Winter 12-2017

Do Learners Have Insight into the Levels of Processing Effect? Exploring Unresolved Levels of Processing Phenomena with Judgments of Learning

Elif Eylul Tekin
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Psychological & Brain Sciences

Do Learners Have Insight into the Levels of Processing Effect? Exploring Unresolved Levels of Processing Phenomena with Judgments of Learning

by

Elif Eylul Tekin

A thesis presented to
The Graduate School
of Washington University in
partial fulfillment of the
requirements for the degree
of Masters of Arts

December 2017
St. Louis, Missouri

Table of Contents

List of Figures iv
List of Tables v
Acknowledgments vi
Abstract vii
Chapter 1: Introduction 1
1.1 Transfer-Appropriate Processing 5
1.2 Problems with the LOP Framework 8
Chapter 2: Experiment 1 11
2.1 Method 18
2.1.1 Subjects 18
2.1.2 Materials 18
2.1.3 Design 20
2.1.4 Procedure 20
2.2 Results 23
2.2.1 JOL condition 26
2.2.2 Intentional and Delay conditions 30
2.3 Discussion 31
Chapter 3: Experiment 2 33
3.1 Method 36
3.1.1 Subjects 36
3.1.2 Materials 37
3.1.3 Design 37
3.1.4 Procedure 37
3.2 Results 39
3.2.1 Standard-JOL condition 42
3.2.2 Reversed-JOL condition 45
3.2.3 Explicit Instructions condition 48
3.2.4 Pleasantness condition 50
Chapter 4: General Discussion 51
4.1 The LOP Effect and Intentional Learning 51

4.2 Congruency effect and JOLs 55
4.3 Diagnosticity of JOLs 55
4.4 JOLs as Memory Modifiers 58
4.5 Conclusion 60
References 62
Appendix A 66
Appendix B 68

List of Figures

Figure 1.1: The LOP Effect ……...……...……...….4

Figure 2.1: Study Trials ……...……...……...……...……22

Figure 2.2: Corrected Recognition Scores across Orienting Tasks for Experiment 1 ……..…25

Figure 2.3: Corrected Recognition Scores for *Yes/no* Responses for Experiment 1 ……...……26

Figure 2.4: Corrected Recognition Scores and JOLs for the JOL condition ……..…..…27

Figure 2.5: Z-scores for the JOL condition ….……....……....……29

Figure 3.1: The LOP Effect under The Reversed Order Paradigm ….……..…...……34

Figure 3.2: Corrected Recognition Scores across Orienting Tasks for Experiment 2 ……..…40

Figure 3.3: Corrected Recognition Scores for *Yes/no* Responses for Experiment 2 ……...……41

Figure 3.4: Corrected Recognition Scores and JOLs for the Standard-JOL Condition ...……43

Figure 3.5: Z-scores for the Standard-JOL Condition ….…….…….....……44

Figure 3.6: Corrected Recognition Scores and JOLs for the Reversed-JOL Condition ...……...46

Figure 3.7: Z-scores for the Reversed-JOL Condition ….……...….....……47

Figure 3.8: Corrected Recognition Scores for the Explicit Instructions Condition ...……..….49

List of Tables

Table 2.1: Overall Hit, False Alarm Rates and Corrected Recognition Scores ……..…..…..…24

Table 2.2: Mean Gamma Correlations and Standard Deviations ……....30

Table B.1: Confidence Judgments for Orienting Tasks across Study Groups ……..…….....70

Table B.2: Confidence Judgments for Orienting Tasks across Response Types ……..…....70

Acknowledgments

I am deeply grateful to my advisor, Roddy Roediger, for his insight and mentorship throughout this project. I further thank the members of my defense committee, Kathleen McDermott and Mark McDaniel, for their interest, direction and advice. I also want to recognize members of the Memory Lab, Wenbo Lin and Jeremy Yamashiro, for their comments. I am particularly grateful to my labmate and roommate, Oyku Uner, who always supported me both in the lab and at home throughout this process. I would like to thank Reshma Gouravajhala, Francis Anderson, Samuel Chung and Marina Gross for the numerous work parties and motivational speeches.

I offer my sincerest gratitude to my former advisors, Aysecan Boduroglu and Esra Mungan, for their mentorship and for encouraging me to apply to Washington University in St. Louis. Lastly, I am genuinely thankful to my family and friends, especially Kuzey Nalbant, who never stopped supporting me from a long distance.

Funding was provided by a James S. McDonnell Foundation.

Elif Eylul Tekin

Washington University in St. Louis

December 2017

ABSTRACT OF THE THESIS

Do Leaners Have Insight into the Levels of Processing Effect? Exploring Unresolved Levels of Processing Phenomena with Judgments of Learning

by

Elif Eylul Tekin

Master of Arts in Psychological & Brain Sciences

Washington University in St. Louis, 2018

Professor Henry L. Roediger, III

The levels of processing (LOP) effect shows that semantic processing leads to better retention than other types of processing. The effect is routinely obtained on many types of tests, yet, to this day, its mechanisms are still debated and it is poorly understood. In two old/new recognition experiments, I investigated potential explanations as to why the LOP effect occurs under intentional learning instructions. I asked a) whether subjects were aware of the LOP effect while they were studying the material, b) whether explicitly encouraging subjects to study the words with their idiosyncratic strategies would eliminate the effect, and c) whether the shallow orienting tasks impaired future performance after deep encoding of the material. I employed the standard LOP paradigm in which the orienting question appeared before the word (Experiment 1) and a reversed order paradigm in which the word appeared before the orienting question (Experiment 2). In both experiments, a group of subjects made judgments of learning (JOLs). The results indicated that subjects did not accurately predict the LOP effect, even though they were somewhat aware of it. The LOP effect still occurred under the reversed order paradigm with explicit instructions to study during delay and under the reversed order paradigm with JOLs, though it was attenuated or eliminated between some levels. In addition, the act of making JOLs enhanced performance for the shallow orienting tasks, adding to evidence that JOLs are

reactive measures. Thus, giving JOLs can promote semantic processing to some extent and attenuate the LOP effect. Not being able to predict the LOP effect accurately and not engaging in spontaneous semantic processing under the shallow orienting questions might be potential explanations for the recurring LOP effect under intentional learning instructions.

Chapter 1: Introduction

Craik and Lockhart (1972) proposed the levels of processing (LOP) framework that described memory trace as a by-product of perceptual analyses. These analyses vary from surface levels of features (i.e., structural or phonemic) to deeper (i.e., semantic) processes and create a hierarchy or a continuum of "levels of analysis," where the analyses follow structural to semantic levels (Craik, 1973). With deeper processing, retention improves, enhancing recall and recognition. When the required level of analysis is reached for the task at hand, further analysis does not occur. An assumption of the LOP framework is that we tend to try to extract meaning from various stimuli or events in our everyday lives, and therefore frequently engage in semantic processing. We do not process peripheral details of an event deeply, and thus they are more likely to be forgotten than information processed semantically. For example, we would likely remember the gist and main points of a story through semantic analysis, but we would not be able to remember perceptual structures such as the exact words and structures of the sentences (e.g. Jenkins, 1974; Sachs, 1967). In short, the forgetting rate depends on the levels of processing. In early theoretical papers, depth of processing was associated with longer processing time. Hence, the LOP theory predicted longer processing times for semantic processing and shortest processing time for shallow processing.

Before the LOP theory was fully developed, studies using different orienting tasks during learning showed consistent findings with the predictions of the LOP theory, and laid the groundwork for its empirical support. For instance, Hyde and Jenkins (1969) manipulated learning instructions and orienting tasks. Subjects either received incidental or intentional learning instructions before study. In the incidental learning condition, subjects did not know that

their memories would be tested, whereas in the intentional learning condition subjects knew that they would be tested for their memory of the words. During learning, subjects engaged in one of three orienting tasks while studying the words: counting the number of letters in the word, counting occurrences of the number of letter E's (both shallow processing tasks) or rating the pleasantness of the word (a deep processing task). There was also a control group without any orienting task under intentional learning instructions, and thus a total of seven different groups. Following the learning phase, subjects were asked to recall the words.

Their results were later interpreted within the LOP framework in which different orienting tasks were thought to instigate different levels of processing (Craik & Lockhart, 1972). Groups that studied the material with the deep processing task (pleasantness) recalled more words than the groups with the shallow processing tasks. The control group also performed at the same level as the pleasantness rating groups, supporting the idea that people normally process words in meaningful (i.e., deep) ways. Interestingly, the incidental and intentional learning instructions did not lead to different outcomes in memory performance. Thus, even though subjects given intentional learning instructions knew that they had to learn the material (and were motivated to remember the words), they still performed poorly on the memory test following shallow orienting tasks. This may lead one to ask: 1) Why did the shallow processing groups with intentional learning not use semantic processing to help them remember the words? 2) If the control group can perform at the same level as the semantic processing group, why did the shallow processing groups with intentional learning not perform at the same level?

To gain further understanding of levels of processing, Craik and Tulving (1975) explored the claims of the LOP framework in ten experiments. They introduced the standard LOP paradigm that would be used in most of the LOP experiments from that point forward. In this

paradigm, orienting tasks were presented in the form of *yes/no* questions before each word is presented. Questions were intended to induce different levels of processing, and all subjects were presented with all levels, thus employing a within-subject design. After seeing each word, subjects responded either *yes* or *no*. In most of Craik and Tulving's experiments, during the study phase subjects were first presented with questions about either the appearance (e.g., "*Is the word in capital letters?")*, rhyme (*"Does the word rhyme with BEAGLE?")*, or category of a word ("*Is the word a type of bird?*") and then were presented with the word (e.g., *EAGLE)* that the question was addressing. They varied the type of learning instructions (i.e., incidental or intentional) and the type of explicit memory test (i.e., recognition or cued recall) across experiments.

Figure 1.1 shows the LOP findings from Craik and Tulving's Experiment 9. In a classroom setting, they employed the standard LOP paradigm explained above with the intentional learning instructions, 6 sec presentation rate, and a recognition test. Their results revealed that the category orienting task led to the highest performance and the case orienting task led to the lowest performance. In addition, *yes* answers led to higher performance than *no* answers under the rhyme and category tasks. They obtained similar LOP effects throughout all ten experiments both under recognition and recall tests. They also discredited some predictions from the early LOP framework. For instance, their results revealed that the LOP effect was greater for questions to which the answer was *yes*, but, *yes* and *no* answers had similar processing times. Thus, if processing time was an index of how deeply an item was processed based on the LOP framework, *yes/no* questions should have been processed with the same depth and there should not be any retention difference between the two. In Experiment 6, Craik and Tulving explored this issue by creating questions that would make both *yes* and *no* answers

congruent units of the question. In other words, the *yes/no* answers were equated on their elaborateness (e.g., for targets *HOUSE* and *MOUSE*, the question "*Is the object bigger than a chair?"*) and under these conditions there was no retention difference between *yes* and *no* answers. Therefore, they concluded that congruity of the answer with the question was another important factor that affected retention.

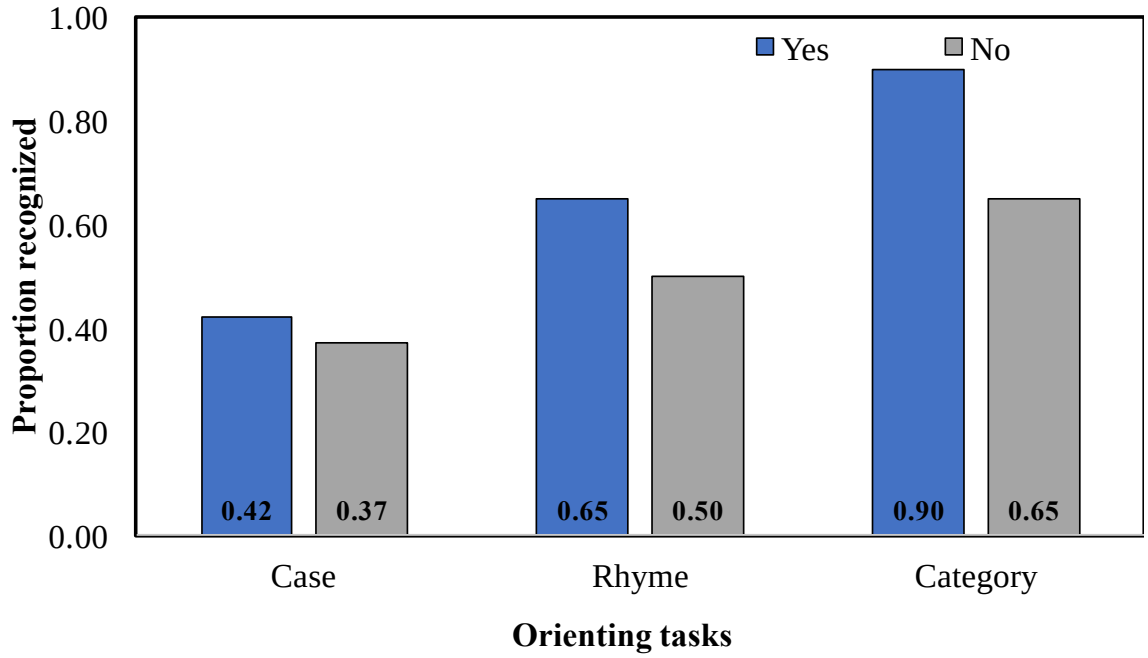

Figure 1.1 The LOP effect from Craik and Tulving (Experiment 9, 1975). It was not possible to estimate error bars from the statistics provided.

Further, in Experiment 5, Craik and Tulving compared a complex structural orienting task that took considerable time with a standard semantic task (a sentence task) and found that even though the structural task created longer processing times during encoding, the semantic task led to better retention (see also Walsh & Jenkins, 1973). Therefore, "depth of processing" cannot be explained by processing time, contradicting the empirical measure of "depth" proposed in the earlier theory. Also in line with Hyde and Jenkins (1969), the LOP effects occurred under intentional learning conditions, which should not be the case according to the

theory: Under the intentional learning instructions, subjects should engage in full processing of stimuli despite the type of orienting tasks as do control groups without orienting tasks. Subjects should not have processed stimuli at shallow levels given that they knew their memory would be tested later. In short, Craik and Tulving's series of experiments revealed shortcomings of the 1972 framework and also extended empirical findings: The complexity and the context of the stimuli (e.g., congruity) also affected later retention, a point not captured by the original framework. Therefore, they proposed that depth of processing could correspond to elaborateness (i.e., richness) of the encoding.

1.1 Transfer-Appropriate Processing

Further research following Craik and Tulving's (1975) also did not fit the LOP framework, and in fact, some strongly challenged it. The strongest challenge came from studies that manipulated retrieval in addition to encoding. During the study phase, Morris, Bransford and Franks (1978) presented subjects with phonemic (e.g., "*EAGLE rhymes with legal"*) and semantic sentences (e.g., "*A EAGLE has feathers"*) to engage differentiated processing of *EAGLE*. During the subsequent test phase, half of the subjects took a standard *yes/no* recognition test, whereas the other half took a rhyme recognition test. The rhyme recognition test consisted of non-studied words, half of which rhymed with target words (e.g., "*regal"* for "*EAGLE*") and other half did not. Subjects were instructed to respond *yes* if the presented word rhymed with a target word (a rhyme recognition test). For the standard semantic recognition test, the results revealed the standard LOP effect: Subjects performed better for semantically encoded words than phonemically encoded words. For the rhyme recognition test, however, the results were inverted: For *yes* answers, subjects accurately recognized phonemically encoded words more frequently than semantically encoded words. Morris et al. claimed that it was neither "depth" nor

"elaborateness" of the encoding that determined retention, but rather whether processes at test matched those at encoding. They named this idea "transfer appropriate processing (TAP)" and claimed that the TAP framework could account for results that the LOP theory could not.

Similar results have been found when different test types or test cues were used. For instance, McDaniel, Friedman and Bourne (1978) compared retention for perceptual and conceptual problems using both auditory and visual recognition tests. During the study phase, subjects were presented with two-by-two matrices of words that differed both conceptually and perceptually (e.g., "HIPPO", "tank", "*mouse*", "*PISTOL*") and asked to name a common value either based on conceptual grouping (a size dimension or their membership in a category) or based on perceptual grouping (font type or upper - lowercase letter). For the auditory recognition test, subjects had to respond whether or not they had seen the presented word in the matrices. For the visual recognition test, subjects had to choose the correct perceptual presentation of the same word (e.g., "HIPPO", *"HIPPO*", "hippo", "*hippo*"). The results showed the standard LOP effect for auditory recognition; for visual recognition, however, retention following perceptual grouping was higher compared to conceptual grouping, again contradicting the LOP framework, and supporting the TAP framework.

As another example, Fisher and Craik (1977) examined whether memory performance reflected the durability of a trace or compatibility between the trace and retrieval cue. In their second experiment, they presented subjects with word pairs during the study phase. Target words were paired with either associate or rhyme cues (e.g., for the target word *CAT*, *dog* or *hat*). For the cued recall test, subjects were shown either the same level cue from the study phase (if the pair was *hat – CAT,* they were shown *hat - ?*) or the other level cue (*dog - ?*). Their results demonstrated that when the cues matched between encoding and retrieval, retention was higher.

In other words, target words paired with rhyme cues and tested with rhyme cues led to better recall compared to target words paired with associative (i.e., semantic) cues and tested with rhyme cues. They acknowledged that these results could not be explained by the LOP framework alone: “The retention levels associated with a particular type of encoding were not fixed, but depended heavily on the type of retrieval cue used” (p. 709). Again their results supported greater retention when conditions at retrieval matched those at encoding, as predicted by the TAP framework or the encoding specificity principle (Morris et al., 1978; Tulving & Thomson, 1973).

