Original Article

Reactivity of Judgments of Learning in a Levels-of-Processing Paradigm

Eylul Tekin and Henry L. Roediger III

Department of Psychological & Brain Sciences, Washington University in St. Louis, MO, USA

Abstract: Recent studies have shown that judgments of learning (JOLs) are reactive measures in paired-associate learning paradigms. However, evidence is scarce concerning whether JOLs are reactive in other paradigms. In *old/new* recognition experiments, we investigated the reactivity effects of JOLs in a levels-of-processing (LOP) paradigm. In Experiments 1 and 2, for each word, subjects saw a *yes/no* orienting question followed by the target word and a response. Then, they either did or did not make a JOL. The *yes/no* questions were about target words' appearances, rhyming properties, or category memberships. In Experiment 3, for each word, subjects gave a pleasantness rating or counted the letter "e". Our results revealed that JOLs enhanced recognition across all orienting tasks in Experiments 1 and 2, and for the e-counting task in Experiment 3. This reactive effect was salient for shallow tasks, attenuating – but not eliminating – the LOP effect after making JOLs. We conclude that JOLs are reactive in LOP paradigms and subjects encode words more effectively when providing JOLs.

Keywords: judgments of learning, reactivity, metacognition, levels of processing

Judgments of learning (JOLs) have become one of the most frequently used metacognitive measures since the revival of interest in metacognition in the 1980s (Arbuckle & Cuddy, 1969; Lovelace, 1984; Shaughnessy, 1981; Vesonder & Voss, 1985). JOLs assess the likelihood that an individual will remember a studied item on a future test. More importantly, JOLs assess people's monitoring accuracy. This ability (or inability) to monitor one's own learning is critical because monitoring partially determines how people control their learning behavior (Bjork et al., 2013; Nelson & Narens, 1990). For this reason, extensive theoretical and practical research on JOLs has appeared since the 1980s. The plethora of studies examined variables that affect JOLs (Dunlosky & Matvey, 2001; Koriat & Ma'ayan, 2005); the underlying bases of JOLs (Koriat, 1997; Koriat et al., 2004); when JOLs are accurate or inaccurate (Koriat & Bjork, 2005; Rhodes & Castel, 2009); how timing of JOLs (immediate vs. delayed) affects their accuracy (Kimball & Metcalfe, 2003; Nelson & Dunlosky, 1991); and the role of JOLs in study time allocation (Metcalfe & Kornell, 2005; Metcalfe & Finn, 2008).

One issue that has not received much attention until recently is the reactive effects of immediate JOLs (see Rhodes & Tauber, 2011; Spellman & Bjork, 1992, for some research on reactivity of delayed JOLs). That is, asking people to monitor their learning through JOLs might intentionally or unintentionally change the learning outcome or behavior itself (Ericsson & Simon, 1980; Harris & Lahey, 1982). Most researchers, however, have assumed no reactivity of JOLs and did not have control conditions without JOLs to test for JOL reactivity. Unfortunately, findings and interpretations in many studies would change if JOLs are reactive, and thus it is important to address the issue of reactivity.

A limited number of studies that included control conditions (i.e., no-JOL conditions) found mixed results regarding JOL reactivity. Some studies suggested that people performed better when they made JOLs in paired-associate learning and single-word learning paradigms (Dougherty et al., 2005; Yang et al., 2015; Zechmeister & Shaughnessy, 1980), whereas other studies did not report any differences between JOL and no-JOL conditions using similar paradigms and a foreign language learning paradigm (Benjamin et al., 1998; Dougherty et al., 2018; Kelemen & Weaver, 1997; Kornell & Bjork, 2008; Tauber & Rhodes, 2012). In these studies, the goal was not to investigate JOL reactivity, and thus presentation time of items was not always equivalent across JOL and no-JOL conditions.

More recently, two studies directly examined JOL reactivity in a paired-associate learning paradigm controlling for the time confound. In both studies, subjects studied related (easy) and unrelated (difficult) word pairs and half of subjects made immediate JOLs after each study trial. In Soderstrom et al. (2015), each pair was presented for 8 s in JOL and no-JOL conditions (experimenter-paced), with subjects in the JOL condition making JOLs in the last half

https://doi.org/10.1027/2151-2604/a000425

(4 s) of each trial. In Mitchum et al. (2016), subjects paced their study of pairs (self-paced). The word pairs were not presented during the period of making JOLs to ensure that the JOL condition did not have extra study time. Both studies reported that on the final cued recall test, subjects who made JOLs showed a greater recall difference between related and unrelated pairs compared to subjects who did not make JOLs. Some differences did occur between studies, however. Soderstrom et al. (2015) obtained greater recall for related pairs in the JOL condition (positive reactivity) and no recall difference between JOL and no-JOL conditions for unrelated pairs. On the other hand, Mitchum et al. (2016) found poorer recall for unrelated pairs in the JOL condition (negative reactivity) and no recall difference between JOL and no-JOL conditions for related pairs. Soderstrom et al. (2015) concluded that JOLs strengthened information used in making JOLs and when the final test is sensitive to such information, making JOLs would enhance performance on judged items. More specifically, in their study, JOLs enhanced cued recall solely for related pairs because for such pairs, subjects used the strong relatedness between words as a basis for their JOLs, whereas for unrelated pairs, there was no relationship to be used. In contrast, Mitchum et al. (2016) concluded that JOLs made subjects aware that unrelated pairs were not as memorable as related ones, and thus changed their mastery learning goal to a more pragmatic one (i.e., the changed-goal hypothesis, for a more detailed discussion, see Janes et al., 2018). Although their interpretations of reactivity differed, these studies as well as a meta-analysis conducted by Double et al. (2018) suggested that JOLs were reactive measures, at least for related word pairs. Furthermore, Witherby and Tauber (2017) found positive JOL reactivity on a 2-day delayed cued recall test. These conclusions, however, are limited because in all these studies a paired-associate learning paradigm was employed and stimuli consisted of either mixed or pure lists of related and unrelated word pairs. JOLs are used in many other paradigms, and thus it is important to investigate the generalizability of JOL reactivity. We report three experiments examining reactive effects of JOLs in a levels-of-processing (LOP) paradigm.

The LOP effect refers to the robust finding that deep levels of processes (semantic tasks) produce higher and long-lasting retention than shallow levels of processes (orthographic or phonemic tasks) (Craik & Lockhart, 1972; Hyde & Jenkins, 1969). We used the standard LOP paradigm introduced by Craik and Tulving (1975), in which subjects are presented with words like "chair" along with *yes/no* questions about the word's appearance (e.g., "Is the word in lowercase letters?"), phonemic properties (e.g., "Does the word rhyme with hair?"), or its category membership (e.g., "Does the word refer to a type of a furniture?"). The questions induce different levels of processing, and half of the questions are congruent with the target (the answer is *yes*) and the other half is incongruent (the answer is *no*). In the standard LOP paradigm, subjects first see a *yes/no* orienting question, followed by the target word to which they respond *yes* or *no,* and repeat the same procedure for all target words. In ten experiments, Craik and Tulving consistently demonstrated that the category task produced highest recall and recognition, whereas the case task produced lowest, with the rhyme task intermediate. Using the LOP paradigm, we examined whether making JOLs affected the robust LOP effect. The LOP paradigm is similar to single-word learning paradigm; however, the latter does not include orienting tasks to induce different levels of processing. Furthermore, the LOP paradigm is quite different than a paired-associate learning paradigm since target words are presented alone and subjects answer *yes/no* orienting questions, yet it still encompasses varying levels of relatedness between orienting questions and target words. In this case, the congruent category questions are most semantically related to targets (e.g., "chair" is a type of "furniture"), whereas case questions and incongruent questions are less related (e.g., "chair" is not a type of "fruit").

