Solving a Mystery in Levels-of-Processing Research: The Word-First Paradigm

EYLUL TEKIN and HENRY L. ROEDIGER, III
Washington University in St. Louis

In the standard levels-of-processing (LOP) paradigm, subjects are presented with a question directing them to attend to surface features (case), phonetic features (rhyme), or semantic features (category) of an upcoming word. Answering the question greatly affects recall and recognition for the word, and this outcome is attributed to differing LOP of the word. Craik (1977) reasoned that if the words appeared before the questions (i.e., the word-first paradigm) under intentional learning instructions, the LOP effect would disappear. However, when he conducted experiments using the word-first paradigm, the effect remained. We conducted 2 experiments to try to explain this puzzle. In the first experiment, we replicated Craik's findings in one condition, and in other conditions we asked subjects to make judgments of learning (JOLs) in an attempt to eliminate the LOP effect because JOLs seem to promote deep processing. However, the LOP effect still occurred, albeit in attenuated form. In Experiment 2, we tested 2 new conditions to overcome and explain the LOP effect in the word-first paradigm. In one case, subjects were instructed to use idiosyncratic study strategies to memorize the word; the second condition required subjects to perform pleasantness judgments on the words. The pleasantness rating condition eliminated the LOP effect, but the first did not. We conclude that in the word-first paradigm, subjects typically keep the word in mind without semantic processing using rote rehearsal, despite intentional learning instructions.

KEYWORDS: levels of processing, intentional learning, word-first paradigm, judgments of learning

The levels-of-processing (LOP) framework has been one of the most widely discussed in cognitive psychology in the last 50 years; the original Craik and Lockhart (1972) article has been cited 13,580 times as of March 12, 2021. Similarly, the Craik and Tulving (1975) paradigm that has become the canonical method for studying the LOP effect has been cited 5,489 times. Despite the huge volume of research on this topic, puzzles and mysteries about it remain (Roediger & Gallo, 2002). We describe one such puzzle in this article and attempt to solve it in two experiments.

The LOP effect is straightforward yet robust: When subjects study a set of words (e.g., *apple*) that follow questions, semantic tasks (e.g., category questions, *Is the word type of fruit?*) result in greater retention than phonemic or orthographic tasks (e.g., rhyme or case questions such as *Does the word rhyme with chapel?* or *Is the word lowercase?*). Although a few studies reported similar findings (Hyde & Jenkins, 1969; Tresselt & Mayzner, 1960) before the LOP framework was developed, the two articles by Craik, Lockhart, and Tulving propelled such studies to the forefront of memory research. Ac-

American Journal of Psychology
Summer 2021, Vol. 134, No. 2 pp. 129–141 •

cording to the LOP framework, the memory trace is a byproduct of perceptual analyses that vary from surface processes (orthographic or phonemic) to deeper (semantic) processes. Semantic processing produces greater retention and less forgetting than shallow processing. The LOP framework assumes that people spontaneously engage in semantic processing in their everyday lives to make sense of their environments. Accordingly, intentional learning instructions (i.e., learning the material for a later test) should also promote semantic processing, because knowing that they will be tested, people should engage in their natural semantic processing regardless of the presence or absence of orienting tasks (Craik & Lockhart, 1972).

As predicted by the LOP framework, intentional learning instructions without orienting tasks yield similar levels of retention to encoding with most semantic processing tasks (Hyde & Jenkins, 1969, 1973; Walsh & Jenkins, 1973). For instance, in Hyde and Jenkins's (1969) research, control groups given intentional learning instructions performed similarly to other deep processing groups who rated the pleasantness of words across three experiments. However, the LOP effect was also routinely obtained under intentional learning instructions when subjects were presented with orienting tasks (Chow, Currie, & Craik, 1978; Craik, 1973; Craik & Tulving, 1975). Given that subjects were able to process all the words meaningfully under intentional learning instructions without orienting tasks, why did they fail to do so with orienting tasks?

Importantly, in the majority of studies that found the LOP effect under intentional learning instructions, target words were always preceded by the orienting questions (or tasks). This standard procedure, referred to as "question-first" henceforward, might have hindered spontaneous deep processing of these words that may have otherwise occurred under intentional learning instructions. That is, when orienting questions precede target words, subjects might find it difficult to switch from experimentally assigned shallow processing tasks to spontaneous deep processing during the study of each word. If this is the case, allowing subjects to engage in their own natural deep processing before the experimentally provided shallow task should eliminate the LOP effect.

Craik (1977) examined just this possibility in an exploratory experiment and developed the word-first paradigm. Unlike the question-first paradigm, in which the orienting questions preceded target words, in the word-first paradigm target words preceded the orienting questions. Each study trial consisted of a target word followed by a 5-s delay and then the orienting question. The idea was that without the orienting task preceding the target word and with the long delay of 5 s, subjects would process words meaningfully, as in other conditions without orienting tasks (e.g., Hyde & Jenkins, 1969). In Craik (1977), subjects received intentional learning instructions, studied target words followed by questions asking about case, rhyme, and category, and later took a recognition test. In addition to this word-first paradigm, Craik used another condition with the question-first LOP paradigm for comparison. Although retention after the shallower questions in the word-first paradigm was slightly greater than in the question-first paradigm, the category task still yielded greater retention than case and rhyme tasks in the word-first paradigm. Thus, even when subjects had 5 s to process target words meaningfully without preceding orienting tasks, the LOP effect still occurred. This finding, described as "mystifying" by Craik (1977, p. 689), is in clear opposition to the original LOP framework's predictions and remains an unresolved conundrum in the LOP literature (Roediger & Gallo, 2002).

In two experiments, we examined why the LOP effect occurs in this neglected word-first paradigm under intentional learning instructions. In Experiment 1, we aimed to replicate the original findings of Craik (1977) and to provide conditions that should promote further spontaneous deep processing in the word-first paradigm. We also tested subjects' awareness of the LOP effect. In Experiment 2, we addressed other explanations of why the LOP effect was not eliminated in the word-first paradigm.

