-JOL group (1.65). Although this pattern failed to reach conventional significance, *F*(2, 105) = 2.74, *MSE* = 0.61, *p* = .07, *ηp*2 = .05, *p*BIC = .40, a set of post-hoc analyses were conducted. These analyses revealed a significant difference in discriminability between the item-level and no-JOL groups, *t*(69) = 2.08, *SEM* = 0.20, *p* = .04, *d* = 0.49. The comparisons between the global and no-JOL groups and item-level and global JOL groups, however, did not reach significance (*t*s ≤ 1.99, *p*s *≥* .05,  *p*BICs ≥ .59, *d*s ≤ 0.47). Separately, response criterion differed between encoding groups, *F*(2, 105) = 6.59, *MSE* = .15, *p* < .01, *ηp*2 = .11. Mean *c* was greatest for participants in the no-JOL group (.23), followed by global JOLs (.11) and item-level JOLs (-.10). All comparisons differed significantly (*t*s ≥ 2.15, *p*s ≤ .03, *d*s ≥ 0.51), except for the comparison between the global JOL and no-JOL groups, *t*(70) = 1.42, *SEM* = .09, *p* = .16, *d* = 0.32, *p*BIC = .75.

Regarding false recognition, *d′* was numerically highest for the global JOL group (1.35), followed by the item-level (1.04) and no-JOL groups (1.04). Differences between encoding groups, however, were non-significant, *F*(2, 105) = 1.92, *MSE* = 0.62, *p* = .15, *ηp*2 = 0.03, *p*BIC = .94. Similarly, no significant differences in *c* emerged between groups (item-level JOL = .18, global JOL = -.04, no-JOL = .16; *F*(2, 105) = 2.24, *MSE* = .26, *p* = .11, *ηp*2 = 0.04, *p*BIC = .92.).

**Discussion**

Findings in Experiment 2 are clear. Consistent with both Experiment 1B and an item-order account of reactivity, making item-level JOLs at study improved correct recognition compared to participants in the no-JOL control group. Furthermore, like Experiment 1B, global JOLs were non-reactive on correct recognition. Signal detection analyses similarly revealed that discriminability was greater for participants in the item-JOL group compared to the no-JOL control group. Regarding the DRM illusion, only global JOLs increased false recognition of critical lures. For item JOLs, no changes in the false recognition were detected relative to the no-JOL control. Because the DRM illusion is often inflated by relational encoding tasks, the finding that global JOLs increased false recognition of critical lures further suggests that global JOLs likely encourage relational encoding to a greater extent relative to silent reading and item-level JOLs. Separately, although tasks which emphasize item-specific encoding generally reduce the DRM illusion, the finding that item-level JOLs did not decrease this effect suggest that this JOL type may still encouraged some degree of relational encoding, albeit to a lesser extent compared to global JOLs. This may be particularly likely, given that pre-existing item relations are a highly salient cue which participants use to inform the magnitude of their JOLs (see Koriat, 1997). However, when considered alongside findings from the previous experiments, there is growing evidence that item-level JOL reactivity primarily reflects increased item-specific encoding.

**General Discussion**

The present study provided an additional test of the item-order account of JOL reactivity, which proposes that eliciting JOLs for individual words promotes item-specific encoding while impairing relational processing. In doing so, memory was compared between participants making item-level JOLs and a non-JOL control task for free-recall (Experiment 1A) and recognition (Experiments 1B and 2) testing contexts. Additionally, Experiment 2 investigated whether positive reactivity on recognition of categorized/uncategorized word lists would extend to DRM lists. Each experiment also included a separate group of participants who made global JOLs following each list. Thus, all experiments provided a comparison between JOL tasks which differentially emphasized item-specific and relational encoding. Finally, because participants in Experiment 2 studied DRM lists, this experiment also allowed for an assessment of JOL reactivity effects on the DRM illusion. Critically, because the DRM illusion on recognition memory has been shown to be differentially affected by item-specific and relational processes (e.g., Huff & Bodner, 2013), Experiment 2 directly tested the item-order account’s central claim that item-level JOL reactivity on word lists is primarily driven by item-specific rather than relational processes.