Although targets were not presented in pairs in the LOP paradigm, one possible outcome would be positive reactivity effects for congruent category questions similar to those obtained in paired-associative learning, because previous findings suggested that JOLs enhanced learning of related items (Janes et al., 2018; Soderstrom et al., 2015), and these orienting questions are related to targets semantically. In line with this reasoning, we would predict no reactivity effects for shallow case tasks because there are only two question types (lowercase or uppercase). The same prediction would be made for incongruent category and rhyme questions, because the targets are unrelated with these orienting questions. If these results were observed, we would expect JOLs to boost the LOP effect, especially for congruent questions. If we obtained negative reactivity for case questions as suggested by Mitchum et al. (2016), that would also augment the LOP effect. Another more speculative outcome, briefly discussed by Janes et al. (2018) and Dougherty et al. (2005), is that making JOLs might force people to process stimuli more attentively (for a review, see Double & Birney, 2019). If this is true, case and rhyme tasks might benefit more from this attention since they involve shallow processing. Then we would expect JOLs to attenuate (or possibly even eliminate) the LOP effect observed in previous studies. A third possibility is observing no reactivity effects in the LOP paradigm.

In all experiments, we also examined whether subjects predicted the LOP effect. Such awareness has been debated in the LOP literature. Shaw and Craik (1989) examined subjects' awareness of the LOP effect in a cue-target paradigm

(e.g., letter cues, "starts with ic: ice"; rhyme cues, "rhymes with dice: ice"; category cues, "something slippery: ice"). Their results revealed the usual LOP effect on cued recall, but subjects overestimated their performance for letter cues and underestimated their performance for category cues. Although category and rhyme cues had slightly higher JOLs than letter cues, Shaw and Craik concluded that subjects were "largely insensitive to the recall differences associated with different types of processing" (p. 134). There is, however, another possible explanation for this finding. It is possible that subjects were, in fact, sensitive to JOL instructions and they believed that, given the specific cue, they would be able to recall the target word. That is, the presence of a plausible cue in all conditions (e.g., starts with ic:) might have overshadowed their awareness of the LOP effect. Thus, we aimed to replicate Shaw and Craik's findings using the standard LOP procedure. Of course, if JOLs are reactive and they change the LOP effect, answering the question of whether subjects are aware of the LOP effect becomes more difficult, because the typical effect may not be obtained.

The primary goal of the present study is to examine the generalizability of JOL reactivity in a LOP paradigm. Thus, in Experiment 1, we investigated whether making JOLs affected recognition in the standard LOP paradigm. In Experiment 2, we aimed to replicate our findings from Experiment 1 using a wider JOL rating scale. In Experiment 3, we employed orienting tasks without *yes/no* questions to assess JOL reactivity. As a secondary goal, we examined people's awareness of the LOP effect in all three experiments.

Experiment 1

In Experiment 1, we asked the following questions: (1) Do JOLs affect recognition in the standard Craik-Tulving LOP paradigm? (2) Do JOLs augment or attenuate the LOP effect? and (3) Do people have insight into their learning processes under various orienting tasks? All subjects studied the material under one of three between-subject conditions: (1) the standard LOP paradigm; (2) the standard LOP paradigm with item-by-item JOLs; and (3) the standard LOP paradigm with a delay after each study trial (to equate the time difference created by JOLs).

Method

Subjects

A priori power analysis was conducted to determine a sufficient sample size using an α of .05, and a power of 0.80. For a medium effect size (f = 0.25) for a between factors main effect in a repeated measures analysis of variance (ANOVA), the minimum required sample size was 108. In total, we recruited 143 subjects (M_{age} = 35.5 years, SD_{age} = 9.51 years) from Amazon's Mechanical Turk (MTurk) because we expected some attrition. Subjects were located in the United States and had been respondents in a high number of studies (above 500) and with high approval rates (above 90%). They were paid $7. Subjects were randomly assigned to one of three conditions. Six subjects were replaced either due to their primary language not being English or experimental problems (e.g., losing Internet connection), and 11 were replaced as outliers.[1] Out of 126 remaining subjects, 42 subjects were in each condition.

Materials

Sixty nouns were selected from Van Overschelde et al. (2004) and Nelson et al. (2004). The words had a concreteness level above 2.5 out of 5 (Brysbaert et al., 2014). Their logarithm of HAL frequency in the English Lexicon Project ranged from 6.75 to 11.79 (Balota et al., 2007). Six questions and a rhyming noun were generated for each word, pertaining to the word's physical appearance, its sound and its meaning. Thus, each word had two questions from three types of questions (case, rhyme, category), one with a *yes* response (congruent) and one with a *no* response (incongruent). The incongruent rhyme and category questions were chosen randomly from the rhyme and category question pool, respectively. One of the six questions served as the orienting task during the study phase for each word for each subject. The materials are deposited in https://osf.io/g7jud/.

Design

A 3 × 2 × 3 mixed factorial design was used, such that orienting task (case, rhyme, category) and congruency (congruent, incongruent) were manipulated within subjects. The study phase was manipulated between subjects: One group simply performed the standard LOP paradigm (Standard), one performed the standard LOP paradigm

[1] In a preliminary analysis, correct *yes/no* responses during study, and average reaction times for recognition decisions, confidence judgments and judgments of learning (for the JOL condition) were calculated for each subject. Outliers were detected based on 3 standard deviations (*SD*) above and below the sample average for reaction times and 3 *SD* below the sample average for correct *yes/no* responses. Number of correct *yes/no* responses of five subjects were lower than the sample average for following reasons: not responding to the orienting question in 3 s and/or incorrectly responding to them. Given that this is the main LOP manipulation of the study; their data were not used. The remaining six subjects were detected as outliers based on either their recognition decisions' or confidence judgments' reaction times. Examining each subject's trials individually revealed that some of these responses had reactions times as long as 5 min. Given that it is not possible to control for what subjects were doing during that period, these subjects were replaced.

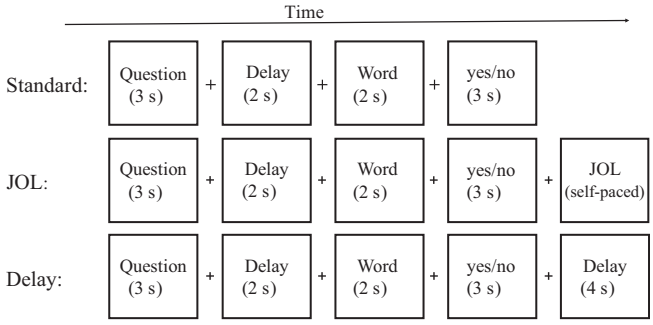

Figure 1. Study trials for each condition in Experiments 1 and 2. The Delay condition was excluded from Experiment 2.

with JOLs (JOL), and the third group had a delay after each trial (Delay). The same 60 words were used for all three groups, with 10 words comprising each of the 6 within-subject conditions (Orienting Task × *yes/no*). Each target word served in each of the 6 within-subject conditions counterbalanced across subjects. That is, for seven participants a target word served in the case-*yes* condition, for another seven participants, the same target word served in the case-*no* condition and so on.