EXPERIMENT 1

While Craik (1977) found that the LOP effect still emerged in the word-first paradigm, he also demonstrated that it was lower than that of the standard question-first paradigm because of greater retention in the shallow orienting tasks. This outcome suggests that at least some elaboration on target words occurred in the word-first paradigm when the targets were presented before the shallow tasks. A possible reason for observing the LOP effect in the word-first paradigm is that subjects might not fully elaborate on the target words because they wait for the orienting task. That is, knowing that there will be a specific orienting task for each target word, subjects may be more likely to truncate deep processing or engage in rote rehearsal than they would under intentional learning instructions without orienting tasks. If that is the case, promoting further elaboration should eliminate the LOP effect observed in the word-first paradigm. In Experiment 1, we aimed to boost subjects' elaboration by asking them to monitor their learning through judgments of learning (JOLs) after the orienting task.

JOLs are the predictions subjects make about how well they will remember an item on a later test (Arbuckle & Cuddy, 1969; Nelson & Narens, 1990). They are also reactive measures that influence retention (Janes, Rivers, & Dunlosky, 2018; Soderstrom, Clark, Halamish, & Bjork, 2015; Tekin & Roediger, 2020). For instance, Soderstrom et al. demonstrated that making JOLs increased retention for related word pairs, whereas it did not affect unrelated word pairs, in both cases relative to conditions without JOLs. They proposed that making JOLs strengthened the already existing relationship between words in related pairs, yielding greater retention for the pairs. Because target words are related to questions half the time (for "yes" responses) in the LOP paradigm, we reasoned that having subjects perform JOLs on the targets might cause the LOP effect to be eliminated for "yes" responses (and possibly for "no" responses, too).

Another related reason for obtaining the LOP effect in the word-first paradigm with intentional learning instructions might be subjects' lack of insight into how the orienting tasks affect learning. That is, although intending to learn all items, subjects might not have realized that shallow orienting tasks would not do the job. Because of this unawareness, they might have truncated their natural deep processing and waited for the orienting question rather than using more effective learning strategies. In fact, in a cue–target paradigm, Shaw and Craik (1989) found that subjects overestimated their retention after the shallow task and underestimated it after the category task. Although subjects reported somewhat lower JOLs for the shallow task than the category task, these findings suggested that subjects might be insensitive to the LOP manipulation. In Shaw and Craik's study, however, subjects saw a credible cue for each target (e.g., "starts with *ic*" is the shallow cue for "ice"). These plausible cues in turn might have overridden subjects' awareness of the LOP effect in the cue–target paradigm. Therefore, we also examined subjects' awareness of the LOP manipulation in both the question-first and word-first paradigms with JOLs. We included the question-first paradigm to investigate subjects' awareness of the standard LOP effect.

In Experiment 1, subjects received intentional learning instructions and studied target words in one of the following conditions before taking a recognition test: the word-first paradigm (word-first), the word-first paradigm with JOLs (word-first JOL), and the question-first paradigm with JOLs (question-first JOL). We asked whether JOLs increase retention in the word-first paradigm and eliminate the LOP effect and whether subjects show awareness of the LOP effect via JOLs. The original findings of Craik (1977) have not been replicated, and therefore we also asked whether we could replicate Craik's LOP effect in the word-first condition.

METHOD

Subjects

A priori power analysis was conducted to determine a sufficient sample size with an α of 0.05 and a power of 0.80. For a medium effect size (f = 0.25) for a between-factor main effect in a repeated-measures ANOVA, the minimum required sample size was 108. In total, 144 subjects (M_{age} = 36.45, SD_{age} = 11.10) were recruited on 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 excluded because they reported either cheating or taking breaks, 11 subjects were replaced because they were identified as outliers,[1] and 1 was replaced because of experimental error. Out of 126 remaining subjects, 42 were in each condition.

Materials and Design

Sixty concrete nouns were selected from norms collected by Van Overschelde, Rawson, and Dunlosky (2004) and Nelson, McEvoy, and Schreiber (2004). The words had a concreteness level above 2.5 out of 5 according to Brysbaert, Warriner, and Kuperman (2014). The logarithm of Hyperspace Analogue to Language frequency in the English Lexicon Project (Balota et al., 2007) ranged from 6.75 to 11.79. 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). Each word appeared together with one of the six questions during the study phase equally often across subjects.

A 3 × 2 × 3 mixed factorial design was used, such that question type (case, rhyme, category) and congruency (congruent, incongruent) were manipulated within subjects and the study condition (word-first, word-first JOL, question-first JOL) was manipulated between subjects.

Procedure

Figure 1a shows study trials for each condition. The order of study trials was random for each subject. In the word-first and word-first JOL conditions, the target word was presented at the beginning of the study trial for 2 s, and a delay of 5 s followed. Subjects then saw the orienting question for 3 s and had 3 s to make a "yes/no" response to the question. This sequence of study trials in the word-first condition totaled 13 s per trial. For the word-first JOL condition, there was an additional 5 s at the end of each study trial to make a JOL rating, totaling 18 s per trial. For the question-first JOL condition, the orienting question was presented first for 3 s, followed by a 5-s delay. The word was then presented for 2 s. Subjects had 3 s to make a "yes/no" response and 5 s to give a JOL rating, totaling 18 s per study trial. Subjects made their JOLs on a 5-point scale, where they indicated with 1 that *I will definitely not remember the word* and with 5 that *I will definitely remember the word*. After the study phase, subjects completed a 10-min president recognition test in which they tried to identify presidents of the United States from plausible lures (Roediger & DeSoto, 2016). For the test phase, subjects took