Consistent with an item-order account, Experiment 1A found that item-level JOLs were non-reactive on free-recall, while global JOLs produced positive reactivity on free-recall of categorized but not uncategorized lists. Importantly, this null reactivity pattern is consistent with patterns reported by Zhao et al. (2023) who similarly showed that item-level JOLs did not improve free-recall of study lists. Experiment 1B then found that when recognition testing was used, the inverse pattern was observed on categorized lists. As such, item-level JOLs improved recognition while global JOLs were non-reactive on this test type. Finally, item-level JOLs also improved recognition of uncategorized lists. Taken together, findings from Experiments 1A and 1B suggest a dissociation between JOL type and test format.

Next, Experiment 2 replicated reactivity patterns observed in Experiment 1B as again item-level but not global JOLs were reactive on correct recognition. Because participants in Experiment 2 studied DRM lists, this experiment also tested whether each JOL task differentially affected false recognition of critical lures. Importantly, because the DRM illusion on recognition memory is differentially affected by item-specific and relational encoding tasks (Huff & Bodner, 2013), the use of DRM lists in Experiment 2 provided an additional test of whether item-level and global JOLs encouraged item-specific and relational encoding. Consistent with this account, only the requirement to provide global JOLs increased the DRM illusion, suggesting that the enhanced relational nature of the global JOL task increased activation of the critical lure (see Roediger et al., 2001). As such, findings in Experiment 2 suggest that compared to item-level JOLs, global JOLs placed a greater emphasis on list-wise relations.

Overall, the finding that JOL reactivity patterns are moderated by test format is consistent with previous research. For example, studies investigating JOL reactivity with cue-target word pairs have commonly demonstrated that item-level reactivity is moderated by test type, with these JOLs being reactive when memory is assessed via cued-recall and recognition but not free-recall testing (Myers et al., 2020; see also Soderstrom et al., 2015). Based on a cue-strengthening account of reactivity, this dissociation likely arises due to recall and recognition tests emphasizing different cues at retrieval. Consistent with this account, Zheng et al. (2024) recently proposed that item-level JOLs enhance both familiarity and recollection processes which underly recognition memory (see Yonelinas, 2002). Because recognition tests require participants to correctly discriminate between old and new items, this test format likely places greater emphasis on familiarity relative to free-recall testing. Findings from the present study align with this account, as item-level JOLs benefited recognition regardless of the context in which items were studied (e.g., categorized vs. uncategorized lists in Experiment 1B, DRM lists in Experiment 2) but were non-reactive when free-recall testing was used in Experiment 1A.

Separately, because global JOLs are more likely to emphasize list-wise relations relative to item-level JOLs, the global JOL task may be particularly beneficial whenever these relations are highly salient (e.g., categorized lists). However, because of procedural differences, caution is needed when making direct comparisons between reactivity patterns in Experiments 1A/1B. Specifically, while participants in Experiment 1A were tested following each list, participants in Experiment 1B encoded all items prior to completing a single 96 item old/new recognition test. While this change in Experiment 1B occurred to account for potential ceiling effects (i.e., participants otherwise would have completed four short, 24 item recognition tests), global JOLs may have been reactive in Experiment 1A but not 1B due to the category exemplar being more easily retrieved within Experiment 1A’s testing context. However, given that findings for item-level JOLs in Experiment 1A replicated Zhao et al. (2023), it is unlikely that procedural differences alone can fully explain the diverging reactivity patterns reported in Experiments 1A and 1B.

While the present study assessed item-level JOL reactivity for different test formats and list compositions, future research may wish to provide additional tests of the item-order account using serial position analysis or by directly assessing participants’ strategy use at encoding (e.g., Rivers et al., 2023). For example, if item level-JOLs indeed impair relational encoding of word lists, item-level JOLs would be also expected to impair serial position accuracy in free-recall. Separately, participants providing item-level JOLs would be expected to report engaging in fewer relational strategies at encoding. However, note that although item-level JOLs are likely to bias participants towards item-specific encoding, this JOL type likely still encourages some degree of relational encoding, albeit to a lesser extent versus global JOLs. For example, studies utilizing cue-target word pairs have repeatedly demonstrated that reactivity reflects a relational process when recall testing is used (e.g., Halamish & Undorf, 2023; Maxwell & Huff, 2024; Rivers et al., 2023) and, additionally, this pattern approximates memory benefits observed when participants encode pairs via other tasks which similarly encourage relational encoding (e.g., frequency of co-occurrence judgments, judgments of associative memory; Maxwell & Huff, 2022). Thus, whether JOLs are more likely to encourage item-specific or relational encoding appears to be strongly dependent on both the type of stimuli that participants study and the specific framing of the JOL task.

