Memory & Cognition (2021) 49:955–967
https://doi.org/10.3758/s13421-020-01133-2

# Making judgments of learning enhances memory by inducing item-specific processing

Olesya Senkova[1] · Hajime Otani[1]

Accepted: 18 December 2020 / Published online: 4 January 2021


## Abstract

A judgment of leaning (JOL) has been investigated to understand self-regulated learning. However, asking participants to make JOLs may increase memory by creating a reactivity effect. In two experiments, we examined whether making JOLs would enhance memory by inducing item-specific processing. We compared a JOL task with two other tasks that are known to induce item-specific processing: pleasantness rating (Experiment 1) and single imagery (Experiment 2; creating vivid mental images). Participants learned a categorized or uncategorized list of words. Memory should be enhanced when the list promotes relational processing and the task induces item-specific processing. As expected, when the list was categorized, recall was higher in the JOL and item-specific processing conditions (pleasantness rating and single imagery) than in the control condition. Furthermore, recall was similar between the JOL and item-specific processing conditions. When the list was uncategorized, there was no difference in recall among the JOL, item-specific processing, and control conditions. Making JOLs enhances memory by inducing item-specific processing. We concluded that researchers need to carefully consider how making a JOL influences memory when investigating self-regulated study behaviors.

**Keywords** Judgments of learning · Item-specific processing · Relational processing

A judgment of leaning (JOL) is one's judgment about how well a given item is learned and how likely it is that this item will be successfully retrieved on a future memory test. JOLs are metacognitive judgments that reflect monitoring that takes place during the encoding phase of learning. In a typical JOL experiment (e.g., Nelson & Dunlosky, 1991), participants are presented with a list of word pairs (e.g., OCEAN–TREE) and asked to make a prediction on each pair as to how likely it is that they will be able to recall the target word (e.g., TREE) when the cue word (e.g., OCEAN–???) is presented on a memory test. The accuracy of JOLs is assessed by comparing JOL ratings against actual performance on a criterion test such as recall and recognition (see Eakin & Moss, 2019, for a review of methodology). Numerous studies have shown that JOLs are predictive of actual performance, particularly when judgments are made after a period of delay (i.e., delayed JOL effect) compared with when judgments are made immediately after studying an item (e.g., Dunlosky & Nelson, 1992; Nelson & Dunlosky, 1991). Furthermore, studies have shown that JOLs are associated with study behaviors such as which items one would choose for further study and how long one would persist in studying (e.g., Metcalfe & Finn, 2008; Metcalfe & Kornell, 2005; Thiede & Dunlosky, 1999; see Metcalfe, 2009, for a review). These results indicate that JOLs are a critical component of self-regulated study behaviors, consistent with the notion that metacognition plays an important role in monitoring and control of cognition (Nelson & Narens, 1994).

Since the seminal work by Nelson and Dunlosky (1991), the measure of JOLs has attracted considerable attention from many researchers, and since then, many important discoveries have been made (see Metcalfe, 2009). However, there has been an ongoing debate as to whether asking participants to make JOLs would influence memory performance by creating a reactivity effect (see Soderstrom et al., 2015). The reactivity effect occurs when the measure itself alters behaviors (see Ericsson & Simon, 1993; Fox et al., 2011). In the case of JOLs, it may be the case that asking participants to make JOLs may alter their memory and subsequently modify their study behaviors. The issue of reactivity has been discussed since the early days of JOL research (e.g., Dunlosky & Nelson, 1992; Nelson & Dunlosky, 1991; Spellman & Bjork, 1992). However, thus

✉ Olesya Senkova
o.senkova@gmail.com

[1] Department of Psychology, Central Michigan University, Mount Pleasant, MI 48859, USA

far, only a handful of studies have directly investigated the issue, and the results have been mixed (see Double et al., 2018; Rhodes & Tauber, 2011, for a review).

For example, Benjamin et al. (1998) showed that making JOLs does not influence memory. In this study, participants were asked to learn lists of unrelated words and took an immediate free recall test after studying each list. For half of the lists, participants were asked to make a prediction for each recalled word as to whether they would be able to recall this item on a future recall test. When participants completed all the lists, they received the final recall test in which they were asked to recall the words from all the lists. The final recall test showed that there was no difference in recall between the lists for which participants made recall predictions and the lists for which they did not make recall predictions. These results were consistent with the notion that making JOLs does not influence memory.

In contrast, other studies have shown that making JOLs indeed influences memory. For example, Soderstrom et al. (2015) showed that the act of making JOLs enhances memory performance when study materials consist of strongly related cue–target word pairs. Furthermore, these researchers showed that providing JOLs makes the cue–target relationship salient to participants, similar to when participants are asked to generate a target word using a word fragment (e.g., ORCHID–FL_W_R). Note that the reactivity effect such as this is not a temporary phenomenon. In three experiments, Witherby and Tauber (2017) showed that the reactivity effect was still present on a test that was administered two days later. Other researchers also showed that asking participants to make JOLs modifies how participants study the material by changing the study goal of participants. In support of this notion, Mitchum et al. (2016) reported that presenting a probability scale for making a JOL informs participants that some items are more difficult to remember than others, which then leads participants to abandon the mastery study goal and, as a result, put less effort in learning difficult items.

Researchers investigating the delayed JOL effects also showed that making JOLs may influence memory. According to the self-fulfilling prophecy (SFP) hypothesis (Spellman & Bjork, 1992), delayed JOLs tend to be more accurate than immediate JOLs because making a delayed judgment tends to increase the likelihood of recalling the target on a subsequent memory test. They proposed two mechanisms that may influence memory when participants are asked to make delayed JOLs: covert recall attempts and the spacing effect. First, when making a JOL, participants would covertly attempt to recall the target, and if successful, they assign a high JOL rating. The successful retrieval of the target, in turn, acts as an additional study opportunity, increasing the likelihood of recalling the target on a subsequent memory test. Second, the difference between immediate and delayed judgments can be explained by the spacing effect or the effect showing that memory performance is higher when there is a spacing between repeated study trials. Based on this notion, it can be explained that the benefit of retrieving the target would be greater for the delayed judgment than for the immediate judgment because for the former, there is a spacing between the initial presentation and the second exposure (when the retrieval attempt is successful). The SFP hypothesis has received some support over the years. For instance, Kimball and Metcalfe (2003) showed that for items that received high JOL ratings, recall was similar regardless of whether a delayed JOL was made with a cue only or a delayed JOL was made with a cue only, followed by an exposure to both the cue and target. These results showed that for these high JOL items, making a delayed JOL was similar to receiving an extra exposure to the study material. For low JOL items, receiving an extra exposure to the cue and target enhanced memory because for these items, the likelihood of spontaneously retrieving the targets was low.

