Memory & Cognition
https://doi.org/10.3758/s13421-020-01025-5

Judgments of learning (JOLs) selectively improve memory depending on the type of test

Sarah J. Myers[1] · Matthew G. Rhodes[1] · Hannah E. Hausman[1]

Abstract

JOL reactivity refers to the finding that making judgments of learning (JOLs) while studying material influences later memory for that material. Findings of JOL reactivity have been mixed, with some experiments reporting changes to memory when participants make JOLs and others finding no influence of JOLs. Soderstrom, Clark, Halamish, and Bjork (*Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(2), 553–558, 2015) proposed that JOL reactivity will only occur if the final test is sensitive to the same cues used to inform JOLs. The current study evaluated this account by manipulating the type of final test. In four experiments, participants studied mixed lists of related and unrelated word pairs and either made JOLs or did not make JOLs. Making JOLs generally enhanced memory for related word pairs when a cued-recall test was administered. However, during free recall, JOLs had no influence on memory for target information, likely because cue–target associations (which are used to inform JOLs) are less beneficial in the absence of cues. JOLs improved item recognition memory for words that were studied in related pairs, although the effect was small. Collectively, data from a meta-analysis of these experiments indicate that JOL reactivity depends on the type of final test, with reactivity most likely to occur when the final test is sensitive to the same cues used to inform JOLs. Future work should continue examining different tests and study materials in order to develop a comprehensive theory of JOL reactivity.

Keywords Metamemory · Judgments of learning (JOLs) · JOL reactivity

To fully understand memory, it is important to understand how people assess their memory. For example, participants typically choose to restudy material they believe has not yet been learned and to stop studying material they judge as learned (Dunlosky & Thiede, 1998). One common method of examining self-assessments of memory is to solicit judgments of learning (JOLs), whereby participants indicate the likelihood of remembering a studied item on a later test (for a review, see Rhodes, 2016).

JOLs are typically regarded as neutral measurements of memory monitoring (see T. O. Nelson, 1990), reflecting an individual's assessment of learning without affecting memory for the material being judged. However, some research indicates that the act of making JOLs may influence later memory (e.g., Arbuckle & Cuddy, 1969; King, Zechmeister, & Shaughnessy, 1980; Witherby & Tauber, 2017), a finding referred to as "JOL reactivity." The possibility of JOLs being a reactive measurement has substantial implications, potentially distorting conclusions about the role of memory monitoring in learning. Accordingly, this paper explores a theoretical account proposing that JOL reactivity reflects the combination of cues used to make JOLs and cues present on later tests.

JOL reactivity: Data and theory

Prior research has yielded mixed results regarding whether making JOLs for items during study (i.e., immediate JOLs) affects learning (for an examination of effects on learning when JOLs are made after a delay, see Rhodes & Tauber, 2011; Tauber, Dunlosky, & Rawson, 2015). Several studies have found that making immediate JOLs improves memory compared with not making JOLs during study (e.g., Arbuckle & Cuddy, 1969; Janes, Rivers, & Dunlosky, 2018; King et al., 1980; Soderstrom, Clark, Halamish, & Bjork, 2015; Witherby & Tauber, 2017). For example, Soderstrom et al. (2015) found that JOLs selectively improved memory of related, but not

✉ Sarah J. Myers
Sarah.Jean.Myers@colostate.edu

[1] Colorado State University, Behavioral Sciences Building, 410 W. Pitkin St., Fort Collins, CO 80523, USA

Published online: 02 March 2020

unrelated, word pairs. Specifically, participants in their experiments studied mixed lists of related (e.g., *Railroad–Train*) and unrelated (e.g., *Practice–Tree*) cue–target word pairs; some participants made JOLs while studying, and others did not make JOLs. All participants then received a cued-recall test whereby they were given the cue and supplied the target (e.g., *Railroad–?*). On this test, JOLs improved cued recall of related word pairs but did not influence recall of unrelated pairs. Mitchum, Kelley, and Fox (2016) used a similar procedure and also observed that JOLs improved cued recall of related word pairs. However, in contrast to Soderstrom et al. (2015), they found that making JOLs resulted in poorer recall of unrelated word pairs.

Still other studies have detected no differences in later memory between participants who made JOLs and those who did not (e.g., Begg, Martin, & Needham, 1992; Keleman & Weaver, 1997; Tauber & Rhodes, 2012). For example, Tauber and Rhodes (2012) found that making immediate JOLs for single-item word lists had no effect on free recall performance. Thus, prior work provides inconsistent evidence of whether and how immediate JOLs influence memory performance. Indeed, in a meta-analysis of 19 experiments from eight independent studies, Double, Birney, and Walker (2018) found no overall effect of immediate JOLs on memory. However, results were moderated by the type of material such that JOL reactivity was evident for related word pairs and single-item word lists but absent for unrelated pairs. Therefore, JOL reactivity may only occur for specific types of materials.

Soderstrom et al. (2015) accounted for this material-specific reactivity by suggesting that making JOLs strengthens memory for the cues that inform JOLs. If memory on a later criterion test depends on the same cues that are used to inform JOLs, then making JOLs should improve performance on that test. For example, learners attend to relatedness when making JOLs, giving related word pairs higher JOLs than unrelated pairs (Arbuckle & Cuddy, 1969; see Mueller, Tauber, & Dunlosky, 2013, for a review). Soderstrom et al. (2015) proposed that when participants attend to relatedness to inform their JOLs, they strengthen encoding of the relationship between the items in related pairs. However, this relational processing does little to enhance encoding for unrelated pairs, which have no semantic relationship. Because a final cued-recall test requires participants to recall the target given the cue, operations that strengthen cue–target relationships (such as making JOLs) would enhance performance. Soderstrom et al.'s (2015) findings supported this hypothesis, whereby JOLs elevated cued recall of related pairs, but showed no influence on unrelated pairs. However, a key prediction of this theory remains to be tested. Namely, JOL reactivity should only occur if the criterion test relies on the same cues that inform JOLs (e.g., pair relatedness). We investigated this prediction by examining JOL reactivity across different types of criterion tests.

The current study

In the experiments reported, participants studied mixed lists of related and unrelated word pairs and either made JOLs or did not make JOLs (i.e., no-JOL condition). We then examined performance on tests that should be sensitive to cue–target relatedness (cued recall) and tests that should be less sensitive to cue–target relatedness (free recall and item recognition). Based on Soderstrom et al.'s (2015) account, JOL reactivity should be most potent when participants can use the relationship between words in related pairs to retrieve the target when given the cue. Therefore, we hypothesized that JOLs would improve memory for related pairs on a cued-recall test. On a free recall or item-recognition test, because participants are not provided with the cue, pair relatedness should be less useful when identifying target information. Thus, no differences in performance were expected between JOL and no-JOL conditions during free recall or item recognition.

Prior work provides tentative support for these hypotheses. For example, one experiment failed to find JOL reactivity on a free-recall test (Tauber & Rhodes, 2012; but see Begg, Duft, Lalonde, Melnick, & Sanvito, 1989). However, prior work did not systematically manipulate the type of item or test for participants who made or did not make JOLs. We did so in the current experiments and tested the hypothesis that JOL reactivity depends on the overlap between cues used to inform JOLs and cues used on a final test.

In the current study, we focus on the effect of JOLs on related word pairs because prior evidence indicates that the relationship between words serves as a dominant cue to inform JOLs (Mueller et al., 2013), and relatedness is known to influence cued recall of related pairs (Soderstrom et al., 2015). Although relatedness is a diagnostic cue between related and unrelated word pairs (i.e., participants give higher ratings to related pairs than to unrelated pairs), participants may also incorporate other cues into their judgments, particularly among the same type of pair (Undorf, Sollner, & Broder, 2018). Soderstrom et al. (2015) argued that making JOLs based on relatedness should have "little or no effect" (p. 554) on the memorability of unrelated word pairs because there is no relatedness between these pairs. However, more information is needed to predict how other cues may influence memorability of unrelated word pairs. Accordingly, whereas relatedness is a dominant cue when judging mixed lists of related and unrelated pairs, judgments are still most likely multifaceted and may rely upon other cues as well (e.g., word familiarity, imageability). Given the indeterminate nature of how these cues may influence memorability of unrelated word pairs across various criterion tests, we focus mostly on the effect of JOLs on later memorability of related pairs across multiple test types. Effects of JOLs on unrelated pairs are reported, but we remain agnostic with regard to the basis for those JOLs.