This notion may partially explain why item-level JOLs did not reduce the DRM illusion in Experiment 2 as initially anticipated. Because JOLs are particularly sensitive to pre-existing item relations (e.g. Halamish & Undorf, 2023; Maxwell & Huff, 2024, Soderstrom et al., 2015; Rivers et al., 2023), these judgments may have still encouraged participants to process list-wise relations to some degree, even when they were elicited for individual items (e.g., item-level JOLs). This additional relational processing likely mitigates any potential benefits of item-specific encoding that would otherwise reduce the DRM illusion, and which are typically with observed using other item-specific encoding tasks (e.g., McCabe et al., 2004; Huff & Bodner, 2013). Moreover, when participants are instructed to apply item-specific tasks while studying related word lists (e.g., categorized lists, DRM lists), they likely also engage in some degree of relational encoding due to the inherent relational properties of the study lists (Hunt & Seta, 1984). Thus, although both JOL tasks were designed to encourage one specific type of processing over the other, it is unlikely that either JOL task was truly process-pure (see Huff & Bodner, 2019). However, given that the present study is the first to explore JOL reactivity within the context of the DRM lists, more research will be needed to fully explore the differential effects of both JOL types on correct and false recognition.

Finally, given the benefits of item-level JOLs on recognition testing that were observed in Experiments 1B and 2 as well as similar benefits reported on recognition of cue-target word pairs (e.g. Maxwell & Huff, 2024; Myers et al., 2020), future research may wish to explore whether this effect extends to educationally relevant stimuli. Although previous research has found that JOLs are non-reactive when participants’ study general knowledge facts (Schäfer & Undorf, 2024) and text passages (Ariel, Karpicke, Witherby, & Tauber, 2021), these studies have primarily assessed memory using recall testing, rather than recognition (though see Schäfer & Undorf, who found no positive reactivity on general knowledge when using a multiple-choice paradigm). However, as previously noted, recognition testing is generally more sensitive to familiarity cues relative to free-recall testing. As such, if item-level JOLs consistently enhance both familiarity and recollection as proposed by Zheng et al. (2024), these judgments would be expected to improve memory for educationally relevant study materials, so long as learning is assessed using recognition-based approaches. Ultimately, however, more research is needed to fully explore the extent to which JOLs can improve memory within applied contexts.

**Conclusion**

In summary, the present study evaluated JOL reactivity effects on word list learning. By comparing between item-level and global JOLs, the present study provided an additional test of the item-order account of JOL reactivity, which posits that item-level JOL reactivity observed with word lists primarily reflects enhanced item-specific encoding. Overall, the present study provides further support for an item-order account of JOL reactivity, as item-level JOLs were reactive on recognition but not free-recall testing. Importantly, when the JOL tasks was reframed to encourage relational encoding, JOLs did not improve recognition and free-recall was only improved when list items contained pre-existing relations (e.g., categorized lists). Additionally, global JOLs also increased the DRM illusion in Experiment 2. Taken together, findings from the present study suggest that while item-level JOLs are likely to encourage item-specific encoding, the potential memorial benefits of JOLs and the specific processes underlying reactivity are influenced by the type of stimuli participants study, the method of testing, and the manner in which JOLs are framed.

**Declarations**

**Data Availability Statement**

Study materials, data files, and *R* code used for analyses have been made available via OSF (https://osf.io/t453a/).

**Compliance with Ethical Practices**

The reported studies were approved by the Institutional Review Board at Midwestern State University (Protocol #24022005). All participants provided informed consent prior to participating in the experiments. The author reports no conflicts of interest.

**Funding Information**

Data collection was partially supported by an intramural research grant awarded to the author by Midwestern State University.

**References**

Ariel, R., Karpicke, J. D., Witherby, A. E., & Tauber, S. K. (2021). Do judgments of learning directly enhance learning of educational materials? *Educational Psychology Review, 33*, 693-712.

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.

Burns, D. J., Jenkins, C. L., & Dean, E. E. (2007). Falsely recalled items are rich in item-specific information. *Memory & Cognition, 35*, 1630–1640.