Procedure

All subjects received intentional learning instructions. That is, they were informed that there would be a recognition test at the end, and were asked to learn the words for the test. Figure 1 shows the procedure of study trials for each condition. The initial part of each study trial was identical for all three conditions: The orienting question was presented for 3 s, and a delay of 2 s followed. After the delay, the target word was presented for 2 s. Subjects then had 3 s to make a *yes/no* response, thus totaling to 10 s per study trial. This part of the study trial constituted a standard LOP experiment, and the target words and their orienting questions were presented in a random order during the study phase. Subjects in the JOL condition provided a self-paced JOL for each word after each *yes/no* response by rating how likely it is that they would recognize the word they just studied on a scale of 1 (*Would definitely not recognize*) to 5 (*Would definitely recognize*). The target words were not presented during JOLs. In the Delay condition, subjects were given a 4 s unfilled delay after making a *yes/no* response to control for a possible time confound in the JOL condition. In a pilot study, subjects spent approximately 4 s making JOLs, and thus the time frame for the Delay condition was determined to be 4 s. Craik and Tulving (1975) showed that the presentation time of a word in the standard LOP paradigm did not affect retention, but the Delay condition served as another check. Therefore, we assumed that any performance differences between the JOL and the Standard and Delay conditions would arise from making JOLs.

After the study phase, subjects completed an approximately 10-minute president recognition test in which they identified the US presidents (Roediger & DeSoto, 2016). Then, subjects took an *old/new* recognition test consisting of 60 old and 120 new items (180 in total). After each recognition decision, they made a confidence rating on a 5-point scale.

Results

We first consider the recognition results of all three groups, then analyze JOL ratings separately. The upper part of Table 1 provides overall hit and false alarm rates, corrected recognition scores (hits – false alarms) and d' scores[2] for each group. Corrected recognition scores[3] were the primary dependent variable. Analyses using d' lead to the same conclusions. For all the pairwise comparisons, Sidak comparison was used unless indicated otherwise. The results did not change when the Bonferroni correction was used. The results for confidence ratings are in https://osf.io/g7jud/ because they were not of primary interest.

[2] One of the subjects had a perfect hit rate of 1.00 and another subject had a perfect false alarm rate of 0.00. To calculate *d'* scores for these subjects, corrected hit rates and false alarm rates were calculated. For perfect hit rates, half a hit was subtracted from the total number of hits to calculate corrected hit rate (59.5/60), and for perfect false alarm rates, half a false alarm was added to perfect false alarm to calculate corrected false alarm rate (0.5/120) (Macmillan & Creelman, 2005).

[3] To conduct the statistical analyses, six different hit rates (case/*yes,* case/*no,* rhyme/*yes,* rhyme/*no,* category/*yes,* category/*no*) were calculated for each subject. Corrected recognition scores were used as the dependent variable instead of *d'* scores because 64 subjects had perfect hit rates of 1.00 in at least one of the Congruency × Orienting Task combinations. Calculating *d'* scores for perfect hit rates required correction of perfect hit rates described above (in this case 9.5/10) and this artificially lowered the performance for those subjects who had more perfect hit rates.

Table 1. Overall hit, false alarm rates, corrected recognition scores, and *d*-primes across study conditions

	Hits		False alarms		Hits – False alarms		*d*′	
Study condition	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
				Experiment 1				
Standard	0.69	0.17	0.29	0.19	0.41	0.23	1.31	0.92
Delay	0.69	0.17	0.31	0.15	0.38	0.19	1.13	0.66
JOL	0.77	0.16	0.23	0.15	0.54	0.27	1.75	1.09
				Experiment 2				
Standard	0.67	0.22	0.19	0.13	0.48	0.19	1.57	0.66
JOL	0.81	0.14	0.09	0.08	0.72	0.14	2.49	0.67
				Experiment 3				
No-JOL	0.63	0.16	0.45	0.15	0.18	0.15	0.49	0.40
JOL	0.69	0.13	0.41	0.15	0.28	0.19	0.78	0.57

Corrected Recognition Scores

Figure 2A shows that corrected recognition scores increased across orienting tasks, with the JOL condition having the highest scores. A 2 (congruency) × 3 (orienting task) × 3 (study condition) repeated measures ANOVA confirmed a main effect of the study condition, $F(2, 123) = 5.87$, $p = .004$, $\eta^2_p = .09$. Overall, subjects in the JOL condition had higher recognition scores than subjects in the Delay and Standard conditions, ($p = .005$, $p = .031$, respectively). A main effect of orienting tasks revealed that, collapsed across all groups, the LOP effect was obtained, $F(1.89, 232.63) = 91.76$, $p < .001$, $\eta^2_p = .43$. The case task yielded lowest performance ($M = 0.34$, $SE = 0.02$), the category task yielded highest performance ($M = 0.54$, $SE = 0.02$), and the rhyme task was intermediate ($M = 0.45$, $SE = 0.02$). Although not shown in Figure 1, congruent words ($M = 0.48$, $SE = 0.02$) resulted in higher corrected recognition scores than incongruent words ($M = 0.41$, $SE = 0.02$), a recurring finding in the LOP literature, $F(1, 123) = 45.39$, $p < .001$, $\eta^2_p = .27$. The Orienting Tasks × Congruency interaction was also reliable, $F(2, 246) = 8.12$, $p < .001$, $\eta^2_p = .06$, showing similar recognition scores for congruent ($M = 0.34$, $SE = 0.03$) and incongruent words ($M = 0.33$, $SE = 0.03$) in the case task, $p = .327$. In the other two tasks, congruent words had higher recognition scores than incongruent words (with differences of .08 and .10 for the rhyme and category tasks, respectively, $ps < .001$).

The Orienting Tasks × Study condition interaction did not reach conventional levels of statistical significance, $F(3.78, 232.63) = 2.15$, $p = .079$, $\eta^2_p = .03$. Planned pairwise comparisons were still conducted to examine the nature of reactivity. For the case task, the JOL condition yielded significantly higher recognition scores than the Standard condition, $p = .011$, and the Delay condition, $p = .001$. For the rhyme and category tasks, the JOL condition had significantly higher recognition scores than the Delay condition, $ps < .05$. The Standard and Delay conditions did not differ from one another on any tasks ($ps > .05$), meaning that the extra 4 s at the end of each trial did not improve performance for the Delay condition. Tukey's *HSD* tests further revealed that the difference scores between the case and category tasks were greater for the Delay condition compared to the JOL condition, $p = .045$. Although these differences did not reach conventional levels of significance for the Standard condition ($ps > .05$), they were in the same direction. Thus, making JOLs during study enhanced performance especially for the case task, and attenuated

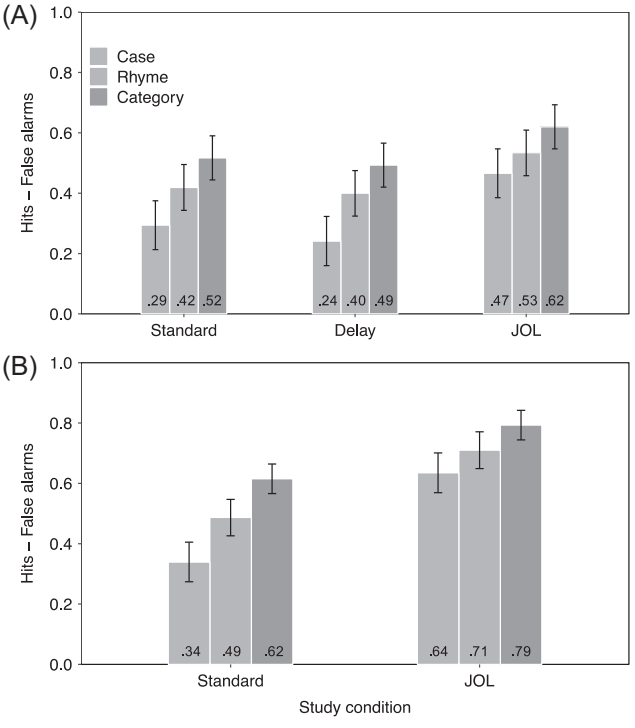

Figure 2. Corrected recognition scores across orienting tasks for each study condition in Experiment 1 (A) and Experiment 2 (B). Error bars indicate 95% confidence intervals.