Other researchers (Akdoğan et al., 2016; Jönsson et al., 2012; Kelemen & Weaver, 1997) also showed that making delayed JOLs enhanced memory for these items on a delayed test. However, Tauber et al. (2015) showed that retrieval attempts associated with making delayed JOLs are not as effortful as when participants are explicitly tested because these retrieval attempts are often terminated when the cue word is not familiar. Also, Son and Metcalfe (2005) showed that when participants make delayed JOLs, they try to retrieve only some items and terminate retrieval attempts for other items, which is different from retrieval attempts during a test when participants tend to try to retrieve each item.

Because the results of the previous studies have been mixed as to whether asking participants to make JOLs would create a reactivity effect, there have been two papers that published the results of a meta-analysis on this topic (Double et al., 2018; Rhodes & Tauber, 2011). Rhodes and Tauber (2011) examined whether delaying judgments would indeed increase JOL accuracy and whether delaying judgments would also increase memory performance (reactivity effect). Double et al. (2018) focused on immediate JOLs, directly comparing memory performance between conditions with a JOL task and conditions without a JOL task. The analysis by Rhodes and Tauber included 45 studies, and the analysis by Double et al. included 17 studies. Rhodes and Tauber showed that there was a robust beneficial effect of delaying judgments on JOL accuracy; however, the effect of delaying judgments on memory was much smaller than the effect on accuracy. The results showed that delaying judgments increased memory under the following conditions: when both the cue and target were presented for making a judgment (as opposed to a cue only), when the materials were paired associates, when the delayed judgments were made with a delayed interval of 1 min or less, when the cues used for judgments did not match the cues used for the test (e.g., cue–target pairs for judgments

and cue only for test), when a within-subjects manipulation was used (i.e., participants making both immediate and delayed judgments), and when children as opposed to adults were tested. These results therefore indicated that delaying judgments can create a reactivity effect under some conditions. The results of Double et al. showed that a reactivity effect was present in 6 out of 17 studies. The reactivity was present when the study materials were word pairs that consisted of related cue–target as well as when the study materials were a word list consisting of single words.

Based on these findings, it is evident that making JOLs can modify memory; however, it is also evident that the reactivity effect is not always present. Therefore, we conducted two experiments to examine how making JOLs would influence memory performance. In particular, we investigated the type of processing a JOL task would induce. Our hypothesis was that making JOLs would induce item-specific processing because when one makes a JOL, one would focus attention on a particular item, thereby enhancing the distinctiveness of each item (Hunt, 2006, 2012). In other words, we propose that asking participants to make a JOL would be similar to asking participants to perform an encoding task (Craik & Tulving, 1975; Hyde & Jenkins, 1969), which is designed to induce item-specific processing (Einstein & Hunt, 1980; Hunt & Einstein, 1981; Hunt & McDaniel, 1993; Hunt & Seta, 1984). This notion has been proposed before (Mitchum et al., 2016; Schmidt & Schmidt, 2017); however, as far as we know, there has not been a direct test of this hypothesis.

In the present experiments, we presented participants with a list consisting of single words and asked them to perform a JOL task or a well-established item-specific processing task (a pleasantness rating task in Experiment 1, and a single imagery task in Experiment 2). Subsequently, memory performance in these conditions was compared with memory performance in an intentional learning control condition. We hypothesized that both JOL and item-specific processing tasks would enhance memory performance relative to the control condition, and that the enhancement would be similar between these two conditions. In Experiment 1, we selected a pleasantness rating task because this task was most notably used by Hunt and colleagues (e.g., Einstein & Hunt, 1980; Hunt & Einstein, 1981; Hunt & Seta, 1984) in their investigation of the effect of relational and item-specific processing on memory (see Hunt & McDaniel, 1993, for an extensive review). According to Hunt and colleagues, optimal memory requires both types of processing such that highlighting the uniqueness of each item (item-specific processing) is most beneficial to memory when one is also paying attention to the similarity among the items (relational processing). In their experiments, Hunt and colleagues used a pleasantness rating task to induce item-specific processing and a category sorting task to induce relational processing. In Experiment 2, we selected a single imagery task, in which participants were asked to create a vivid mental image of each word, because this task has been used to induce item-specific processing in several experiments (e.g., Burns & Schoff, 1998; Burns et al., 2007; Huff & Bodner, 2014; Otani & Hodge, 1991). It is also important to note that Hodge and Otani (1996) showed that memory performance was comparable between the pleasantness rating and single imagery tasks in both free recall and recognition, indicating that these tasks are similar in inducing item-specific processing.

To detect the effect of item-specific processing, we additionally manipulated the list type, such that half of the participants received a categorized list whereas the other half of the participants received an uncategorized list. According to Hunt and colleagues (e.g., Einstein & Hunt, 1980; Hunt, 2006, 2012; Hunt & Einstein, 1981; Hunt & McDaniel, 1993; Hunt & Seta, 1984), the structure of the study list is also important because a categorized list has a tendency to induce relational processing, and an uncategorized list has a tendency to induce item-specific processing. Thus, the processing induced by the list structure can interact with the processing induced by the encoding task such that optimal memory would result when both relational and item-specific processing are simultaneously present. Consistent with this notion, Hunt and colleagues (e.g., Hunt & Einstein, 1981) showed that when the list was categorized, recall was higher when participants engaged in an item-specific processing task (i.e., pleasantness rating) whereas when the list was uncategorized, recall was higher when participants engaged in a relational processing task (i.e., category sorting). Based on these previous findings, in the present experiments, we expected that an item-specific processing task such as JOLs would enhance memory when a list was categorized more so than when a list was uncategorized.

Note that in most JOL studies, study lists consist of word pairs (see Nelson et al., 2004). However, in the present experiments, we decided to use a list consisting of single words because manipulating encoding tasks is more established with a list of single words than a list of word pairs (e.g., levels of processing; Craik & Tulving, 1975). Furthermore, there have been studies in which participants were asked to make JOLs on single items (e.g., Dunlosky et al., 2000; Otani et al., 2014; Schmidt & Schmidt, 2017). Notably, a meta-analysis by Double et al. (2018) included three studies that used a word list that consisted of single words, and of these three studies, two (Yang et al., 2015; Zechmeister & Shaughnessy, 1980) showed a reactivity effect, whereas one (Tauber & Rhodes, 2012) did not.