Experiment 1

In Experiment 1, participants studied two blocks of related and unrelated word pairs and either made JOLs during study or did not make JOLs. Participants were administered a cued recall or free-recall test after the first block and the other test type after the second block, ensuring that all participants took each test. We anticipated that participants who made JOLs would correctly recall more targets from related pairs than those who did not make JOLs on the cued-recall test. However, given the absence of cues, no differences between the two conditions were expected for free recall.

Method

Participants

A power analysis using G*Power (Faul, Erdfelder, Lang, & Buchner, 2007) indicated that a sample size of 34 participants per condition was required to detect an effect size of $d = 0.69$ (the effect size for the difference between JOL and no-JOL conditions for related items reported in Soderstrom et al., 2015, Experiment 1b), assuming α =.05, power of .80, and a two-tailed test. Sample size was increased to 40 participants per condition to ensure equal sample sizes across counterbalances.

Participants were 86 (46 JOL, 40 no JOL) students from Colorado State University who received course credit for participation. Four participants were removed from the JOL condition for not providing JOLs for at least 80% of the study trials in both lists, and two were removed from the JOL condition for technical malfunctions. Therefore, data from 40 participants in the JOL and 40 in the no-JOL condition were included in analyses. Participants (26 men, 54 women) were 17 to 27 (M = 18.99, SD = 1.78) years old.

Materials

Sixty related cue–target word pairs (forward strength 0.400–0.739, M = 0.537) selected from the University of South Florida Free Association Norms (USF-FAN; D. L. Nelson, McEvoy, & Schreiber, 1998) were used in Experiment 1, frequency: 8.013–12.253 (M = 10.025), concreteness: 259–637 (M = 529.7), target word length: 3–8 letters (M = 4.417). Pairs were divided into four lists of 15 pairs that were closely matched in average forward association, frequency, concreteness, and target length. An unrelated version of each of the four lists was created by randomly pairing the targets with different, unrelated cues. Four lists of 15 related and 15 unrelated pairs were then created and counterbalanced so that target words were paired equally often with a related or unrelated cue. For example, the target *Bee* was paired with a related cue (*Buzz*) for half the participants, and an unrelated cue (*Clever*) for the remaining participants. Twelve other related word pairs were used as buffers. Half the buffer pairs were randomly re-paired to make unrelated buffers. Data for buffers were not included in any analyses.

Design and procedure

A 2 (judgment: JOL, no JOL) × 2 (test type: cued recall, free recall) × 2 (pair type: related, unrelated) mixed-factor design was used. Judgment was manipulated between participants, whereas test type and pair type were manipulated within participants. The experiment was run in E-Prime Version 2.0 (Schneider, Eschman, & Zuccolotto, 2002).

After providing consent, participants were informed that they would study word pairs and be asked to remember the word on the right of each pair (i.e., the target) on a later test. Participants were not told what type of test they would receive prior to studying the pairs, although they were told they may or may not be given the word on the left (i.e., the cue). Participants then studied 30 pairs (15 related, 15 unrelated) at a 12-second rate, presented in a unique random order for each participant. In addition to the 30 pairs, three buffer pairs were included at the beginning and end of each study block to account for primacy and recency effects. Half the participants provided JOLs while studying both lists (JOL condition), and half did not provide JOLs (no-JOL condition). Both conditions were shown each pair for the entire 12 seconds. In the JOL condition, the JOL prompt appeared after 5 seconds and was displayed for 7 seconds with each pair, equating exposure time between conditions. For the JOL rating, participants indicated from 0% to 100% how likely it was that they would remember the target word on a later test.

Following 5 minutes of adding sums, participants then took a cued or free-recall test, with test order counterbalanced across participants.[1] For cued recall, participants were given each of the 30 cues and had 10 seconds to type each corresponding target word. For the free-recall test, participants had 3 minutes to type as many target words as they could. No feedback was provided for either test. After completing this first block, participants studied a second list of 30 word pairs, added sums for 5 minutes, and then took the other test type (cued or free recall).

Scoring and analysis

Minor spelling mistakes were marked as correct provided the response was not a different word (e.g., *For* instead of *Fog* would be marked as incorrect). Plurals of target words were also marked as correct. Data were analyzed using SPSS

[1] Test order did not interact with the judgment condition in any experiment (see results in the supplemental materials, available at osf.io/ew5z2).

Version 24 (IBM Corp., 2016) and R Version 3.4.2 (R Core Team, 2014).

We employed both frequentist and Bayesian methods of analysis. For the focal analyses, we report the p value, a standardized effect size measure (Cohen's d or η_p^2), and the Bayes factor (BF). Bayes factors quantify the strength of the evidence in favor of the alternative hypothesis (JOL reactivity) relative to the null hypothesis (no JOL reactivity; see Kruschke, 2013, for a discussion of Bayes factors).

The Bayes factor is a ratio of the likelihood of the data given the alternative hypothesis to the likelihood of the data given the null hypothesis (BF_{10}). A Bayes factor of 1 means that the data are equally likely under the alternative and null hypotheses. Unlike null hypothesis significance testing, Bayes factors can indicate that the null hypothesis is more probable than the alternative hypothesis (i.e., when $BF_{10} < 1$), and is often reported as the reciprocal BF_{01}. We interpret Bayes factors using recommendations from Wagenmakers (2007), whereby Bayes factors provide weak ($1 < BF \leq 3$), positive ($3 < BF \leq 20$), strong ($20 < BF \leq 150$), or very strong ($BF > 150$) evidence in favor of one hypothesis over the other. Following Rouder, Speckman, Sun, Morey, and Iverson (2009), we used the JZS prior because it requires the fewest prior assumptions about the range of the true effect size. All Bayes factors were calculated using the BayesFactor R Package (Morey & Rouder, 2018).

Results

Data for JOL magnitude and resolution are reported in the supplemental materials available on the Open Science Framework (osf.io/ew5z2). The key prediction for Experiment 1 was that providing JOLs, relative to not providing JOLs, would selectively benefit performance for related items on the cued-recall but not the free-recall test. Therefore, we implemented a set of planned analyses to compare performance for those who made JOLs and did not make JOLs for related items and unrelated items separately. We also report the 2 (pair type: related, unrelated) × 2 (judgment: JOL, no JOL) mixed-factor ANOVAs[2] for completeness. We expected an ordinal interaction, whereby JOLs influence performance for related pairs if the test is sensitive to cues of relatedness, but have little to no effect on unrelated pairs. Thus, we predicted that JOLs would consistently benefit cued recall of related pairs, but the effect of JOLs on unrelated pairs would vary. The alpha level for all analyses was set to 0.05.

Cued recall

Overall, cued recall (see Fig. 1a) was superior for related items ($M = 72.583$, $SE = 2.169$) compared with unrelated items ($M = 21.083$, $SE = 2.299$), $F(1, 78) = 486.162$, $p < .001$, $\eta_p^2 = .862$. Participants who made JOLs ($M = 50.083$, $SE = 2.695$) also exhibited numerically, but not significantly, higher recall than participants who did not make JOLs ($M = 43.583$, $SE = 2.695$), $F(1, 78) = 2.908$, $p = .092$, $\eta_p^2 = .036$. The Pair Type × Judgment interaction was not significant, $F(1, 78) = 2.037$, $p = .158$, $\eta_p^2 = .025$. Figure 1a suggests that JOLs boosted cued recall of related pairs and only slightly enhanced cued recall of unrelated items, consistent with an ordinal interaction. These interactions are difficult to detect using null-hypothesis significance testing (Bobko, 1986; Keppel, 1982; see also the General Discussion). Because of this, planned comparisons were conducted to compare recall of related and unrelated items separately, regardless of the results of the ANOVA.