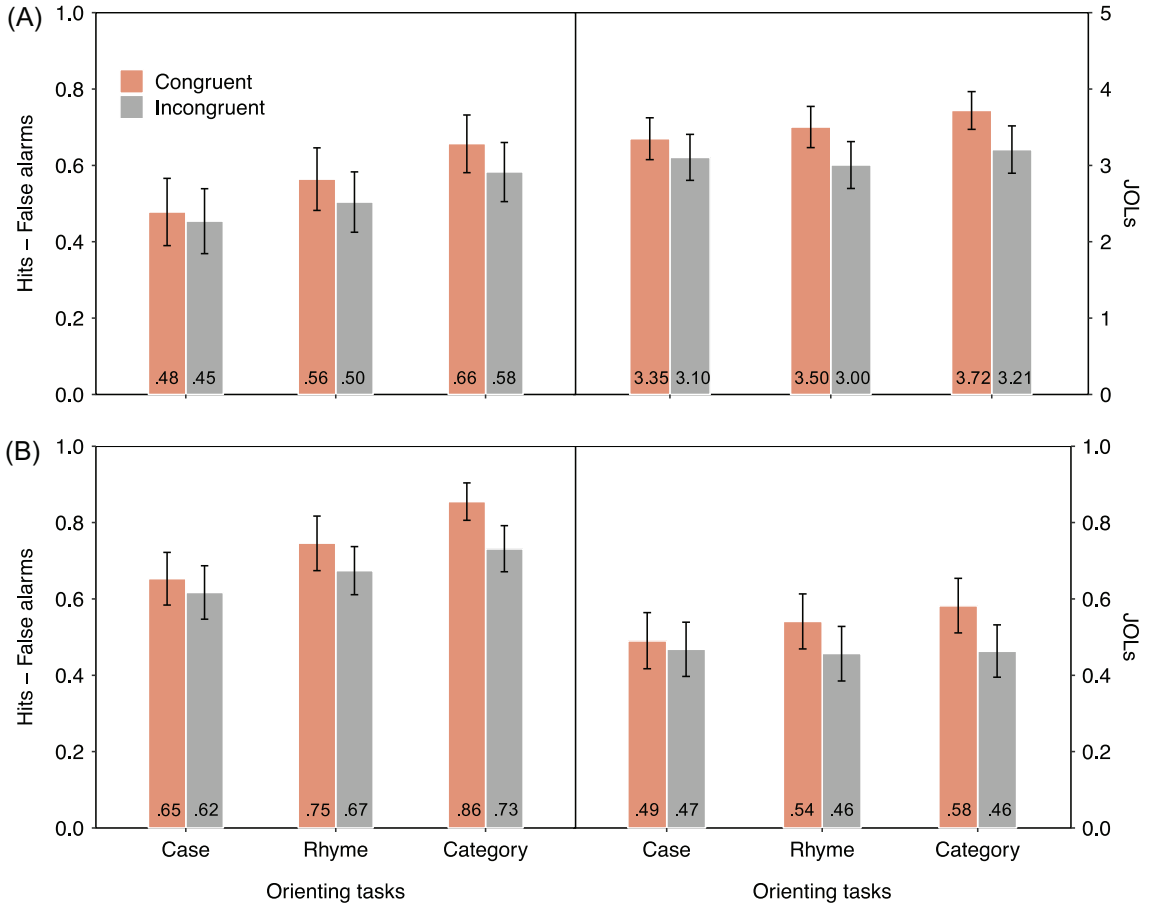

Figure 3. Corrected recognition scores across orienting tasks and congruent/incongruent words (left panel) and judgments of learning across orienting tasks and congruent/incongruent words for the JOL condition (right panel) for Experiment 1 (A) and for Experiment 2 (B). Error bars indicate 95% confidence intervals.

recognition differences between some orienting tasks, at least relative to the Delay condition.

The standard LOP effect was observed within Standard and Delay conditions: Recognition scores for each orienting task differed from one another significantly, *p*s < .001 (Figure 2). In the JOL condition, the category task had higher recognition scores compared to the rhyme task, $p = .001$, and the case task, $p < .001$, whereas the case and the rhyme tasks differed marginally, $p = .051$, with means of 0.47 for the case and 0.53 for the rhyme tasks. Other interactions were not reliable (*p*s > .05).

Judgments of Learning

Figure 3A shows corrected recognition scores (left panel) and corresponding JOL ratings (right panel) for the JOL condition. On average, subjects spent 2.99 s making JOLs. For JOLs, a 2 (congruency) × 3 (orienting tasks) repeated measures ANOVA showed a main effect of orienting tasks, $F(2, 82) = 10.43$, $p < .001$, $\eta^2_p = .20$. The category task ($M = 3.46$, $SE = 0.13$) led to higher JOL ratings compared to the case ($M = 3.23$, $SE = 0.14$) and rhyme tasks ($M = 3.25$, $SE = 0.14$), *p*s = .001. However, unlike corrected recognition scores, JOL ratings in the case task did not differ from JOL ratings in the rhyme task, $p = .951$; the corrected recognition score in the case task was lower than the recognition score in the rhyme task, although the difference was marginal. These results indicated that subjects understood the LOP effect slightly (by giving higher JOL ratings for the category task), but they did not accurately predict the LOP effect.

A main effect of congruency revealed that JOL ratings for congruent/incongruent words followed a similar pattern as in corrected recognition scores, $F(1, 41) = 52.58$, $p < .001$, $\eta^2_p = .56$. Congruent words ($M = 3.52$, $SE = 0.12$) led to higher JOL ratings than incongruent words ($M = 3.11$, $SE = 0.15$). The interaction was also reliable, $F(1.64, 67.40) = 3.76$, $p = .036$, $\eta^2_p = .08$. For congruent words, the category task led to higher JOL ratings than the case and the rhyme tasks, *p*s < .05, yet the case and rhyme tasks did not differ, $p = .204$. For incongruent words, JOL ratings at the category task were only significantly higher than JOL ratings in the rhyme task, $p = .008$. Other

pairwise comparisons were not significant, *p*s > .05. The results indicated that the effect in JOLs was carried mainly by congruent words; JOLs for subjects' incongruent words did not follow the pattern present in recognition. Thus, subjects showed limited awareness of the full LOP effect.

Discussion

Experiment 1 demonstrated JOL reactivity using a LOP paradigm. Subjects who made JOLs had higher recognition scores than subjects who did not make JOLs (the Standard and Delay conditions). Providing JOLs especially enhanced recognition scores for the case task, and in comparison to the Delay condition, reactivity was observed for all tasks. Although the act of monitoring diminished the LOP effect, it did not eliminate it.