In sum, in the present experiments, we assumed that making JOLs would induce item-specific processing similar to when one performs other item-specific processing tasks, such as rating pleasantness and creating a mental image of each word. Based on this notion, we predicted that making JOLs would enhance recall when the list was categorized more so

than when the list was uncategorized. Furthermore, the enhancement would be similar between the JOL and other item-specific processing conditions (pleasantness rating and single imagery). In addition, we predicted that when the list was uncategorized, the JOL condition would show minimal memory enhancement, again similar to the pleasantness rating and single imagery conditions.

## Experiment 1

In Experiment 1, making JOLs was compared with a well-known encoding task of rating pleasantness. Half of the participants studied a categorized list of words, whereas the other half of the participants studied an uncategorized list of words. We expected that relative to the control condition, memory enhancement would be similar between the JOL and pleasantness rating conditions, and that memory enhancement would be more likely to occur when the list was categorized as opposed to uncategorized.

### Method

**Participants** Participants were 32 male and 172 female undergraduate students attending introductory psychology courses at a public university in the Midwest region of the United States. They participated to earn extra course credit. An equal number ($n = 34$) of participants were randomly assigned to six between-subjects conditions, which were created by a 3 (encoding task: JOL, pleasantness rating, control) × 2 (list type: categorized, noncategorized) factorial design. We determined that 34 participants per condition would be sufficient based on an analysis using G*Power software (Faul et al., 2007). According to this analysis, assuming a medium effect size $f = 0.25$ with power $(1 - \beta)$ of .80, the minimum sample size would be 26 participants per condition. However, we acknowledge that detecting an interaction effect may require a larger sample size (see http://shiny.ieis.tue.nl/anova_power/ by Lakens & Caldwell, 2019). In fact, as indicated below, we did not have sufficient power to detect the interaction effect, even though there was enough power for a priori follow-up analyses. Nevertheless, previous researchers used a smaller sample size than ours when they manipulated a pleasantness rating task. In particular, Einstein and Hunt (1980) used 18 per condition, Hunt and Einstein (1981) used 19 per condition, and Hodge and Otani (1996) used 28 per condition. The experiment was conducted with approval given by the Institutional Review Board (IRB) where data were collected.

**Materials** Two lists, categorized and uncategorized, were constructed from English words selected from the Van Overschelde et al. (2004) category norms. The categorized list included 32 words from four different categories, and the noncategorized list included 32 words from 16 different categories (see Appendix 1). These words were chosen from the middle ranking for each category with the proportion of participants producing a particular word given a category cue ranging from .07 to .77. Furthermore, the length of words varied from five to eight letters. A PowerPoint presentation was used to present the words, one at a time, in the middle of the computer screen in lowercase letters at the rate of one word per 5 s. The order of the words was randomized once, and the same order was used for all participants. Each slide presenting a word was followed by an instruction slide, presented for 7 s. The instruction slide presented a scale from 0% to 100%, which was used for performing the encoding tasks. In addition, a sheet with randomly generated two-digit numbers was prepared for a filler task. A blank sheet of paper was used for a free recall test.

**Procedure** Participants, who were tested in small groups up to four individuals, were told that they would be presented with a list of words, and their task would be to remember as many of these words as possible. In addition, after each word was presented, participants in the JOL condition were asked to rate their JOL, indicating how likely they would be able to recall this word later using a scale from 0% (*definitely will not recall*) to 100% (*definitely will recall*). Participants in the pleasantness condition were asked to rate the pleasantness of the word using a scale from 0% (*definitely not pleasant*) to 100% (*definitely pleasant*). Participants in the control condition were asked to choose and write an arbitrary number from 0% to 100%. Following the study phase, the participants were asked to perform a filler task for 2 min, crossing out the numbers divisible by three. Then, participants completed a self-paced free recall test in which they were asked to recall and write as many of the study words as possible.

### Results

The dependent measure was the proportion of correctly recalled words. Table 1 shows the means across encoding task and list type. As shown, both the JOL and pleasantness rating

**Table 1** Mean proportion of correct recall as a function of encoding task and list type in Experiment 1

| | | Encoding task | | |
|---|---|---|---|---|
| | | JOL | Pleasantness rating | Control |
| Categorized list | *M* | .50 | .50 | .40 |
| | *SD* | .14 | .14 | .12 |
| Uncategorized list | *M* | .38 | .37 | .33 |
| | *SD* | .13 | .12 | .11 |

conditions resulted in higher recall than the control condition for both the categorized and uncategorized lists. However, the difference was much smaller for the uncategorized list than for the categorized list.

To compare the proportion of correctly recalled words across the conditions, we conducted a 3 (encoding task: JOL, pleasantness rating, control) × 2 (list type: categorized, uncategorized) analysis of variance (ANOVA). The results indicated that the main effect of encoding task was significant, $F(2, 198) = 7.98$, $MSE = 0.02$, $p < .001$, $\eta_p^2 = .08$. Least significant difference (LSD) tests showed that recall was higher in the JOL ($M = .44$, $SD = .15$, $p < .001$) and pleasantness rating conditions ($M = .43$, $SD = .14$, $p = .001$) than in the control condition ($M = .36$, $SD = .12$). No difference was found between the former two conditions ($p = .87$). The main effect of list type was also significant, $F(1, 198) = 35.72$, $MSE = 0.02$, $p < .001$, $\eta_p^2 = .15$. Recall was higher for the categorized list ($M = .46$, $SD = .14$) than for the uncategorized list ($M = .36$, $SD = .12$). The interaction was not significant, $F(2, 198) = 1.06$, $MSE = 0.02$, $p = .35$, $\eta_p^2 = .01$. Note, however, that the observed power for the interaction was only .23.