The planned comparisons showed that, for related items, participants who made JOLs recalled significantly more targets than did participants who did not make JOLs, $t(78) = 2.267$, $p = .026$, $d = 0.507$, $BF_{10} = 2.09$. For unrelated items, there was no difference in recall between participants who

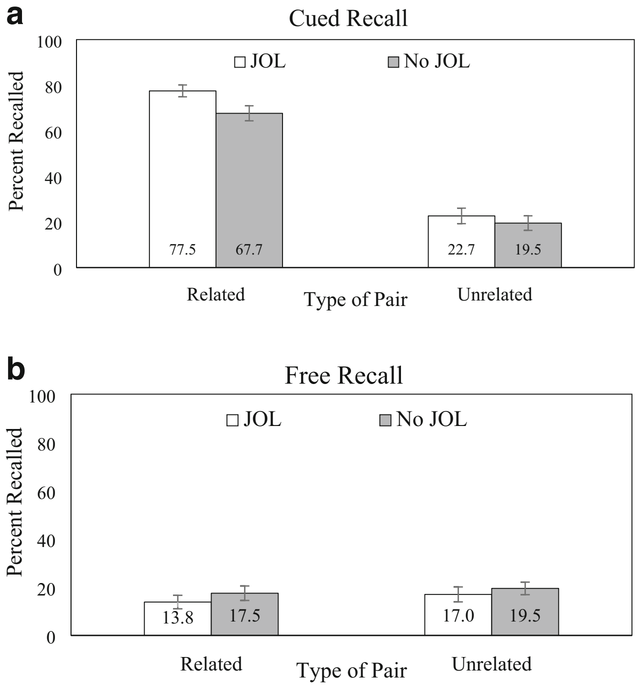

Fig. 1 **a** Average percentage recalled during cued recall of related and unrelated word pairs for participants who did (JOL) or did not (no JOL) provide JOLs during study in Experiment 1. Error bars reflect one standard error of the mean. **b** Average percentage recalled on a free-recall test for related and unrelated word pairs in Experiment 1 for participants who did (JOL) or did not (no JOL) provide JOLs during study. Error bars reflect one standard error of the mean

[2] We did not deem it appropriate to report omnibus analyses including different test types, as these tests involve data on different measurement scales. Interested readers can view the 2 (test type: cued recall, free recall/recognition) x 2 (pair type: related, unrelated) x 2 (judgment: JOL, no JOL) mixed-factor omnibus ANOVA for Experiments 1-3 at osf.io/ew5z2.

made JOLs and those who did not, $t(78) = 0.689$, $p = .493$, $d = 0.154$, $BF_{01} = 3.50$.

Free recall

Overall, free recall (see Fig. 1b) was numerically, but not significantly, higher for unrelated ($M = 18.250$, $SE = 2.002$) than for related targets ($M = 15.667$, $SE = 2.105$), $F(1, 78) = 2.853$, $p = .095$, $\eta_p^2 = .035$. There was no main effect of judgment, $F(1, 78) = 0.654$, $p = .421$, $\eta_p^2 = .008$, and no Pair Type × Judgment interaction, $F(1, 78) = 0.145$, $p = .704$, $\eta_p^2 = .002$. Planned analyses indicated that there was no significant difference in free recall between participants who made JOLs and those who did not for either related, $t(78) = -0.871$, $p = .386$, $d = -0.171$, $BF_{01} = 3.10$, or unrelated word pairs, $t(78) = -0.624$, $p = .534$, $d = -0.110$, $BF_{01} = 3.63$.

Discussion

Results from Experiment 1 showed that JOLs selectively improved memory for related word pairs on a cued-recall test. In contrast, on the free-recall test, for which targets must be recalled in the absence of cues, JOLs did not influence memory for either type of word pair. Bayesian analyses supported these conclusions, providing positive evidence in favor of the null hypothesis (i.e., no JOL reactivity) for both types of pairs on the free-recall test. Thus, Experiment 1 suggests that JOL reactivity does not occur when a criterion test is not sensitive to the same cues used to inform JOLs (Soderstrom et al., 2015). However, we note that recall on the free-recall test was low (*M*s = 13.83%–19.50%), leaving open the possibility that these data reflect scaling artifacts due to floor effects. Experiment 2 thus sought to replicate Experiment 1 under conditions that improved free-recall performance.

Experiment 2

In Experiment 2, we again compared performance between JOL and no-JOL conditions for related and unrelated word pairs. To enhance free recall, participants had an extra study opportunity for each list before completing either a cued-recall or free-recall test.

Method

Participants

A power analysis assuming a moderate effect size ($d = 0.588$) indicated that 47 participants per condition, assuming $\alpha = .05$, power of .80, and a two-tailed test, were necessary to reliably detect a difference between the JOL and no-JOL condition for cued recall of related pairs. This was the mean effect size (weighted by sample size) of Experiments 1 and 3[3] between the JOL and no-JOL condition for cued recall of related pairs. To equate participants per counterbalance, 48 participants were tested in each condition.

Participants were 102 (52 JOL, 50 no JOL) students who received course credit for participation. Four participants were removed from the JOL condition: three for not providing JOLs for at least 80% of the trials and one for not completing the experiment. Two participants were removed from the no-JOL condition due to technical errors. Therefore, data from 96 participants (34 men, 62 women) were analyzed. Participants were 18 to 43 years old ($M = 19.900$, $SD = 2.918$).

Materials and procedure

Experiment 2 was identical to Experiment 1, with the exception that participants in Experiment 2 received an extra study trial. Specifically, participants began each block by studying 30 word pairs, presented at a 12-second rate. After studying the entire list once, participants were then shown the list a second time, in a new randomized order. Participants in the JOL condition were only prompted to provide JOLs during the second study trial.

Results

As in Experiment 1, cued recall and free recall performance were analyzed separately in 2 (pair type: related, unrelated) × 2 (judgment: JOL, no JOL) mixed-factor ANOVAs.

Cued recall

Overall, cued recall (see Fig. 2a) was superior for related items ($M = 79.931$, $SE = 2.148$) compared with unrelated items ($M = 33.472$, $SE = 2.892$), $F(1, 94) = 321.725$, $p < .001$, $\eta_p^2 = .774$. The main effect of judgment was not significant, $F(1, 94) = 0.704$, $p = .404$, $\eta_p^2 = .007$. Although it followed the predicted pattern, the pair type × judgment interaction was not significant, $F(1, 94) = 2.675$, $p = .105$, $\eta_p^2 = .028$. Follow-up tests were conducted given our a priori predictions. Participants who made JOLs recalled numerically more related targets on the cued-recall test than participants who did not make JOLs, although this difference was not significant, and the Bayes factor indicated that both hypotheses were equally likely, $t(94) = 1.842$, $p = .069$, $d = 0.376$, BF_{01}

[3] Experiment 2 was completed after Experiment 3. To aid organization, we grouped experiments with a free-recall test (Experiments 1 and 2) and experiments with a recognition test (Experiments 3 and 4) in separate sections.

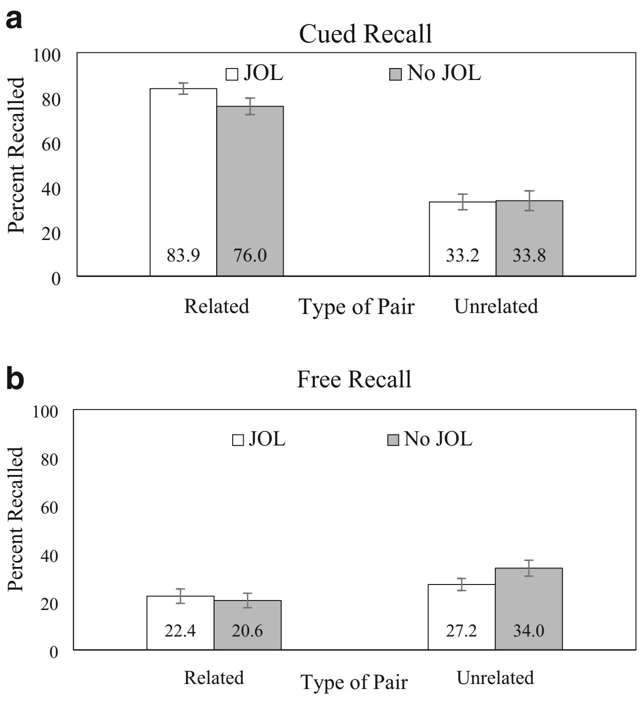

Fig. 2 **a** Average percentage recalled on a cued-recall test in Experiment 2 for participants who did (JOL) or did not (no JOL) provide JOLs during study. Error bars reflect one standard error of the mean. **b** Average percentage recalled on a free-recall test in Experiment 2 for participants who did (JOL) or did not (no JOL) provide JOLs during study. Error bars reflect one standard error of the mean

= 1.05. For unrelated items, there was no significant difference between the judgment conditions, $t(94) = -0.096$, $p = .924$, $d = -0.020$, $BF_{01} = 4.64$.