According to the original LOP framework, phonemic or structural processing should never lead to greater retention than semantic processing, hence, the findings discussed above greatly challenged the primary tenet of the theory. Fisher and Craik (1977) and Craik (1977) tried to update the LOP theory, and admitted that the LOP framework lack consideration of the retrieval phase in the learning/remembering process. In light of later findings (McDaniel et al., 1978; Morris et al., 1978; Moscovitch & Craik, 1976), they proposed that along with the orienting task (i.e., levels of processing), retrieval conditions such as cue type and distinctiveness also influence retention. They suggested that semantic encoding allows greater potential for later retention and that cue uniqueness and retrieval conditions help to achieve that potential. They based their claim on the findings that regardless of the interactions of encoding and retrieval, semantic encoding still led to the highest performance, on average. These modifications, however, did not end criticism against the LOP framework (Baddeley, 1978; Eysenck, 1978; Morris et al., 1978; Nelson, 1977).

Tulving (1979) claimed that the LOP theory is circular in its explanations for current findings: Retention is better for deeply processed items, yet, we can only know that they are

deeply processed because they are remembered better. Further, he asserted that the LOP framework cannot empirically define "depth" and even more importantly it neglected the importance of the retrieval stage and retrieval cues. He concluded that the LOP theory is unnecessary to explain the findings in the literature and that the encoding specificity principle is sufficient by itself. Although occasionally mentioned by its defenders (Craik, 2002; Lockhart & Craik, 1978; 1990), not many studies have directly addressed the LOP framework after the 1980s, although the task itself has been frequently used.

1.2 Problems with the LOP Framework

The strong criticism of the LOP framework, based on the absence of a measure of depth and the empirical findings that supported the TAP framework (or the encoding specificity principle), led to neglect of other problems within the framework. There are still inconsistent findings that remain unexplored besides these major problems (Roediger & Gallo, 2001). In particular, these inconsistent findings showed the LOP effect, and yet violated the LOP theory, because according to the theory, the effect should be eliminated. One such finding is that under intentional learning instructions subjects given the shallow tasks had lower performance than subjects in control groups (and in deep processing groups). In other words, shallow processing groups were not able to engage in as much deep processing as control groups (Hyde & Jenkins, 1969; Walsh & Jenkins, 1973). Nevertheless, the findings from control groups confirmed that subjects spontaneously engaged in deeper (or more semantic) processing when left to their own devices. For this reason, it is not clear why subjects given the shallow orienting tasks do not opt for deep processing under intentional learning instructions.

As the original LOP theory predicted, intentional learning instructions without orienting tasks lead to similar levels of retention as deep processing, confirming its assumption that people

naturally engage in deep processing. The persistent finding that intentional learning instructions with orienting tasks do not eliminate the LOP effect, however, is inconsistent with the LOP theory. That is, under intentional learning instructions, the LOP theory predicted that subjects would engage in deep or meaningful processing regardless of the orienting task. Nonetheless, the LOP effect is difficult to eliminate even with motivational manipulations or by explaining the effect to subjects. Craik and Tulving (1975, Experiment 10) manipulated the amount of reward subjects received for remembering words, giving higher rewards to words learned under the shallow orienting tasks. They thus expected subjects to pay more attention to items under the shallow tasks, which should have led to retention comparable to that following semantic processing. Instead, the results again revealed the standard LOP effect. Chow, Currie and Craik (1978) explained the LOP effect to subjects before the study phase. Further, after a primary shallow orienting task, subjects either performed a semantic task on their own (create adjectives for each word) or performed a semantic task explicitly (whether the word fit in a sentence frame or not). The results revealed that explaining the LOP effect to subjects beforehand did not eliminate the LOP effect. The explicit semantic task following the shallow orienting task, on the other hand, eliminated the effect.

One possible reason the LOP effect does persist despite intentional learning instructions is that subjects may not be aware of the effects of the orienting tasks. That is, although intending to learn all items, they may fail to spend additional time and attention encoding shallowly processed items more deeply. For instance, they may think that an extra task on the material would lead to better retention or would have no effect regardless of the nature of the task. Another reason is that even if subjects attempt to use deep processing under shallow orienting tasks, they might have difficulties in switching from experimentally assigned shallow processing

to spontaneous deep processing. If so, manipulations that promote spontaneous deep processing can help them engage in deep processing even when performing shallow orienting tasks; hence, the LOP effect may be eliminated. In the current thesis, I investigated these possible explanations as to why the LOP effect occurs. Experiment 1 examined whether subjects were aware of the LOP effect using judgments of learning, whereas Experiment 2 addressed whether it is possible to promote spontaneous deep processing under shallow orienting tasks in the reversed order LOP paradigm. I will explain and review the reversed order paradigm and previous findings in the introduction to Experiment 2.

Chapter 2: Experiment 1

One way to investigate subjects' awareness of the LOP effect would be to ask subjects to make metamemory judgments about their learning. If subjects are not aware that different orienting tasks lead to different retention levels, their metamemory judgments should be similar across orienting tasks when their performances across orienting tasks show the LOP effect. Little prior research has asked this question, but a few studies are relevant. Seamon and Virostek (1978) demonstrated that subjects understood that certain orienting tasks led to greater processing than others. In their first experiment, they defined depth of processing as "amount of processing or degree of difficulty associated with each question" (p. 283), and asked subjects to rank 13 classification questions from low-to-high based on this definition. The questions varied from "Is it printed in capital letters?" to "Can you use it in a sentence 'A ___ fell down?'" The results revealed that subjects reliably ranked semantic questions as better and the agreement among people was high (coefficient of concordance was .60). In a second experiment, Seamon and Virostek showed that when a different sample of subjects were given the same classification questions to study words, their recall performance was correlated (.64) with subject-defined depth rankings from the previous experiment.

Even though Seamon and Virostek (1978) showed that subjects were able to differentiate between different amounts of processing in the classification questions, there are several limitations to their study. First, the relation between rankings and performance was correlational. In addition, measurements on depth rankings and recall performance came from two different groups of subjects. Thus, there is no way of knowing whether subjects would be aware of processing differences while they were studying the material. Finally, subjects were asked to

rank the depth of processing required for the questions, not to estimate future memory performance on the questions.

Only two LOP studies have experimentally examined subjects' predictions about their future performance on a memory test. Cutting (1975) asked half of his subjects to study 24 words with a shallow orienting task (i.e., checking letter E's in each word) and other half with a deep orienting task (i.e., rating the pleasantness of each word) under incidental learning instructions. At the end of the study phase, he asked subjects to rate, on a scale of 1 (very poor recall) to 10 (very good recall), how well they would perform if a free recall test for all the words occurred. Even though subjects recalled more of the deeply encoded items, their performance ratings did not differ between words studied under deep and shallow encoding. This indicated that subjects were not able to predict their future performance at an aggregate (or global) level in a between-subjects design.

Similarly, Shaw and Craik (1989) used subjects' predictions to assess monitoring differences among types of processing and age groups, but unlike Cutting (1975) they asked for item-by-item predictions and manipulated the types of processing within subjects. During the study phase, older and younger adults studied 60 words with either letter cues (e.g., "*starts with ic: ice*"), rhyme cues (e.g., "*rhymes with dice: ice*") or category cues (e.g., "*something slippery: ice*"). After each word, subjects were told to "rate their own memory" to predict the "likelihood of recalling a word they studied, given the cue" on a 10-point scale (p. 132). During the test phase, subjects saw the same cue for each word and were asked to recall the target word. The results revealed the standard LOP effect for actual performance (i.e., semantic cues led to better recall than rhyme cues which led to better recall than letter cues), but both older and younger adults were insensitive to performance changes amongst different types of cues. Subjects

overestimated their performance for letter cues and underestimated their performance for category cues. Although category and rhyme cues led to slightly higher predictions, Shaw and Craik (1989) concluded that subjects were largely insensitive to the LOP manipulation and that subjects' abilities to distinguish between recallable items from non-recallable items did not change based on the cue type.

In Experiment 1, I investigated the persistence of the LOP effect under intentional learning instructions, using a procedure similar to that of Shaw and Craik (1989): taking judgments of learning (JOL) after each item. JOLs are metamemory judgments about one's knowledge or future performance and are customarily assessed after each item is presented during the study phase (Nelson & Narens, 1990). JOLs tap into subjects' metacognitive monitoring about how well they think they have learned the item and the probability of remembering the item on a future test. Previous research on JOLs suggest that people are moderately accurate in predicting their future performances, and that JOLs can be diagnostic of the likelihood of remembering items in the future (Arbuckle & Cuddy, 1969; Dunlosky & Nelson, 1994). JOLs are derived from subjects' beliefs or theories about their learning, which in turn may be based on a variety of heuristic or deliberate inferences of (Koriat,1997; Koriat, Bjork, Sheffer, & Bar, 2004). For this reason, I used JOLs as a tool to explore subjects' beliefs about the LOP effect. In addition, Koriat et al. (2004) found that the retention interval impacted subjects' JOLs when the experiment had a within-subjects design (i.e., subjects heard about all of the retention intervals) whereas the retention interval did not have an effect on JOLs in between-subject designs. They concluded that JOLs were comparative in nature and captured relative memorability. Therefore, I used a within-subjects design to allow subjects to compare across different levels of processing.

Experiment 1 served as a partial replication to Shaw and Craik's (1989) study with some critical variations to their procedure. Shaw and Craik did not use orienting questions and their cues were always congruent with target words. Instead, I used the standard LOP paradigm, and thus Experiment 1 included three orienting tasks (i.e., questions) and *yes/no* answers. Further, the JOL instructions in the current study were more detailed than those of Shaw and Craik's and instead of cued recall, subjects took a recognition test. One possible explanation for Shaw and Craik's findings that JOLs did not predict the LOP effect is that subjects were in fact sensitive to the prediction 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 the conditions may have overshadowed their metamemory sensitivity to the LOP manipulation.

In his cue utilization approach to JOLs, Koriat (1997) accounted for JOLs' sensitivity differences by distinguishing between intrinsic and extrinsic cues. Intrinsic cues are related to items' a priori characteristics (e.g., frequency, processing fluency) whereas extrinsic cues are characteristics of the learning conditions that can affect overall performance (e.g., serial position, encoding operations). JOLs tend to be based only on intrinsic cues and are rarely (if at all) sensitive to extrinsic cues. This difference explains why JOLs sometimes fail to predict actual performance; both types of cue contribute to actual performance, not just intrinsic cues Koriat (1997) considered that the "extrinsic factors… include encoding operations performed by the learner such as levels of processing" (p. 350), and proposed that Shaw and Craik (1989) only found mild differences in JOLs because of this.

Nevertheless, I propose that the orienting tasks in a LOP manipulation could be considered as either extrinsic cues, because subjects perform encoding operations on the words, or as intrinsic cues, because the orienting questions are related to item characteristics such as

item's phonemic or semantic properties. Then, if subjects consider the relatedness of the orienting question to the word while giving JOLs, the orienting question should act as an intrinsic cue. For instance, when a pair of words are semantically related, the presentation of one of the words offers an intrinsic cue for the other. Dunlosky and Matvey (2001) revealed that related word pairs led to higher JOLs compared to unrelated word pairs, and relatedness also improved performance on the recall test. On the contrary, in Shaw and Craik (1989), even though the category cues were semantically related cues to the targets, subjects' predictions for category pairs were similar to those for rhyme and letter pairs.

Dunlosky and Matvey (2001) and Shaw and Craik (1989) taken together present a contradiction; on the one hand, relatedness between a cue and a target seems to predict both higher JOLs and performance relative to unrelated pairs. On the other hand, it does not seem to be the case that JOLs are reliably sensitive to the LOP manipulation. This quandary reveals that the process by which individuals generate their JOLs matters. That is, JOLs should predict later recall if they are derived from judgments of the intrinsic relations between cues and cued material. In the current study, words were not presented in pairs, yet, the orienting questions could still serve as cues while making JOLs. In this case, the category questions should be the most related cues, whereas the case questions should be the least related cues given they are not item-specific. If different orienting tasks lead subjects to attend to these intrinsic relations, then I would expect the magnitude of JOLs to mirror the LOP effect.

Furthermore, the standard LOP paradigm employs *yes/no* questions as orienting task questions and previous results consistently showed that *yes* questions led to higher retention than *no* questions (Craik, 1977; Craik & Tulving, 1975). Given that the target word is inherently congruent with the presented orienting question or not, I predicted that *yes/no* question type

would serve as an intrinsic cue. Thus, if *yes/no* retention differences were present, JOLs would also reflect these differences.

Another potential issue that merits discussion is that making JOLs might produce a reactive effect. That is, forcing subjects to monitor their learning through JOLs might influence the learning process itself and can affect retention under different orienting tasks. For instance, Mitchum, Kelley and Fox (2016, Experiment 5) found that subjects who made JOLs showed larger recall difference between related (easy) items and unrelated (difficult) items compared to subjects who did not make JOLs. Thus, they concluded that subjects who made JOLs became aware that some items were not as memorable and changed their mastery goal to a more pragmatic one. Similarly, Soderstrom, Clark, Halamish, and Bjork (2015) examined whether giving JOLs affected the process of learning itself while controlling for study time confounds. During the study phase, subjects studied word pairs with different semantic relatedness and half of them gave item-by-item JOLs whereas the other half did not. Both groups saw the pairs for 8-sec and the JOL group was prompted to give JOLs halfway through the presentation (i.e., after 4-sec). Later both groups took a cued recall test. Their results indicated that JOLs enhanced retention only for the targets of strongly related pairs (*loaf* - *bread*) but not for targets of weakly related (*mercy* - *justice*) or unrelated pairs (*sack* - *flag*). Both of these studies used related and unrelated word pairs as material and in both cases JOLs enhanced retention for related pairs, leading to conclusion that JOLs further strengthened the cue – target relation only when the relation was already strong.

In their second experiment, Soderstrom et al. (2015) presented related word pairs to subjects and had them either read the intact pair or generate the target from the cue. In addition, half of the subjects gave JOLs for read word pairs whereas the other half did not give JOLs. The

results revealed that, for the JOL condition, the generation effect was attenuated because retention for read word pairs increased compared to the no-JOL condition. It is important to note that the subjects in this experiment did not give JOLs for the generated pairs, whereas in current experiments subjects gave JOLs for every item, including those in the category orienting task. Another major difference is that, instead of word pairs, I used single words as the study material. Other studies using recall of single word lists did not find differences between groups who gave JOLs and did not give JOLs. For instance, in their Experiment 2A, Benjamin, Bjork and Schwartz (1998) presented subjects with word lists during learning and gave them an immediate recall test and a final recall test. They asked half of the subjects to make JOLs for each recalled item during the immediate free recall test and examined whether giving JOLs altered initial or final recall performance and concluded that giving JOLs did not affect recall performance. Therefore, it is not clear whether JOLs would have any impact on retention. To explore this possibility, Experiment 1 included a control group without JOLs.

Experiment 1 examined the persistence of the LOP effect under intentional learning using the standard LOP paradigm. Adding JOLs to the standard procedure allowed us to answer the following questions: 1) Do people have insight into their learning processes under various orienting tasks? and 2) Do people have awareness that *yes/no* responses might lead to different levels of retention? 3) Do JOLs enhance retention and attenuate the LOP effect? Answers to these questions would provide a first step towards a possible explanation for the empirical results under intentional learning instructions. All subjects were given intentional learning instructions and studied the material under one of following conditions: 1) giving JOLs after each study trial, 2) having a delay after each study trial to account for the time confound created by JOLs, 3) no JOLs or delay after study trials. If subjects were not aware of the effects of orienting tasks on

their memory, they would have similar JOLs across different orienting tasks, and this would explain the discordance between the LOP framework and the empirical findings.

2.1 Method

2.1.1 Subjects

An a priori power analysis was conducted through G*Power software to determine a sufficient sample size using a large effect size (f = 0.4), an alpha of 0.05, and a power of 0.80. Based on this, the required total sample size was 72, 24 subjects in each condition. In Amazon's Mechanical Turk (MTurk), it was not possible to control for what subjects were doing during the experiment. To account for this uncontrolled nature of MTurk, the sample size was increased by 75%, gathering 42 subjects in each condition. Participation was restricted to people who were located in United States and who had high completion and performance rates in other studies (>90%). In total, 143 subjects (M = 35.5, SD = 9.51) were recruited from MTurk. Seventeen of them were replaced due to one of the following reasons: English was not their native language (n = 1), making phone calls during the experiment (n = 2), they requested to be excluded (n = 3) or they were outliers (n = 11). Outlier criteria will be discussed in more detail below. Thus, the remaining 126 subjects were randomly assigned to one of three conditions, with 42 subjects in each condition. The study was approved by the Washington University IRB.

2.1.2 Materials

Sixty concrete nouns were selected from norms collected by Van Overschelde, Rawson and Dunlosky (2004) and Nelson, McEvoy and Schreiber (2004), controlling for concreteness (Brysbaert, Warriner, & Kuperman, 2014; Nelson et al., 2004) and for frequency (Balota et al., 2007). The words had a concreteness level above 3.5 out of 7 according to Nelson et al. (2004) and above 2.5 out of 5 according to Brysbaert et al. (2014). The logarithm of HAL frequency in

the English Lexicon Project (Balota et al., 2007) ranged from 6.75 to 11.79. Words consisted of four to nine letters. For each word a rhyming noun was generated to be used in the rhyme orienting task condition. For example, a target word might be *apple* and its rhyme could be *chapel.* Materials are in Appendix A.

Six questions were created for each word, pertaining to the word's physical appearance (a letter case judgment), its sound (a rhyme judgment) or its meaning (a semantic judgment). Thus, each word had two questions from three types of questions (case, rhyme, category), one with a *yes* response and one with a *no* response. For example, for *yes* responses, the word *apple* had a case question, "Is the word in lowercase letters?"; a rhyme question, "Does the word rhyme with CHAPEL?"; and a category question, "Is the word a type of fruit?". For *no* responses, the questions were "Is the word in uppercase letters?"; "Does the word rhyme with SLAM?"; "Is the word a type of measurement?" for the case, rhyme and category questions, respectively. One of the six questions served as the orienting question during the study phase.