Although the interaction was not significant, we conducted further analyses based on the a priori hypothesis that the effect of making JOLs and rating pleasantness would be greater for the categorized list than for the uncategorized list because the effect of performing an item-processing task should be greater when the list encourages relational processing by emphasizing the similarity among items than when the list does not. To test this hypothesis, we conducted a separate one-way ANOVA on each list. For the categorized list, the results indicated that the difference among the encoding conditions was significant, $F(2, 99) = 6.80$, $MSE = 0.02$, $p = .002$, $\eta_p^2 = .12$. LSD tests showed that recall was higher in the JOL ($M = .50$, $SD = .14$, $p = .002$) and pleasantness rating conditions ($M = .50$, $SD = .14$, $p = .002$) than in the control condition ($M = .40$, $SD = .12$). No difference was found between the former two conditions ($p = .93$). For the uncategorized list, the results indicated that the difference among the encoding conditions was not significant, $F(2, 99) = 1.78$, $MSE = 0.01$, $p = .17$, $\eta_p^2 = 0.04$. As expected, these results showed that the effect of the JOL and pleasantness rating tasks was greater for the categorized list than for the uncategorized list.

In addition to these conventional analyses, we also computed Bayesian factors in order to model the data under the null and alternative hypotheses. A conventional analysis based on $p$ values only considers the null hypothesis, whereas a Bayesian analysis considers both the null and alternative hypotheses at the same time. The latter approach is superior to the former approach because it allows for an estimate of how likely the observed data would occur when the null hypothesis is true or when the alternative hypothesis is true (see Jarosz & Wiley, 2014). We specifically compared the JOL and pleasantness rating conditions for each list to show that the observed data fit the null hypothesis better than the alternative hypothesis. We computed Bayesian factors with the Rouder's method using IBM SPSS Statistics for Windows (Version 26.0; IBM Corp., Armonk, NY, USA). The results showed that for the categorized list, an estimated Bayesian factor (null/alternative) indicated that the observed data fit the null hypothesis 5.43 times better than the alternative hypothesis. For the uncategorized list, an estimated Bayesian factor (null/alternative) indicated that the observed data fit the null hypothesis 5.39 times better than the alternative hypothesis. These results provided strong evidence that as predicted, recall was similar between the JOL and pleasantness rating conditions for both the categorized and uncategorized lists.[1]

Next, we examined whether there was a difference in JOL ratings and accuracy between the categorized and uncategorized lists. Note that for these analyses, only two groups of participants (JOL conditions) were included. We consider JOL ratings first. Given that the categorized list was easier to learn than the uncategorized list, JOL ratings should be higher for the categorized list than for the uncategorized list. However, an independent-samples $t$ test on JOL ratings showed the difference was not significant, $t(66) = 0.61$, $p = .54$, indicating that there was no difference in JOL ratings between the categorized ($M = 47.75$, $SD = 17.48$) and uncategorized lists ($M = 45.10$, $SD = 18.13$). Accordingly, JOL ratings did not reflect the fact that the categorized list was easier to learn than the uncategorized list.

In terms of JOL accuracy, two types of accuracy need to be considered: relative and absolute. Relative accuracy indicates whether higher JOL ratings are associated with a higher likelihood of recalling items regardless of the actual numbers. In other words, recall should be higher for an item that was rated 80% than 40%, and this relationship should still hold even when the ratings are 60% and 20%. Absolute accuracy indicates whether the mean JOL rating matches actual recall. In other words, if the average JOL rating is 80%, the actual recall should be 80%. It is difficult to predict the effect of list type on relative accuracy. However, given that making JOLs increased actual recall when the list was categorized, the relative accuracy may have been lower for the categorized list than for the uncategorized list. There are several measures of relative accuracy, but we chose Goodman–Kruskal gamma because in the JOL literature, gamma is the most common. Gamma ranges from −1 to +1, with +1 indicating perfect accuracy and 0 indicating no association between JOL ratings and recall performance. The result of an independent-samples $t$ test showed that the difference was not significant, $t(66) = 1.12$, $p = .27$, indicating that there was no difference in relative accuracy between the categorized ($M = .19$, $SD = .31$, range: −.35 to .66) and uncategorized lists ($M = .27$, $SD = .26$, range:

[1] We thank a reviewer for suggesting Bayesian analyses.

−.28 to .63). These results indicated that relative accuracy was similar between the two lists even though making JOLs increased recall for the categorized list and not for the uncategorized list.

Absolute accuracy can be examined in several ways, but the most intuitive way is to compute *signed difference scores* for each participant (see Dunlosky & Metcalfe, 2009) by averaging JOL ratings across items and comparing the average rating with the actual recall (see Van Overschelde & Nelson, 2006). If the average JOL rating matches actual recall, the difference should be zero. If the rating is higher than actual recall, participants are said to be overconfident, whereas if the rating is lower than actual recall, participants are said to be underconfident. The effect of list type on absolute accuracy is also difficult to predict. However, given that making JOLs increased recall when the list was categorized, the absolute accuracy may have been lower for the categorized list than for the uncategorized list. Contrary to this expectation, the results showed that there was no significant difference between the categorized ($M = -2.16$, $SD = 18.37$) and uncategorized lists ($M = 7.41$, $SD = 24.18$), $t(66) = 1.84$, $p = .07$. These results showed that absolute accuracy was similar between the lists even though making JOLs increased recall for the categorized list and did not increase recall for the uncategorized list. However, note that the results showed a trend indicating that accuracy was higher when the list was categorized than uncategorized. As shown below, this trend was consistent with the results of Experiment 2, which showed that accuracy was significantly higher for the categorized list than for the uncategorized list.

### Discussion

Experiment 1 investigated whether making JOLs would influence memory performance by inducing item-specific processing. We tested this hypothesis by comparing a JOL task with a well-established task of rating pleasantness because the pleasantness rating task has been shown to induce item-specific processing (e.g., Einstein & Hunt, 1980; Hunt & Einstein, 1981; Hunt & Seta, 1984). We predicted that both the JOL and pleasantness rating tasks would produce memory enhancement relative to a control condition, and that the enhancement would be similar between the JOL and pleasantness rating conditions. Furthermore, we predicted that the enhancement would be greater when the list was categorized than uncategorized because it has been shown that the effect of item-specific processing is stronger when the list encourages relational-processing by emphasizing similarities among study items (Hunt & Einstein, 1981).

The results indicated that recall was higher in the JOL and pleasantness rating conditions than in the control condition, with the former two conditions showing similar performance. These results were in agreement with the assumption that making JOLs influences memory by encouraging item-specific processing similar to rating pleasantness. Although the Encoding Task × List Type interaction was not significant, further analyses comparing the encoding conditions on each list showed that the difference among the conditions was significant for the categorized list but not for the uncategorized list. These results are similar to the results reported by Soderstrom et al. (2015) that making JOLs enhanced memory performance for strongly related word pairs but not for unrelated word pairs, and further support the assumption that making JOLs promotes item-specific processing (e.g., Hunt & Einstein, 1981).