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

Chang, M. & Brainerd, C. J. (2024). Judgments of learning reactivity on item-specific and relational processing. *Journal of Intelligence, 12*(1), 1–14.

Chang, M. & Brainerd, C. J. (2022). Association and dissociation between judgments of learning and memory: A meta-analysis of the font-size effect. *Metacognition and Learning, 17*, 443–476.

Deese, J. (1959). On the prediction of occurrence of particular verbal intrusions in immediate recall. *Journal of Experimental Psychology, 58*(1), 17–22.

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

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

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., Buchner, A, & Lang, A. G. (2009). Statistical power analyses using G\*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*, 1149–1160.

Gallo, D. A. (2006). *Associative illusions of memory: False memory research in DRM and related tasks.* Psychology Press.

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

Halamish, V. & Undorf, M. (2023). Why do judgments of learning modify memory? Evidence from identical pairs and relatedness judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 49*(4), 547–556.

Hautus, M. J. (1995). Corrections for extreme proportions and their biasing effects on estimated values of *d′*. *Behavior Research Methods, Instruments, & Computer, 27*, 46–51.

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

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

Huff, M. J. & Hutchison, K. A. (2011). The effects of mediated word lists on false recall and recognition. *Memory & Cognition, 39*, 941-953.

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

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

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

Kaya, S. & Mulligan, N. W. (in press). Metamemory judgments and design effects: Judgment of learning (JOL) reactivity in free recall is affected by study list structure. *Memory & Cognition*, 1–15.

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. & Goldsmith, M. (1996). Monitoring and control processes in the strategic regulation of memory accuracy. *Psychological Review, 103*(3), 490–517.

Makowski, D. (2018). The *psycho* package: An efficient and publishing-oriented workflow for psychological science. *Journal of Open Source Software*, *3*(22), 470.

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. (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.

Maxwell, N. P. & Huff, M. J. (2023). Is discriminability a requirement for reactivity? Comparing the effects of mixed vs. pure list presentations on judgment of learning reactivity*. Memory & Cognition, 51*(5), 1198–1213.

Maxwell, N. P. & Huff, M. J. (2024). Judgment of learning reactivity reflects enhanced relational encoding on cued-recall but not recognition tests. *Metacognition and Learning, 19*, 189-213.

Maxwell, N. P., Huff, M. J., & Buchanan, E. M. (2022). The *lrd* package: An *R* package and Shiny application for processing lexical data. *Behavior Research Methods, 54*, 2001-2024.

McCabe, D. P., Presmanes, A. G., Robertson, C. L., & Smith, A. D. (2004). Item-specific processing reduces false memories. *Psychonomic Bulletin & Review, 11*, 1074–1079.

McDaniel, M. A. & Bugg, J. M. (2008). Instability in memory phenomena: A common puzzle and a unifying explanation. *Psychonomic Bulletin & Review, 15,* 237-255.

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.

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.

Rawson, K. A. & Zamary, A. (2019). Why is free recall practice more effective than recognition practice for enhancing memory? Evaluating the relational processing hypothesis. *Journal of Memory and Language, 105*, 141–152.

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. & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: Evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*(4), 615–625.

Rivers, M. L., Dunlosky, J., Janes, J. L., Witherby, A. E., & Tauber, S. K. (2023). Judgments of learning enhance recall for category-cued but not letter-cued items. *Memory & Cognition, 51*, 1547–1561.

Roediger, H. L. III, Balota, D. A., & Watson, J. M. (2001). Spreading activation and arousal of false memories. In H. L. Roediger III, J. S. Nairne, I. Neath, & A. M. Surprenant (Eds.), *The nature of remembering: Essays in honor of Robert G. Crowder* (pp. 95–115). American Psychological Association.

Roediger, H. L., & McDermott, K. B. (1995). Creating false memories: Remembering words not presented lists. J*ournal of Experimental Psychology: Learning, Memory, and Cognition, 21*(4), 803–814.

Roediger, H. L., Watson, J. M., McDermott, K. B., & Gallo, D. A. (2001). Factors that determine false recall: A multiple regression analysis. *Psychonomic Bulletin & Review, 8*, 385–407.