The test phase consisted of an old/new recognition test, with 180 words (60 targets, 120 lures), each presented separately. The large number of lures served to eliminate possible ceiling effects and to follow the procedure used by Craik and Tulving (1975). A president recognition test (Roediger & DeSoto, 2016), which lasted about 10 minutes, was used as a filler task between the study phase and the test phase, to clear short-term memory and reduce recognition performance. Briefly, 41 presidents were intermixed with 82 lures (mostly names of former vice presidents and other famous Americans). Subjects gave a *yes/no* answer regarding whether or not the person was a U.S. president and indicated their confidence level for each answer, on a 100-point scale.

2.1.3 Design

A 3 x 2 x 3 mixed factorial design was used, such that question type (case, rhyme, category) and response type (*yes, no*) were manipulated within-subjects. The study phase (Intentional, Delay, and JOL) was manipulated between-subjects. Studied material was same for all three groups, with 10 words comprising each of the six within-subject conditions (three levels x *yes-no*). Each word appeared in each of the six within-subjects conditions equally across subjects, hence, every six subjects represented a counterbalanced iteration of the experiment.

2.1.4 Procedure

Subjects were tested using Collector (Garcia & Kornell, n.d.) through MTurk and were randomly assigned to conditions. All three conditions employed intentional learning instructions: "Please try to answer each question as quickly as possible, but you should also make sure your answer is correct. In addition, we will be testing your memory for the words on a later test, so try to learn each word in addition to making a response." In addition to this instruction, subjects in the JOL condition had the following instruction: "After you make a response, you will be asked to rate how likely is it that you will remember the word you just studied during a later recognition memory test where you will be asked to indicate whether the presented word is OLD (i.e., you saw it during study phase) or NEW (i.e., you did not see it during study phase). Please enter a number between 1 and 5, where 1 indicates *I would definitely not recognize the word* and 5 indicates *I would definitely recognize the word*".

Figure 2.1a shows an outline for study trials for each condition. The initial part of each study trial was identical for all three conditions: The encoding question (case, rhyme or category) was presented first for 3-sec, and a delay of 2-sec followed. After the delay, the target word was presented for 2-sec. Subjects then had 3-sec to make a *yes/no* response for the orienting question,

thus totaling to 10-sec per study trial. This part of the study trial constituted a standard LOP experiment, and the study phase of the Intentional condition consisted of 60 standard trials. Additionally, subjects in the JOL condition gave a self-paced judgment of learning rating for each word after making a *yes/no* response, which lasted about 4 sec. In a third condition, to control for a possible time confound in the JOL condition, subjects in the Delay condition were given a 4-sec unfilled delay after making a *yes/no* response, leading to 14-sec per study trial. Prior research has shown that time allocated to a word in the standard LOP paradigm does not affect the LOP effect (Craik & Tulving, Experiment 5, 1975; Walsh & Jenkins, 1973), so the assumption is that any differences in later recognition performance between the JOL condition and the Intentional condition and between the JOL condition and the Delay condition would be due to making JOLs.

After the study phase, subjects completed the 10-minute president recognition test before beginning the test phase for the words. For the test phase, subjects responded as to whether or not they saw the presented word during the study phase by clicking OLD or NEW on the screen. After making a recognition decision, they made a confidence judgment about their response on a 5-point scale, in which 1 indicated *Not confident at all* and 5 indicated *Definitely confident*. The recognition test was self-paced. After completing this procedure for all 180 words (60 targets, 120 lures), subjects completed a final questionnaire about the experiment in which they reported 1) any problems, 2) what else they may have been doing during the experiment, and 3) whether they would like to be excluded from the study. The experiment lasted 60 to 90 minutes, depending on subject's pace.

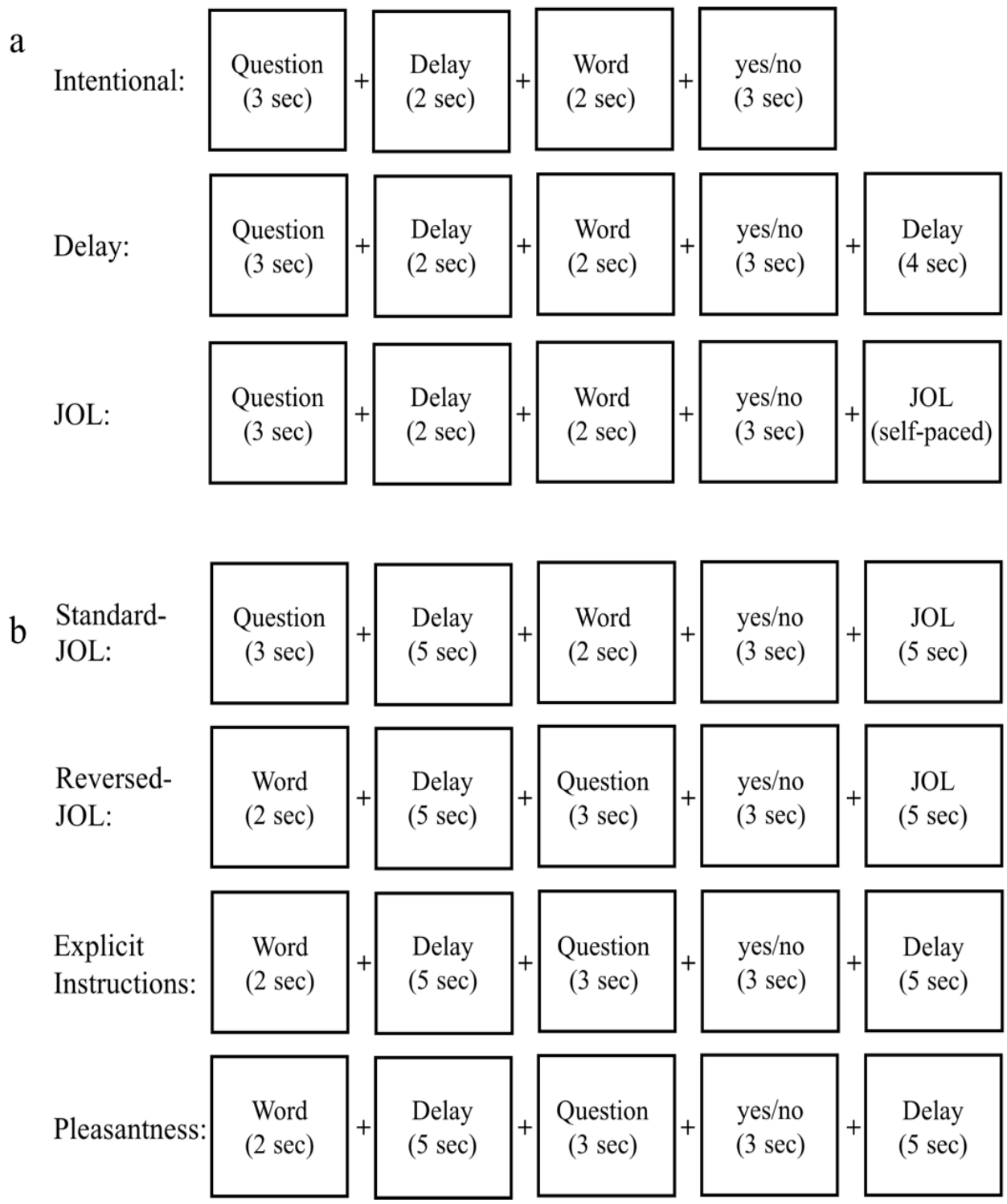

Figure 2.1 Study trials for each condition in Experiment 1 (a). Study trials for each condition in Experiment 2 (b).

2.2 Results

In a preliminary analysis, hit rates, correct *yes*/*no* responses, and average reaction times for recognition decisions, confidence judgments and judgments of learning (for the JOL group) were calculated for each subject. Number of correct *yes*/*no* responses during the study phase was calculated to see whether subjects were paying attention to orienting questions and the words. Outliers were detected based on 3 standard deviations (SD) above and below the sample average for reaction times and 3 STD below the sample average for hit rates and correct *yes*/*no* responses[1]. Eleven subjects were detected as outliers on at least one measure and were replaced First I consider the recognition results of all three groups, then I turn to each group separately for purposes of clarity. The results from confidence judgments are reported in Appendix B because they are not the primary interest of the study.

The upper part of Table 2.1 provides overall hit rates, false alarm rates, corrected recognition scores (hits minus false alarms) and *d'* scores[2] for each group. Corrected recognition scores were used as the primary dependent variable for analysis to take false alarms into account.[3]

[1] Number of correct *yes*/*no* responses for five of 11 outliers were lower than the sample average for following reasons: not responding to the orienting question in 3-sec and/or incorrectly responding to them. 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 either their recognition decisions' or confidence judgments' reaction times. Looking at each subject's trials individually revealed that some of these responses had reactions times around as long as 5-min. Given that it is not possible to control for what subjects were doing during that period in MTurk, these subjects were replaced.

[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 be able 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 and corrected hit rate was calculated after this correction (59.5/60), and for perfect false alarm rates, half a false alarm was added to perfect false alarm of 0 and corrected false alarm rate was calculated after this correction (.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 many subjects (N = 64) had perfect hit rates of 1.00 at (at least) one of response type x 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.

Measures	Hits		False Alarms		Hits – False Alarms		*d'*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Group	Experiment 1							
Intentional	.69	.17	.29	.19	.41	.23	1.31	.92
Delay	.69	.17	.31	.15	.38	.19	1.13	.66
JOL	.77	.16	.23	.15	.54	.27	1.75	1.09
	Experiment 2							
Standard-JOL	.81	.12	.16	.15	.65	.20	2.14	.86
Reversed-JOL	.88	.10	.16	.11	.72	.17	2.50	.90
Pleasantness	.90	.11	.17	.15	.73	.21	2.67	1.07
Explicit Instructions	.83	.13	.18	.14	.66	.20	2.26	1.01

Table 2.1 Overall Hit, False Alarm Rates and Corrected Recognition Scores for Experiment 1 and 2

Figure 2.2 shows that corrected recognition scores (Hits – False alarms) increased across three orienting tasks, with the JOL group having highest scores. To test whether these differences were statistically significant, a 2 (response type) x 3 (orienting task) x 3 (group) repeated measures ANOVA was conducted. There was a main effect of group, $F(2,123) = 5.85$, $p = .004$, $\eta^2_p = .09$. Overall, subjects in the JOL condition had higher recognition scores than subjects in the Delay and Intentional conditions, ($p = .005$, $p = .032$, respectively). A main effect of orienting tasks revealed that, collapsed across all groups, the LOP effect was replicated, $F(1.89,232.68) = 91.62$, $p < .001$, $\eta^2_p = .43$. The case task led to lowest performance ($M = .33$, $SE = .02$), the category task led to highest performance ($M = .54$, $SE = .02$), and the rhyme task was intermediate ($M = .45$, $SE = .02$). Also, there was a main effect of response type, $F(1,123) = 45.60$, $p < .001$, $\eta^2_p = .27$: *yes* responses ($M = .48$, $SE = .02$) resulted in higher corrected recognition scores than *no* responses ($M = .41$, $SE = .02$), a standard and recurring finding in the LOP literature. In addition, the orienting tasks x group interaction was marginally reliable, $F(3.78,232.68) = 2.15$, $p = .079$, $\eta^2_p = .03$. For the case task, the JOL group had significantly

higher recognition scores than the Intentional group, p = .011, and the Delay group, p = .001, and at the rhyme and category tasks the JOL group had significantly higher recognition scores than the Delay group (p = .042, p = .048, respectively). Furthermore, the difference scores between the case and rhyme tasks and between the case and category tasks were higher for the Delay group compared to the JOL group, (p = .064, p = .045, respectively). Thus, making JOLs during the study attenuated standard retention difference between the two tasks, differences that are observed in Intentional and Delay groups.

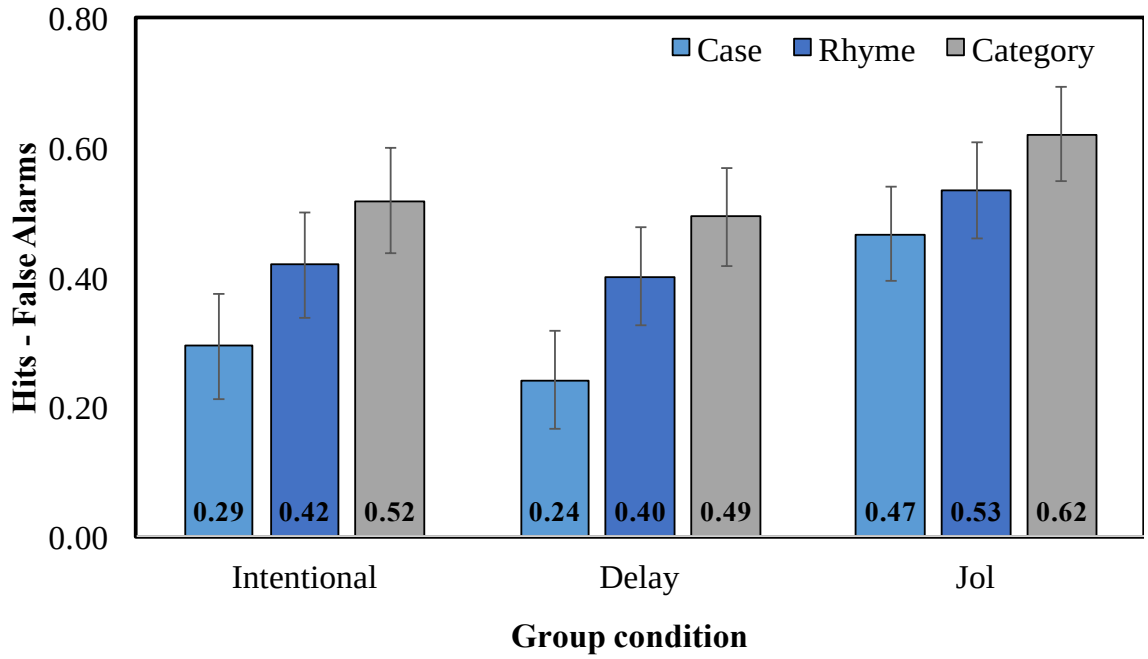

Figure 2.2 Corrected recognition scores across orienting tasks for each condition in Experiment 1. Error bars indicate 95% confidence intervals.

Figure 2.3 shows corrected recognition scores for both *yes* and *no* responses demonstrated the LOP effect. The orienting tasks x response type interaction was reliable, $F(2,246) = 8.13$, $p < .001$, $\eta^2_p = .06$, showing similar recognition scores of *yes* and *no* responses at the case task, p = .323, whereas at other two tasks *yes* responses led to significantly higher recognition scores than no responses, ps < .001. Other interactions were not statistically

significant (*p*s > .05).

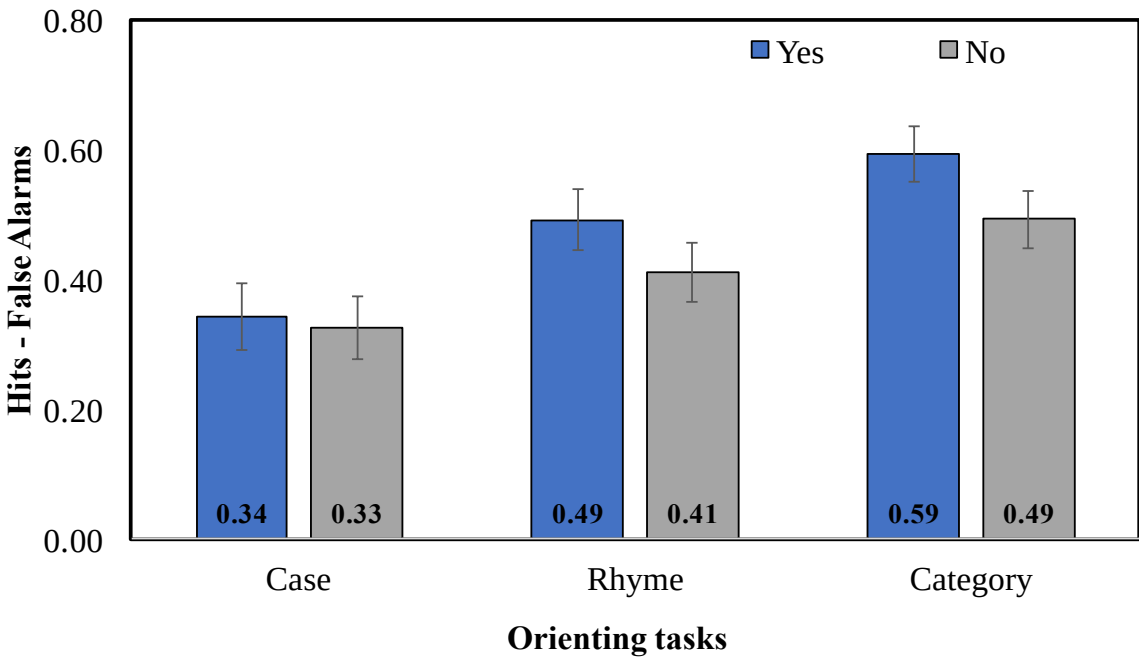

Figure 2.3 Corrected recognition scores for yes/no responses across orienting tasks for Experiment 1. Error bars indicate 95% confidence intervals.

2.2.1 JOL condition

The results within the JOL condition will be discussed first because they are of primary interest. Figure 2.4a shows corrected recognition scores collapsed across different orienting tasks and response types for the JOL group, and Figure 5b shows corresponding JOL ratings that illustrates a slight increase across orienting tasks, mostly for *yes* responses. For corrected recognition scores, pairwise comparisons revealed that the category task led to significantly higher recognition scores compared to the rhyme task, $p = .001$, and to the case task, $p < .001$, whereas the case task and the rhyme task only differed marginally, $p = .052$, with means of .47 for the case and .53 for the rhyme tasks. Thus, JOLs eliminated part of the standard LOP effect.