Another interesting finding was that there was no difference in JOL ratings as well as accuracy (both relative and absolute) between the categorized and uncategorized lists. If list type influences recall, it would be expected that JOL ratings and/or JOL accuracy would be affected by list type. Contrary to this expectation, list type did not make any difference in JOL ratings or JOL accuracy. Note that the ratings and accuracy were compared across groups of participants, and there are several problems associated with intergroup comparisons of JOL accuracy (see Dunlosky & Metcalfe, 2009; Schwartz & Metcalfe, 1994). Thus, further research is needed to examine JOL accuracy and the reactivity issue. Nevertheless, if making JOLs modifies memory, JOLs would no longer truly capture the natural ways individuals regulate their own study behaviors. Therefore, researchers need to exercise caution when using JOLs to investigate such behaviors.

In sum, the results of Experiment 1 supported the hypothesis that making JOLs enhances memory by inducing item-specific processing. However, rating pleasantness is only one of several tasks that have been shown to induce item-specific processing (see Hodge & Otani, 1996). In order to test whether the results were task-specific or process-specific, in Experiment 2, another well-known task was used to induce item-specific processing.

## Experiment 2

The purpose of Experiment 2 was to test further the hypothesis that making JOLs would be similar to engaging in an item-specific processing task. The design and the procedure of Experiment 2 were the same as those in Experiment 1, but in Experiment 2, making JOLs was compared with a *single imagery task* or a task of creating a vivid mental image of each word, which is another well-known item-specific processing task (see Hodge & Otani, 1996). Similar to Experiment 1, we hypothesized that both the JOL and single imagery conditions would produce similar memory enhancement relative to the control condition, and that memory enhancement would be greater when the list was categorized as opposed to when the list was uncategorized.

## Method

**Participants** Participants were 68 male and 172 female undergraduate students in introductory psychology courses at a public university in the Midwest region of the United States. They participated to earn extra course credit. An equal number ($n$ = 40) of participants were randomly assigned to six between-subjects conditions in a 3 (encoding task: JOL, single imagery, control) × 2 (list type: categorized, uncategorized) factorial design. We attempted to increase statistical power by increasing the sample size to 40 per condition. The experiment was conducted in accordance with the approval given by the IRB where data were collected.

**Materials and procedure** The materials were the same as those in Experiment 1. The procedures for the JOL and control conditions were the same as those in Experiment 1. The only difference between Experiments 1 and 2 was that in Experiment 2, participants in the single imagery condition were asked to create a mental image of each word as vividly as possible. In this condition, each slide presented a word and was followed by an instruction slide that showed a scale from 0% (*not very vivid*) to 100% (*very vivid*), and participants were asked to rate the vividness of the image they created using this scale and write down their ratings on the response sheet (see Appendix 2 for the instructions for the single imagery condition).

## Results and discussion

The dependent measure was the proportion of correctly recalled words. Table 2 shows the means across encoding task and list type. As shown, both the JOL and single imagery conditions showed higher recall than the control condition for both the categorized and uncategorized lists, and as expected, the difference was greater for the categorized list than for the uncategorized list.

To compare the proportion of correctly recalled words across the conditions, we conducted a 3 (encoding task: JOL, single imagery, control) × 2 (list type: categorized, uncategorized) ANOVA. The results indicated that the main effect of encoding task was significant, $F(2, 234) = 11.39$, $MSE = 0.01$, $p < .001$, $\eta_p^2 = .09$. LSD tests showed that recall was higher in the JOL ($M = .44$, $SD = .15$, $p < .001$) and single imagery conditions ($M = .45$, $SD = .14$, $p < .001$) than in the control condition ($M = .37$, $SD = .11$). No difference was found between the JOL and single imagery conditions ($p = .43$). The main effect of list type was also significant, $F(1, 234) = 71.52$, $MSE = 0.01$, $p < .001$, $\eta_p^2 = .23$. Recall was higher for the categorized list ($M = .48$, $SD = .14$) than for the noncategorized list ($M = .36$, $SD = .10$). Lastly, the Encoding Task × List Type interaction was significant, $F(2, 234) = 3.49$, $MSE = 0.01$, $p = .03$, $\eta_p^2 = .03$.

Because the interaction was significant, we conducted a separate one-way ANOVA on each list. For the categorized list, the results indicated that the difference among the encoding conditions was significant, $F(2, 117) = 10.67$, $MSE = 0.02$, $p < .001$, $\eta_p^2 = .15$. LSD tests showed that recall was higher in the JOL ($M = .50$, $SD = .15$, $p = .001$) and single imagery conditions ($M = .54$, $SD = .13$, $p < .001$) than in the control condition ($M = .41$, $SD = .11$). No difference was found between the JOL and single imagery conditions ($p = .23$). For the uncategorized list, the results indicated that the difference among the encoding conditions was not significant, $F(2, 117) = 1.88$, $MSE = 0.01$, $p = .16$, $\eta_p^2 = 0.03$.

The results of Bayesian analyses also confirmed that there was no difference in recall between the JOL and single imagery conditions. For the categorized list, an estimated Bayesian factor (null/alternative) showed that the observed data fit the null hypothesis 3.25 times better than the alternative hypothesis. For the uncategorized list, an estimated Bayesian factor (null/alternative) showed that the observed data fit the null hypothesis 5.64 times better than the alternative hypothesis. These results provided strong evidence that as predicted, recall was similar between the JOL and single imagery conditions.

Next, we examined JOL ratings and accuracy by comparing the JOL conditions between the two lists. An independent-samples $t$ test on JOL ratings showed the difference was not significant, $t(78) = 0.67$, $p = .50$, indicating that there was no difference in JOL ratings between the categorized ($M = 49.68$, $SD = 16.01$) and uncategorized lists ($M = 47.30$, $SD = 15.65$), despite the fact that recall was easier for the categorized list than for the uncategorized list.