Schäfer, F. & Undorf, M. (2024). On the educational relevance of immediate judgment of learning reactivity: No effects of predicting one’s memory for general knowledge facts. Journal of Applied Research in Memory and Cognition, 13(1), 113-123..

Schwartz, B. L. & Metcalfe, J. (2017). Metamemory: An update of critical findings. In J. H. Bryne (Ed.), *Learning and memory: A comprehensive reference* (2nd ed., pp. 423–432). Academic Press.

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

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.

Undorf, M., Schäfer, F., & Halamish, V. (2024). Making judgments of learning either enhances or impairs memory: Evidence from 17 experiments with related and unrelated word pairs. *Collabra:* Psychology, 10(1): 117108

Van Overschelde, J. P., Rawson, K. A., & Dunloskey, J. Category Norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language, 50*(3), 289-335.

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

Yonelinas, A. P. (2002). The nature of recollection and familiarity: A review of 30 years of research. *Journal of Memory and Language, 46*(3), 441–517.

Zhao, W., Li, J., Shanks, D. R., Li, B., Hu, X., Yang, C., & Luo, L. (2023). Metamemory judgments have dissociable reactivity effects on item and interitem relational memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 49*(4), 557-574.

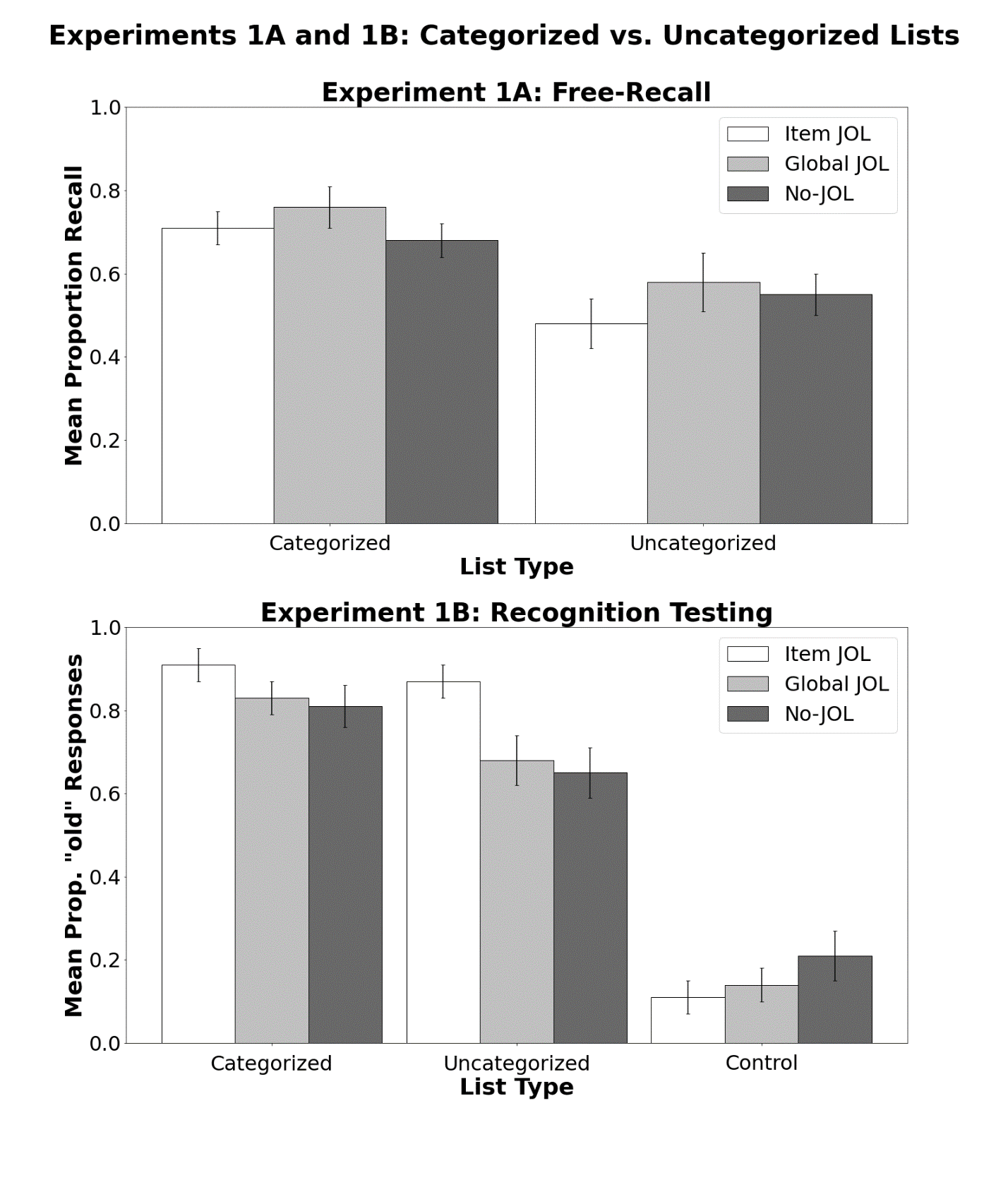
Zheng, J., Baike, L., Zhao, W., Su, N., Fan, T., Yin, Y., Hu, Y., Hu, X., Yang, C., & Luo, L. (2024). Soliciting judgments of learning reactivity facilitates both recollection- and familiarity- based recognition memory. *Metacognition and Learning, 19,* 609-633.