Free recall

Overall, free recall (see Fig. 2b) was significantly higher for unrelated ($M = 30.625$, $SE = 2.239$) than for related items ($M = 21.458$, $SE = 1.941$), $F(1, 94) = 24.082$, $p < .001$, $\eta_p^2 = .204$. There was no main effect of judgment (JOL: $M = 24.792$, $SE = 2.653$; No JOL: $M = 27.292$, $SE = 2.653$), $F(1, 94) = 0.444$, $p = .507$, $\eta_p^2 = .005$, but the pair type × judgment interaction was significant, $F(1, 94) = 5.313$, $p = .023$, $\eta_p^2 = .053$. Follow-up tests indicated that, for related items, recall did not differ between those who made JOLs and those who did not, $t(94) = 0.465$, $p = .643$, $d = 0.095$, $BF_{01} = 4.23$. For unrelated items, recall was numerically but not significantly higher for those who did not make JOLs than for those who did, $t(94) = -1.520$, $p = .132$, $d = -0.310$, $BF_{01} = 1.69$.

Discussion

Experiment 2 replicated Experiment 1, once again detecting an effect of JOL reactivity (albeit not substantial or significant) for related items. JOLs provided no memory advantage for either type of pair on a free-recall test, even when performance was elevated compared with Experiment 1. Bayes factors provided weak or positive evidence for the null for both types of items on free recall, but was inconclusive for cued recall of related items. Although the Bayesian evidence was sometimes inconclusive in this and other experiments, it is important to note that Bayes factors become more informative as sample size increases (Wagenmakers, Wetzels, Borsboom, van der Maas, & Kievit, 2012). We refer readers to a mini meta-analysis reported following Experiment 4 for a broader view of the evidence from Bayes factors.

In all, the pattern of results in Experiments 1 and 2 suggest that reactivity depends on the overlap between cues used to make JOLs and to later retrieve answers on a final test. Specifically, only tests sensitive to cue–target relationships (related items in cued recall) led to elevated memory performance after making JOLs.

Experiment 3

In Experiment 3, we sought to further explore Soderstrom et al.'s (2015) theory by considering another type of criterion test: item recognition. In particular, participants in Experiment 3 made either JOLs or did not make JOLs while studying related (*Buzz–Bee*) and unrelated (*Table–King*) word pairs. They then completed a cued-recall test or item-recognition (*Did you study BEE?*) test for the target of each pair. We anticipated that reactivity would be evident during cued recall for related pairs, consistent with previous experiments. In contrast, we expected that JOL reactivity would not be evident for either type of pair on the item-recognition test. Indeed, previous research suggests that during item recognition, participants must rely on item-level information (i.e., remembering specifically seeing *BEE* during study) and not relational information between the studied pair (cf. Hockley & Consoli, 1999). In Experiment 3, participants were charged with recognizing the target (*Bee*) in the absence of the original encoding context (*Buzz–Bee*). Therefore, item recognition should be insensitive to the cue–target relationships between word pairs, and thus reactivity should not occur.

Previous research provides inconclusive evidence for whether JOLs would influence later memory on a recognition test. For example, Begg et al. (1989, Experiments 1 and 4) included conditions with and without JOLs using recognition tests for lists of unrelated words. Their Experiment 1 appeared to result in positive reactivity, whereas Experiment 4 suggested no reactivity and negative reactivity (i.e., making JOLs harmed recognition compared with not making JOLs). Yang et al. (2015) and Halamish (2018) found positive reactivity on a recognition test when participants made JOLs for a list of unrelated words. From these studies, it is unclear what effect JOLs would have on an item-recognition test for targets from related and unrelated word pairs.

Method

Participants

One hundred thirty-eight (69 JOL, 69 no JOL) students received course credit for participating in Experiment 3. Nine participants from the JOL condition and eight from the no-JOL condition were removed because they did not respond to at least 90% of the recognition test items. Therefore, a total of 121 participants (54 men, 67 women) were included in analyses of Experiment 3: 60 in the JOL condition, 61 in the no-JOL condition. A sensitivity analysis indicated that with a sample size of 121 participants, α =.05, power of .80, and a two-tailed test, we could detect an effect size of d = 0.514 or higher.

Materials

Stimuli for Experiment 3 consisted of six lists that each contained 15 related pairs (forward strength 0.400–0.739, M = 0.506) and 15 unrelated pairs, frequency: 6.397–13.552 (M = 10.015), concreteness: 250–637 (M = 525.6), target word length: 3–8 letters (M = 4.644). These pairs consisted of those used in Experiment 1 and 30 new related pairs selected from the USF-FAN (Nelson et al., 1998). To form unrelated pairs, target words from one list were matched with unrelated cue words from another list. The six lists were counterbalanced so that target words were equally likely to appear with a related cue, an unrelated cue, or as a lure (i.e., item not studied) on the recognition test.

Design and procedure

A 2 (judgment: JOL, no JOL) × 2 (test type: cued recall, recognition) × 2 (pair type: related, unrelated) × 2 (item status: studied, lure) mixed-factor design was used. Judgment was manipulated between participants, with the remaining variables manipulated within participants. Participants studied two lists of 30 word pairs (15 related, 15 unrelated) and either made JOLs during study (JOL condition) or studied pairs without making JOLs (no-JOL condition). Participants completed a cued-recall test for one of these two lists in the same manner as in prior experiments. For the other list, they completed an item-recognition test.

The recognition test consisted of 60 items presented in a random order one at a time: 30 studied targets from the cue–target pairs and 30 lures that were not presented in the study phase. For each word, participants were instructed to select "yes" if they had studied the word in the previous list or "no" if the word was new. Because of an experimenter error, the test was not forced response. Each item appeared on the screen for 5 seconds, and if participants did not respond, the program automatically advanced to the next item. Because of this error, we removed participants who did not respond to at least 90% of the recognition test items, and recognition accuracy was adjusted for the total number of items to which participants provided a response.

Results

Cued recall

The percentage of targets correctly recalled (see Fig. 3a) was analyzed in a 2 (judgment: JOL, no JOL) × 2 (pair type: related, unrelated) mixed-factor ANOVA. Overall, recall was significantly higher for related (M = 76.194, SE = 1.738) than for unrelated items (M = 25.754, SE = 2.161), $F(1, 119)$ = 720.809, $p < .001$, η_p^2 = .858. On average, participants making JOLs (M = 55.500, SE = 2.445) recalled significantly more targets than participants who did not make JOLs (M = 46.448; SE = 2.425), $F(1, 119)$ = 6.910, p = .010, η_p^2 = .055. Although results followed the predicted pattern, the pair type ×

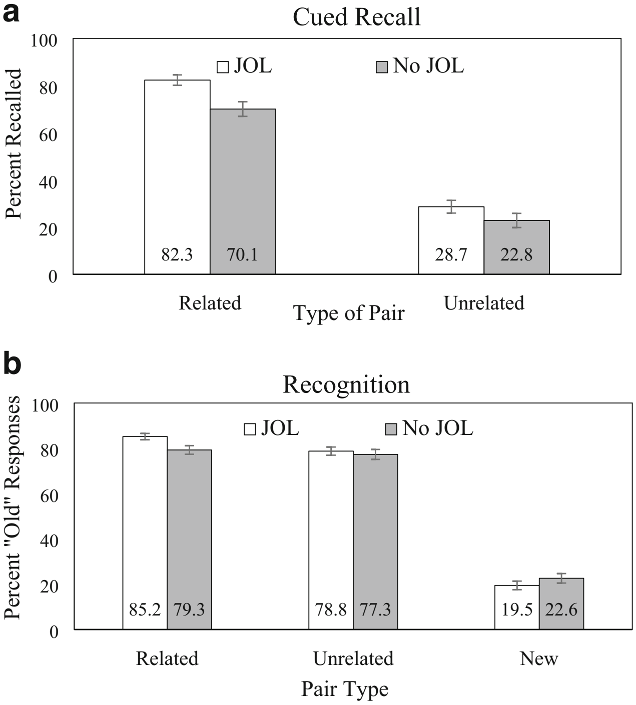

Fig. 3 **a** Average percentage recalled on a cued-recall test in Experiment 3 for participants who did (JOL) or did not (no JOL) provide JOLs during study. Error bars reflect one standard error of the mean. **b** Average percentage of "old" responses for items studied in related and unrelated pairs and new items (i.e., lures). Participants did (JOL) or did not (no JOL) provide JOLs during study. Error bars reflect one standard error of the mean

judgment interaction was not significant, $F(1, 119) = 2.950$, $p = .088$, $\eta_p^2 = .024$. Follow-up tests indicated that for related items, recall was significantly better for those who made JOLs relative to those who did not, $t(119) = 3.532$, $p = .001$, $d = 0.642$, $BF_{10} = 45.56$. For unrelated items, there was no difference between the JOL and no-JOL conditions, $t(119) = 1.347$, $p = .180$, $d = 0.245$, $BF_{01} = 2.29$.