To examine the relative accuracy of JOLs, we computed Goodman-Kruskal gamma scores. An independent-samples $t$ test showed that there was no difference between the categorized ($M = .24$, $SD = .25$, range: −.32 to .68) and uncategorized lists ($M = .30$, $SD = .27$, range: −.71 to .69), $t(78) = 1.0$, $p = .32$, indicating that relative accuracy was similar between the two lists despite the fact that making JOLs increased recall for the categorized list and did not increase recall for the uncategorized list. To examine absolute accuracy, we computed signed difference scores comparing the average JOL rating

**Table 2** Mean proportion of correct recall as a function of encoding task and list type in Experiment 2

| | | Encoding task | | |
|---|---|---|---|---|
| | | JOL | Single imagery | Control |
| Categorized list | *M* | .50 | .54 | .41 |
| | *SD* | .15 | .13 | .11 |
| Uncategorized list | *M* | .37 | .37 | .33 |
| | *SD* | .12 | .07 | .10 |

with actual recall for each participant. An independent-samples $t$ test showed that there was a significant difference between the categorized ($M$ = −.72, $SD$ = 22.06) and uncategorized lists ($M$ = 9.95, $SD$ = 19.56), $t(78) = 2.29$, $p$ = .03, indicating that absolute accuracy was higher for the categorized list than for the uncategorized list. These results therefore indicated that absolute accuracy of JOLs can be influenced by the type of processing that the study material would promote.

Experiment 2 investigated whether making JOLs and creating a single image would influence memory performance in a similar way. This prediction was based on the assumption that both JOL and single imagery tasks would promote item-specific processing. If this assumption is correct, both the JOL and single imagery conditions would produce memory enhancement for the categorized list. In contrast, for the uncategorized list, memory enhancement by these conditions would be minimal because the uncategorized list would naturally promote item-specific processing, which is the same processing type that the JOL and single imagery tasks would promote. Based on the notion that optimal memory requires a combined effect of relational and item-specific processing, memory enhancement should occur when the study material promotes relational processing (categorized list), whereas the encoding task promotes item-specific processing (JOL and single imagery). The results were consistent with these expectations. There was a significant interaction between list type and encoding task such that for the categorized list, recall was higher in the JOL and single imagery conditions than in the control condition whereas for the uncategorized list, recall was similar among the three conditions.

With regard to JOL ratings and relative accuracy, there was no difference between the categorized and uncategorized lists, replicating the results of Experiment 1. However, absolute accuracy was different between the two lists, such that absolute accuracy was higher for the categorized list than for the uncategorized list. These results therefore showed that JOL accuracy can be influenced by the type of processing that the study material would promote. Nevertheless, it is difficult to explain this finding. One possibility is that participants were naturally overconfident, and therefore, by increasing recall, their ratings became more accurate by reducing the difference between their ratings and actual recall. Obviously, other possibilities need to be explored in the future studies.

## General discussion

The present two experiments examined whether the act of making JOLs would enhance memory performance by inducing item-specific processing. As mentioned in the Introduction, there has been an ongoing debate as to whether asking participants to make JOLs would create a reactivity effect because some studies have shown that making JOLs would influence memory (e.g., Soderstrom et al., 2015; Zechmeister & Shaughnessy, 1980), whereas other studies have shown that making JOLs would not influence memory (e.g., Benjamin et al., 1998; Tauber & Rhodes, 2012). In fact, two papers that published a meta-analysis showed that a JOL task would create a reactivity effect under some conditions, but not in other conditions (Double et al., 2018; Rhodes & Tauber, 2011).

In the present experiments, we hypothesized that making JOLs would enhance memory by inducing item-specific processing because the task of making JOLs would direct one's attention to a particular item and enhance the distinctiveness of each item in memory (Hunt, 2006, 2012). Our assumption was that a JOL task is similar to other encoding tasks that are designed to induce a particular type of processing by asking participants to make a particular type of judgment on a given item (Craik & Tulving, 1975; Hyde & Jenkins, 1969). These tasks have been extensively used to study the effect of encoding processing on memory at least since the beginning of the levels of processing approach (Craik & Lockhart, 1972). Hunt and colleagues expanded on the levels of processing approach and proposed that there are two types of processing that are particularly important to memory—item-specific processing and relational processing. Their proposal was that when these two types of processing are combined, memory is optimized by taking advantage of two important principles of memory—organization and distinctiveness (see Hunt, 2006, 2012; Hunt & McDaniel, 1993). It is also important to note that encoding tasks are not the only source that would influence the type of processing. Hunt and colleagues showed that the type of study materials is also important because when the study materials promote relational processing, performing a task that would induce item-specific processing becomes beneficial, whereas when the study materials promote item-specific processing, performing a task that would induce relational processing becomes beneficial (see Hunt & Einstein, 1981; Hunt & McDaniel, 1993). Our assumption was that making JOLs would induce item-specific processing and increase the distinctiveness of each item. If this is the mechanism of how JOLs would enhance memory, we should be able to detect it by presenting a categorized list, which would promote relational processing. Accordingly, in the present experiments, we manipulated the encoding task and the study material. The tasks were a JOL task and two other tasks that have been shown to induce item-specific processing: the pleasantness rating and single imagery tasks (see Hodge & Otani, 1996). The study material was a categorized list and an uncategorized list. If making JOLs would induce item-specific processing similar to other item-specific processing tasks, recall should be enhanced when the list is categorized more so than when the list is uncategorized. Furthermore, the enhancement should be similar between the JOL task and the

other tasks that induce item-specific processing (i.e., pleasantness rating and single imagery; see Hodge & Otani, 1996). In addition, when the list is uncategorized, the enhancement should be minimal because the uncategorized list would promote item-specific processing, which is the same processing type as these tasks would promote.

The results were consistent with these predictions. The results of Experiment 1 showed that for the categorized list, recall was higher in the JOL and pleasantness rating conditions than in the control condition. Furthermore, the former two conditions produced similar recall performance. These results were replicated in Experiment 2 with the single imagery task. The results of Experiment 2 showed that for the categorized list, recall was higher in the JOL and single imagery conditions than in the control condition, with the former two conditions showing similar recall performance. The results were different for the uncategorized list. In both Experiments 1 and 2, recall was similar among all these conditions when the list was uncategorized.[2] These results therefore support the hypothesis that making JOLs promotes item-specific processing similar to rating pleasantness or creating a single mental image of each word. These results are also in agreement with the results obtained by Soderstrom et al. (2015), which showed that making JOLs enhanced memory performance for strongly related cue–target pairs, but not for weakly related or unrelated cue–target pairs. Their explanation was that making JOLs would make the cue–target relationship salient. This explanation is similar to the notion that memory is enhanced when both similarities and differences are emphasized.