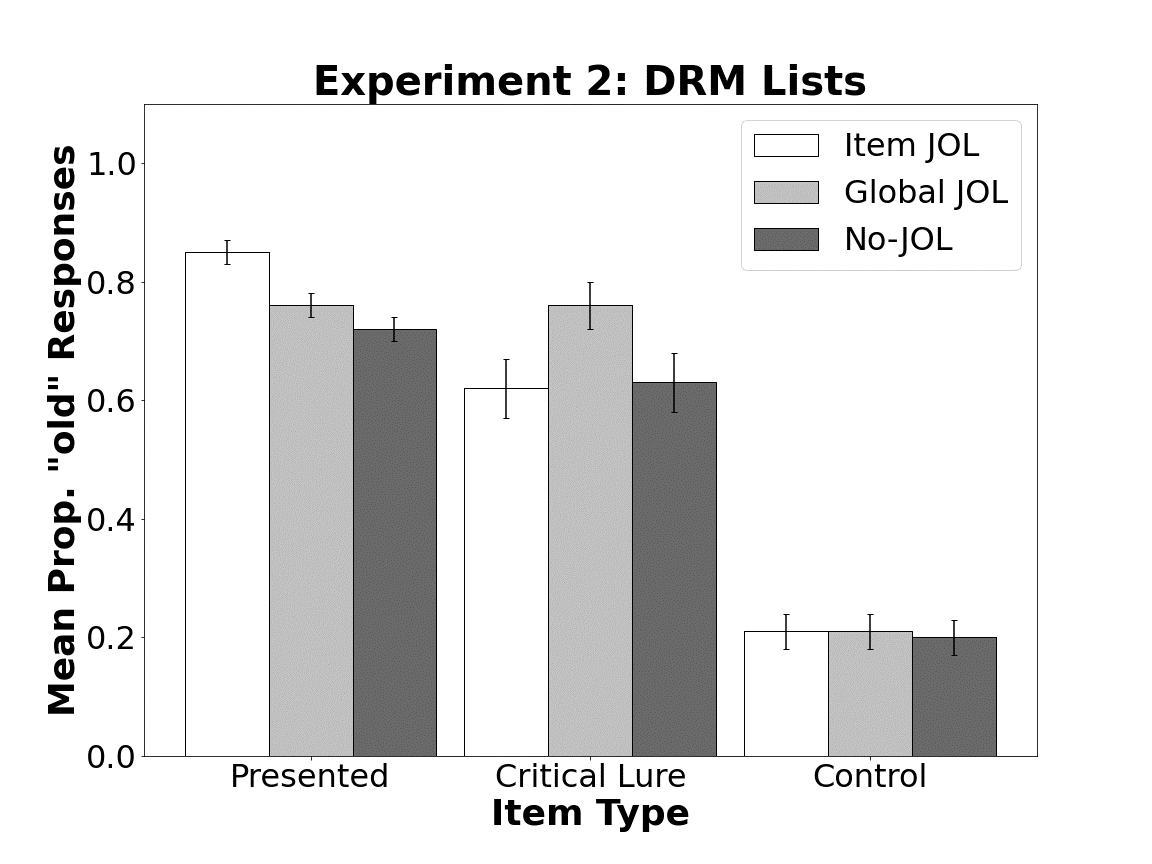
Table 1.