Recognition

Analysis of hits and false alarms We first considered the proportion of studied items correctly called "old" (i.e., hits; see Fig. 3b). Hits were analyzed in a 2 (pair type: related, unrelated) × 2 (judgment: JOL, no JOL) mixed-factor ANOVA. Overall, hits were significantly more likely for items studied in related pairs ($M = 82.223$, $SE = 1.119$) than for items studied in unrelated pairs ($M = 78.026$, $SE = 1.466$), $F(1, 119) = 9.651$, $p = .002$, $\eta_p^2 = .075$. On average, those who made JOLs ($M = 81.966$, $SE = 1.585$) also had a numerically higher hit rate relative to those who did not make JOLs ($M = 78.283$, $SE = 1.572$), although this difference was not significant, $F(1, 119) = 2.722$, $p = .102$, $\eta_p^2 = .022$.

The judgment × pair type interaction was not significant, $F(1, 119) = 2.682$, $p =.104$, $\eta_p^2 = .022$, although those who made JOLs appeared to have a higher hit rate for related items than did those who did not make JOLs. Follow-up tests indicated that for items studied in related pairs, hits were significantly more likely for participants who made JOLs than for participants who did not, $t(119) = 2.633$, $p = .010$, $d = 0.479$, $BF_{10} = 4.22$. For items studied in unrelated pairs, there was no difference in hits between those who did or did not make JOLs, $t(119) = 0.501$, $p = .617$, $d = 0.091$, $BF_{01} = 4.61$.

False alarms (i.e., mistakenly endorsing lures) did not differ between the JOL and no-JOL condition, $t(119) = 1.110$, $p = .269$, $d = 0.202$, $BF_{01} = 2.960$. Because new items were never studied in pairs, there were not separate categories of related and unrelated lures.

Signal detection analyses We conducted independent samples *t* tests to analyze differences in discriminability and response criterion. Discriminability (d') did not differ between the JOL ($M = 2.023$, $SE = 0.098$) and no-JOL conditions ($M = 1.913$, $SE = 0.138$), $t(119) = 0.647$, $p = .519$, $d = 0.118$, $BF_{01} = 4.27$. Response criterion (C) also did not differ between the JOL ($M = -.010$, $SE = .054$) and no-JOL conditions ($M = .042$, $SE = .067$), $t(119) = -0.607$, $p = .545$, $d = -0.109$, $BF_{01} = 4.37$.

Discussion

Results for the cued-recall test replicated Experiment 1. That is, relative to participants who did not make JOLs, participants who provided JOLs exhibited significantly better memory for related items. Bayesian analyses provided strong evidence of JOL reactivity for related items and weak evidence of no reactivity for unrelated items during cued recall. Results for the recognition test were contrary to our hypotheses. JOLs were associated with significantly elevated hit rates, but only for targets that had been studied in related pairs (with positive Bayesian evidence). Because this contradicted our a priori hypotheses, we conducted a fourth experiment to replicate the finding.[4]

Experiment 4

In Experiment 4, we further investigated whether JOLs influence item recognition. Specifically, participants studied two lists of related and unrelated word pairs while providing JOLs for one list and not providing JOLs for the other (i.e., judgment was manipulated within participants). Participants completed an item-recognition test for each list.

Method

Participants

A sample size of 156 was determined by a power analysis using an effect size of $d = 0.226$, $\alpha = .05$, power of .80, and a two-tailed test. This was the mean effect size (weighted by sample size) of Experiment 3 and another experiment (reported at osf.io/ew5z2) for the difference in hit rates of related items between JOL and no-JOL conditions. To equate the number of participants per counterbalance, sample size was increased to 160, and one extra participant completed the experiment.

Two hundred seventeen participants were recruited from Amazon Mechanical Turk and were compensated $5 each for completing the study. Demographic data were not collected. Twenty-five participants were removed because they did not provide JOLs for at least 80% of the trials. Twenty-three others were removed because they responded in less than 500 ms to at least six of the 60 items (10%) on at least one of the recognition tests, responses we deemed too rapid to truly consider as a response. Eight others were removed because they reported

[4] In another experiment (Experiment 4b), we had participants identify whether an item was studied and subsequently indicate whether they recollected details about the item or whether the item was merely familiar (results are presented at osf.io/ew5z2). In summary, differences were evident between the JOL and no-JOL conditions for recollection and familiarity judgments, but when recognition was collapsed across these responses, no overall JOL reactivity occurred. This failure to replicate JOL reactivity on a recognition test also motivated Experiment 4, and data from Experiment 4b are included in the reported meta-analysis.

technical difficulties.[5] Therefore, a total of 161 participants were included in analyses.

Design and procedure

A 2 (judgment: JOL, no JOL) × 2 (pair type: related, unrelated) × 2 (item status: studied, lure) within-participants design was used. Stimuli for Experiment 4 consisted of eight lists that each contained 15 related pairs (forward strength 0.400–0.739, $M = 0.499$) and 15 unrelated pairs, frequency: 6.397–13.552 ($M = 9.944$), concreteness: 250–670 ($M = 531.7$), target word length: 3–8 letters ($M = 4.750$). Thirty new related word pairs from the USF-FAN database (Nelson et al., 1998) were added to the 90 pairs from Experiment 3. The unrelated pairs and mixed lists were created using methods similar to Experiment 3.

The procedure for Experiment 4 was identical to Experiment 3, with the following exceptions. First, instead of randomly assigning participants to the JOL or no-JOL condition, all participants provided JOLs during one study block and did not provide JOLs during the other block (i.e., judgment was manipulated within participants). Order was counterbalanced so that half the participants provided JOLs for the first block and half provided JOLs for the second block.[6] Participants completed an item-recognition test for both word lists. Tests in Experiment 4 were forced response, whereby participants were required to respond to each item on the test.

Results

Analysis of hits and false alarms Hits (see Fig. 4) were analyzed in a 2 (pair type: related, unrelated) × 2 (judgment: JOL, no JOL) repeated-measures ANOVA. Overall, hits were significantly more likely for items studied in related pairs ($M = 73.333$, $SE = 1.238$) than in unrelated pairs ($M = 69.689$, $SE = 1.326$), $F(1, 160) = 15.550$, $p < .001$, $\eta_p^2 = .089$. On average, hits were also significantly greater when participants made JOLs ($M = 73.375$, $SE = 1.401$) compared with when they did not ($M = 69.648$, $SE = 1.416$), $F(1, 160) = 6.285$, $p = .013$, $\eta_p^2 = .038$. The judgment × pair type interaction was not significant, $F(1, 160) = 0.543$, $p = .462$, $\eta_p^2 = .003$. Follow-up tests indicated that for items studied in related pairs, the hit rate was significantly greater after making JOLs compared with not making JOLs, $t(160) = 2.698$, $p = .008$, $d = 0.230$, $BF_{10} = 2.91$, although this effect was small. For items studied in unrelated pairs, hits were greater when participants made JOLs than when they did not, although this difference was not significant, and the Bayes factor favored the null, $t(160) = 1.785$, $p = .076$, $d = 0.156$, $BF_{01} = 2.42$. False alarms did not differ between JOL and no-JOL conditions, $t(160) = -0.363$, $p = .717$, $d = -0.029$, $BF_{01} = 10.67$.