It must be noted that in the meta-analysis by Double et al. (2018), there were three studies that used a list of single words. Among these, two showed a reactivity effect (Yang et al., 2015; Zechmeister & Shaughnessy, 1980), whereas one did not show a reactivity effect (Tauber & Rhodes, 2012). It is important to note that none of these studies used a categorized list, and thus the reason that memory enhancement did or did not occur in these studies is not clear. However, it is reasonable to assume that a JOL task is an encoding task (Craik & Tulving, 1975; Hyde & Jenkins, 1969) and could induce a deep level of processing, and if so, would enhance memory performance. In fact, in the present experiments, when the list was uncategorized, memory performance was slightly higher for the JOL condition (albeit nonsignificant) than for the control condition in both experiments. One additional note on encoding tasks is that most of the past studies that manipulated encoding tasks used an incidental learning instruction in an attempt to keep processing as pure as possible (e.g., Craik & Tulving, 1975; see Otani et al., 2019, for the history of memory research methodology). Obviously, in a JOL study, it is impossible to use an incidental learning instruction. Thus, the lack of difference among the conditions with the uncategorized list is understandable. Compared with the past studies that used an incidental learning instruction, the effect of the processing manipulation may not have been strong enough in the present experiments, perhaps due to mixing of item-specific processing and intentional learning strategies. Nevertheless, further studies are needed to investigate the reason that the uncategorized list did not show a difference among the conditions.

We also acknowledge that similar performance does not necessarily means that similar processes are responsible for the performance across conditions. However, there is no direct measure of type of processing underlying memory performance. This has been the main difficulty of the processing approach to memory since the beginning of the levels of processing approach (Craik, 2002). Nevertheless, there have been attempts to detect item-specific and relational processing using a repeated-measures paradigm (e.g., Burns, 1993; Burns et al., 2007; Burns & Schoff, 1998; Mulligan, 2000, 2002). In this paradigm, the same recall test is repeated multiple times without providing study trials between tests. The patterns of item gains and loss across tests have been used as an index of relational and item-specific processing (see also Hunt & McDaniel, 1993). Thus, this approach should be used in the future to provide converging evidence that making JOLs induces item-specific processing.

In terms of JOL ratings, both experiments showed that there was no difference between the categorized and uncategorized lists. This finding was surprising because the categorized list was easier to learn than the uncategorized list. A possible explanation is that we used common words from various categories (see Appendix 1), and therefore participants may not have viewed these as particularly easy or difficult to remember. In fact, the JOL ratings were about 50% for both lists in both experiments. Furthermore, the order of the words was randomized. This means that the categorical nature of the list may not have been particularly salient in the categorized list. In the future, the categorized list should be presented with the words from each category in a block. Additionally, each word could be presented with a category cue (e.g., fruit–apple) to emphasize the categorical nature of the list.

Regarding accuracy, in Experiment 1, list type did not influence relative accuracy or absolute accuracy. Because

[2] We also conducted an additional analysis by combining the data from Experiments 1 and 2 in order to increase statistical power. The aim of this analysis was to further examine whether making JOLs as well as performing an item-specific processing task had increased recall when the list was uncategorized because there was a slight increase in recall in these conditions relative to the control condition for both Experiments 1 and 2. Because memory performance was comparable for the pleasantness rating (Experiment 1) and single imagery (Experiment 2) conditions, we combined these conditions as an item-specific processing condition, such that the analysis was based on a 3 (encoding task: JOL, item-specific processing, control) × 2 (list type: categorized, uncategorized) ANOVA. The results of this analysis did not change the conclusion of the present experiments.

making JOLs increased recall for the categorized list, we expected that JOL ratings would not reflect actual recall, thereby resulting in lower accuracy. This expectation was not confirmed. In Experiment 2, relative accuracy was similar between the two lists, replicating Experiment 1. The result of absolute accuracy was different in Experiment 2 such that absolute accuracy was higher for the categorized list than for the uncategorized list. This finding was also contrary to the expectation because we expected that for the categorized list, JOL ratings would not match actual recall because the act of making JOLs itself would result in increased recall. A possible explanation of this unexpected result is that participants were naturally overconfident, and therefore, when actual recall was increased, accuracy was increased by reducing overconfidence, such that the difference between the ratings and actual recall became smaller. However, Experiment 1 did not show this effect, even though there was a trend showing that participants were slightly more accurate when the list was categorized than uncategorized. These results therefore indicated that absolute accuracy of JOLs can be altered when the study material promotes relational processing and making JOLs modifies memory. Obviously, these speculations need to be examined more directly in future studies, particularly using within-participant comparisons (for the difficulty of comparing accuracy between groups, see Dunlosky & Metcalfe, 2009; Schwartz & Metcalfe, 1994). Furthermore, it has been well established that immediate JOLs are less accurate than delayed JOLs (Rhodes & Tauber, 2011). Accordingly, immediate JOLs may not be sensitive enough to show the effect of processing types on relative and absolute accuracy. Therefore, future studies should use delayed JOLs to investigate this issue.

In sum, the results of the current experiments showed that the relationship between JOLs and recall performance is complex. In fact, it is not always the case that a variable that influences memory also influences JOLs. Such dissociations have been reported in the past (see Schwartz & Efklides, 2012, for an extensive review). For example, studying a list multiple times has been shown to increase cued-recall performance; however, Kornell and Bjork (2009) reported that participants did not increase JOL ratings to match the increased performance.

Although the present experiments did not show a straightforward effect that memory enhancement has on JOL ratings and accuracy, the results of the present experiments make it clear that researchers need to carefully consider how making a JOL influences memory when investigating self-regulated study behaviors. Making JOLs can increase memory and may modify self-regulated study behaviors (see, e.g., Dunlosky & Connor, 1997; Metcalfe & Finn, 2008; Metcalfe & Kornell, 2005; Mitchum et al., 2016; Thiede & Dunlosky, 1999, for how JOLs can influence study behaviors). Note that these results are reminiscent of the difficulty associated with introspection, which Wundt and Titchener used to investigate subjective experience (see Fox et al., 2011; Murray, 1983). Throughout the history of psychology, it has been shown that when one introspects on one's own internal experience, the act of introspection itself may modify the experience (see Fox et al., 2011, for a meta-analysis showing when introspection does and does not become reactive). However, the results of the present experiments did not show that making JOLs always modifies memory. The results showed that when the list was uncategorized, memory enhancement did not occur. These results are another reminder that attention to detail is called for when designing a study. Any method of measuring behavior (particularly subjective experience) needs to be tested before implementing it in an experiment, as it may influence the behavior itself. Furthermore, whenever possible, an appropriate control condition needs to be included in order to assess whether the measurement itself is modifying the behavior (Mitchum et al., 2016). It is also advisable that other nonreactive measures be explored (Double et al., 2018; Mitchum et al., 2016).