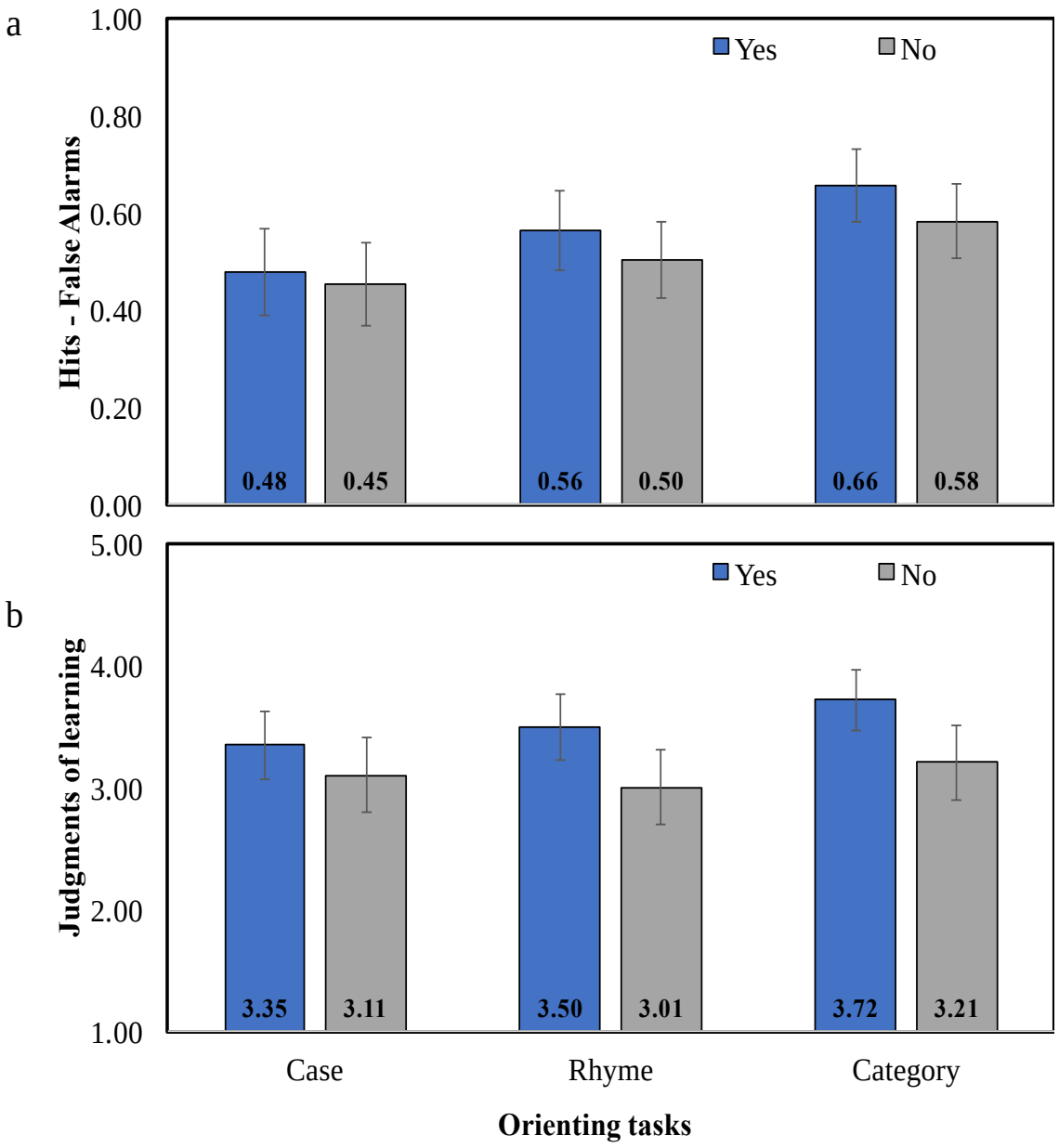

Figure 2.4 Corrected recognition scores across orienting tasks and yes/no responses for the JOL group (a). Judgments of learning across orienting tasks and yes/no responses for the JOL group (b). Error bars indicate 95% confidence intervals.

A 2 (response type) x 3 (orienting tasks) repeated measures ANOVA was conducted for JOLs (Figure 2.4b). There was a main effect of orienting tasks, $F(2,82) = 10.43$, $p < .001$, $\eta^2_p = .20$. Pairwise comparisons revealed that the category task ($M = 3.46$, $SE = .13$) led to higher JOL ratings compared to the case ($M = 3.23$, $SE = .14$) and rhyme tasks ($M = 3.25$, $SE = .14$), ps = .001. However, unlike corrected recognition scores, JOL ratings at the case task did not differ

significantly from JOL ratings at the rhyme task, $p = .951$; although of marginal significance, the recognition score at the case task was lower than the recognition score at the rhyme task. These results indicated that subjects did understand the LOP effect slightly (by giving higher JOL ratings for the category task), but did not accurately predict the LOP effect or the magnitude of the effect.

A main effect of response type in JOL scores revealed that JOL ratings for *yes/no* response followed a similar pattern with corrected recognition scores, $F(1,41) = 52.58$, $p < .001$, $\eta^2_p = .56$. *Yes* responses ($M = 3.52$, $SE = .12$) led to higher JOL ratings than *no* responses ($M = 3.11$, $SE = .15$). The interaction was also reliable, $F(1.64,67.40) = 3.76$, $p = .036$, $\eta^2_p = .08$. For *yes* responses, the category task led to higher JOL ratings than the case task and the rhyme task, ps $< .05$, yet, the case and rhyme tasks did not differ, $p = .204$. For *no* responses, JOL ratings at the category task were only significantly higher than JOL ratings at the rhyme task, $p = .008$. Other pairwise comparisons were not significant, ps $> .05$. As can be observed in Figure 5b, the effect in JOLs was carried mainly by *yes* responses; subjects showed no awareness of the retention difference between case and category tasks for *no* responses.

Another way to compare subjects' actual performance with their predictions is to examine whether subjects' JOL ratings reflected subjects' hit rates within given orienting task. In other words, were subjects' estimations accurate for each orienting task? For this comparison, each subject's hit rates and JOL ratings were transformed into z-scores within each subject. For each subject, z-scores of hit rates and z-scores of JOL ratings were calculated for all three orienting tasks (totaling to six z-scores per subject) using subject's aggregate hit rate and JOL rating averages. Figure 2.5 shows transformed hit rates and JOL ratings at each orienting tasks: Hit rates are indeed steeper than JOL ratings. A 2 (measure) x 3 (orienting tasks) repeated

measures ANOVA revealed a main effect of orienting tasks, $F(2,82) = 28.15$, $p < .001$, $\eta^2_p = .41$, and no main effect of measure, $F(1,41) = .80$, $p = .377$, $\eta^2_p = .02$. The interaction was marginally reliable, $F(2,82) = 2.60$, $p = .080$, $\eta^2_p = .06$, and driven mostly by differences between hit rates and JOL ratings at the case task. In the case task, subjects overestimated their future learning (JOL ratings were higher than hit rates), $p = .043$.

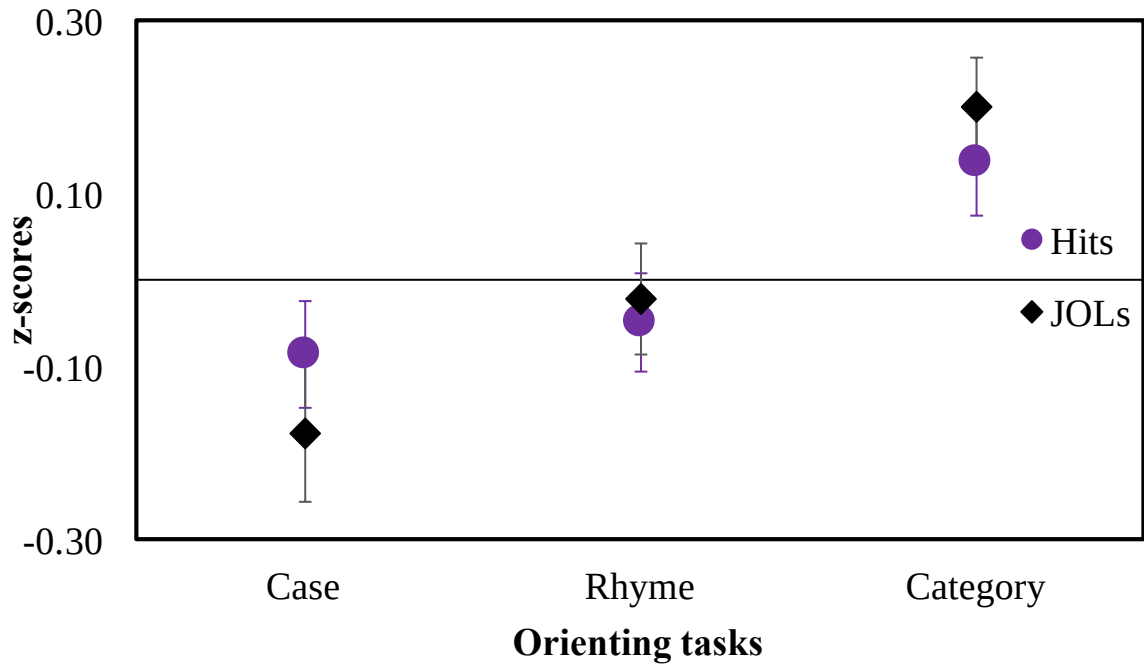

Figure 2.5 Change in z-scores for hit rates and JOLs across orienting tasks for the JOL group. Error bars indicate 95% confidence intervals.

Besides looking at the correspondence between mean JOLs and mean hit rates across the orienting tasks, another way to assess predictive accuracy of JOLs is by looking at resolution or relative accuracy of JOLs. Relative accuracy refers to how accurate JOLs are at discriminating between items that are more likely to be recalled and items that are less likely to be recalled on a later memory test. Through relative accuracy, I tested whether subjects could predict which items they were going to recognize (discriminating between hits and misses) and whether this ability changed based on the orienting task. Because gamma correlations are the most common measure

of resolution in the metamemory literature, I also used gamma correlations (Nelson, 1984). They provide a measure of variability present in memory performance that can be explained by variability in JOLs, in other words, the degree that JOLs can predict performance.

Four different gamma correlations were computed for each subject: one for items under the case task, one for items under the rhyme task, one for items under the category task and one for all items. The upper part of Table 2 reports averages across subjects for each correlation. First, the analyses revealed that the relation between JOL ratings and hit rates were greater than 0 for all items, $t(39) = 3.20$, $p = .003$, suggesting subjects were able to predict which items they were going to recognize. However, for orienting tasks the relation was only greater than 0 for the items under the rhyme task, $t(34) = 2.08$, $p = .045$, and not for the items under the case task or the category task, ps > .05. Second, the values did not differ from one another in terms of resolution (or predicting recognition). One-way repeated measures ANOVA ($N = 24$) amongst four averages revealed that there was no main effect of correlation type, $F(2,69) = 1.25$, $p = .296$, $\eta^2_p = .05$. Thus, subjects were overall sensitive to interitem differences in recognition, but their sensitivity was not affected by the LOP manipulation.

Orienting Tasks	Case		Rhyme		Category		All Items	
Group	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
JOL (E-1)	0.14	0.53	0.21*	0.60	0.13	0.59	0.22**	0.43
Standard-JOL (E-2)	0.17*	0.44	0.28**	0.44	0.28*	0.61	0.29**	0.35
Reversed-JOL (E-2)	0.30**	0.57	0.20†	0.56	0.34**	0.64	0.27**	0.45

Note: † $p < .10$, * $p < .05$, ** $p < .01$

Table 2.2 Mean Gamma Correlations and Standard Deviations from Experiment 1 and 2

2.2.2 Intentional and Delay conditions

Figure 3a shows corrected recognition scores for the Intentional and the Delay groups. Within

each group, corrected recognition scores for each orienting task differed from one another significantly, showing the standard LOP effect, *p*s < .001. The two groups did not differ from one another on any tasks, meaning that the extra 4-sec at the end of each trial did not improve performance for the Delay group, confirming results of Craik and Tulving (1975).

2.3 Discussion

Experiment 1 replicated the robust LOP effect under intentional instructions for the Intentional and the Delay conditions. The main interest was the findings from the JOL condition that answered three questions raised by Experiment 1. First, subjects in the JOL condition were somewhat (but not fully) aware of the retention differences across orienting tasks. Overall, they correctly predicted that the category task would lead to higher performance than the case and rhyme tasks, but they did not predict the (marginal) difference between the case and the rhyme tasks. More specifically, however, this effect was only present for *yes* responses. For *no* responses, predictions of the case and the category tasks did not differ from one another. Furthermore, subjects overestimated their performance under the case task, showing inaccurate insight. Interestingly, though, subjects predicted that *yes* responses would to better retention than *no* responses for all orienting tasks. Thus, they were aware that they would remember words that were congruent with the orienting question better.

Lastly, providing JOLs enhanced retention for the shallow tasks compared to the shallow tasks of the Intentional and Delay conditions. Thus, JOLs promoted more effective processing on the shallow tasks by forcing subjects to monitor for their future learning. Even though the act of monitoring did not eliminate the LOP effect, it diminished it. This finding is in line with Soderstrom et al.'s (2015) finding that giving JOLs enhanced retention of read word pairs, and this in return attenuated the generation effect. In short, the results from Experiment 1 suggested

that not having accurate insight to the LOP effect might be one explanation as to why the effect occurs under intentional learning instructions, but it also suggested that giving JOLs promoted semantic processing to some extent for the shallow orienting tasks.

Chapter 3: Experiment 2

Experiment 2 explored another possible explanation for the LOP effect under intentional learning instructions: Subjects might find it difficult to switch from the demands of shallow processing to spontaneous deep processing that occurs in control groups. If this is the case, allowing subjects to engage in their own natural deep processing before the experimenter provided shallow processing might eliminate the LOP effect. Thus, the aim of Experiment 2 was to investigate whether it is possible to promote natural deep processing under shallow orienting tasks. With a similar goal, Craik (1977) developed a variation of the LOP paradigm, called the reversed order or the word-first paradigm, as a way to induce subjects into performing better at the shallow orienting tasks by promoting semantic processing of the words. In the standard LOP paradigm, words are always preceded by the orienting task questions, and subjects are required to give an answer to the question after seeing the word. In the reversed order paradigm, Craik first presented the word during the study phase, followed by a delay and then the orienting question, which subjects then answered. The basic idea was that without the orienting task preceding the word and with the relatively long delay (5-sec) following it, the word should always be processed meaningfully (i.e., at a deep level) as in control groups without orienting tasks. He compared the reversed order and the standard paradigms under both incidental and intentional learning conditions, using case, rhyme and category questions as orienting tasks.

Figure 3.1 reveals the results: Although retention following the shallower tasks increased in the reversed order paradigm compared to the standard LOP paradigm, it was still lower than retention following a deep orienting task. Thus, even when the words were presented before the questions under intentional learning conditions, subjects given the shallow orienting tasks did not engage in spontaneous deep processing as control groups naturally did. Basically, under the

reversed order paradigm, the LOP effects still emerged, conflicting with the original LOP framework. Craik described his results as "mystifying" (p. 689). Yet, similar to the results of Experiment 1, Craik's (1977) results showed that under the reversed order paradigm the LOP effect diminished amongst orienting tasks due to increased retention at the shallow orienting tasks, hinting at some semantic processing during these tasks. Combining the results from the reversed order paradigm and Experiment 1 of the current thesis, Experiment 2 explored whether it is possible to fully attenuate the LOP effect by forcing subjects to monitor via JOLs and promoting spontaneous deep processing via the reversed order paradigm.

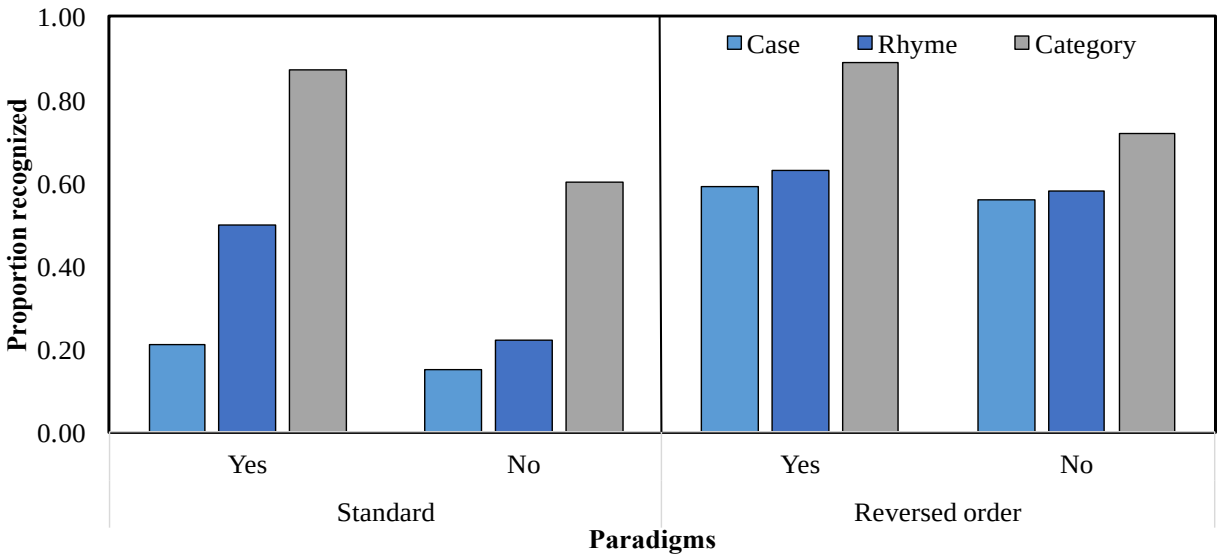

Figure 3.1 The LOP effect for the reversed order paradigm (left panel) and for standard paradigm (right panel) under the intentional learning instructions. Estimates are from Craik's (1977) Experiment 2, Figure 2.