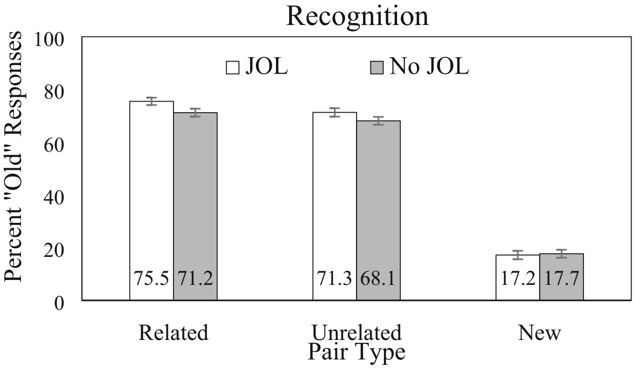

Fig. 4 Average percentage of "old" responses in Experiment 4 for related, unrelated, and new items (i.e., lures). Participants made JOLs during study for one block (JOL) and did not make JOLs for the other block (no JOL). Errors bars reflect one standard error of the mean

Signal-detection analyses Discriminability (d') was significantly greater when participants made JOLs ($M = 2.183$, $SE = 0.111$) than when they did not ($M = 1.946$, $SE = 0.102$), $t(160) = 2.180$, $p = .031$, $d = 0.175$, $BF_{01} = 1.14$, although the effect was small, and Bayesian evidence was inconclusive. Response criteria (C) did not differ between the JOL ($M = .323$, $SE = .047$) and no-JOL conditions ($M = .365$, $SE = .044$), $t(160) = -0.819$, $p = .414$, $d = -0.072$, $BF_{01} = 8.19$.

Discussion

JOLs again increased hit rates on a recognition test (replicating Experiment 3), as well as significantly enhancing discriminability. In contrast to Experiment 3, JOLs slightly elevated hits for items studied in unrelated pairs (although this difference was not significant and the Bayes factor favored no difference) as well as related pairs. We did not predict this finding a priori and, as it did not occur in Experiment 3, consider the finding tentative. Most importantly, we detected small but significant reactivity for hits for related word pairs on a recognition test.

Meta-analysis of experiments

The reported experiments tested the premise that providing JOLs selectively enhances memory, compared with not making JOLs, when the final criterion test relies on cues participants consider when making their JOLs. However, results did not strongly support this hypothesis in each experiment. In order to provide greater purchase on these data, we report a

[5] Patterns of results did not differ when all participants were included in analyses.

[6] Because judgment was manipulated within participants, we examined potential order effects and report these analyses at osf.io/ew5z2. Results did not indicate any consistent order effects.

small-scale, fixed-effects meta-analysis of these experiments, broken down by test type (free recall, cued recall, item recognition) and pair type (related, unrelated). To ensure complete inclusion of available data and thus more precise point estimates, we also included Experiment 4b (reported at osf.io/ew5z2) when estimating effect sizes. The meta-analysis was conducted after all data had been collected (Ueno, Fastrich, & Murayama, 2016).

For each experiment, we input an effect size (Cohen's d) representing the standardized difference in performance for items given JOLs and items not given JOLs (differences in hit rates were considered for item recognition). For the repeated-measures design used in Experiments 4 and 4b, we also accounted for the correlation between the two measures, using Cohen's d_{rm} (Lakens, 2013). Aggregate effect sizes are reported weighted by sample size (cf. Hedges & Olkin, 1985). All analyses were conducted using Comprehensive Meta-Analysis Version 2.0 (Borenstein, Hedges, Higgins, & Rothstein, 2005). A Bayesian meta-analysis was also conducted, and Bayes factors for the effect sizes are reported in the proceeding sections. Given that each mean weighted effect size represents data from only 2–3 experiments, these data should be treated with caution and viewed largely as descriptions of the cumulative pattern of results rather than a basis for strong inferences.

In aggregate, making JOLs conferred a small benefit to memory performance relative to not making JOLs ($d = 0.17$, $p < .001$, 95% CI: [0.10, 0.25]). Test type significantly moderated the JOL reactivity effect, $Q = 11.48$, $p = .003$. JOL reactivity was significantly larger on cued-recall and recognition tests than on free-recall tests (ps = .008, .006, respectively), but did not differ significantly between cued recall and recognition ($p = .15$). Pair type (collapsed across all tests) also moderated the JOL reactivity effect such that JOLs benefited learning of related pairs more than unrelated pairs, although this effect did not reach the alpha threshold, $Q = 3.76$, $p = .052$.

The key question for the present research was whether the effect of pair type on JOL reactivity depended on the test type. We could not determine whether there was a pair type × test type interaction in the meta-analysis because there were only 2–3 means per condition (see k in Table 1). Williams (2012) recommended a minimum of five means per condition in a moderator analysis. Therefore, we examined JOL reactivity across the five experiments for related and unrelated items as two separate meta-analyses and examined test type as a moderator in each meta-analysis.

Related items

The upper panel of Table 1 displays the mean weighted effect size, 95% confidence interval, inferential statistical tests, number of experiments contributing (k), and total number of participants for each test type for related items. Collapsed across all test types, there was a small, statistically significant JOL reactivity effect for related items, $d = 0.25$, $p < .001$, 95% CI [0.14, 0.35], which was moderated by test type, $Q = 8.81$, $p = .01$. Consistent with Soderstrom et al. (2015), participants making JOLs demonstrated better cued recall than participants who did not make JOLs ($d = 0.518$, $BF_{10} = 1384.89$), with this difference characterized as a medium effect with very strong Bayesian evidence. There was no reactivity evident for free recall ($d = -0.036$, $BF_{01} = 5.96$). However, an advantage in favor of JOLs was present for hits on item-recognition tests ($d = 0.228$, $BF_{10} = 72.49$), with the Bayes factor providing strong evidence.

Unrelated items

The lower panel of Table 1 displays meta-analytic data for each test type for unrelated items. Collapsed across all test types, there was a very small JOL reactivity effect for unrelated items, $d = 0.10$, $p = .05$, 95% CI [−0.002, 0.21], that did not meet conventional significance. In addition, the effect was numerically, but not significantly, moderated by type of test, $Q = 5.54$, $p = .06$. A small but significant difference was evident favoring items given JOLs for item recognition ($d = 0.158$, $BF_{01} = 2.04$), but Bayesian evidence was inconclusive and favored the null (i.e., no JOL reactivity). Therefore, there is not enough evidence to determine whether JOL reactivity occurs for unrelated items on a recognition test. A small, nonsignificant benefit of JOLs was found for cued recall ($d = 0.135$, $BF_{01} = 4.12$), but Bayesian evidence again provided moderate evidence of no effect. For free recall, JOLs appeared to confer a slight disadvantage ($d = -0.232$, $BF_{10} = 1.61$), as performance was better when participants did not provide JOLs (cf. Mitchum et al., 2016). However, this disadvantage was not significant, and the Bayes factor was inconclusive.

In all, consistent with our predictions, the effect of test type on JOL reactivity was numerically larger for related pairs than for unrelated pairs and was only statistically significant for related pairs. However, this conclusion should be interpreted with caution, as we could not directly test the pair type × test ype interaction because of the small number of effect sizes in each condition.

General discussion

Soderstrom et al. (2015) posited that JOL reactivity occurs only if the criterion test is sensitive to the same cues used to inform JOLs (e.g., relatedness). By this account, while studying related and unrelated word pairs, participants use relatedness as a cue to inform their JOLs of related items (cf. Arbuckle & Cuddy, 1969). On a later cued-recall test, that prior attention to relatedness enhances memory for the target

Table 1 Meta-analysis of memory performance for items given JOLs versus no JOLs

Test type	Effect size	CI95_Lower	CI95_Upper	Z	p	k	N
Related Items							
Free recall	−0.036	−0.332	0.259	−0.242	.809	2	88
Cued recall	0.518	0.287	0.750	4.388	<.001	3	148
Item recognition	0.228	0.095	0.360	3.372	.001	3	270
Unrelated Items							
Free recall	−0.232	−0.529	0.064	−1.535	.125	2	88
Cued recall	0.135	−0.093	0.363	1.158	.247	3	148
Item recognition	0.158	0.032	0.284	2.466	.014	3	270

Note. Effect size = mean weighted effect size, Cohen's *d*; CI95 = 95% confidence interval; *k* = number of effect sizes; *N* = number of participants contributing to the effect size

when provided with the cue, leading to JOL reactivity for related pairs. The current study investigated a key prediction of this explanation by examining performance on tests that should be sensitive to relatedness (i.e., cued recall) and tests that should be less sensitive to relatedness (i.e., free recall, item recognition) when participants made JOLs compared with when they did not make JOLs.