These results suggested that relatedness with the orienting question cannot fully drive the reactivity effects in the standard LOP paradigm, nor can the changed-goal hypothesis (i.e., allocating resources toward easier items). If these were true, the LOP effect should have been augmented through either positive reactivity for the category task or negative reactivity for the shallow task. Rather, these results hinted that especially for the shallow orienting tasks, making JOLs promoted more effective processing to some extent.

Lastly, subjects in the JOL condition were somewhat (but not fully) aware of the retention differences across orienting tasks. More specifically, this effect was only present for congruent words. For incongruent words, predictions of the case and the category tasks did not differ from one another. Interestingly, subjects predicted that congruent words would lead to better retention than incongruent words, showing awareness of the congruency effect.

Experiment 2

In Experiment 2 we aimed to replicate our findings, given the importance of replication, and because (to our knowledge) Experiment 1 provided the first demonstration of reactivity effects in a LOP paradigm. In Experiment 2, instead of a 5-point scale we employed a 100-point scale in the JOL condition to make our procedure more comparable to previous studies (Mitchum et al., 2016; Soderstrom et al., 2015). The change of scale in JOLs may matter, but in making confidence ratings after recognition decision, scale changes (e.g., from a 5-point scale to a 100-point scale) had no effect (Tekin & Roediger, 2017). Because the results from the Delay condition were similar to the Standard condition, we did not include the Delay condition in Experiment 2.

Method

Subjects

Ninety-five subjects (M_{age} = 36.86, SD_{age} = 11.92) were recruited from MTurk. Participation criteria and compensation were same as Experiment 1. Eleven subjects were excluded because they were identified as outliers in one or more of the outlier criteria explained before. Remaining 84 subjects were randomly assigned to one of the two conditions, with 42 subjects in each condition.

Materials and Design

The same materials were used in Experiment 2 as in Experiment 1. Experiment 2 employed a 3 × 2 × 2 mixed factorial design. The orienting tasks (case, rhyme, category) and congruency (congruent, incongruent) were manipulated within subjects, and the study phase was manipulated between subjects (Standard, JOLs).

Procedure

The procedure was identical to that of Experiment 1. The only difference was the scale used by the JOL condition. In Experiment 2, instead of a 5-point scale, subjects in the JOL condition made JOLs on a 100-point scale.

Results

The middle part of Table 1 provides overall hit and false alarm rates, corrected recognition scores and *d'* scores[4] for each group. Corrected recognition scores were again used as the dependent variable.

Corrected Recognition Scores

Figure 2B shows that corrected recognition scores increased across three orienting tasks, demonstrating the LOP effect. A 2 (congruency) × 3 (orienting task) × 2 (study condition) repeated measures ANOVA confirmed a main effect of the study condition, $F(1, 82) = 6.80$, $p < .001$, $\eta^2_p = .34$. Overall, subjects in the JOL condition had higher recognition scores than subjects in the Standard condition (positive reactivity). A main effect of orienting tasks confirmed the LOP effect, $F(1.84, 151.10) = 70.03$, $p < .001$, $\eta^2_p = .46$. The congruency effect was replicated, $F(1, 82) = 39.46$, $p < .001$, $\eta^2_p = .33$. Furthermore, the Orienting Tasks × Congruency interaction was reliable, $F(2, 164) =$, $p < .001$, $\eta^2_p = .11$, replicating

[4] Two subjects had perfect hit rates of 1.00 and another two subjects had false alarm rates of 0.00. To be able to calculate *d'* scores for these subjects, corrected hit rates and false alarm rates were calculated (Macmillan & Creelman, 2005).

the results of Experiment 1 (and most studies in the LOP literature). Recognition scores did not differ between congruent and incongruent words for the case task, $p = .185$, but they did for the rhyme and category tasks (differences of .07 and .13, respectively, ps < .01).

The Orienting Tasks × Study condition interaction was also reliable, $F(1.84, 151.10) = 5.24$, $p = .008$, $\eta^2_p = .06$. For all orienting tasks, the JOL condition had significantly higher recognition scores than the Standard condition, ps < .001. Tukey's *HSD* tests further revealed that the difference scores between the case and category tasks were higher for the Standard condition compared to the JOL condition, $p = .004$, indicating that making JOLs attenuated the LOP effect between these two tasks. The Standard and JOL conditions both demonstrated the LOP effect, ps < .001. Other interactions were not reliable (ps > .05). These results replicated Experiment 1.

Judgments of Learning

Figure 3B shows corrected recognition scores (left panel) and corresponding JOL ratings (right panel) for the JOL condition. JOL ratings made on a scale of 0–100 were transferred to proportions (i.e., divided by 100) to match the scale of corrected recognition. JOL ratings illustrate a slight increase across orienting tasks, mostly for congruent words. On average, subjects spent 3.60 s making JOLs. A 2 (congruency) × 3 (orienting tasks) repeated measures ANOVA on JOLs showed a main effect of orienting tasks, $F(2, 82) = 7.91$, $p = .001$, $\eta^2_p = .16$. The category task ($M = 52.21$, $SE = 3.33$) led to higher JOL ratings compared to the case ($M = 47.81$, $SE = 3.52$) and rhyme tasks ($M = 49.80$, $SE = 3.44$), ps < .05. However, unlike corrected recognition scores, the case task did not differ significantly from the rhyme task, $p = .224$.

A main effect of congruency revealed that JOL ratings followed a similar pattern as that in corrected recognition scores, $F(1, 41) = 39.62$, $p < .001$, $\eta^2_p = .49$. Congruent words ($M = 53.71$, $SE = 3.44$) led to higher JOL ratings than incongruent words ($M = 46.17$, $SE = 3.41$). The interaction was also reliable, $F(1.65, 67.60) = 10.47$, $p > .001$, $\eta^2_p = .20$. For congruent words, the category task yielded higher JOL ratings than the case and the rhyme tasks, ps < .05, and the rhyme task produced higher JOL ratings than the case task, $p = .019$. For incongruent words, however, JOL ratings did not differ across any tasks, ps > .05, indicating that the main effect of orienting task was driven by the congruent words. Subjects were unaware of the LOP effect when they answered *no* to the orienting question. Furthermore, as with the corrected recognition scores, JOLs for congruent and incongruent words did not differ significantly in the case task, $p = .110$, but differed in the rhyme and category tasks, ps < .001. These results indicate that JOLs are sensitive to the congruency effect, but not as sensitive to the LOP effect.

Discussion

Experiment 2 replicated the results from Experiment 1. JOLs enhanced retention across all three orienting tasks, especially for the case task, and attenuated the LOP effect compared to not making JOLs. Nonetheless, JOLs did not eliminate the LOP effect. Furthermore, as with Experiment 1, subjects revealed mild awareness of the LOP effect: JOL ratings were higher in the category task, but did not differ between the case and rhyme tasks. In Experiment 2, this effect was solely driven by the congruent words; JOL ratings for incongruent words did not differ across orienting tasks. Unlike the LOP effect, subjects again predicted the congruency effect.