Further, Experiment 2 tested several possible explanations as to why the LOP effect was not eliminated in the reversed order paradigm in Craik (1977). First, similar to the reasoning in Experiment 1, subjects might not have insight about their learning under different orienting tasks, and hence do not allocate enough resources to encode the material at the shallow levels. Second, subjects might process the word deeply as control groups would do, but the following shallow orienting questions might be impairing performance or deflecting attention, resulting in

the LOP effect. For instance, Cermak, Schnorr, Buschke and Atkinson (1970) gave subjects word pairs to study and asked them to concentrate on either the dictionary meaning of the pairs, the sound of the pairs, or both the sound and dictionary meaning. Their results revealed that subjects in the semantic group had higher performance on the following recognition test compared to the phonetic group but also compared to the semantic *and* phonetic group. Clearly, concentrating on the sound of words impaired recognition performance even when the instructions included focusing on the meaning along with the sound. Similarly, under the reversed order paradigm, the rhyme or the case tasks might decrease attention from the meaning of the word, hence inducing the LOP effect. If this is the case, I would expect the shallower tasks to impair performance even after the word was processed semantically.

Lastly, even under intentional learning instructions, subjects might not be processing or attending to the presented word in the following delay period because they know that an orienting question will follow. This knowledge might stop them from processing the word deeply as expected under intentional learning instructions without an orienting task. A possible way to account for this is to directly encourage subjects to study the items with their idiosyncratic strategies during the delay period. McDaniel and Kearney (1984) showed that when subjects were not instructed to use a certain strategy (i.e., unstructured strategy), they were able to adapt their strategies to the task demands and perform similarly to or higher than groups with structured strategies. Therefore, for the reversed order paradigm, explicitly instructing subjects to study the words during the delay without any specific strategy might promote meaningful processing. It would also eliminate the possibility of waiting for the orienting question without attending to the word. To test these potential explanations, as in Experiment 1, I added JOLs to the procedure to investigate how the reversed order paradigm affected subjects' learning

judgments. In addition, to ensure semantic processing of the words, I replaced the delay period following each word with pleasantness ratings or explicitly asked subjects to study the presented word with their own idiosyncratic strategy during the delay period.

In Experiment 2, I aimed to answer the following questions: 1) Do JOLs combined with the reversed order paradigm attenuate the LOP effect? 2) Do JOLs reflect actual performance under the reversed order paradigm? 3) Do shallower levels of processing impair performance by interfering with semantic processing? 4) Do explicit instructions to study the words eliminate the LOP effect under the reversed order paradigm? All subjects were given intentional learning instructions and studied the material under one of following conditions: 1) under the standard LOP paradigm with JOLs; 2) under the reversed order paradigm with JOLs; 3) under the reversed order paradigm with pleasantness ratings; and 4) under the reversed order paradigm with explicit study instructions. As in Experiment 1, I predicted that if subjects are not aware of the effects of orienting tasks on their memory, then the magnitude of JOLs will be comparable across different orienting tasks. Further, I examined whether shallow tasks impaired previous semantic processing by introducing a deep processing task (pleasantness ratings) before shallow tasks. Lastly, I aimed to eliminate the possibility that subjects might not attend to the word during the delay by encouraging them to study the material during delay.

3.1 Method

3.1.1 Subjects

An a priori power analysis through G*Power software using an effect size of 0.4, an alpha of 0.05, and a power of 0.80 revealed that a sufficient sample size is 96, or 24/condition. As in Experiment 1, the sample size was increased by 75% to 42 subjects per each condition. Thus, 168 subjects (M = 37.34, SD = 11.09) were recruited from MTurk. Participation criteria were

same as Experiment 1. Fourteen subjects[1] were replaced because they were identified as outliers in one or more of the outlier criteria. Ten subjects were replaced due to one of following answers in the final questionnaire: taking breaks (n = 2), writing the words down (n = 1), reporting an experiment error (n = 1), or restarting the experiment after internet connectivity problem (n = 6). With replacements, a total of 192 subjects were recruited through MTurk. Subjects were randomly assigned to one of the four conditions, with 42 subjects in each condition.

3.1.2 Materials

The same materials were used in Experiment 2 as in Experiment 1.

3.1.3 Design

A 3 x 2 x 4 mixed factorial design was used, such that question type (case, rhyme, category) and response type (*yes*, *no*) were manipulated within-subjects and the type of study phase (Standard-JOL, Reversed-JOL, Pleasantness, Explicit Instructions) was manipulated between-subjects.

3.1.4 Procedure

Subjects were randomly assigned to one of the four between-subjects conditions. All conditions had same intentional learning instructions as in Experiment 1. Subjects in the Standard-JOL and the Reversed-JOL conditions had the additional JOL instructions as in Experiment 1. Unlike Experiment 1, in Experiment 2, subjects had 5-sec to make JOL ratings, hence the instructions also emphasized giving a rating within the 5-sec window. Subjects in the Pleasantness condition had additional instructions about pleasantness ratings: "After you see a word, you will be asked to rate how pleasant that word is to you on a scale of 1 to 5. Please enter a number between 1 and

[1] As in Experiment 1, number of correct *yes*/*no* responses, hit rates, and reaction times for recognition decisions and confidence judgments were used as measures to detect outliers. An additional measure of JOL scores (for the Standard-JOL group and the Reversed-JOL group) was calculated in Experiment 2 because JOL ratings were experimenter-paced (5-sec), and some of the subjects did not give JOL ratings within that time period. JOL scores were the total number of JOL ratings (out of 60) a subject gave during the study phase. As with correct *yes*/*no* responses and hit rates, if a subjects' JOL score was 3 STD below the sample average, subjects were detected as outliers. Four of 14 subjects were detected as outliers based on their JOL scores, one based on number of correct *yes*/*no* responses, two based on their hit rates and remaining seven based on their average reaction times.

5, where 1 indicates *Not pleasant at all* and 5 indicates *Very pleasant.* You will have 5 seconds to give each rating." Lastly, based on McDaniel and Kearney's (1984) "uninstructed" condition, subjects in the Explicit Instructions condition had following instructions: "After you see a word, you will be asked to study that word for 5 seconds. During this period, try hard to study the previously presented word using any strategy you think would be useful."

For Experiment 2, I adopted Craik's (1977) procedure with addition of JOLs and pleasantness ratings. Figure 2b shows study trials for each condition. They were as follows: For the Standard-JOL condition, the orienting question (case, rhyme or category) was presented first for 3-sec, and a delay of 5-sec followed. After the delay the target word was presented for 2-sec. Subjects then had 3-sec to make a *yes/no* response for the question and 5-sec to give a JOL rating, totaling to 18-sec per study trial. For the Reversed-JOL condition, the target word was presented at the beginning of the trial, and after a 5-sec delay, subjects saw the orienting question. Besides, the change of order of the word and the question, the trial was identical to the Standard-JOL condition. For the Explicit Instructions condition, the target word was presented at the beginning of the trial for 2-sec and then subjects were instructed to study the previously presented word during subsequent 5-sec period using their idiosyncratic strategies. After the 5-sec period, the orienting question was presented for 3-sec and was followed by 3-sec *yes/no* response. At the end of the trial, there was a 5-sec delay to equate trial time to the Standard-JOL and the Reversed-JOL conditions. For the Pleasantness condition, procedures were identical to the Explicit Instructions condition, except instead of studying the target word during 5-sec period, subjects gave a pleasantness rating. Thus, for all conditions, a study trial lasted for 18-sec and every subject spent equal time, 10-sec, on a given trial before answering the orienting question.

After the study phase, the filler task and the test phase followed. These parts were identical to Experiment 1. After completing the recognition test, subjects again completed a final questionnaire, this time with an additional question: "Please explain the strategy or strategies you used during the study time after words" was presented for the Explicit Instructions group and "Did you use any strategy to study the words? If so, can you explain your strategy" for the remaining groups.

3.2 Results

The lower part of Table 2.1 provides overall hit rates, false alarm rates, corrected recognition scores (hits – false alarms) and *d'* scores[2] for each group in Experiment 2. As with Experiment 1, corrected recognition scores were used as dependent variable[3].

Figure 3.2 shows that besides the Pleasantness group, each group shows the standard LOP pattern, but to a lesser extent compared to data from Experiment 1. A 2 (response type) x 3 (orienting tasks) x 4 (group) repeated measures ANOVA was conducted for corrected recognition scores. There was a main effect of orienting tasks, $F(1.93,315.75) = 38.62$, $p < .001$, $\eta^2_p = .19$. As in Experiment 1, overall results confirmed the LOP effect: The case task led to lowest performance ($M = .65$, $SE = .02$), and the category task led to highest performance ($M = .73$, $SE = .02$) with the rhyme task in the middle ($M = .69$, $SE = .02$). In addition, the main effect of response type showed that *yes* responses ($M = .71$, $SE = .02$) led to higher corrected recognition scores than *no* responses ($M = .67$, $SE = .02$), $F(1,164) = 31.55$, $p < .001$, $\eta^2_p = .16$.

[2] As in Experiment 1, some subjects had perfect hit rates ($n = 10$) or false alarm rates ($n = 6$). To calculate *d'* scores, Macmillan and Creelman's correction (2005) was used.

[3] As in Experiment 1, corrected recognition scores were used as the dependent variable instead of *d'* scores because many subjects ($N = 138$ out of 168) had perfect hit rates of 1.00 at (at least) one of response type x 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.

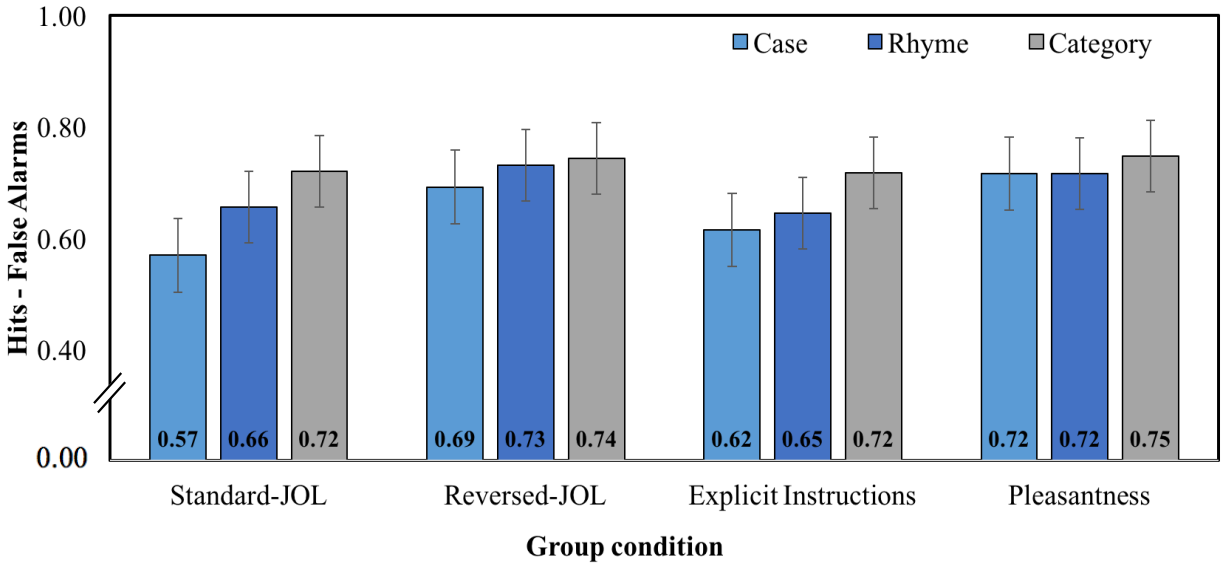

Figure 3.2 Corrected recognition scores across orienting tasks for each condition in Experiment 2. Error bars indicate 95% confidence intervals.

There was no main effect of group, $F(3,164) = 1.85$, $p = .140$, $\eta^2_p = .03$. Further, the orienting tasks x group interaction was reliable, $F(5.78,315.75) = 4.58$, $p < .001$, $\eta^2_p = .08$, and was driven by the differences amongst groups at the case task: The Standard-JOL group had significantly lower recognition scores than the Pleasantness group, $p = .013$, and marginally lower recognition scores than the Reversed-JOL group, $p = .060$. The groups did not differ after performing the rhyme or category tasks, $ps > .05$. Further analyses were computed to compare difference scores between tasks across groups. The pairwise comparisons revealed that giving pleasantness ratings led to significantly lower difference scores between the case and the rhyme tasks compared to the Standard-JOL condition, $p = .005$. In addition, the Pleasantness and the Reversed-JOL conditions had significantly lower difference scores between the case and the category tasks compared to the Standard-JOL condition, ($p < .001$, $p = .005$, respectively). The Pleasantness condition had somewhat lower difference scores between the case and the category tasks compared to the Explicit Instructions condition, $p = .084$, and the Reversed-JOL condition had somewhat lower difference score between the rhyme and the category tasks compared to the

Explicit Instructions condition, $p = .090$. Therefore, for the Pleasantness and Reversed-JOL conditions, the LOP effect was attenuated compared to the other two conditions.

Figure 3.3 shows corrected recognition scores for each orienting task separately for *yes* and *no* responses. The orienting tasks x response type interaction was marginally reliable, $F(2,328) = 2.37$, $p = .096$, $\eta^2_p = .01$, indicating that *yes* responses showed the LOP effect, yet, for *no* responses, the category task led to highest performance, but the case task and the rhyme task did not differ from another, $p = .122$. Also, pairwise comparisons revealed that for the case task *yes* responses had marginally higher recognition scores than *no* responses, $p = .069$, whereas at other two tasks *yes* responses had significantly higher recognition scores, ps $< .001$. Lastly, the response type x group interaction was also reliable, $F(3,164) = 7.43$, $p < .001$, $\eta^2_p = .12$, and results will be discussed within each group. The three-way interaction was not reliable.

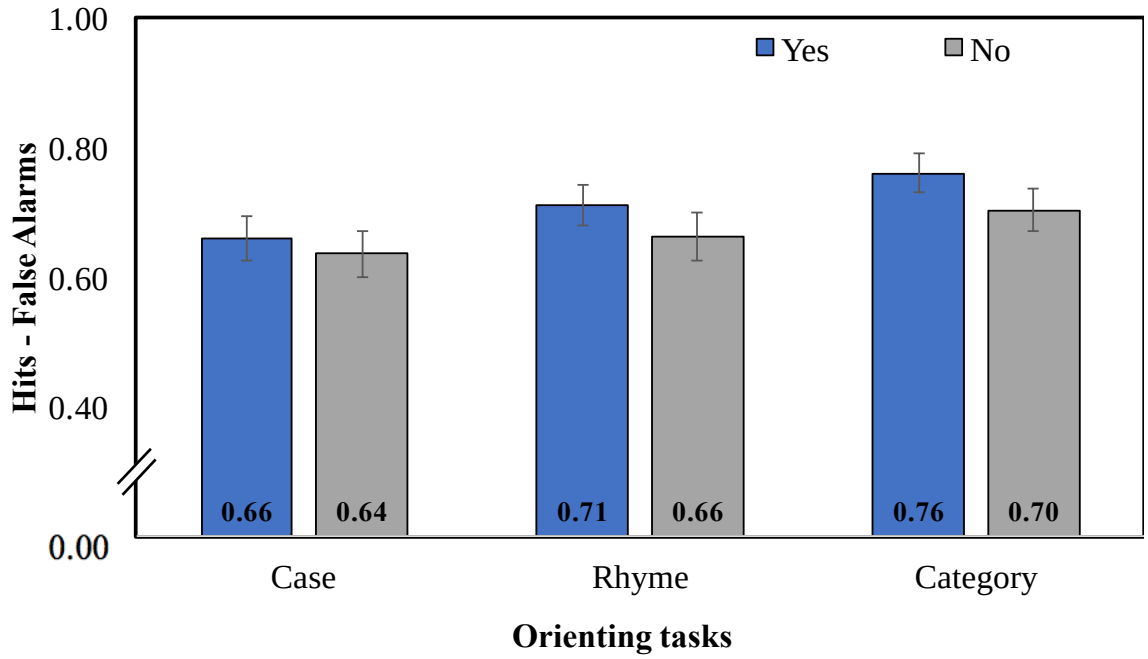

Figure 3.3 Corrected recognition scores for yes/no responses across orienting tasks for Experiment 2. Error bars indicate 95% confidence intervals.

3.2.1 Standard-JOL condition

Figure 3.4a shows corrected recognition scores for the Standard-JOL group. The results for corrected recognition scores replicated the JOL condition in Experiment 1: Performance in all three orienting tasks differed from one another in terms of corrected recognition scores, *p*s < .05. In addition, *yes* responses (*M* = .70, *SE* = .03) had significantly higher recognition scores than *no* responses (*M* = .60, *SE* = .03), *p* < .001. Thus, I again obtained the standard LOP effect using the prototypical Craik and Tulving (1975) paradigm along with JOLs.

Figure 3.4b shows that the JOL ratings show a slight increase across orienting task, mostly for *yes* responses as in Experiment 1. To test this pattern statistically, a 2 (response type) x 3 (orienting tasks) repeated measures ANOVA was conducted for JOLs. Showing the same pattern with corrected recognition scores, *yes* responses (*M* = 3.40, *SE* = .10) led to higher JOL ratings than *no* responses (*M* = 2.90, *SE* = .11), $F(1,41) = 48.32$, $p < .001$, $\eta^2_p = .54$. There was a also main effect of orienting tasks, $F(2,82) = 8.35$, $p = .001$, $\eta^2_p = .17$; 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 compared to the rhyme task (*M* = 3.21, *SE* = .11) and category task (*M* = 3.22, *SE* = .10), *p*s < .05, but unlike recognition scores, the rhyme task and the category task did not lead to statistically different JOL ratings, $p = .991$.