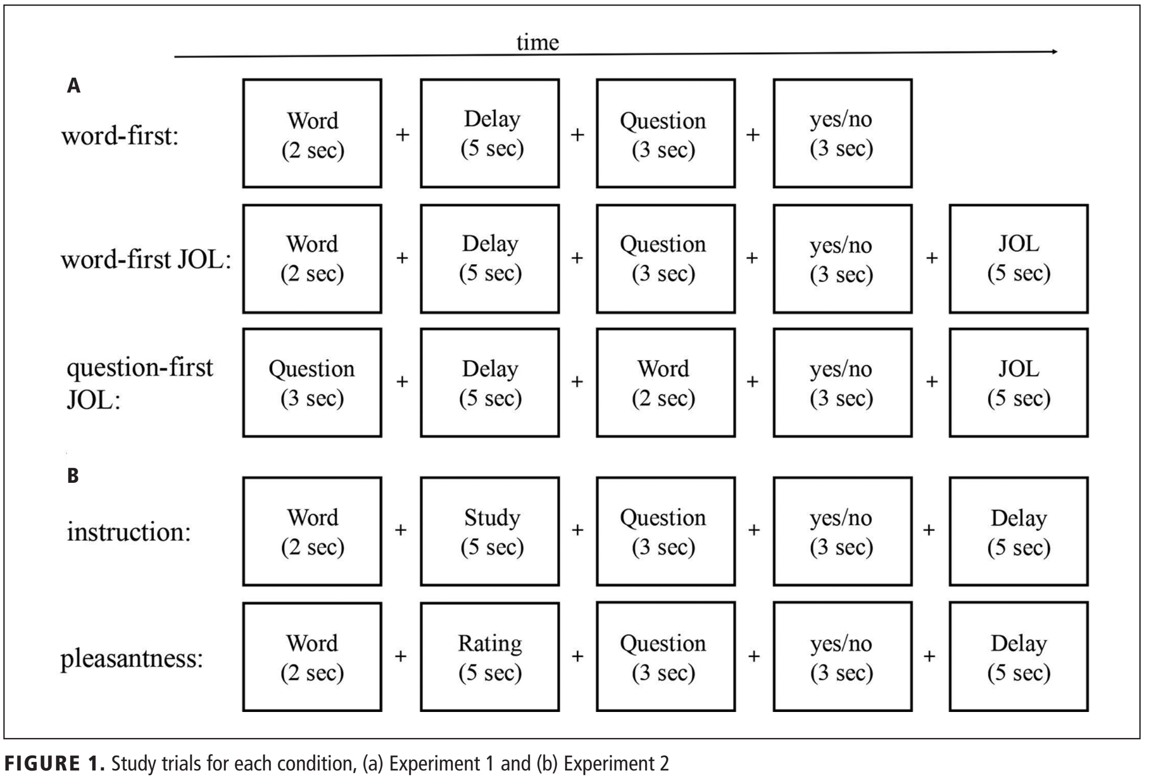

FIGURE 1. Study trials for each condition, (a) Experiment 1 and (b) Experiment 2

a self-paced old/new recognition test that consisted of 180 words (60 targets, 120 lures) and indicated whether they had studied each word separately and made a confidence rating on a 5-point scale.

RESULTS

The upper part of Table 1 provides overall hit rates, false alarm rates, corrected recognition scores (hits – false alarms), and d' scores[2] for each condition. Corrected recognition scores[3] were used as the primary dependent variable. For all the pairwise comparisons, Sidak comparison was used unless indicated otherwise.

Corrected Recognition Scores

Figure 2 shows the corrected recognition scores for the three conditions, with the word-first condition having the lowest scores for all orienting tasks. Thus, requiring JOLs increased recognition scores between the word-first and word-first JOL condition, confirming again that JOLs are reactive measures. A 2 (congruency) × 3 (orienting tasks) × 3 (study condition) repeated measures ANOVA showed a main effect of study condition, $F(2, 123) = 11.84, p < .001, \eta_p^2 = .16$. As noted, subjects in the word-first condition had lower corrected recognition scores than subjects in the word-first JOL and question-first JOL conditions, ps < .01, but the latter two conditions did not differ from one another, $p = .277$.

The main effect of orienting tasks was also significant, $F(1.84, 226.79) = 38.96, p < .001, \eta_p^2 = .24$, and was moderated by a reliable interaction with the study condition, $F(3.69, 226.79) = 3.13, p = .018, \eta_p^2 = .05$. See Figure 2 for the results. For the word-first and question-first JOL conditions, the results for corrected recognition scores replicated the LOP effect: All three orienting tasks differed from one another in terms of corrected recognition scores, ps < .05. For the word-first JOL condition, however, this was not the case. The difference between the category and case tasks did not reach conventional levels of significance, $p = .058$, and other pairwise comparisons were not significant, ps > .05, indicating that the LOP effect has been minimized. To determine whether this .05 recognition score difference between case and category tasks was meaningful, we conducted Bayes factor analysis. There was strong evidence for the alternative hypothesis (Bayes factor_{10} =12.48), suggesting that the two tasks produced different recognition scores (Wagenmakers, Morey, & Lee, 2016). Thus, we probably did not eliminate the LOP effect by using JOLs.

For the case task, subjects in the word-first JOL condition had higher corrected recognition scores (.69) than those in the control word-first condition (.46), p < 001. The difference in recognition after the case task between the word-first (.46) and question-first JOL (.57) conditions did not reach statistical significance, $p = .080$, nor did the difference in recognition between the word-first JOL (.69) and the question-first JOL (.57) conditions, $p = .053$. For the rhyme and category tasks, both the question-first JOL

TABLE 1. Overall Hits, False Alarm Rates, and Corrected Recognition Scores

	Hits		False alarms		Hits – False alarms		*d′*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Study condition				**Experiment 1**				
Word-first	.80	.17	.29	.21	.51	.23	1.64	0.95
Word-first JOLs	.88	.10	.16	.11	.72	.17	2.50	0.90
Question-first JOLs	.81	.12	.16	.15	.65	.20	2.14	0.86
				Experiment 2				
Instructions	.83	.13	.18	.14	.66	.20	2.26	1.01
Pleasantness	.90	.11	.17	.15	.73	.21	2.67	1.07

Note. JOLs = judgments of learning.