Experiment 3

The results of Experiment 1 and 2 consistently revealed that JOLs were reactive measures in the standard LOP paradigm. However, the relatedness between target words and orienting questions might have still contributed to observed reactive effects. That is, we still observed reactive effects for target-specific category and rhyme tasks which might be in part driven by relatedness between orienting questions and target words. In Experiment 3, we aimed to eliminate this possibility. Instead of using *yes/no* questions, we employed a pleasantness rating task for deep processing and an e-counting task for shallow processing (Hyde & Jenkins, 1969). Furthermore, in Experiments 1 and 2, subjects were aware of the congruency effect but were not fully aware of the LOP effect. However, it is possible that subjects were not able to incorporate both congruency and LOP effects into their JOLs. The congruency manipulation in Experiments 1 and 2 might have overshadowed subjects' awareness of the LOP manipulation. Therefore, Experiment 3 also examined whether subjects can predict the LOP effect without the congruency manipulation.

In Experiment 3, subjects studied a list of words under pleasantness rating and e-counting tasks: (1) without making JOLs; or (2) while making JOLs. A pilot study showed that 12 out of 36 subjects performed at ceiling with a hit rate higher than .95 on an immediate recognition test. Because Witherby and Tauber (2017) previously demonstrated that reactive effects of JOLs were still observed after two days, the recognition test in Experiment 3 occurred 2 days after the study phase to avoid ceiling effects.

Method

Subjects

One hundred eleven subjects were recruited from MTurk with the same participation and compensation criteria as

in Experiments 1 and 2. Ninety-seven of them completed both sessions. Ten out of 97 subjects them did not follow instructions for Session 2 and five were identified as outliers based on previously discussed criteria. The remaining 82 subjects were randomly assigned to one of the two conditions, with 40 subjects in the no-JOL condition and 42 in the JOL condition.

Materials

Eighty concrete nouns were selected from norms collected by Nelson et al. (2004). The words were 5–8 letters long, and had either 0, 1, or 2 letter "e"s. Their logarithm of HAL frequency ranged from 4.34 to 13.67 (Balota et al., 2007). The target words are deposited in https://osf.io/g7jud/.

Design

Experiment 3 employed a 2 × 2 × 2 mixed factorial design. The orienting tasks (e-counting, pleasantness rating) were manipulated within subjects, and the study condition was manipulated between subjects (No-JOL, JOL). Lastly, the tasks were presented in two blocks. That is, subjects completed one of the tasks (e.g., pleasantness) in the first block and the other task (e.g., e-counting) in the second block. The order of the tasks (pleasantness first, e-counting first) were counterbalanced across subjects.

Procedure

Subjects were randomly assigned to a counterbalancing condition for the task order and received intentional learning instructions. For both the pleasantness task and the e-counting task, subjects were first presented with instructions regarding the task. Then the target word was presented for 2 s at the beginning of each study trial. For the pleasantness task, subjects rated pleasantness of each word on a scale from 0 to 4, and for the e-counting task, they counted the letter "e"s in the word and chose a number from 0 to 4. For both tasks, subjects had 5 s. For the initial block, subjects repeated the same assigned task for 40 target words (i.e., 40 study trials). Then subjects switched to the other task for another 40 target words. Half of the subjects made JOL ratings on a 100-point scale at the end of each study trial. Two days after the initial study phase, subjects took an *old/new* recognition test that consisted of 80 old items and 80 new items.

Results

The lower part of Table 1 provides overall hit rates, false alarm rates, corrected recognition scores and d' scores. As with previous experiments, corrected recognition scores were used as the dependent variable.

Table 2. Corrected recognition scores and JOL ratings across study and counterbalance conditions, and orienting tasks for Experiment 3

	E-counting		Pleasantness	
Corrected recognition scores	*M*	*SE*	*M*	*SE*
Study condition				
No-JOL	0.13	0.03	0.23	0.03
JOL	0.26	0.03	0.30	0.03
Counterbalance				
E-counting first	0.23	0.03	0.27	0.03
Pleasantness first	0.17	0.03	0.27	0.03
JOL ratings				
JOL	55.96	3.17	57.60	2.76

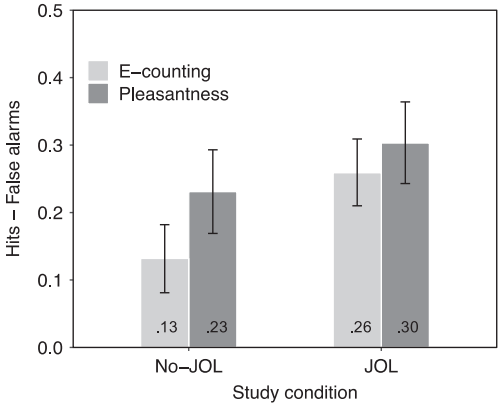

Figure 4. Corrected recognition scores orienting tasks for each study condition in Experiment 3. Error bars indicate 95% confidence intervals.

Corrected Recognition Scores

The upper part of Table 2 shows the corrected recognition scores across study and counterbalance conditions and orienting tasks. Figure 4 shows that corrected recognition scores were higher for the JOL condition in both orienting tasks. For corrected recognition scores, a 2 (study condition) × 2 (orienting tasks) × 2 (counterbalance) repeated measures ANOVA confirmed a main effect of study condition, $F(1, 78) = 7.06$, $p = .010$, $\eta^2_p = .08$, indicating the JOL condition showed reactive effects. The main effect of orienting tasks revealed the LOP effect, $F(1, 78) = 30.81$, $p < .001$, $\eta^2_p = .28$, and there was no main effect of counterbalancing, $F < 1$, *ns*. The Orienting Tasks × Study condition interaction was reliable, $F(1, 78) = 4.63$, $p = .035$, $\eta^2_p = .06$. Interestingly, the Orienting Tasks × Counterbalance interaction was also reliable, $F(1, 78) = 4.66$, $p = .034$, $\eta^2_p = .06$, with the three-way interaction reaching marginal significance, $F(1, 78) = 3.68$, $p = .059$, $\eta^2_p = .05$. The Study × Counterbalance interaction did not reach significance, $F < 1$, *ns*.

To understand the nature of the two-way interactions, pairwise comparisons were conducted. For both JOL and

No-JOL conditions, there was a LOP effect, ps < .05, though the effect was much smaller for the JOL condition. Subjects in the JOL condition recognized more words after the e-counting task than did subjects in the No-JOL condition, p = .001, though the two groups' recognition scores on the pleasantness task did not differ statistically, p = .102. Furthermore, for both counterbalancing conditions, there was a LOP effect, ps < .05, though subjects who completed the e-counting task first showed smaller difference between the orienting tasks (i.e., a smaller LOP effect) compared to subjects who completed the pleasantness task first. We did not perform pairwise comparisons on the marginal three-way interaction because we did not have any a priori predictions.

Judgments of Learning

The lower part of Table 2 shows JOL ratings across tasks for the JOL condition. On average, subjects spent 3.08 s making JOLs. A 2 (orienting tasks) × 2 (counterbalance) repeated measures ANOVA did not reveal any main effects of orienting tasks or counterbalance, Fs < 1, *ns*, indicating that subjects were not aware of the LOP effect. The interaction was not reliable either, F < 1, *ns*.