The response x orienting task interaction was also reliable, $F(2,82) = 12.75$, $p < .001$, $\eta^2_p = .24$, and was driven by differences in *yes* and *no* responses. For *yes* responses, the case task led to lower JOL ratings than the rhyme task and the category task, *p*s < .001, the rhyme and category tasks did not differ, $p = .214$. For *no* responses, JOL ratings for the orienting tasks did not differ from one another, *p*s > .05. Thus, even though the main effect of orienting tasks was

statistically significant for JOLs, both for *yes* and *no* responses subjects did not accurately predict retention difference between the rhyme and category levels that was evident in their performance.

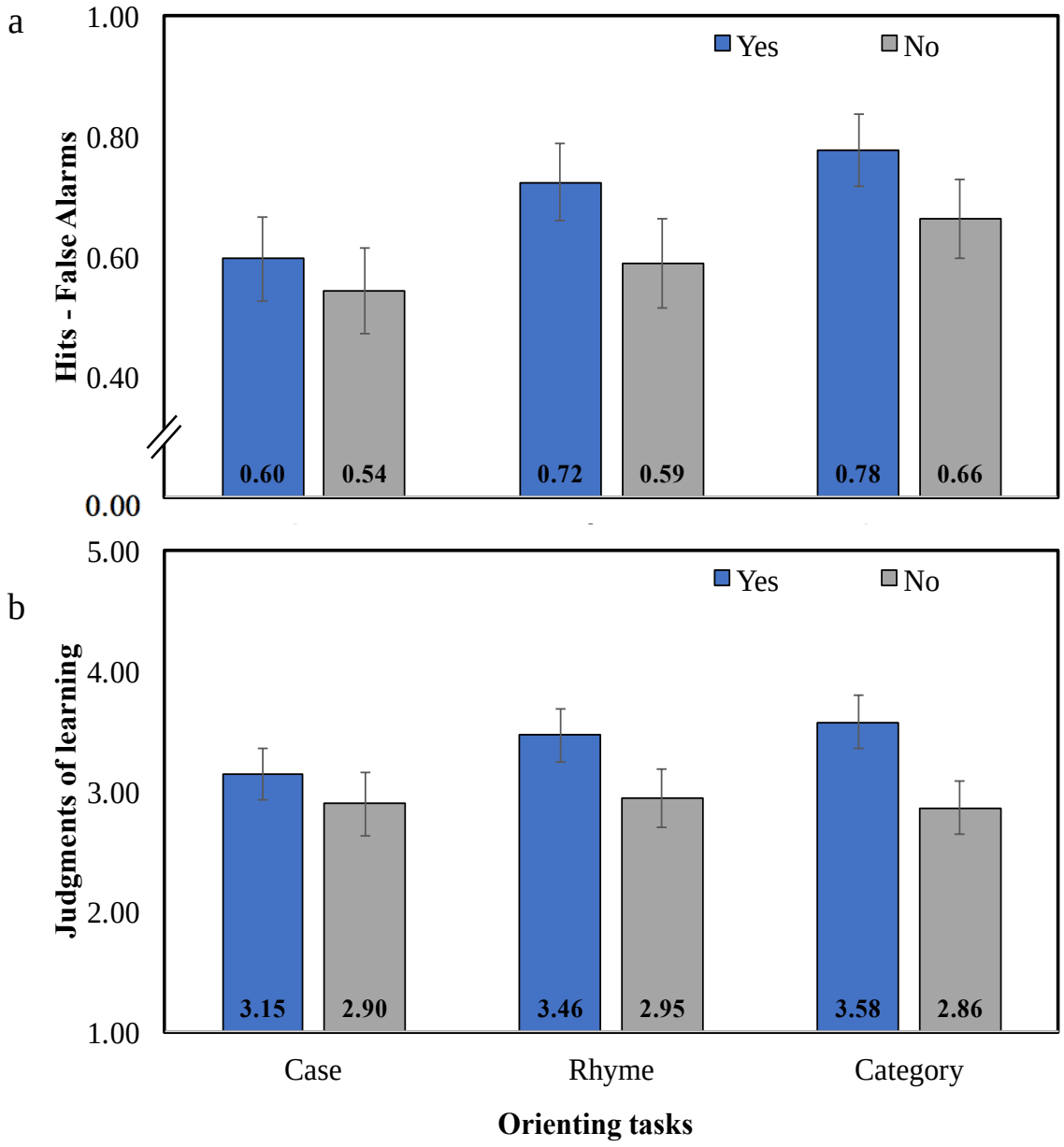

Figure 3.4 Corrected recognition scores across orienting tasks and yes/no responses for the Standard-JOL group (a). Judgments of learning across orienting tasks and yes/no responses for the Standard-JOL group (b). Error bars indicate 95% confidence intervals.

To compare hit rates and JOL ratings further, each subject's hit rates and JOL ratings were transformed into z-scores for all three orienting tasks. Figure 3.5 shows a cross-over interaction between hits and JOLs at the rhyme task after the transformation. A 2 (measure) x 3 (orienting tasks) repeated measures ANOVA was conducted to see whether JOLs yielded similar levels of hit rates across the three orienting tasks. There was a main effect of orienting tasks, $F(2,82) = 21.94$, $p < .001$, $\eta^2_p = .35$, and no main effect of measure, $F(1,41) = .25$, $p = .618$, $\eta^2_p = .01$. More interestingly, the interaction was reliable, $F(2,82) = 5.83$, $p = .004$, $\eta^2_p = .13$, and was driven by differences between hit rates and JOL ratings at the case and the category tasks. At the case task, subjects overestimated their future recognition (JOL ratings were higher than hit rates), $p = .018$, whereas at the category task, subjects underestimated their future recognition (hit rates were higher than JOL ratings), $p = .001$. Likewise, this comparison also revealed that subjects attenuated the LOP effect in their predictions in the standard paradigm.

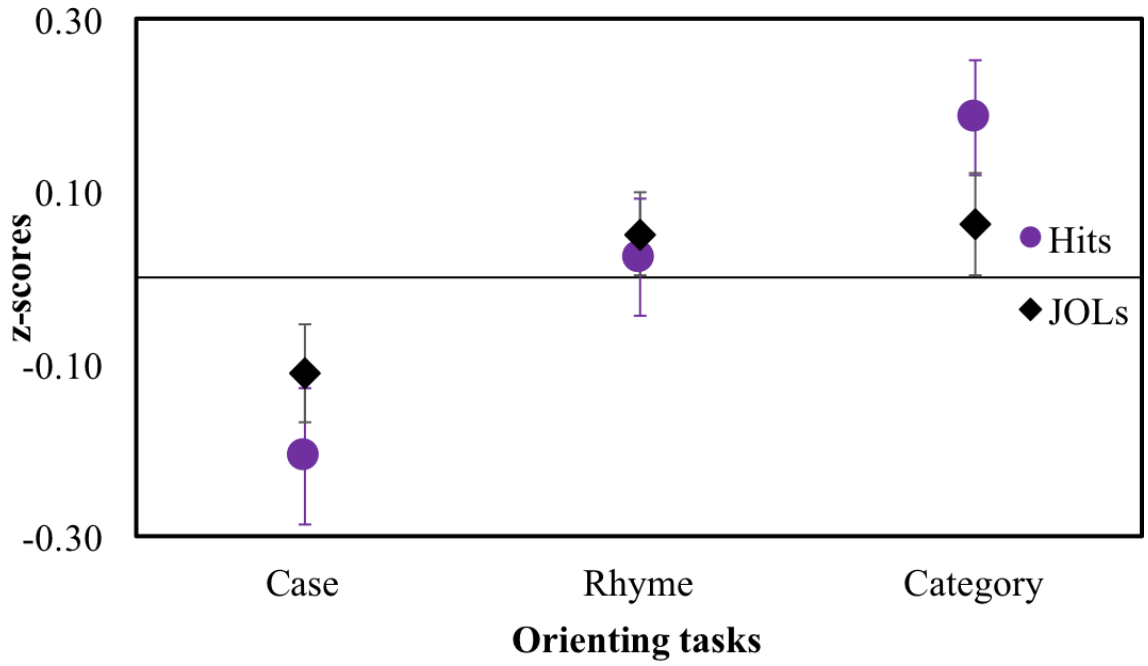

Figure 3.5 Change in z-scores for hit rates and JOLs across orienting tasks for the Standard-JOL group. Error bars indicate 95% confidence intervals.

The middle part of Table 2.2 reports averaged gamma correlations for four different gamma correlations for each subject. Subjects were able to discriminate between the items they would recognize and the items they would not recognize under all orienting tasks: All of the average correlations were significantly higher than 0 for the case, rhyme and category tasks and all items, $t(36) = 2.40$, $p = .022$, $t(33) = 3.75$, $p = .001$, $t(29) = 2.53$, $p = .017$, $t(39) = 5.25$, $p < .001$, respectively. As in Experiment 1, one-way repeated measures ANOVA ($N = 25$) revealed that the correlations did not differ from one another statistically, $F(1.78,72) = .73$, $p = .471$, $\eta^2_p = .03$, showing that the LOP manipulation did not change discriminability of JOLs.

3.2.2 Reversed-JOL condition

Figure 3.6 shows corrected recognition scores and JOL ratings for the Reversed-JOL group. Pairwise comparisons amongst orienting tasks revealed that corrected recognition scores for the case task was significantly lower than for the category task, $p = .044$, and marginally lower than the rhyme task, $p = .097$. Surprisingly, recognition scores in the rhyme task did not significantly differ from scores in the category task, $p = .880$, and, thus the original LOP effect was not completely obtained. Moreover, *yes* responses ($M = .73$, $SE = .03$) did not differ from *no* responses ($M = .71$, $SE = .03$), $p = .149$. These results differ from Craik's (1977) findings in the reversed order paradigm, which revealed the LOP effect; however, the current experiment also included the additional JOL ratings. Hence, it might be that the combination of the reversed order paradigm and making JOLs attenuated the difference between the case level and the rhyme level and eliminated the difference between the rhyme level and the category level.

To see whether JOL ratings (Figure 3.6b) showed a similar pattern to subjects' actual performance, a 2 (response type) x 3 (orienting tasks) repeated measures ANOVA was conducted and revealed a marginal effect of orienting tasks, $F(2,82) = 2.49$, $p = .089$, $\eta^2_p = .06$,

but none of the pairwise comparisons amongst orienting tasks for JOLs were significant, *p*s > .05, revealing a different pattern from subjects' actual performance. In addition, unlike corrected recognition scores, overall *yes* responses (M = 3.31, SE = .11) led to higher JOL ratings than *no* responses (M = 3.13, SE = .11), $F(1,41) = 7.62$, $p = .009$, $\eta^2_p = .16$.

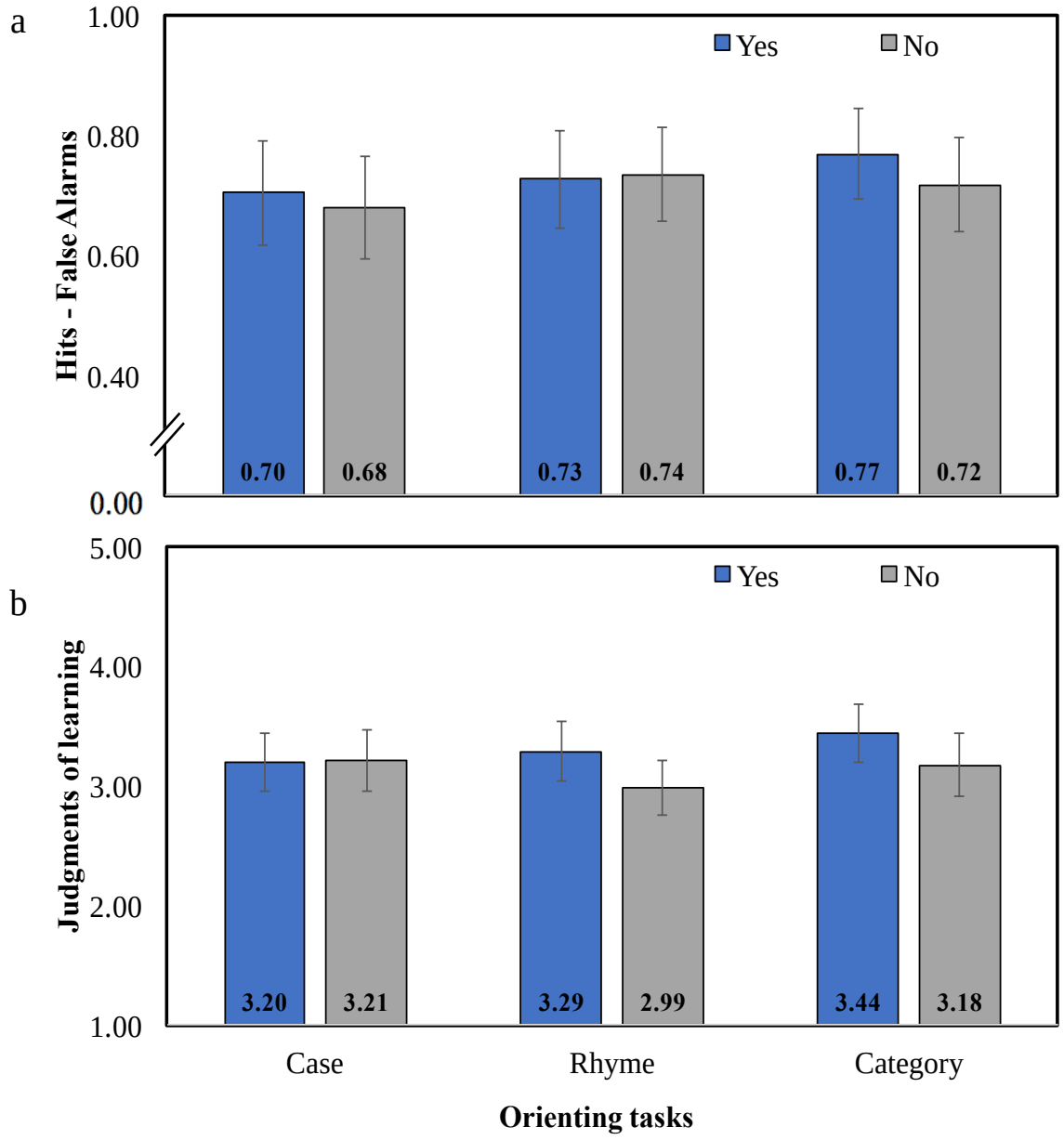

Figure 3.6 Corrected recognition scores across orienting tasks and yes/no responses for the Reversed-JOL group (a). Judgments of learning across orienting tasks and yes/no responses for the Reversed-JOL group (b). Error bars indicate 95% confidence intervals.

The interaction was also reliable, $F(2,82) = 3.13$, $p = .049$, $\eta^2_p = .07$, and was driven by different JOL ratings for *yes* and *no* responses in the rhyme and category tasks, *p*s < .05. In the case task, JOL ratings for *yes* and *no* responses did not differ, $p = .882$, in agreement with the recognition results. These JOL results differ markedly from results coming from corrected recognition scores in Figure 3.6a.

Figure 3.7 shows hit rates and JOL ratings after the z-score transformation. A 2 (measure) x 3 (orienting tasks) repeated measures ANOVA revealed a main effect of measure, a main effect of orienting tasks, $F(2,82) = 5.37$, $p = .006$, $\eta^2_p = .12$. There was no main effect of measure, $F(1,41) = .88$, $p = .354$, $\eta^2_p = .02$, nor a reliable interaction, $F(2,82) = 1.80$, $p = .174$, $\eta^2_p = .04$. Under the reversed order paradigm subjects' hit rates and JOL ratings did not change from one another at any orienting task. In other words, subjects correctly estimated their hit rates.

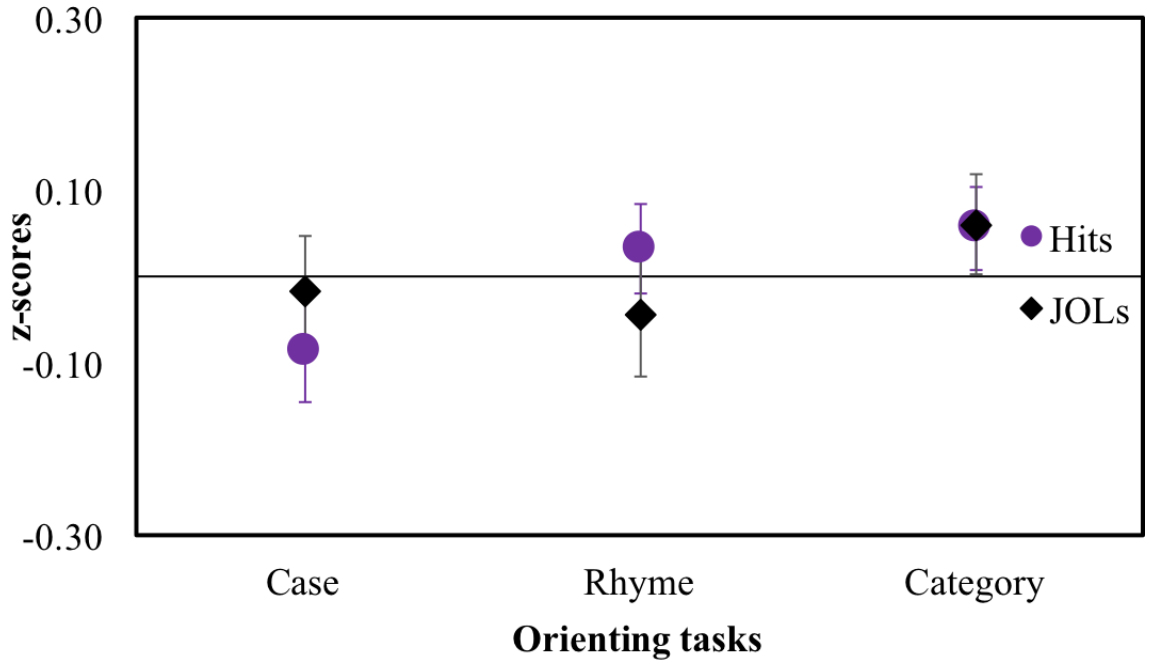

Figure 3.7 Change in z-scores for hit rates and JOLs across orienting tasks for the Reversed-JOL group. Error bars indicate 95% confidence intervals.