In Experiment 1, JOLs significantly improved memory for related pairs, but did not significantly influence memory for unrelated word pairs in a cued-recall test, similar to Soderstrom et al. (2015). However, JOLs did not influence performance for either pair type on a free-recall test, even when free-recall performance was elevated (Experiment 2). Experiment 3 replicated the cued-recall test results of Experiment 1, and JOLs also improved the hit rate of items studied in related pairs on an item-recognition test, contrary to our predictions. Accordingly, we conducted Experiment 4 to replicate results for item recognition. Our findings were consistent with results from Experiment 3: JOLs elevated hits for items studied in related pairs. A marginal benefit on hits was also found for unrelated pairs given JOLs.

Collectively, results from these experiments provide somewhat mixed support for the patterns of JOL reactivity that we anticipated would occur across different types of tests. This may reflect some imprecision in the effect sizes that informed power analyses, as several of the test types examined (recognition, free recall) had little prior precedent in the literature. Thus, some experiments may be underpowered with respect to the true population effect size. Furthermore, we did not have sufficient power to support our arguments suggestive of an interaction between making JOLs and the type of word pair due to the expectation that this is an ordinal interaction (i.e., JOLs influence memory for related pairs more than for unrelated pairs). An ordinal interaction such as this (particularly with the moderate effect sizes detected) would require hundreds of participants to detect. Future research could devote more subjects to similar experiments and further develop materials to reduce variability in performance in order to more effectively estimate the size of this interaction.

To provide better purchase on the pattern of findings, we conducted a meta-analysis of all experiments completed. Overall, the most robust JOL reactivity was evident for related items subjected to a cued-recall test ($d = 0.518$), with Bayesian analyses providing very strong evidence for this effect. Item recognition of related items was also characterized by significant JOL reactivity and strong Bayesian evidence, although the effect size was small ($d = .228$). Unrelated item recognition yielded a small effect of reactivity on hit rates, but inconclusive Bayesian evidence. Free recall yielded little reactivity with a trend for JOLs to harm free recall of unrelated items ($d = -0.232$), although Bayesian evidence was inconclusive. The Bayes factors for unrelated items were most likely inconclusive because, as noted previously, JOLs may have differing effects depending on what cues participants used to inform their JOLs.

Accounts of JOL reactivity

Soderstrom et al. (2015) reported that making JOLs selectively benefitted memory for related items during cued recall. They explained these data by proposing that JOLs encourage participants to attend to specific cues, such as relatedness, that may support performance on a subsequent test. Our results generally comport with this account and offer an important corollary: The direction and strength of JOL reactivity depends on both the study material and type of final test. For example, reactivity was evident for related items given JOLs tested via cued recall, but not when tested with free recall.

Although not anticipated, our finding that JOLs may influence hits in item recognition is consistent with models suggesting that recognition decisions reflect broad, global access to memory (e.g., Hintzman, 1988; but see also Shiffrin & Steyvers, 1997), as well as accounts holding that the rememberer interrogates memory by reinvoking the processes instantiated at encoding (e.g., Jacoby, Shimizu, Daniels, & Rhodes, 2005). That is, the item on the recognition test may encourage participants to retrieve other elements of the encoding context when making their recognition decision. If JOLs increase the likelihood that participants retrieve elements of the encoding context (e.g., the studied word pair), a recognition advantage might accrue. However, additional data collection is necessary to fully examine the mechanisms that account for JOL reactivity during item recognition.

Results also align with the item-specific–relational account of encoding strategies (see, e.g., Mulligan & Peterson, 2015; Peterson & Mulligan, 2013). Similar to other encoding effects (e.g., generation effects, bizarreness effects), JOLs may selectively strengthen memory of information specific to each item. With cue–target word pairs, this would entail strengthening the relationship between the words in the pair and also possibly strengthening item-specific features of the target words. Importantly, JOLs should not encourage interitem processing (i.e., relationships between different word pairs). On later tests that rely heavily on item-specific information (cued recall and item recognition), JOLs may improve performance. However, JOLs may harm performance on tests that rely on interitem processing (such as certain free-recall tests). In the current experiments, there were no structured relationships between the various targets in each study list, so there was not necessarily any interitem processing that could be disrupted by making JOLs. Future research is needed to systematically determine the effect of JOLs on item-specific and interitem processing separately.

As an alternative to Soderstrom et al.'s (2015) account, Mitchum et al. (2016) proposed that the act of making JOLs indirectly influences memory by drawing attention to the difficulty of items, thus affecting participants' study decisions. Specifically, their account holds that, when making JOLs, participants devote more time and effort to items judged easy to learn (e.g., related pairs) and less effort to items judged difficult to learn (e.g., unrelated pairs). Consequently, JOLs improve memory for easy items, but harm memory for difficult items. For example, under experimenter-paced study (e.g., Mitchum et al., 2016, Experiment 5), making JOLs benefitted cued recall of related items, but harmed recall of unrelated items. Our results are generally inconsistent with this account of JOL reactivity. Indeed, the only evidence of JOLs harming memory evident in our meta-analysis was for unrelated items on free-recall tests, yielding a small effect size ($d = -0.232$). Moreover, our Experiments 1–3 used a very similar procedure as Mitchum et al.'s (2016) Experiment 5, but did not detect significant differences between the JOL and no-JOL conditions for cued recall of unrelated items.

A similar explanation suggested by Double et al. (2018) holds that JOL reactivity is driven by task difficulty. Specifically, Double et al. (2018) argued that making JOLs draws participants' attention to their confidence in mastering items. When participants feel confident they will learn material (e.g., for related items), JOLs confer a memory benefit, but this does not occur for difficult material (e.g., unrelated items; see Double & Birney, 2017). Although promising, this explanation cannot fully account for the pattern of results reported in the current study. In particular, if related items uniformly confer high confidence in mastery, JOL reactivity should be evident for related items regardless of the criterion test, in contrast to our results. We note that our experiments cannot entirely rule out a difficulty-based explanation and suggest that future work would benefit by competitively testing these accounts. Indeed, the present results cannot be accommodated by any single framework, suggesting that a comprehensive account may need to invoke multiple processes.

Implications and conclusions

JOL reactivity presents an important challenge to research on metamemory. That is, if the act of making JOLs influences memory performance, then researchers must not only account for reactivity but also adjust theory to understand when reactivity occurs. The current experiments indicate that accounts must consider the type of final test in conjunction with the type of study material. This perspective is consistent with Jenkins' (1979; see also Roediger, 2008) tetrahedral model of memory experiments, which holds that any conclusions about memory represent a combination of four factors: participants, retrieval (i.e., type of test), events (i.e., type of study stimuli), and encoding (i.e., instructions, activities at encoding). The present experiments reflect only a two-dimensional combination of these factors—materials and retrieval—and even within those categories many possibilities remain to be exhausted. For example, future research must also consider other types of cues used to make JOLs and how those cues may influence different tests. Likewise, other stimuli (e.g., single words, pictures, faces) may produce different patterns of performance than word pairs (see, e.g., Double et al., 2018; Tauber & Rhodes, 2012). Accordingly, the tetrahedral model sets a useful agenda for understanding JOL reactivity and developing theory by considering the combinations of stimuli, participants, tests, and encoding conditions that predict its occurrence.

For the present, the experiments reported in this paper indicate that JOL reactivity is influenced by the overlap between cues used to make JOLs and cues used on a final criterion test. Future research is needed to more carefully identify what cues participants use to make JOLs, and how these cues would influence later test performance.

Open practices statement

Supplemental materials are available at osf.io/ew5z2. No experiments were preregistered.

Author Note This work was supported by the National Science Foundation Graduate Research Fellowships under Grant No. 006784-00002 to the first author and Grant No. DGE-1321845 to the third author. The authors report no conflicts of interest.