Discussion

Experiment 3 revealed JOL reactivity in a LOP paradigm after eliminating relatedness between target words and orienting questions as a possible reason for reactivity. Making JOLs improved recognition especially for the shallow tasks and attenuated the LOP effect. In Experiment 3, JOLs statistically only benefited the e-counting task; however, even for the pleasantness task, the effect was in the same direction (.07 corrected recognition score difference). These findings suggest that making JOLs might promote more effective processing, and thus enhance retention. Surprisingly, we found a reliable interaction between orienting tasks and counterbalance conditions. Given that we observed reactivity effects for both counterbalancing conditions with JOLs, and the two counterbalancing conditions yielded similar recognition overall, we do not further discuss this interaction. Lastly, we demonstrated that subjects were not aware of the LOP effect even after the congruency manipulation was removed. Given that we observed a smaller LOP effect in the JOL condition (.05 corrected recognition score difference), it is not surprising that subjects were not aware of the LOP effect. There was not much of an effect to be aware of.

General Discussion

The primary purpose of the current study was to investigate reactive effects of JOLs in a LOP paradigm and, more specifically, to see if making JOLs influenced the LOP effect. All experiments demonstrated JOL reactivity. Of course, we also replicated the LOP effect, but it was smaller for conditions in which subjects made JOLs. The secondary goal of the experiments was to examine subjects' awareness of the LOP effect. We first review the results for JOL reactivity and then consider additional findings of the LOP effect.

Judgments of Learning as Memory Modifiers

Recent research has demonstrated that JOLs are often reactive measures that influence learning in paired-associate paradigms (Janes et al., 2018; Mitchum et al., 2016; Soderstrom et al., 2015). We examined reactivity effects of JOLs in a LOP paradigm to test the generalizability of these findings. In Experiments 1 and 2, JOLs enhanced retention compared to not making JOLs in a standard LOP paradigm (Craik & Tulving, 1975), and in Experiment 3, we replicated this finding in a different LOP paradigm. In Experiment 2, subjects in the JOL condition recognized more words in all the LOP tasks than subjects in the Standard condition, whereas in Experiments 1 and 3, this was true only for shallow LOP tasks. Nonetheless, the effects were numerically in the same direction for category tasks. Furthermore, in all experiments, making JOLs benefited shallow tasks more than deep tasks, attenuating the LOP effect. Thus, our findings of reactivity were in line with recent studies that demonstrated JOL reactivity.

How do JOLs influence future performance? Mitchum et al. (2016) proposed the changed-goal hypothesis to explain JOL reactivity in related and unrelated pairs. They hypothesized that making JOLs prompted subjects to change their learning goal from mastering all items to the more pragmatic goal of just mastering easy items. Thus, the difference in recall increases due to focus on related pairs, whereas recall may actually become worse for unrelated pairs due to their neglect (negative reactivity). They found that, relative to subjects in the no-JOL condition, subjects in the JOL condition spent less time studying difficult unrelated items and performed worse on those items. Therefore, Mitchum et al. concluded that subjects became aware that they could not master difficult items by providing JOLs, and thus switched their resources away from difficult items to easy ones. Our results were inconsistent with the account proposed by Mitchum et al (2016). If people had changed their goals to pragmatic ones, then making JOLs should have impaired retention on case tasks. However, our results undercut this explanation, because (1) JOLs enhanced rather than harmed recognition on shallow tasks, and (2) subjects did not show

accurate awareness of the LOP effect to adjust their learning goal.

Soderstrom et al. (2015), on the other hand, proposed that JOLs enhance retention for the information used as a basis of JOLs and that reactive effects of JOLs would be observed for judged items when the final test is sensitive to such information. They showed that JOLs bolstered cued recall of the easy related pairs, but had no impact on the difficult unrelated pairs. Given that relatedness was a basis for JOLs only for related pairs, and the cued recall test was sensitive to relatedness between words, making JOLs only emphasized the already existing cue-target relationship between related pairs (positive reactivity). We also observed positive reactivity for JOLs in all orienting tasks, but especially in shallow tasks. Our results might broaden Soderstrom et al.'s account by indicating that making JOLs improve retention especially if JOLs strengthen information that is not strengthened otherwise. That is, in shallow tasks that do not readily promote semantic processing, making JOLs might have strengthened semantic information and because the recognition test was sensitive to such information, larger reactive effects were obtained. In category tasks that already promote semantic processing, however, the reactive effects were not as large.

Why did our results occur? We can only speculate. One idea is that the act of making JOLs forces subjects to think about the inherent memorability of the word, perhaps more so than the type of processing just performed. That is, after making a case judgment (lowercase or uppercase) on targets (Experiments 1 and 2), subjects were asked to rate how likely it was that they would recognize the target later. Performing this judgment probably makes subjects think about the meaning of the word, or to perform at least some semantic/elaborative processing. Such elaboration might have enhanced retention following the shallow tasks, because shallow tasks do not benefit from semantic processing in the standard LOP paradigm. Furthermore, because retention is typically lower for the shallow tasks, they have more room to improve. However, elaboration did not enhance retention as much for items processed in the category or pleasantness tasks, because they have already benefited from semantic processing. Dougherty et al. (2005) also suggested that "One interesting hypothesis is that making a metacognitive judgment forces participants to process the to-be-remembered item more thoroughly than they would if no judgment was made. Thus, the act of making the judgment may affect how well the item is stored in memory" (p. 1110). This explanation is also in line with previous findings that reported JOL reactivity in single-word learning paradigms (Double et al., 2018). However, future research should examine reactive effects in different paradigms to see if this hypothesis withstands scrutiny in accounting for JOL reactivity.

Subjects' Awareness of the Lop Effect

In the present experiments, we also examined whether subjects were able to predict the LOP effect. The answer seems to be mixed, in large part because of the reactivity of JOL judgments. For the standard LOP paradigm in Experiment 1, JOLs showed some differences among the orienting tasks that corresponded to recognition scores. Thus, subjects were not oblivious to the LOP manipulation and understood that the orienting tasks would affect future performance differently, at least to some extent. Nevertheless, subjects only predicted the LOP effect for congruent words but not for incongruent words in Experiment 2, and differences in their JOL ratings did not reflect differences in recognition among all orienting tasks in both experiments. One possibility for some of the differences is that the congruency effect might have overshadowed the LOP effect. In Experiment 3, we manipulated orienting tasks without a *yes/no* decision and we observed that subjects' JOLs did not differ across tasks. These results all together suggest that subjects are somewhat, but not greatly, aware of the LOP effect, as also found in prior work (Shaw & Craik, 1989).

This relative insensitivity of subjects' JOLs to the LOP effect is in line with Koriat's (1997) cue utilization framework, in which he accounted for JOLs' sensitivity differences by distinguishing between intrinsic and extrinsic cues. Intrinsic cues are related to items' a priori characteristics (e.g., frequency) whereas extrinsic cues are characteristics of learning conditions that can affect overall performance (e.g., encoding operations). JOLs are sensitive to intrinsic cues and rarely sensitive to extrinsic cues. Since the LOP manipulation is an extrinsic cue, subjects might have been less aware of its influence on their learning. It is important to note, however, reactive effects observed in all experiments might have also affected JOLs' sensitivity to the LOP manipulation. That is, since making JOLs weakened the LOP effect by increasing recognition for shallow tasks, subjects might have not been able to accurately predict the small recognition differences. Because of JOLs' reactivity, they cannot be easily used to assess subjects' knowledge of some experimental manipulations.