The lower part of Table 2.2 reports means of four gamma correlations under the reversed

order paradigm. Under the reversed order paradigm, subjects could predict recognizability above chance. For items in the case task, items in the category task and all items, the relation between JOL ratings and hit rates were greater than 0, $t(33) = 3.09$, $p = .004$, $t(31) = 2.96$, $p = .006$, $t(37) = 3.67$, $p = .001$, respectively, whereas for items in the rhyme task, the relation was only marginally higher than 0, $t(28) = 1.91$, $p = .067$. In addition, similar to previous results, the gamma correlations did not differ from one another in terms of resolution, $F(2.04,69) = .15$, $p = .864$, $\eta^2_p = .01$, ($N = 24$).

3.2.3 Explicit Instructions condition

Pairwise comparisons revealed that the category task led to significantly higher corrected recognition scores compared to the rhyme and case tasks, ps < .001, whereas the case task and the rhyme task did not differ from one another, $p = .285$, (Figure 8a). Also, *yes* responses ($M = .68$, $SE = .03$) led to significantly higher recognition scores than *no* responses ($M = .64$, $SE = .03$), $p = .004$.

The purpose of this condition was to make sure that subjects studied the words during the delay instead of waiting for the orienting questions. In the final questionnaire subjects were asked to report the strategy they used, and based on their reports, they were divided into two groups. Subjects who reported repeating the words during the delay were identified as using rote rehearsal as a strategy ($N = 17$), whereas subjects who reported associations or imagery were identified as using elaboration as a strategy ($N = 13$). (Twelve subjects did not report any specific strategy). The assumption was that the group who used elaboration as a method would not show the LOP effect or would show it to a lesser degree, because of their prior semantic processing of the word in the case and rhyme orienting tasks.

Figure 3.8 shows that the group that used elaboration performed better than the group that

used rote rehearsal at each orienting task. A 2 (strategy) x 2 (response type) x 3 (orienting tasks) repeated measures ANOVA revealed a main effect of orienting tasks, replicating the results from all subjects in the Explicit Instructions group $F(1.65,46.27) = 13.36$, $p < .001$, $\eta^2_p = .32$. The category task differed from the other orienting tasks, ps < .05; however, the case task (M = .64, SE = .04) did not differ significantly from rhyme task (M = .68, SE = .04) in terms of corrected recognition scores, p = .106. As expected subjects who used elaboration (M = .79, SE = .05) performed better than subjects who used rote rehearsal (M = .59, SE = .05), $F(1,28) = 7.89$, p = .009, $\eta^2_p = .22$. There was no main effect of *yes/no* response type, $F(1,28) = 2.62$, $p = .117$, η^2_p = .09, and none of the interactions were reliable, ps > .05, thus, no further analyses were conducted. A potential drawback is that the small sample sizes of two groups might not have not provided enough power to detect effects of some variables.

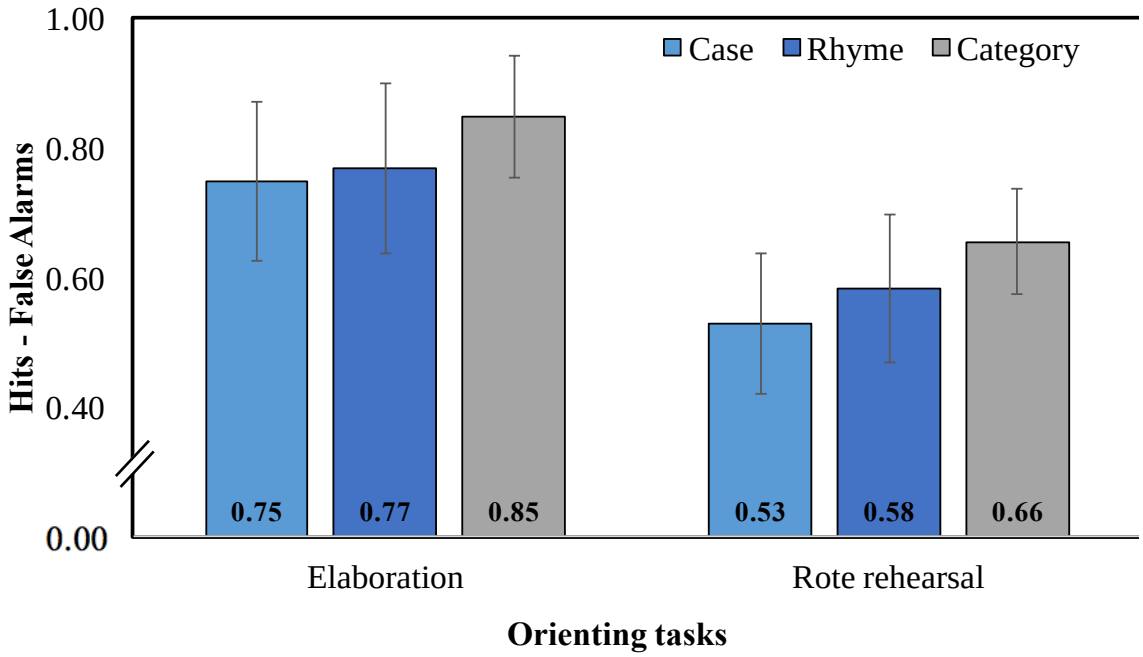

Figure 3.8 Corrected recognition scores across orienting tasks for the Elaboration strategy (N = 13) and the Rote rehearsal group (N = 17). Error bars indicate 95% confidence intervals.

3.2.4 Pleasantness condition

None of the recognition scores for the three orienting tasks differed from one another significantly, *p*s > .05. See Figure 8a. Likewise, *yes* responses (*M* = .73, *SE* = .03) did not differ from *no* responses (*M* = .72, *SE* = .03), *p* = .756. Clearly, giving pleasantness ratings led to semantic processing of all items and the following case and rhyme orienting questions did not hurt performance. These results showed that the LOP effect reported by Craik (1977) under the reversed order paradigm is not due to the shallow tasks impairing retention of semantically processed words. Thus, the LOP effect in the reversed order paradigm is driven by the fact that subjects did not semantically processed the words fully during the 5-sec delay between the word and the orienting question.

Chapter 4: General Discussion

4.1 The LOP Effect and Intentional Learning

The current experiments revisited two findings in the LOP literature where the LOP effect persisted even though Craik and Lockhart's (1972) framework predicted that the effect should be eliminated. The main purposes of Experiment 1 and Experiment 2 were to investigate the recurring LOP effect under the intentional learning instructions and the reversed order paradigm, respectively. Overall, the experiments replicated the LOP effect and retention differences of *yes/no* responses. In Experiment 1, which employed the standard LOP paradigm under intentional learning instructions, I obtained the LOP effect with the control (Intentional), Delay and JOL groups. In Experiment 2, which employed the reversed order paradigm under intentional learning instructions, I again obtained the LOP effect in three instructional conditions (though it was eliminated between some orienting tasks). The Pleasantness group, on the other hand, did not show the effect. After reviewing results and their implications for the LOP framework, I then consider additional findings.

First, are subjects aware that the different orienting tasks lead to different retention levels? The answer seems to be somewhat mixed. For the standard LOP paradigm with self-paced JOLs (Experiment 1) and with experimenter-paced JOLs (Experiment 2), the magnitude of JOLs showed some differences amongst the orienting tasks that correspond to corrected recognition score. 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, the magnitude of their JOL ratings did not reflect performance differences among all of the orienting tasks and were not consistent across the two experiments: Corrected recognition scores showed the LOP effect for both experiments (although the difference between

the case and rhyme tasks was marginal in Experiment 1), whereas JOL magnitudes did not differ for the case and rhyme tasks in Experiment 1 and for the rhyme and category tasks in Experiment 2. Furthermore, for the reversed order paradigm (Experiment 2), the magnitude of JOLs did not differ across any orienting tasks, even though the case task led to lower performance than the rhyme and category tasks. In short, recognition performance was much more affected by orienting tasks than JOL ratings were.

Conclusions about calibration of JOLs with hit rates could not be assessed because both experiments employed a 5-point scale rather than a 100-point scale, but the comparison between mean hit rates with mean JOL ratings through z-score transformation demonstrated that in the standard LOP paradigm, subjects overestimated their performance under the case task. In addition, JOLs were sensitive to recognizability differences of items, and this did not change across orienting tasks. According to Koriat (1997), JOLs are determined more by intrinsic cues relative to extrinsic cues whereas actual performance is determined by both cue types. In the current study JOLs were sensitive in discriminating recognizable items based on intrinsic cues, whereas they were not as sensitive when predicting the effect of the LOP manipulation. This pattern indicates that the orienting tasks in the current study could have served as extrinsic cues that affected overall performance and perhaps overshadowed intrinsic cues. Therefore, JOLs were relatively insensitive to orienting tasks, and subjects underestimate the tasks' effect on performance. These results offered a plausible explanation for the often obtained LOP effect under the intentional learning instructions: Even when subjects know they would be tested for their memory, they do not try to encode the words presented with the shallow tasks "deeply" or allocate their resources accordingly, because they do not predict retention differences accurately. They do not seem to realize fully that shallow levels of processing lead to poorer memory.

Experiment 2 explored another possible explanation for the LOP effect under intentional learning instructions, viz., subjects might have difficulties in switching from shallow processing to natural deep processing even if they try to do so. To prevent such difficulties and promote spontaneous deep processing, Experiment 2 employed several versions of the reversed order paradigm. One of them was combined with JOL ratings, thus it promoted both spontaneous deep processing and monitoring. Although the retention difference between the case and the category tasks was not eliminated but rather was sharply attenuated (a 5% effect), the retention difference between the rhyme and the category tasks was eliminated. This latter outcome is surprising, because the retention differences between semantic processing and shallower processing are often quite persistent. Nevertheless, by providing conditions under which subjects would engage in deep processing naturally and were told to monitor their learning processes, it was possible to eliminate the robust LOP effect in the explicit test of recognition. Thus, under the standard LOP paradigm, subjects' natural deep processing seems to be truncated due to the demands of shallow processing, thus leading to the LOP effect.

Experiment 2 also investigated possible reasons for the persistent LOP effect observed in Craik (1977) under the reversed order paradigm. One of the questions was whether shallow questions following the word impaired performance of already deeply processed words. If so, this would explain the persistence of the LOP effect under the reversed order paradigm since the orienting questions follow the word. However, this hypothesis was not supported. Once semantically processed through pleasantness ratings, the later shallow questions did not lead to lower retention. In the reversed order paradigm, giving pleasantness ratings after each word eliminated the LOP effect and the retention differences caused by *yes/no* responses. This finding might seem in contradiction with Cermak et al. (1970), who reported impaired recognition

performance when subjects concentrated on both the sound and dictionary meaning of the words. However, in the current experiment, subjects first processed the words semantically and then shallow processing followed. Thus, it might be when the task concurrently demands two different types of processing, semantic processing is hindered, but not when the tasks are performed sequentially. This is also in line with the idea that under the demands of a shallow task subjects might have difficulties at engaging in spontaneous semantic processing.

Another question raised by Experiment 2 focused on subjects' activity during the delay between the word and the orienting task in the reversed order paradigm. Craik (1977) assumed that during the delay subjects were processing the word deeply, thus, the LOP effect should be eliminated. Yet, instead of deeply processing or attending to the word, subjects could have waited for the orienting question since they knew the question would follow. As an attempt to control for this possibility, I asked subjects to study the word during the delay period by encouraging subjects to study the words with their idiosyncratic strategies (without identifying any particular one). Although, there is no way to ensure that all subjects followed the instructions throughout the experiment, introducing explicit study instructions during the delay eliminated retention difference between the case and the rhyme tasks. This was true for subjects who used rote rehearsal and for subjects who used elaboration, even though the elaboration subjects had better recognition overall. Thus, it is probable that in Craik (1977) subjects did not attend to the words during the delay. Although, this can partially explain the LOP effect he obtained under the reversed order paradigm, it cannot fully account for it given that the category task still led to highest performance.

4.2 Congruency effect and JOLs

Unlike the LOP effect, subjects were aware that *yes/no* responses would lead to different retention levels. In both experiments, the congruency effects on actual performance were also observable and in the same direction as for JOLs. In fact, besides the case task of the reversed order paradigm, JOLs for *yes* responses were always higher than JOLs for *no* responses. The congruency of the item with the orienting question is a characteristic of the item, hence is an intrinsic cue based on Koriat's (1997) cue utilization model. As the model predicts, congruency influenced the magnitude of JOLs: For congruent *yes* responses JOLs were higher compared to incongruent *no* responses, even in cases where the actual performance did not differ between *yes* and *no* responses. This finding is quite interesting because the congruency effect was not predicted by Craik and Tulving (1975) for the standard LOP paradigm. Craik and Tulving explained post hoc that the compatibility of the orienting question with the target word should be taken into account for later retention. Nevertheless, subjects were able to predict that when the orienting question was congruent with the target word, they would remember target word better at a later test.

Another interesting finding is that in both Experiment 1 and Experiment 2, in the standard LOP paradigm with JOLs, the orienting task and response type interaction was significant. That is even if there was a main effect of orienting task for JOLs, this effect was driven by *yes* responses (Figures 5b and 10b). This interaction is critical as to whether subjects used orienting questions as cues when they gave JOLs or not. It seems likely that subjects predicted the levels of processing to some small extent only because the semantic orienting task generated higher performance and thus more *yes* responses (for congruent trials). *No* questions were not related or specific to target words, hence it would not be useful for subjects to make

their JOLs based on those. Thus, JOLs for *no* responses did not follow the actual performance pattern at all, and therefore did not show the LOP effect. The data from both experiments can be interpreted as support for this possible hypothesis. Obviously, this explanation is post hoc and further research is needed to examine it. In future experiments, I recommend a focus on *yes/no* responses separately, with increased numbers of items for each within subject condition to increase power (there were 10 items at the current experiments).

4.3 Diagnosticity of JOLs

The JOL literature reveals that JOLs are generally somewhat accurate at predicting future performance, yet JOLs were not reliably diagnostic in the current experiments. One possible way to improve accuracy of JOLs might be using delayed JOLs instead. In both experiments, subjects gave immediate JOLs whereas previous studies revealed that delayed JOLs are more diagnostic of future performance (Dunlosky & Nelson, 1994). This can be due to the clearing of recent items from short-term memory while giving delayed JOL ratings (Nelson & Dunlosky, 1991) or retrieving the target while giving delayed JOL ratings, hence, increasing future memorability of the item (Spellman & Bjork, 1992). Neither of these possibilities were true for immediate JOLs in current experiments, thus, through delayed JOLs it might be possible to make subjects' more aware of the LOP effect.

Furthermore, according to Koriat (1997), JOLs can be either theory-based or experience-based: People can either use a priori beliefs or inferences about memory, or online experiences after encountering the item and accuracy of their judgments can change accordingly. In addition, experience-based JOLs are influenced by intrinsic cues of items more so than extrinsic cues. Koriat et al. (2004) examined whether subjects were aware of the effects of retention interval on memory (i.e., future performance decreases as retention interval increases) and found that item-

by-item predictions made for one's self were less diagnostic compared to global predictions made for others' future performance. They proposed that predictions about self were both experience-based and theory-based whereas global predictions about others were only theory-based. Given that experience-based predictions do not capture effects of extrinsic cues as much as theory-based predictions, self-predictions did not capture the effect of retention interval compared to global predictions about others. Therefore, their diagnosticity was lower.

Similarly, in current experiments, subjects made item-by-item JOLs for their own future performance, thus their JOLs were more experience-based instead of theory-based. Therefore, their JOLs were not necessarily reflective of the orienting tasks, whereas the theory-based JOLs could have reflected the LOP effect. When JOLs are more theory-based, subjects might be able to predict the LOP effect more accurately. For instance, if subjects made the predictions for others' future performances instead of themselves, we might observe a more pronounced LOP effect for JOLs. Indeed, Seamon and Virostek (1978) showed that subjects accurately ranked orienting questions based on amount of processing when the questions were not part of an encoding phase and subjects did not study any material under the questions. Clearly, ranking for amount of processing is not the same as predicting future performance, yet, the rankings were driven by subjects' theories about amount of processing rather than their experiences.

In his series of experiments, Castel (2008) first revealed that students underestimated primacy and recency effects (i.e., extrinsic cues) even with repeated study and test trials. They based their JOLs more on intrinsic cues. To overcome intrinsic cues and to emphasize the serial position information of words, Castel asked students to give pre-JOLs along with the serial position information. He found that students were able to predict the primacy and recency effects in this procedure. Taking a similar approach, one possible way to override the effect of intrinsic

cues and emphasize the LOP effect might be asking subjects to give JOL ratings after the orienting questions but before the presentation of the words with possible repeated study-test trials. In this case, subjects would be more compelled to base their JOLs on the orienting questions, and might predict the LOP effect accurately.

4.4 JOLs as Memory Modifiers

An unexpected yet interesting finding was that JOLs enhanced performance only for the shallow tasks. In line with current findings, recent research has also demonstrated that JOLs might be a reactive measure that influences the act of learning itself. There is preliminary support that robust memory phenomena such as the generation effect can be attenuated through JOLs, because giving JOLs for read pairs enhanced their retention and decreased the retention difference between read and generated pairs (Soderstrom et al., 2015, Experiment 2). Similarly, I found that JOLs attenuated the LOP effect through improving performance at the shallower tasks (Experiment 1) and in some cases even eliminated the effect between the category and the rhyme orienting tasks (Experiment 2).