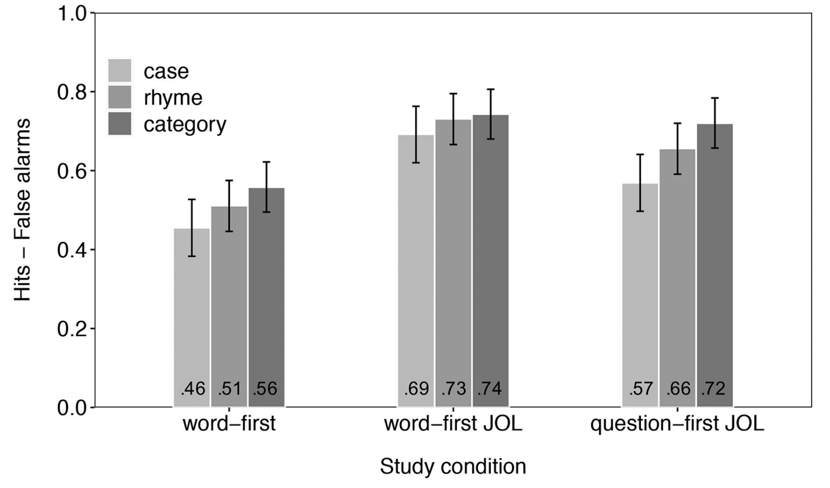

FIGURE 2. Corrected recognition scores across orienting tasks for each condition, Experiment 1. Error bars indicate 95% confidence intervals

(.66 and .72, respectively) and word-first JOL (.73 and .74, respectively) conditions showed higher corrected recognition scores than those in the word-first condition (.51 and .56, respectively), ps < 01. The question-first JOL and word-first JOL conditions did not differ on the rhyme and category tasks, ps > .05.

The main effect of congruency was also significant, $F(1, 123) = 53.13$, $p < .001$, $\eta_p^2 = .30$, with congruent ("yes") responses ($M = .66$, $SE = .02$) recognized more than incongruent ("no") responses ($M = .59$, $SE = .02$). This main effect was qualified by a reliable interaction with the study condition, $F(2, 123) = 6.51$, $p = .002$, $\eta_p^2 = .10$. In the word-first and question-first JOL conditions that showed the LOP effect, congruent words ($M = .55$, $SE = .03$, $M = .70$, $SE = .03$, respectively) produced significantly greater recognition scores than incongruent words ($M = .47$, $SE = .03$, $M = .60$, $SE = .03$, respectively), ps < .01, whereas in the word-first JOL condition, congruent words ($M = .73$, $SE = .03$) did not differ significantly from incongruent words ($M = .71$, $SE = .03$), $p = .160$. Other interactions were not significant, $p > .05$.

JOLs

Figure 3 shows the JOL ratings for the word-first JOL condition and the question-first JOL condition. The ratings show a slight increase across orienting tasks in the question-first JOL condition, mostly for congruent words, whereas JOLs do not seem to differ as much across orienting tasks for the word-first JOL condition. To test this pattern statistically, two separate 2 (congruency) × 3 (orienting tasks) repeated-measures ANOVAs were conducted for the question-first JOL and word-first JOL conditions, respectively. For the question-first JOL condition, there was a main effect of orienting tasks, $F(2, 82) = 8.35$, $p = .001$, $\eta_p^2 = .17$; however, pairwise comparisons revealed that not all orienting tasks differed from one another. Similar to recognition scores, the case task ($M = 3.02$, $SE = .11$) led to lower JOL ratings than the rhyme ($M = 3.21$, $SE = .11$) and category ($M = 3.22$, $SE = .10$) tasks, ps < .05, but unlike recognition scores, JOL ratings of the rhyme and category tasks were not statistically different, $p = .991$.

Overall, the congruent words ($M = 3.40$, $SE = .10$) yielded higher JOL ratings than incongruent words ($M = 2.90$, $SE = .11$), $F(1, 41) = 48.32$, $p < .001$, $\eta_p^2 = .54$, but this main effect was moderated by a reliable congruency × orienting task interaction, $F(2, 82) = 12.75$, $p < .001$, $\eta_p^2 = .24$. The congruent words mimicked the main effect of orienting tasks on JOLs; however, for incongruent words, JOL ratings did not differ from one another, ps > .05. Thus, subjects in the question-first JOL condition were able to predict only the difference between the case and the rhyme and category tasks for congruent words, as well as