*Anticipated Reactivity Patterns Based on the Item-Order Account and Observed Reactivity Patterns for each List and JOL Type in all Experiments.*

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| Experiment | Test Type | Item/List Type | JOL Type | Predicted Reactivity Pattern | Observed Reactivity Pattern |
| Ex. 1A | Free-Recall | Categorized | Item-level | Null | Null |
|  |  |  | Global | -- | Positive |
|  |  | Uncategorized | Item-level | Null | Null |
|  |  |  | Global | -- | Null |
| Ex. 1B | Recognition | Categorized | Item-level | Positive | Positive |
|  |  |  | Global | -- | Null |
|  |  | Uncategorized | Item-level | Positive | Positive |
|  |  |  | Global | -- | Null |
| Ex. 2 | Recognition | DRM Studied Items | Item-level | Positive | Positive |
|  |  |  | Global | -- | Null |
|  |  | DRM Critical Lures | Item-level | Negative | Null |
|  |  |  | Global | -- | Positive |

*Note*: The item-order account makes no specific predictions regarding Global JOLs. As such, predicted reactivity patterns are only provided for Item-level JOLs. For completeness, observed Global JOL patterns are included in the right-most column.

*Figure 1*. Mean proportion of correct free-recall in Experiment 1A (top panel) and mean proportion of “old” responses in Experiment 1B (bottom panel). “Control” columns indicate “old” responses to non-studied control items, which were taken from the non-studied counterbalance. Bars indicate 95% *CI*s.



*Figure 2.* Mean proportion of “old” responses to presented items and critical lures in Experiment 2. Bars indicate 95% *CI*s. “Control” columns indicate “old” responses to non-studied control items, which were taken from the non-presented counterbalance. Control items are collapsed across list controls and critical lure controls.

**Appendix**

Table A1

*Lexical Properties of Categorized and Uncategorized Lists in Experiments 1A and 1B.*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| List Type | Counterbalance | Frequency | Length | Concreteness |
| Categorized | A | 2.91 (0.53) | 5.29 (1.57) | 6.24 (0.41) |
|  | B | 2.85 (0.62) | 5.71 (1.60) | 6.09 (0.44) |
| Uncategorized | A | 2.94 (0.68) | 5.25 (1.39) | 5.75 (0.65) |
|  | B | 3.01 (0.73) | 4.71 (1.08) | 5.53 (0.94) |

*Notes*: Cells denote mean values; parenthesis indicate *SD*s. Frequency ratings were derived from SUBLTEX (Brysbaert & New, 2009). Concreteness ratings were taken from D. L. Nelson et al. (2004). Values are collapsed across individual study lists for each list type/counterbalance.

Table A2

*Comparison of Mean Recall Proportions in Experiment 1A as functions of Encoding Group and List Type*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Encoding Group | List Type | *M* | *± 95% CI* | U |
| Item JOL | Categorized | .71 | .04 |  |
|  | Uncategorized | .48 | .06 | 1.46\* |
| Global JOL | Categorized | .76 | .05 |  |
|  | Uncategorized | .58 | .07 | 0.99\* |
| No-JOL | Categorized | .68 | .04 |  |
|  | Uncategorized | .55 | .05 | 0.86\* |

*Notes*: Right-most column denotes Cohen’s *d* effect sizes for post-hoc comparisons between list types within encoding groups. \* = *p* < .05. U = Uncategorized lists.

Table A3

*Comparison of Mean Hit Rates, False Alarms, d′*, *and c in Experiment 1B as functions of Encoding Group and List Type.*

|  |  |  |  |
| --- | --- | --- | --- |
| Item Type/Measure | Item JOL | Global JOL | No JOL |
| Categorized Items | .91 (.04) | .83 (.04) | .81 (.05) |
| Uncategorized Items | .87 (.04) | .67 (.06) | .65 (.06) |
| Controls | .11 (.04) | .14 (.04) | .21 (.06) |
| *d′* | 2.76 (0.31) | 1.97 (0.27) | 1.64 (0.25) |
| *c* | 0.03 (0.12) | 0.22 (0.10) | 0.16 (0.14) |

*Notes*: Categorized and Uncategorized item rows denote proportion of hits. Control row denotes proportion of false alarms to non-studied items. *d′* and *c* values were computed using hits that were collapsed across list types. Parentheses indicate ± 95% *CI*s.