How do JOLs influence future performance? Both Mitchum et al. (2016) and Soderstrom et al. (2105) defined difficulty through cue – target relatedness: related items were reviewed as easy and unrelated items as difficult. Mitchum et al. (2016) proposed a goal change hypothesis to explain reactivity of JOLs. That is, giving JOLs leads to a goal change from mastering all items to the more pragmatic one of mastering easy items, and this impairs memory for more difficult items. In a series of experiments, they demonstrated that, compared to the control groups, subjects who gave JOLs showed increased retention differences between the easy and difficult items. They concluded that subjects became aware through JOLs that they could not master difficult items, hence, they switched their resources away from difficult items to easy ones.

Somewhat similarly, Soderstrom et al. (2015) proposed that giving JOLs enhanced retention for easy items by emphasizing the already existing cue – target relationship (instead of impairing retention for difficult items). They showed that giving JOLs bolstered performance of easy items, and did not have any impact on difficult items.

Following this logic, if relatedness of the orienting task to the target acted as a cue in the current study, words processed with the category task would be regarded as easy items whereas words processed under the case task would be viewed as difficult items, with words in the rhyme task in the middle. Thus, for the JOL conditions, either retention at the case task should have been impaired based on the goal change hypothesis or retention at the category task should have been enhanced based on strengthened cue – target relationship. Yet, in Experiment 1, the JOL condition demonstrated enhanced retention for shallower tasks, compared to the Intentional and the Delay conditions, a result that was in the opposite direction of suggested explanations by Mitchum et al. (2016) and Soderstrom et al. (2015). Therefore, neither of these explanations were supported.

Dunlosky and Thiede (1998) suggested that people regulate their study behavior to reduce discrepancy between their current learning level and their desired learning level (i.e., discrepancy reduction theory) and they do this through monitoring their learning. Thus, according to this theory, when people have mastery as their goal, they allocate more resources to difficult items because their discrepancy is higher. It is possible that by forcing subjects to monitor their learning by giving JOLs, subjects adopted a mastery goal, and in turn, they allocated more resources to study difficult items, (i.e., items in the shallow orienting tasks). It is true that discrepancy reduction theory is mostly suggested as an explanation for self-regulated studying in which subjects have the opportunity to study items for an unlimited time, whereas in

current experiments subjects were not able to regulate their study and saw the words only for 2 seconds. Thus, in the current study "allocation of study time" cannot explain the findings, but a broader "allocation of resources" possibly can. A potential drawback of this explanation is that item difficulty was not reflected in subjects' JOLs: If subjects realized that the shallow levels were more difficult via JOLs, and if they allocated more resources to master them, (hence attenuating the LOP effect), the magnitude of JOLs should have shown this awareness. They did not, so this idea also cannot account for the results.

Another hypothesis that is also in line with the LOP framework is that while giving JOLs subjects might have elaborated on the meaning of the word to some degree, and this elaboration enhanced retention, but not to the same degree as the category orienting questions did. The results from Experiment 1 are in line with this explanation. In Experiment 2, I did not have any control group to compare the groups with JOLs, but Craik's (1977) reversed order paradigm can serve as a comparison group. In Craik (1977), the rhyme task still led to lower retention than the category task under the reversed order paradigm, whereas in Experiment 2 adding JOLs to this procedure eliminated this difference. Thus, the combined effect of the reversed order paradigm and giving JOLs might have led to deeper and more elaborative encoding of words under the rhyme task which in turn fully attenuated the LOP effect. Further research may more fully explore this possibility. In short, in both experiments, JOLs were reactive measures that affected retention of the shallow tasks.

4.5 Conclusion

In sum, the present study demonstrated that subjects' predictions were relatively insensitive to the LOP manipulation under the standard LOP paradigm and under the reversed order paradigm. This insensitivity might be one of the reasons as to why the LOP effect persists under intentional

learning instructions and the reversed order paradigm. Further examination of conditions under the reversed order paradigm demonstrated that once the words were deeply processed, the shallow levels did not impair later performance. In addition, encouraging subjects to study or attend to the material through their idiosyncratic strategies might attenuate or eliminate the LOP effect (amongst some levels). The present study also displayed that JOLs were reactive measures that enhanced retention for the shallower tasks, and this led to attenuation of the LOP effect for recognition performance. Giving JOLs might be forcing subjects to encode the words under shallow levels somewhat more elaborately, hence diminishing (and rarely eliminating) the effect amongst two orienting tasks.

References

1. Arbuckle, T. Y., & Cuddy, L. L. (1969). Discrimination of item strength at time of presentation. *Journal of Experimental Psychology*, *81*(1), 126.

2. Baddeley, A. D. (1978). The trouble with levels: A reexamination of Craik and Lockhart's framework for memory research. *Psychological Review*, *85*(3), 139.

3. 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*(3), 445-459.

4. 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.

5. 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.

6. Castel, A. D. (2008). Metacognition and learning about primacy and recency effects in free recall: The utilization of intrinsic and extrinsic cues when making judgments of learning. *Memory & Cognition*, *36*(2), 429-437.

7. 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*(2), 79-81.

8. 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*(2), 109-112.

9. Craik, F. I. (1973). A" levels of analysis" view of memory.

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

11. Craik, F. I. (2002). Levels of processing: Past, present... and future?. *Memory*, *10*(5-6), 305-318.

12. Craik, F. I., & Lockhart, R. S. (1972). Levels of processing: A framework for memory research. *Journal of Verbal Learning and Verbal Behavior*, *11*(6), 671-684.

13. Craik, F. I., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General*, *104*(3), 268.

14. Cutting, J. E. (1975). Orienting tasks affect recall performance more than subjective impressions of ability to recall. *Psychological Reports*, *36*(1), 155-158.

15. 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.

16. Dunlosky, J., & Nelson, T. O. (1994). Does the sensitivity of judgments of learning (JOLs) to the effects of various study activities depend on when the JOLs occur?. *Journal of Memory and Language*, *33*(4), 545-565.

17. Dunlosky, J., & Thiede, K. W. (1998). What makes people study more? An evaluation of factors that affect self-paced study. *Acta Psychologica*, *98*(1), 37-56.

18. Eysenck, M. W. (1978). Levels of processing: A critique. *British Journal of Psychology*, *69*(2), 157-169.

19. Fisher, R. P., & Craik, F. I. (1977). Interaction between encoding and retrieval operations in cued recall. *Journal of Experimental Psychology: Human Learning and Memory*, *3*(6), 701.

20. Garcia, M. A., & Kornell, N. (n.d.). Collector: A program for running psychology experiments on the web. Retrieved from https://github.com/gikeymarcia/Collector

21. 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.

22. Jenkins, J. J. (1974). Remember that old theory of memory? Well, forget it. *American Psychologist*, *29*(11), 785.

23. 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.

24. 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.

25. Lockhart, R. S., & Craik, F. I. (1978). Levels of processing: A reply to Eysenck. *British Journal of Psychology*, *69*(2), 171-175.

26. Lockhart, R. S., & Craik, F. I. (1990). Levels of processing: A retrospective commentary on a framework for memory research. *Canadian Journal of Psychology/Revue Canadienne de Psychologie*, *44*(1), 87.

27. Macmillan, N. A., & Creelman, C. D. (2005). Detection theory: A user's guide. Mahwah, NJ: Lawrence Earlbaum Associates.

28. McDaniel, M. A., Friedman, A., & Bourne, L. E. (1978). Remembering the levels of information in words. *Memory & Cognition*, *6*(2), 156-164.

29. McDaniel, M. A., & Kearney, E. M. (1984). Optimal learning strategies and their spontaneous use: The importance of task-appropriate processing. *Memory & Cognition*, *12*(4), 361-373.

30. 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.

31. Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(5), 519-533.

32. Moscovitch, M., & Craik, F. I. (1976). Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall. *Journal of Verbal Learning and Verbal Behavior*, *15*(4), 447-458.

33. 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.

34. Nelson, T. O. (1977). Repetition and depth of processing. *Journal of Verbal Learning and Verbal Behavior*, *16*(2), 151-171.

35. Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*(1), 109.

36. 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.

37. Nelson, T. O. & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *Psychology of Learning and Motivation*, *26*, 125-173.

38. Roediger, H. L., & DeSoto, K. A. (2016). Recognizing the presidents: Was Alexander Hamilton president?. *Psychological Science*, *27*(5), 644-650.

39. Roediger, H. L., & Gallo, D. A. (2002). Levels of processing: Some unanswered questions. In M. Naveh-Benjamin, M. Moscovitch, & H. L. Roediger (Eds.), *Perspectives on Human Memory and Cognitive Aging: Essays in Honour of Fergus I. M. Craik* (pp. 28-47) Philadelphia: Psychology Press.

40. Sachs, J. S. (1967). Recognition memory for syntactic and semantic aspects of connected discourse. *Perception & Psychophysics*, *2*(9), 437-442.

41. Seamon, J. G., & Virostek, S. (1978). Memory performance and subject-defined depth of processing. *Memory & Cognition*, *6*(3), 283-287.

42. Shaw, R. J., & Craik, F. I. (1989). Age differences in predictions and performance on a cued recall task. *Psychology and Aging*, *4*(2), 131.

43. 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

44. 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.

45. Tulving, E. (1979). Relation between encoding specificity and levels of processing. *Levels of Processing in Human Memory*, 405-428.

46. Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review*, *80*(5), 352.

47. 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.

48. 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*(5), 481-488.

Appendix A

Materials for Experiment 1 and Experiment 2

Target	Category question	Rhyme question
apple	fruit	chapel
ballet	type of dance	pray
bear	four-footed animal	tear
beer	alcoholic beverage	gear
biology	science	apology
bitter	type of taste	litter
blue	color	glue
bomb	weapon	calm
book	reading material	shook
brain	organ	chain
brother	relative	mother
canoe	type of ship	screw
carrot	vegetable	parrot
chair	furniture	hair
chest	body	guest
church	building for religious service	search
claw	part of an animal	draw
colonel	military title	journal
cotton	type of fabric	rotten
cross	symbol	sauce
dollar	money	collar
door	part of building	score
eagle	bird	beagle
eraser	school supplies	racer
floor	part of a room	shore
gasoline	fuel	caffeine
gram	type of a measurement	slam
guitar	musical instrument	bazaar
hammer	tool	grammar
hour	unit of time	tower
house	human dwelling	blouse
juice	non-alcoholic beverage	goose
maple	tree/	staple
zinc	chemical element	sink
mile	unit of distance	smile

mint	herb	lint
noun	part of speech	clown
pearl	stone	girl
president	an elective office	resident
priest	clergy	east
puzzle	toy	muzzle
rain	weather phenomenon	pain
rice	type of grain	twice
river	earth formation	liver
rose	flower	nose
salmon	fish	famine
sandal	footwear	candle
shirt	clothing	hurt
soccer	sport	locker
spice	flavoring food	dice
spider	insect	rider
spoon	kitchen utensil	dune
square	shape	stair
state	territorial unit	eight
steel	metal	seal
teacher	occupation	creature
theft	crime	left
train	vehicle	plane
vision	type of sense	precision
water	liquid	hotter

Appendix B

Results from Confidence Judgments

Experiment 1

Upper part of Table B.1 shows confidence judgments for targets for each group across different levels. Overall, confidence judgments follow a similar pattern as hit rates. A 2 (response type) x 3 (orienting tasks) x 3 (group) repeated measures ANOVA was conducted for confidence judgments. As with hit rates, there was a main effect of orienting tasks, $F(2,246) = 63.59$, $p < .001$, $\eta^2_p = .34$, Overall, confidence judgments revealed the LOP effect: The case task led to lowest confidence ($M = 3.78$, $SE = .06$), and the category task led to highest confidence ($M = 4.25$, $SE = .05$) with the rhyme task in the middle ($M = 4.03$, $SE = .06$). In addition, *yes* responses ($M = 4.14$, $SE = .05$) resulted in higher confidence ratings than *no* responses ($M = 3.89$, $SE = .06$), $F(1,123) = 90.84$, $p < .001$, $\eta^2_p = .43$. However, unlike with hit rates, no main effect of group occurred, $F(2,123) = .92$, $p = .403$, $\eta^2_p = .02$. Thus, even though subjects in the JOL condition had best recognition, they were not more confident compared to subjects in the Intentional and the Delay conditions.

Upper part of Table B.2 provides confidence ratings for *yes/no* responses across orienting tasks, revealing similar patterns to corrected recognition scores. The orienting tasks x response type interaction was reliable, $F(2,246) = .17.73$, $p < .001$, $\eta^2_p = .13$, showing that for the case task, *yes* responses and *no* responses did not differ significantly, $p = .113$, whereas for other two tasks *yes* responses had significantly higher confidence than *no* responses, ps $< .001$. The orienting tasks x group interaction was also reliable, $F(4,246) = 3.04$, $p = .018$, $\eta^2_p = .05$, and was driven by the differences across orienting tasks. For all of the groups, confidence ratings across levels were in the expected direction (indicating the LOP effect). However, pairwise comparisons between the rhyme task and the category task of the Intentional group and between the case task and the rhyme task of the JOL group did not reach significance, ($p = .157$, $p = .788$, respectively). Other interactions were not statistically significant. Overall, results from confidence judgments mirrored results coming from corrected recognition scores (with a few exceptions).

Experiment 2

Lower part of Table B.1 provides confidence judgments for each group across different orienting tasks, revealing a similar pattern to hit rates. A 2 (response type) x 3 (orienting tasks) x 4 (group) repeated measures ANOVA was conducted for corrected recognition scores. There was a main effect of orienting tasks, $F(2,328) = 48.70$, $p < .001$, $\eta^2_p = .23$. Overall, confidence judgments revealed the LOP effect: The case task led to lowest confidence ($M = 4.23$, $SE = .05$), and the category task led to highest confidence ($M = 4.53$, $SE = .03$), with the rhyme task in the middle ($M = 4.35$, $SE = .04$), ps $< .05$. As in Experiment 1, *yes* responses ($M = 4.45$, $SE = .04$) led to higher confidence than *no* responses ($M = 4.31$, $SE = .04$), $F(1,164) = 43.61$, $p < .001$, $\eta^2_p = .21$.

The main effect of group was also significant, $F(3,164) = 4.02$, $p = .009$, $\eta^2_p = .07$. Averaged across all orienting tasks, the Standard-JOL group ($M = 4.23$, $SE = .07$) had lower confidence compared to the Pleasantness group ($M = 4.52$, $SE = .07$), $p = .027$, and somewhat lower confidence compared to the Reversed-JOL group ($M = 4.48$, $SE = .07$), $p = .078$. Remaining groups did not differ from one another in terms of confidence, ps > .05.

Lower part of Table B.2 shows confidence judgments for the orienting tasks and *yes/no* responses: The pattern is similar to correct recognition scores shown in Figure 9a. As with corrected recognition scores, the orienting tasks x response type was reliable, $F(2,328) = 8.98$, $p < .001$, $\eta^2_p = .05$. Confidence ratings for both *yes* and *no* responses showed the LOP effect, and *yes* responses yielded higher confidence ratings than *no* responses at each level, ps < .05. In addition, the orienting tasks x group was reliable, $F(6,328) = 3.22$, $p = .004$, $\eta^2_p = .06$. Pairwise comparisons revealed that the orienting tasks x group interaction was driven by confidence differences in the case and rhyme tasks: In the case task, the Standard-JOL group was less confident than the Pleasantness group, $p = .003$, and the Reversed-JOL group, $p = .036$. In addition, the Explicit Instructions group was somewhat less confident than the Pleasantness group, $p = .092$. In the rhyme task, the Standard-JOL group was somewhat less confident than the Reversed-JOL group, $p = .081$ and the Explicit Instructions group was somewhat less confident than the Pleasantness group, $p = .070$. The groups did not differ in their confidence in the category task, ps > .05. Lastly, the response type x group was also reliable, $F(3,164) = 5.12$, $p = .002$, $\eta^2_p = .09$, indicating that subjects in each group except the Pleasantness group, gave higher confidence ratings for *yes* responses, ps < .05, whereas subjects in the Pleasantness group gave similar confidence ratings for both responses, $p = .400$.

Although these results are not identical to those in corrected recognition, they are similar. More interestingly, except the Pleasantness group, confidence judgments mirrored corrected recognition rates within each group. For the Pleasantness group, the category task led to somewhat higher confidence than the case task, $p = .053$, and the rhyme task, $p = .019$, even though, hit rates did not differ amongst orienting tasks.

Orienting task	Case		Rhyme		Category	
	M	*SE*	*M*	*SE*	*M*	*SE*
Group			Experiment 1			
Intentional	3.69	.11	4.03	.11	4.16	.09
Delay	3.65	.11	4.01	.11	4.24	.09
JOL	3.98	.11	4.04	.11	4.34	.09
			Experiment 2			
Standard-JOL	4.03	.09	4.22	.08	4.44	.07
Reversed-JOL	4.38	.09	4.50	.08	4.57	.07
Pleasantness	4.48	.09	4.48	.08	4.62	.07
Explicit Instructions	4.17	.09	4.21	.08	4.47	.07

Table B.1 Confidence Judgments for Orienting Tasks across Study Groups for Experiment 1 (top) and Experiment 2 (bottom)

Orienting task	Case		Rhyme		Category	
	M	*SE*	*M*	*SE*	*M*	*SE*
Response Type			Experiment 1			
Yes	3.81	.07	4.15	.06	4.47	.05
No	3.74	.07	3.90	.07	4.02	.06
			Experiment 2			
Yes	4.30	.05	4.40	.05	4.66	.04
No	4.22	.05	4.30	.04	4.39	.04

Table B.2 Confidence Judgments for Orienting Tasks across Response Types for Experiment 1 (top) and Experiment 2 (bottom)