Final Thoughts

The current study demonstrated that JOLs enhanced retention for shallow tasks in a LOP paradigm and diminished the LOP effect. These results extend on the previous finding that JOLs are reactive measures in paired-associate learning paradigms, and they suggest that JOL reactivity might occur in other paradigms. Therefore, researchers should be cautious when using JOLs and include a control

condition to account for possible reactive effects. Furthermore, a more systematic examination of previous findings is warranted for studies without control conditions, because JOLs might well have affected the processes they were intended to assess. Our findings also suggest that for educational purposes, simply asking students to monitor their learning during study might serve as a method to improve learning. Future research should examine how generalizable reactive effects are for more complex materials and how making JOLs affects learning relative to other study strategies that are known to boost retention. JOLs during study might prove an effective and natural way to enhance retention of studied material.

References

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

Balota, D. A., Yap, M. J., Hutchison, K. A., Cortese, M. J., Kessler, B., Loftis, B., Neely, J. H., Nelson, D. L., Simpson, G. B., & Treiman, R. (2007). The English lexicon project. *Behavior Research Methods, 39*(3), 445–459. https://doi.org/10.3758/BF03193014

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index. *Journal of Experimental Psychology: General, 127*(1), 55–68. https://doi.org/10.1037/0096-3445.127.1.55

Bjork, R. A., Dunlosky, J., & Kornell, N. (2013). Self-regulated learning: Beliefs, techniques, and illusions. *Annual Review of Psychology, 64*, 417–444. https://doi.org/10.1146/annurev-psych-113011-143823

Brysbaert, M., Warriner, A. B., & Kuperman, V. (2014). Concreteness ratings for 40 thousand generally known English word lemmas. *Behavior Research Methods, 46*(3), 904–911. https://doi.org/10.3758/s13428-013-0403-5

Craik, F. I. M., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11*(6), 671–684. https://doi.org/10.1016/S0022-5371(72)80001-X

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104*(3), 268–294. https://doi.org/10.1037/0096-3445.104.3.268

Double, K. S., & Birney, D. P. (2019). Reactivity to measures of metacognition. *Frontiers in Psychology, 10*, Article 2755. https://doi.org/10.3389/fpsyg.2019.02755

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. https://doi.org/10.1080/09658211.2017.1404111

Dougherty, M. R., Scheck, P., Nelson, T. O., & Narens, L. (2005). Using the past to predict the future. *Memory & Cognition, 33*(6), 1096–1115. https://doi.org/10.3758/BF03193216

Dunlosky, J., & Matvey, G. (2001). Empirical analysis of the intrinsic–extrinsic distinction of judgments of learning (JOLs): Effects of relatedness and serial position on JOLs. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 27*(5), 1180–1191. https://doi.org/10.1037/0278-7393.27.5.1180

Dougherty, M. R., Robey, A. M., & Buttaccio, D. (2018). Do metacognitive judgments alter memory performance beyond the benefits of retrieval practice? A comment on and replication attempt of Dougherty, Scheck, Nelson, and Narens (2005). *Memory & Cognition, 46*(4), 558–565. https://doi.org/10.3758/s13421-018-0791-y

Ericsson, K. A., & Simon, H. A. (1980). Verbal reports as data. *Psychological Review, 87*(3), 215–251. https://doi.org/10.1037/0033-295X.87.3.215

Harris, F. C., & Lahey, B. B. (1982). Subject reactivity in direct observational assessment: A review and critical analysis. *Clinical Psychology Review, 2*(4), 523–538. https://doi.org/10.1016/0272-7358(82)90028-9

Hyde, T. S., & Jenkins, J. J. (1969). Differential effects of incidental tasks on the organization of recall of a list of highly associated words. *Journal of Experimental Psychology, 82*(3), 472–481. https://doi.org/10.1037/h0028372

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

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

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition, 31*(6), 918–929. https://doi.org/10.3758/BF03196445

Koriat, A. (1997). Monitoring one's own knowledge during study: A cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370. https://doi.org/10.1037/0096-3445.126.4.349

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. https://doi.org/10.1037/0278-7393.31.2.187

Koriat, A., Bjork, R. A., Sheffer, L., & Bar, S. K. (2004). Predicting one's own forgetting: The role of experience-based and theory-based processes. *Journal of Experimental Psychology: General, 133*(4), 643–656. https://doi.org/10.1037/0096-3445.133.4.643

Koriat, A., & Ma'ayan, H. (2005). The effects of encoding fluency and retrieval fluency on judgments of learning. *Journal of Memory and Language, 52*(4), 478–492. https://doi.org/10.1016/j.jml.2005.01.001

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits – and costs – of dropping flashcards. *Memory, 16*(2), 125–136. https://doi.org/10.1080/09658210701763899

Lovelace, E. A. (1984). Metamemory: Monitoring future recallability during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(4), 756–766. https://doi.org/10.1037/0278-7393.10.4.756

Macmillan, N. A., & Creelman, C. D. (2005). *Detection theory: A user's guide*. Erlbaum.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin & Review, 15*(1), 174–179. https://doi.org/10.3758/PBR.15.1.174

Metcalfe, J., & Kornell, N. (2005). A region of proximal learning model of study time allocation. *Journal of Memory and Language, 52*(4), 463–477. https://doi.org/10.1016/j.jml.2004.12.001

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

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL effect". *Psychological Science, 2*(4), 267–271. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

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. https://doi.org/10.3758/BF03195588

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

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review, 16*(3), 550–554. https://doi.org/10.3758/PBR.16.3.550

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

Roediger, H. L., & DeSoto, K. A. (2016). Recognizing the presidents: Was Alexander Hamilton president? *Psychological Science, 27*(5), 644–650. https://doi.org/10.1177/0956797616631113

Shaughnessy, J. J. (1981). Memory monitoring accuracy and modification of rehearsal strategies. *Journal of Verbal Learning and Verbal Behavior, 20*(2), 216–230. https://doi.org/10.1016/S0022-5371(81)90389-3

Shaw, R. J., & Craik, F. I. (1989). Age differences in predictions and performance on a cued recall task. *Psychology and Aging, 4*(2), 131–135. https://doi.org/10.1037/0882-7974.4.2.131

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. https://doi.org/10.1111/j.1467-9280.1992.tb00680.x

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. https://doi.org/10.1037/a0038388

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

Tekin, E., & Roediger, H. L. (2017). The range of confidence scales does not affect the relationship between confidence and accuracy in recognition memory. *Cognitive Research: Principles and Implications, 2*(1), Article 49. https://doi.org/10.1186/s41235-017-0086-z

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the norms. *Journal of Memory and Language, 50*(3), 289–335. https://doi.org/10.1016/j.jml.2003.10.003

Vesonder, G. T., & Voss, J. F. (1985). On the ability to predict one's own responses while learning. *Journal of Memory and Language, 24*(3), 363–376. https://doi.org/10.1016/0749-596X(85)90034-8

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. https://doi.org/10.1016/j.jarmac.2017.08.004

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*, Article 1699. https://doi.org/10.3389/fpsyg.2015.01699

Zechmeister, E. B., & Shaughnessy, J. J. (1980). When you know that you know and when you think that you know but you don't. *Bulletin of the Psychonomic Society, 15*(1), 41–44. https://doi.org/10.3758/BF03329756

History
Received February 14, 2020
Revision received June 12, 2020
Accepted June 12, 2020
Published online October 30, 2020

Open Data
Data can be accessed on Open Science Framework at https://osf.io/g7jud/

Eylul Tekin
Department of Psychological & Brain Sciences
Washington University in St. Louis
Box 1125
One Brookings Drive
St Louis, MO 63130-4899
USA
elifeylultekin@wustl.edu