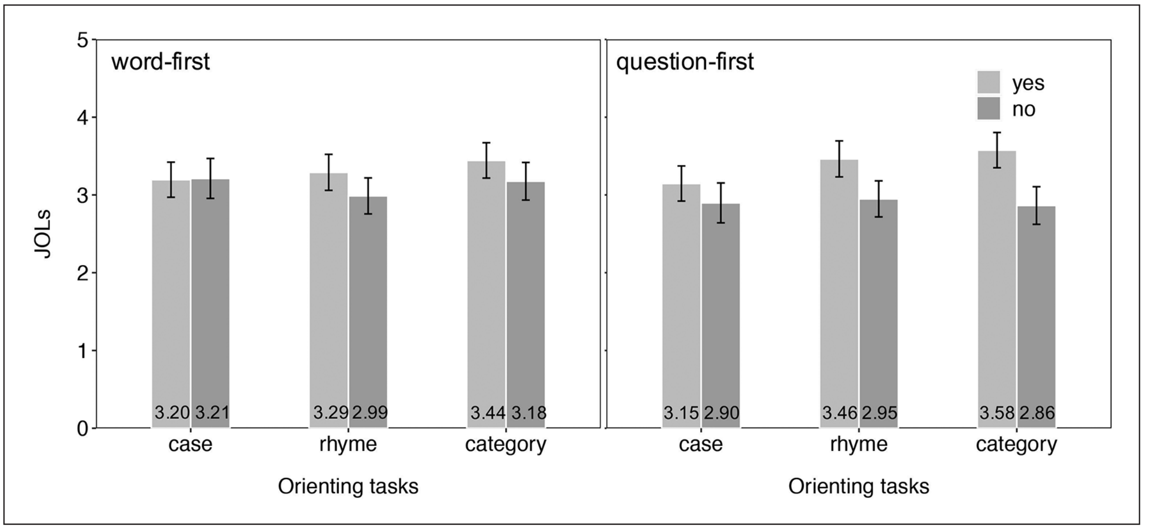

FIGURE 3. JOL ratings across orienting tasks and congruent/incongruent words for the word-first JOL and question-first JOL conditions, Experiment 1. Error bars indicate 95% confidence intervals

the congruency effect. They did not predict the LOP effect for incongruent words.

A 2 (congruency) × 3 (orienting tasks) repeated-measures ANOVA for the word-first JOL condition revealed no main effect of orienting tasks, $F(2, 82) = 2.49$, $p = .089$, $\eta_p^2 = .06$. Unlike corrected recognition scores, congruent words ($M = 3.31$, $SE = .11$) produced higher JOLs than incongruent words ($M = 3.13$, $SE = .11$), $F(1, 41) = 7.62$, $p < .001$, $\eta_p^2 = .16$. The interaction was also reliable, $F(2, 82) = 3.13$, $p = .049$, $\eta_p^2 = .07$, and was driven by different JOL ratings for congruent and incongruent words in the rhyme and category tasks, ps < .05. In the case task, JOL ratings for congruent and incongruent words did not differ, $p = .882$. In short, subjects predicted the congruency effect in their JOLs, although it was not present in their recognition scores.

DISCUSSION

The results of Experiment 1 replicated Craik (1977) in the word-first condition and revealed both the standard LOP and congruency effects. These results (and Craik's findings) differed markedly from the word-first JOL condition in which the LOP effect was minimized, suggesting that the combination of the word-first paradigm and JOLs promoted further elaboration of target words, especially in shallow tasks. However, the Bayes factor analysis showed that the LOP effect may not have been completely eliminated in the word-first JOL condition. In addition, the LOP effect was also obtained in the question-first JOL condition between all orienting tasks. This outcome indicates that in both the question-first and word-first paradigms, additional processing of the words by making JOLs did not eliminate the LOP effect. This finding was in opposition to predictions of the original LOP framework, which would predict no LOP effect after elaboration and thus posed another question for the LOP framework. It is possible that JOLs might not have induced deep processing in Experiment 1 as traditional deep processing tasks (e.g., pleasantness ratings). The results also indicate that providing JOLs is a reactive task because they affect retention in addition to measuring subjects' predictions.

These findings supported the idea that even when the target words were presented before the orienting tasks, subjects did not fully engage in spontaneous semantic processing in the word-first paradigm. Furthermore, subjects showed some awareness of the LOP effect in the question-first JOL condition. That is, for congruent words in the question-first JOL condition, subjects were able to predict that they would recognize fewer words after encoding with the case task than in the category and rhyme tasks. However, for items that required "no" responses (i.e., incongruent items), subjects showed no awareness of the LOP effect. These outcomes suggested that when the LOP effect was larger, as in the standard question-first paradigm, subjects were somewhat aware of the LOP effect for congruent items, and the lack of awareness cannot fully explain the LOP effect. Critically, in the word-first JOL condition, subjects were insensitive to the small LOP effect observed in recognition performance regardless of the congruency of items.

EXPERIMENT 2

Experiment 1 suggested that even in the word-first paradigm, subjects might not be processing or attending to the presented word during the 5-s delay period. That is, instead of processing the word deeply as expected under intentional learning instructions, they might just engage in rote rehearsal and wait for the orienting question without fully processing the word. In Experiment 1, we used JOLs to promote additional elaboration in the word-first paradigm. In Experiment 2, we aimed to promote semantic processing in the word-first paradigm using methods besides JOLs. One way was to directly encourage subjects to study the items using their idiosyncratic mnemonic strategies during the delay period. When subjects are not instructed to use a certain strategy, they can still adapt their strategies to the task demands and perform similarly to or higher than in conditions with structured strategies (McDaniel & Kearney, 1984). Therefore, just explicitly instructing subjects to study the words during the 5-s delay might stop them from waiting for the question and promote natural semantic processing in the word-first paradigm.

We also used a second method intended to facilitate semantic encoding of the word in the word-first paradigm. It may be that in the word-first paradigm, subjects just perform rote rehearsal to remember the word rather than encoding it meaningfully because they need to remember it only until one of the three orienting task instructions appears. To address this question, in the second condition of Experiment 2 we required semantic processing of the word during the 5-s delay. However, even this step may not eliminate the LOP effect if the shallow orienting tasks undercut or impair deep encoding, thereby resulting in poorer retention and the LOP effect. For instance, Cermak, Schnorr, Buschke, and Atkinson (1970) demonstrated that focusing on both phonetic and semantic features of words impaired retention compared with just focusing on semantic features. Similarly, shallow tasks after the 5-s delay in the word-first paradigm might impair meaningful processing even after semantic processing. To address this question, in Experiment 2 we introduced a pleasantness rating task (a deep processing task) during the 5-s delay to ensure semantic processing of the words. This condition was different from the word-first JOL condition of Experiment 1 in that semantic processing was promoted before the orienting task, whereas in the word-first JOL condition, it was promoted after the orienting task. If the LOP effect still emerged under the pleasantness condition, we would conclude that shallow tasks after deep processing tasks impair retention.