Table A4

*Comparison of Mean JOLs in Experiments 1A, 1B, and 2 as Functions of Encoding Group and List Type*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Experiment | Encoding Group | List Type | *M* | *± 95% CI* |
| Ex. 1A | Item JOL | Categorized | 62.01 | 4.96 |
|  |  | Uncategorized | 54.89 | 5.11 |
|  | Global JOL | Categorized | 55.36 | 8.40 |
|  |  | Uncategorized | 43.95 | 7.72 |
| Ex. 1B | Item JOL | Categorized | 67.99 | 5.01 |
|  |  | Uncategorized | 61.66 | 5.50 |
|  | Global JOL | Categorized | 56.14 | 7.49 |
|  |  | Uncategorized | 54.11 | 8.13 |
| Ex. 2 | Item JOL | DRM | 62.48 | 6.71 |
|  | Global JOL | DRM | 62.06 | 7.79 |

Table A5

*Comparison of Mean Response Latencies in Experiments 1A, 1B, and 2 as Functions of Encoding Group and List Type*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Experiment | Encoding Group | List Type | *M* | *± 95% CI* |
| Ex. 1A | Item JOL | Categorized | 4321.59 | 438.17 |
|  |  | Uncategorized | 3769.80 | 410.82 |
|  | Global JOL | Categorized | 2492.04 | 279.88 |
|  |  | Uncategorized | 2629.52 | 301.66 |
|  | No-JOL | Categorized | 2923.10 | 528.57 |
|  |  | Uncategorized | 2835.28 | 466.23 |
| Ex. 1B | Item JOL | Categorized | 4021.13 | 423.43 |
|  |  | Uncategorized | 3784.67 | 423.26 |
|  | Global JOL | Categorized | 2568.60 | 415.10 |
|  |  | Uncategorized | 2322.85 | 478.28 |
|  | No-JOL | Categorized | 2299.74 | 368.69 |
|  |  | Uncategorized | 2213.85 | 321.32 |
| Ex. 2 | Item JOL | DRM | 3323.41 | 499.86 |
|  | Global JOL | DRM | 2103.41 | 212.83 |
|  | No-JOL | DRM | 2198.22 | 366.64 |

*Note*: RTs are displayed in ms.

Table A6

*Lexical Properties for DRM lists in Experiment 2*

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Counterbalance | Frequency | Length | Concreteness | BAS |
| A | 2.76 (0.74) | 5.63 (1.61) | 5.02 (1.22) | .207 (.20) |
| B | 2.89 (0.84) | 5.23 (1.50) | 5.31 (1.08) | .218 (.20) |

*Notes*: Cells denote mean values; parenthesis indicate *SD*s. Frequency ratings were derived from SUBLTEX (Brysbaert & New, 2009). Concreteness ratings were taken from D. L. Nelson et al. (2004). BAS = Backward Associative Strength between item and critical lure. Values are collapsed across individual study lists for each list type/counterbalance.

Table A7

*Comparison of Mean Hit Rates, False Alarms, d′*, *and c in Experiment 2 as functions of Encoding Group and Item Type.*

|  |  |  |  |
| --- | --- | --- | --- |
| Item Type/Measure | Item JOL | Global JOL | No JOL |
| List Items | .85 (.02) | .76 (.02) | .72 (.02) |
| List Item Controls | .20 (.03) | .20 (.03) | .18 (.04) |
| Critical Items | .62 (.04) | .76 (.04) | .63 (.05) |
| Critical Item Controls | .25 (.04) | .27 (.03) | .25 (.03) |
| *d′* | 2.05 (0.24) | 1.71 (0.23) | 1.65 (0.29) |
| *c* | -0.10 (0.13) | 0.11 (0.13) | 0.23 (0.11) |

*Notes*: List Items row denotes proportion of hits on studied items. Control rows indicate false alarms on non-studied items taken from control lists and critical lures corresponding to non-studied lists. False alarms were collapsed across both control item types when computing *d′* and *c.*

1. False alarms were collapsed across categorized and uncategorized list types, given that New items were not associated with a specific list type. As such, *d′* and *c* were computed for all list items rather than separately for each list type. [↑](#footnote-ref-1)