References

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology, 81*(1), 126–131. doi:https://doi.org/10.1037/h0027455

Begg, I., Duft, S., Lalonde, P., Melnick, R., & Sanvito, J. (1989). Memory predictions are based on ease of processing. *Journal of Memory and Language, 28*(5), 610–632. doi:https://doi.org/10.1016/0749-596X(89)90016-8

Begg, I. M., Martin, L. A., & Needham, D. R. (1992). Memory monitoring: How useful is self-knowledge about memory? *European Journal of Cognitive Psychology*, *4*(3), 195–218. doi:https://doi.org/10.1080/09541449208406182

Bobko, P. (1986). A solution to some dilemmas when testing hypotheses about ordinal interactions. *Journal of Applied Psychology, 71*(2), 323.

Borenstein, M., Hedges, L., Higgins, J., & Rothstein, H. (2005). Comprehensive meta-analysis 2.0. Englewood: Biostat. Retrieved from https://www.meta-analysis.com/

Double, K. S., & Birney, D. P. (2017). Are you sure about that? Eliciting confidence ratings may influence performance on Raven's Progressive Matrices. *Thinking & Reasoning, 23*(2), 190–206. doi:https://doi.org/10.1080/13546783.2017.1289121

Double, K. S., Birney, D. P., & Walker, S. A. (2018). A meta-analysis and systematic review of reactivity to judgements of learning. *Memory, 26*(6), 741–750. doi:https://doi.org/10.1080/09658211.2017.1404111

Dunlosky, J., & Thiede, K. W. (1998). What makes people study more? An evaluation of factors that affect self-paced study. *Acta Psychologica, 98*(1), 37–56. doi:https://doi.org/10.1016/s0001-6918(97)00051-6

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. doi:https://doi.org/10.3758/bf03193146

Halamish, V. (2018). Can very small font size enhance memory? *Memory & Cognition, 46*(6), 979–993. doi:https://doi.org/10.3758/s13421-018-0816-6

Hedges, L., & Olkin, I. (1985). *Statistical models for meta-analysis*. San Diego: Academic Press.

Hintzman, D. L. (1988). Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review*, *95*(4), 528–551. doi:https://doi.org/10.1037/0033-295X.95.4.528

Hockley, W. E., & Consoli, A. (1999). Familiarity and recollection in item and associative recognition. *Memory & Cognition, 27*(4), 657–664. doi:https://doi.org/10.3758/BF03211559

IBM Corp. (2016). IBM SPSS Statistics for Windows (Version 24.0) [Computer software]. Armonk: IBM Corp.

Jacoby, L. L., Shimizu, Y., Daniels, K. A., & Rhodes, M. G. (2005). Modes of cognitive control in recognition and source memory: Depth of retrieval. *Psychonomic Bulletin & Review*, *12*(5), 852–857. doi:https://doi.org/10.3758/BF03196776

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. https://doi.org/10.3758/s13423-018-1463-4 .

Jenkins, J. J. (1979). Four points to remember: A tetrahedral model of memory experiments. In L. S. Cermak & F. I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 429–446). Hillsdale: Erlbaum.

Keleman, W. L., & Weaver, C. A. (1997). Enhanced metamemory at delays: Why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(6), 1394–1409. doi:https://doi.org/10.1037//0278-7393.23.6.1394

King, J. F., Zechmeister, E. B., & Shaughnessy, J. J. (1980). Judgments of knowing: The influence of retrieval practice. *American Journal of Psychology, 93*(2), 329–343. doi:https://doi.org/10.2307/1422236

Kruschke, J. K. (2013). Bayesian estimation supersedes the *t* test. *Journal of Experimental Psychology: General*, *142*(2), 573–603. doi:https://doi.org/10.1037/a0029146

Lakens, D. (2013). Calculating and reporting effect sizes to facilitate cumulative science: A practical primer for *t*-tests and ANOVAs. *Frontiers in Psychology, 4*, 863. doi:https://doi.org/10.3389/fpsyg.2013.00863

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. doi:https://doi.org/10.1037/a0039923

Morey, R. D., & Rouder, J. N. (2018). BayesFactor: Computation of Bayes Factors for Common Designs (R Package Version 0.9.12-4.2) [Computer software]. Retrieved from https://CRAN.R-project.org/package=BayesFactor

Mueller, M. L., Tauber, S. K., & Dunlosky, J. (2013). Contributions of beliefs and processing fluency to the effect of relatedness on judgments of learning. *Psychonomic Bulletin & Review, 20*(2), 378–384. doi:https://doi.org/10.3758/s13423-012-0343-6

Mulligan, N. W., & Peterson, D. J. (2015). Negative and positive testing effects in terms of item-specific and relational information. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(3), 859–871. doi:https://doi.org/10.1037/xlm0000056

Nelson, D. L., McEvoy, C. L., & Schreiber, T. A. (1998). *The University of South Florida word association, rhyme, and word fragment norms.* Retrieved from. http://w3.usf.edu/FreeAssociation/

Nelson, T. O. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation, 26,* 125–173. doi:https://doi.org/10.1016/S0079-7421(08)60053-5

Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*(4), 1287–1293. doi:https://doi.org/10.1037/a0031337

R Core Team. (2014). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from http://www.R-project.org/

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

Rhodes, M. G., & Tauber, S. K. (2011). The influence of delaying Judgements of Learning (JOLs) on metacognitive accuracy: A meta-analytic review. *Psychological Bulletin, 137*(1), 131–148. doi:https://doi.org/10.1037/a0021705

Roediger, H. L., III. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology*, *59*, 225–254. doi:https://doi.org/10.1146/annurev.psych.57.102904.190139

Rouder, J. N., Speckman, P. L., Sun, D., Morey, R. D., & Iverson, G. (2009). Bayesian *t* tests for accepting and rejecting the null hypothesis. *Psychonomic Bulletin & Review*, *16*(2), 225–237. doi:https://doi.org/10.3758/PBR.16.2.225

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). E-Prime reference guide. Pittsburgh: Psychology Software Tools, Inc.

Shiffrin, R. M., & Steyvers, M. (1997). A model for recognition memory: REM—Retrieving effectively from memory. *Psychonomic Bulletin & Review*, *4*(2), 145–166. doi:https://doi.org/10.3758/BF03209391

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*(2), 553–558. doi:https://doi.org/10.1037/a0038388

Tauber, S. K., Dunlosky, J., & Rawson, K. A. (2015). The influence of retrieval practice versus delayed judgments of learning on memory: Resolving a memory-metamemory paradox. *Experimental Psychology, 62*(4), 254–263. doi:https://doi.org/10.1027/1618-3169/a000296

Tauber, S. K., & Rhodes, M. G. (2012). Measuring memory monitoring with judgments of retention (JORs). *The Quarterly Journal of Experimental Psychology, 65*(7), 1376–1396. doi:https://doi.org/10.1080/17470218.2012.656665

Ueno, T., Fastrich, G. M., & Murayama, K. (2016). Meta-analysis to integrate effect sizes within an article: Possible misuse and Type I error inflation. *Journal of Experimental Psychology: General*, *145*(5), 643–654. doi:https://doi.org/10.1037/xge0000159

Undorf, M., Söllner, A., & Bröder, A. (2018). Simultaneous utilization of multiple cues in judgments of learning. *Memory & Cognition, 46*(4), 507–519.

Wagenmakers, E. J. (2007). A practical solution to the pervasive problems of *p* values. *Psychonomic Bulletin & Review*, *14*(5), 779–804. doi:https://doi.org/10.3758/BF03194105

Wagenmakers, E. J., Wetzels, R., Borsboom, D., van der Maas, H. L. J., & Kievit, R. A. (2012). An agenda for purely confirmatory research. *Perspectives on Psychological Science, 7*(6), 632–638. doi:https://doi.org/10.1177/1745691612463078

Williams, R. (2012). *Moderator analyses: Categorical models and meta-regression*. Paper presented at the annual Campbell Collaboration Colloquium, Copenhagen, Denmark.

Witherby, A. E., & Tauber, S. K. (2017). The influence of judgments of learning on long-term learning and short-term performance. *Journal of Applied Research in Memory and Cognition, 6*(4), 496–503.

Yang, H., Cai, Y., Liu, Q., Zhao, X., Wang, Q., Chen, C., & Xue, G. (2015). Differential neural correlates underlie judgment of learning and subsequent memory performance. *Frontiers in Psychology, 6,* 1699. doi:https://doi.org/10.3389/fpsyg.2015.01699

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.