In Experiment 2, all subjects were given intentional learning instructions in the word-first paradigm and studied the material under the word-first paradigm with explicit study instructions (instruction) and under the word-first paradigm with pleasantness ratings (pleasantness). We aimed to answer the following questions: Do explicit instructions to study the words eliminate the LOP effect in the word-first paradigm?, Do shallow tasks impair retention by interfering with semantic processing?, and Do pleasantness ratings made before the orienting tasks eliminate the LOP effect in the word-first paradigm?

METHOD

Subjects

Ninety-one subjects ($M_{age} = 37.13$, $SD_{age} = 10.86$) were recruited on MTurk and randomly assigned to one of two conditions. Participation and compensation criteria were the same as in Experiment 1. Seven subjects were replaced because they were identified as outliers in one or more of the previously described outlier criteria. Of the 84 remaining subjects, 42 subjects were in each condition.

Materials and Design

The same materials were used as in Experiment 1. A 3 × 2 × 2 mixed factorial design was used, such that orienting tasks (case, rhyme, category) and congruency (congruent, incongruent) were manipulated within subjects and the study condition (instruction, pleasantness) was manipulated between subjects.

Procedure

Figure 1b shows study trials for each condition. Subjects were randomly assigned to one of the two between-subject conditions under intentional learning instructions. In both conditions, the target word was presented at the beginning of the study trial for 2 s, followed by a delay of 5 s and then the orienting task that consumed 3 s. Subjects then made a "yes/no" response to the question within 3 s and had an additional 5 s at the end of each study trial,[4] totaling 18 s per study trial. Subjects in the pleasantness condition gave a pleasantness rating for each word during the initial 5-s delay, whereas subjects in the instruction condition were told before the study phase to use the 5-s delay to study the words with any strategy they thought would help them remember the words. To remind them of this instruction, the command "Study!" appeared on the screen during the 5-s delay. After the study phase, the filler task and the test phase followed, as in Experiment 1.

RESULTS

The lower part of Table 1 provides overall hit rates, false alarm rates, corrected recognition scores (hits – false alarms), and d' scores for each group. As in Experiment 1, corrected recognition scores were used as the primary dependent variable.

Corrected Recognition Scores

Figure 4 shows the corrected recognition scores for subjects in the two conditions, with those in the instruction condition showing the lower scores than those in the pleasantness condition. A 2 (congruency) × 3 (orienting tasks) × 2 (study condition) repeated-measures ANOVA revealed no main effect of study condition, $F(1, 82) = 2.27$, $p = .135$, $\eta_p^2 = .03$. The main effect of orienting tasks was significant, $F(1.83, 150.32) = 14.03$, $p < .001$, $\eta_p^2 = .15$, and was qualified by a reliable interaction with the study condition, $F(1.83, 150.32) = 3.63$, $p = .033$, $\eta_p^2 = .04$. For the instruction condition, the category task produced significantly greater corrected recognition scores than the rhyme and case tasks, ps < .01, whereas the case task and the rhyme task did not differ, $p = .170$. For the pleasantness condition, none of the orienting tasks differed from one another, ps > .05. Furthermore, subjects in the pleasantness condition outperformed subjects in the instruction condition on the case task, $p = .033$, suggesting that engaging in deep encoding especially boosted retention for shallow processing. The study conditions did not differ on the rhyme and category tasks, ps > .05.

The main effect of congruency was also significant, $F(1, 82) = 5.79$, $p = .018$, $\eta_p^2 = .07$, and qualified by a reliable interaction with the study condition, $F(1, 82) = 3.76$, $p = .056$, $\eta_p^2 = .04$. In the instruction condition, congruent words ($M = .68$, $SE = .03$) produced significantly greater recognition scores than incongruent words ($M = .64$, $SE = .03$), $p = .003$, whereas in the pleasantness condition, recognition of congruent words ($M = .73$, $SE = .03$) did not differ from incongruent words ($M = .72$, $SE = .03$), $p = .743$. Other interactions were not significant, ps > .05.

DISCUSSION

Experiment 2 showed that explicitly instructing subjects to study the words during the 5-s delay eliminated the difference between the shallow case and rhyme tasks but did not entirely eliminate the LOP effect because the category task was superior to the other two conditions. Thus, it is possible that subjects either were not attending to the words during the 5-s delay in Craik's (1977) study or were engaging

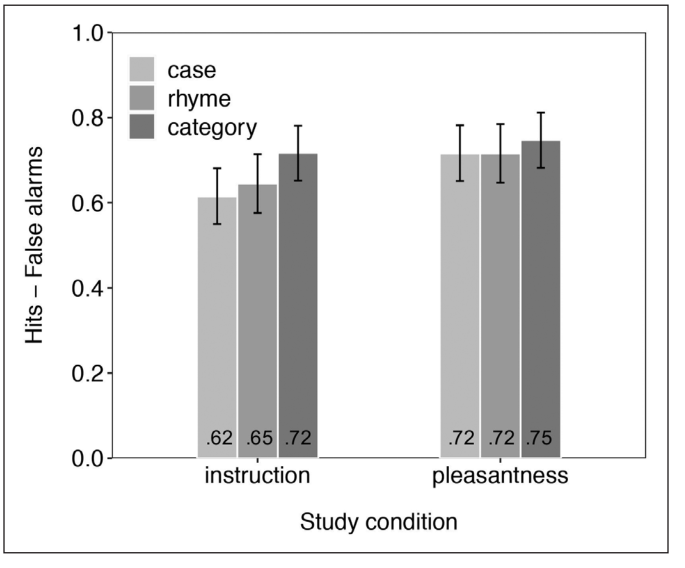

FIGURE 4. Corrected recognition scores across orienting tasks for each condition, Experiment 2. Error bars indicate 95% confidence intervals

only in rote rehearsal. Yet simply instructing subjects to study the words was not enough to eliminate the differences between deep and shallow tasks. On the other hand, when subjects provided pleasantness ratings for the target words, the LOP effect was eliminated in the word-first paradigm. Pleasantness ratings yielded semantic processing of all items, and the case and rhyme orienting questions did not undermine retention. Notably, the effect of making a pleasantness rating on recognition performance was larger for the shallow tasks than for the category task. That is, the additional pleasant ratings increased recognition scores by .10 for the shallow case task relative to the instruction condition, whereas this number was .03 for the category task.

GENERAL DISCUSSION

The two experiments reported here were designed to explain why presenting the word before the orienting question and using intentional learning instructions did not eliminate the LOP effect (Craik, 1977). We examined three possible reasons why this might be the case. We assumed that although subjects were instructed to learn the words when the word occurred before the orienting task, they may have just waited for the question without attending to the word carefully and simply engaged in rote rehearsal until the question appeared. Critically, the word was not presented during the 5-s delay before the orienting task. Therefore, to answer the orienting question correctly, it was necessary for subjects to at least maintain the word in short-term memory while waiting for the question. Only then did they encode the word (on one third of the trials) with deeper processing. A second related possibility was that subjects were not aware of how the orienting tasks would affect their learning (i.e., they were unaware that lower performance would occur after shallow tasks), and thus they did not engage in natural semantic processing before the orienting tasks. The third possibility is that even if subjects processed the word semantically before the orienting tasks, the following shallow tasks might have impaired later retention by undercutting semantic processing. Thus, we used three tasks for subjects to perform on the word to examine these possibilities and to encourage its deeper processing in the word-first paradigm.

In Experiment 1, we replicated Craik's (1977) findings and attempted to eliminate the LOP effect in the word-first paradigm by having subjects make JOLs. This procedure aimed to test our first assumption by promoting deep processing at the end of each study trial. Furthermore, we asked subjects to make JOLs in the question-first paradigm to examine their awareness of the LOP effect. We did not completely eliminate the LOP effect in the word-first JOL condition, although it was minimized. This outcome suggests that subjects did not engage in deep processing before the orienting task in the word-first paradigm. We return to the issue of awareness of the LOP effect later.

In Experiment 2, we used two new conditions in the word-first paradigm. In one, subjects were explicitly told to use idiosyncratic mnemonic strategies to remember the word during the 5 s before the question appeared. This condition aimed to promote natural deep processing and to stop subjects from just waiting for the orienting task; however, this instruction did not eliminate the LOP effect. This finding suggests that even when subjects attend to the word, they might be using an ineffective study strategy. In the second condition, we had subjects make pleasantness judgments on the word during the 5 s before the question appeared to examine whether the following shallow tasks would impair retention after semantic processing. Under this condition, where subjects were forced to access meaning by judging pleasantness, the LOP effect finally disappeared. Thus, we conclude that although they were given intentional learning instructions and saw the target at the beginning of the study trial, subjects did not engage in natural deep processing of the word in the word-first paradigm; instead, they used rote rehearsal and waited for the word.

Another interpretation of our results could be that although intentional learning instructions promoted deep processing of target words in the word-first paradigm, such deep processing was not as effective as deep processing promoted by the categorization task. Some previous studies have shown retention differences across various deep orienting tasks (Nairne, Pandeirada, & Thompson, 2008; Packman & Battig, 1978; Seamon & Virostek, 1978), and thus a similar argument can be made for the word-first paradigm. This interpretation could explain why the category task produced greater retention than other tasks in the word-first paradigm; however, it cannot explain retention differences between shallow case and rhyme tasks. That is, under this argument we would expect similar retention after case and rhyme tasks because both tasks would be performed under intentional learning as a form of less effective deep processing. Nonetheless, the LOP effect across all three tasks was prevalent both in our data and in the Craik (1977) study. Therefore, this interpretation cannot readily account for the LOP effect in the word-first paradigm, although variations in retention within the same level of processing pose another question for the LOP framework (Roediger & Gallo, 2002).

A secondary purpose of our study was to determine whether subjects are aware of the LOP effect, as exhibited by their JOLs, because insensitivity to the LOP effect is a potential reason why subjects did not engage in natural deep processing in the word-first paradigm. The relevant data are shown in Figure 3, which shows JOLs for the question-first paradigm. For words for which the answer to the question was "yes," subjects seemed to show some appreciation of an LOP effect in their JOLs. On the other hand, for the "no" answers, subjects showed no hint of an LOP effect. Although subjects showed limited awareness of the LOP effect in their JOLs, the JOLs did predict (accurately) the difference between items provided "yes" and "no" responses. Thus, we see that the effect predicted by Craik and Tulving (1975), levels of processing, is not well predicted by subjects. However, the effect Craik and Tulving did not predict—the differences in recognition between "yes" and "no" responses—was predicted by subjects (see Tekin & Roediger, 2020). One problem in drawing conclusions about what subjects know about the LOP effect through JOLs is that JOLs are reactive measures, ones that improve performance (Janes et al., 2018; Soderstrom et al., 2015; Tekin & Roediger, 2020). Thus, because JOLs increase recognition, especially for the shallow processing conditions (Tekin & Roediger, 2020), they reduce the LOP effect, as in Experiment 1, and thus it might be difficult for subjects to become fully aware of the reduced effect. Nonetheless, our data show that subjects were aware of the difference caused by answering "yes" or "no" to questions on later performance.

Conclusion

The present study examined why the LOP effect persisted in the word-first paradigm with intentional learning instructions and found that even when the target word preceded the orienting task, subjects did not engage in natural deep processing. Encouraging subjects to study or attend to the material by using their idiosyncratic strategies attenuated the LOP effect but did not eliminate it. Further examination of the word-first paradigm demonstrated that once the words were deeply processed through pleasantness ratings, the shallow tasks that followed did not impair later retention. We conclude that subjects just waited for the orienting task and engaged in ineffective maintenance strategies before the orienting task. The mystery of the word-first paradigm was solved, although others still exist in the LOP literature (Roediger & Gallo, 2002).

NOTES

Address correspondence about this article to Eylul Tekin, 1 Brookings Drive, St. Louis, MO 63130 (e-mail: elifeylultekin@wustl.edu).

1. In a preliminary analysis, correct "yes/no" responses at encoding and average reaction times for JOLs, recognition decisions, and confidence judgments 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. The number of correct "yes/no" responses for five of 11 outliers were lower than the sample average, because they did not respond to the orienting question in a 3-s window or they incorrectly responded to it. Given that this is the main manipulation of the study, their data were not used. The remaining six subjects were detected as outliers based on reaction times to their JOLs, recognition decisions, or confidence judgments. Because it is impossible to control what subjects were doing during these time periods in the MTurk platform, these subjects were replaced.

2. In Experiments 1 and 2 combined, 11 subjects had a perfect hit rate of 1.00, and eight subjects had a perfect false alarm rate of 0.00. To be able to calculate d' scores for these subjects, we calculated corrected hit rates and false alarm rates. For perfect hit rates, half a hit was subtracted from the total number of hits and the corrected hit rate was calculated after this correction (59.5/60), and for perfect false alarm rates, half a false alarm was added to a perfect false alarm of 0 and corrected false alarm rate was calculated after this correction (0.5/120) (Macmillan & Creelman, 2005).

3. To conduct the statistical analyses, we calculated six different hit rates (case/"yes," case/"no," rhyme/"yes,"

rhyme/"no," category/"yes," category/"no") for each subject. Corrected recognition scores were used as the dependent variable instead of *d′* scores because in Experiments 1 and 2 combined, 164 subjects had perfect hit rates of 1.00 in at least one of the six congruency × orienting task combinations. Calculating *d′* scores for perfect hit rates required correction of perfect hit rates described previously (in this case 9.5/10), and this artificially lowered the performance for subjects who had more perfect hit rates. The results remained similar across experiments when we used *d′* scores instead of corrected recognition scores.

4. The additional 5 s was added to the instruction and pleasantness conditions to equate the study trial time with the word-first JOL and question-first JOL conditions in Experiment 1. Previous data from our lab indicate that providing delay at the end of the study trial has no influence on the LOP effect.

REFERENCES

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

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

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

Cermak, G., Schnorr, J., Buschke, H., & Atkinson, R. C. (1970). Recognition memory as influenced by differential attention to semantic and acoustic properties of words. *Psychonomic Science, 19,* 79–81. https://doi.org/10.3758/BF03337430

Chow, P. C., Currie, J. L., & Craik, F. I. (1978). Intentional learning and retention of words following various orienting tasks. *Bulletin of the Psychonomic Society*, *12,* 109–112. https://doi.org/10.3758/BF03329642

Craik, F. I. (1973). A "levels of analysis" view of memory. In P. Pliner, L. Krames, & T. M. Alloway (Eds.), *Communication and affect: Language and thought* (pp. 45–65). Academic Press. https://doi.org/10.1016/B978-0-12-558250-6.50010-9

Craik, F. I. (1977). Depth of processing in recall and recognition. *Attention and Performance, VI,* 679–697.

Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior, 11,* 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,* 268–294. https://doi.org/10.1037/0096-3445.104.3.268

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,* 472–481. https://doi.org/10.1037/h0028372

Hyde, T. S., & Jenkins, J. J. (1973). Recall for words as a function of semantic, graphic, and syntactic orienting tasks. *Journal of Verbal Learning and Verbal Behavior, 12,* 471–480. https://doi.org/10.1016/S0022-5371(73)80027-1

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,* 2356–2364. https://doi.org/10.3758/BF03195588

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

McDaniel, M. A., & Kearney, E. M. (1984). Optimal learning strategies and their spontaneous use: The importance of task-appropriate processing. *Memory & Cognition, 12,* 361–373. https://doi.org/10.3758/BF03198296

Nairne, J. S., Pandeirada, J. N., & Thompson, S. R. (2008). Adaptive memory: The comparative value of survival processing. *Psychological Science, 19,* 176–180. https://doi.org/10.1111/j.1467-9280.2008.02064.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,* 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

Packman, J. L., & Battig, W. F. (1978). Effects of different kinds of semantic processing on memory for words. *Memory & Cognition, 6,* 502–508. https://doi.org/10.3758/BF03198238

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

Roediger III, H. L., & Gallo, D. A. (2002). Levels of processing: Some unanswered questions. *Perspectives on Human Memory and Cognitive Aging: Essays in Honour of Fergus Craik,* 28–47.

Seamon, J. G., & Virostek, S. (1978). Memory performance and subject-defined depth of processing. *Memory & Cognition, 6,* 283–287. https://doi.org/10.3758/BF03197457

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

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. http://dx.doi.org/10.1037/a0038388

Tekin, E., & Roediger, H. L. III. (2020). Reactivity of judgments of learning in a levels-of-processing paradigm. *Zeitschrift für Psychologie, 228,* 278–290. https://doi.org/10.1027/2151-2604/a000425

Tresselt, M. E., & Mayzner, M. S. (1960). A study of incidental learning. *Journal of Psychology, 50,* 339–347. https://doi.org/10.1080/00223980.1960.9916451

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,* 289–335. https://doi.org/10.1016/j.jml.2003.10.003

Wagenmakers, E. J., Morey, R. D., & Lee, M. D. (2016). Bayesian benefits for the pragmatic researcher. *Current Directions in Psychological Science, 25,* 169–176. https://doi.org/10.1177/0963721416643289

Walsh, D. A., & Jenkins, J. J. (1973). Effects of orienting tasks on free recall in incidental learning: "Difficulty," "effort," and "process" explanations. *Journal of Verbal Learning and Verbal Behavior, 12,* 481–488. https://doi.org/10.1016/S0022-5371(73)80028-3