The limitations of the present experiments include the use of lists with single words. As mentioned earlier, most of studies investigating JOLs have used word pairs (see Double et al., 2018). Therefore, the future studies need to examine whether making JOLs would induce item-specific processing using word pairs. Another limitation is that we did not explicitly control the word characteristics such as emotionality. However, the same list was used across conditions, therefore it is unlikely that the word characteristics created a confounding variable.

In conclusion, the present study showed that making JOLs induces item-specific processing and enhances memory when the study material promotes relational processing. Accordingly, it is important to take this into account when using JOLs to investigate self-regulated study behaviors because it is possible that asking participants to make JOLs may change memory and modify study behaviors (e.g., Mitchum et al., 2016). Nevertheless, in terms of practical applications, these results showed that there is a benefit to asking oneself whether a particular study item is learned. By doing so, one would promote item-specific processing of the item, which may in turn lead to enhanced learning and retention of the item.

The data and materials for all experiments are available from the first author. None of the experiments was preregistered.

**Acknowledgements** This paper is based on the first author's dissertation submitted to Central Michigan University. We thank Bennett Schwartz and Abby Knoll for helpful comments on an earlier draft.

## Appendix 1

| Categorized list | Uncategorized list |
|---|---|
| 1. valley | 1. tiger |
| 2. potato | 2. whisk |
| 3. tiger | 3. engineer |
| 4. river | 4. blizzard |
| 5. horse | 5. tennis |
| 6. office | 6. dresser |
| 7. squash | 7. radish |
| 8. canyon | 8. glacier |
| 9. raccoon | 9. sword |
| 10. volcano | 10. jacket |
| 11. stairs | 11. airplane |
| 12. pepper | 12. stomach |
| 13. ocean | 13. moose |
| 14. rabbit | 14. tongs |
| 15. lettuce | 15. bicycle |
| 16. floor | 16. elbow |
| 17. cliff | 17. journal |
| 18. giraffe | 18. thunder |
| 19. lobby | 19. mansion |
| 20. radish | 20. drill |
| 21. ceiling | 21. couch |
| 22. island | 22. gloves |
| 23. elephant | 23. lobby |
| 24. carrot | 24. river |
| 25. moose | 25. stairs |
| 26. window | 26. hockey |
| 27. stream | 27. potato |
| 28. tomato | 28. textbook |
| 29. elevator | 29. shack |
| 30. squirrel | 30. ruler |
| 31. basement | 31. lawyer |
| 32. cabbage | 32. missile |

## Appendix 2

### Single imagery task instruction

Welcome to the experiment. This experiment investigates how people study words. You will be presented with 32 words. Each word will be presented on the screen, one at a time, for 5 seconds, and your task will be to remember as many words as possible. Also, you will be given a sheet of paper, and after the presentation of each word, you will be asked to create as vivid mental image as possible of the object denoted by the word and write down your rating of how vivid this image is. For instance, you may see a word "diamond," and on the next slide you will see a rating scale from 0% (*not at all vivid*) to 100% (*extremely vivid*). Please form a mental picture of diamond, and think about how clear, bright, and detailed this picture is. When the picture is very clear, very bright, and very detailed, rate its vividness as 100%.

## References


Akdoğan, E., Izaute, M., Danion, J., Vidailhet, P., & Bacon, E. (2016). Is retrieval the key? metamemory judgment and testing as learning strategies. *Memory, 24*(10), 1390–1395. https://doi.org/10.1080/09658211.2015.1112812

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

Burns, D. J. (1993). Item gains and losses during hypermnesic recall: Implications for the item-specific-relational information distinction. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 19*(1), 163–173. https://doi.org/10.1037/0278-7393.19.1.163

Burns, D. J., Jenkins, C. L., & Dean, E. E. (2007). Falsely recalled items are rich in item-specific information. *Memory & Cognition, 35*, 1630–1640. https://doi.org/10.3758/BF03193497

Burns, D. J., & Schoff, K. M. (1998). Slow and steady often ties the race: Effects of item-specific and relational processing on cumulative recall. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 24,* 1041–1051. https://doi.org/10.1037/0278-7393.24.4.1041

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

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

Craik, F. I. M. (2002). Levels of processing: Past, present . . . and future? *Memory, 10*(5/6), 305–318. https://doi.org/10.1080/09658210244000135

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

Dunlosky, J., & Connor, L. T. (1997). Age differences in the allocation of study time account for age differences in memory performance. *Memory & Cognition, 25*(5), 691–700. https://doi.org/10.3758/BF03211311

Dunlosky, J., Hunt, R. R., & Clark, E. (2000). Is perceptual salience needed in explanations of the isolation effect? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *26*(3), 649–657. https://doi.org/10.1037/0278-7393.26.3.649

Dunlosky, J., & Metcalfe, J. (2009). *Metacognition.* Los Angeles, CA: Sage.

Dunlosky, J., & Nelson, T. O. (1992). Importance of the kind of cue for judgments of learning (JOL) and the delayed-JOL effect. *Memory & Cognition, 20*, 374–380. https://doi.org/10.3758/BF03210921

Eakin, D. K., & Moss, J. M. (2019). The methodology of metamemory and metacomprehension. In H. Otani & B.L. Schwartz (Eds.), *Handbook of research methods in human memory*. New York: Routledge. https://doi.org/10.4324/9780429439957-8

Einstein, G. O., & Hunt, R. R. (1980). Levels of processing and organization: Additive effects of individual-item and relational processing. *Journal of Experimental Psychology: Human Learning and Memory, 6*(5), 588–598. https://doi.org/10.1037/0278-7393.6.5.588

Ericsson, K. A., & Simon, H. A. (1993). *Protocol analysis: Verbal reports as data* (Rev. ed.). The MIT Press, Cambridge, MA.

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/bf03193146

Fox, M. C., Ericsson, K. A., & Best, R. (2011). Do procedures for verbal reporting of thinking have to be reactive? A meta-analysis and recommendations for best reporting methods. *Psychological Bulletin*, *137*(2), 316–344. https://doi.org/10.1037/a0021663

Hodge, M. H., & Otani, H. (1996). Beyond category sorting and pleasantness rating: Inducing relational and item-specific processing. *Memory & Cognition*, *24*, 110–115. https://doi.org/10.3758/BF03197277

Huff, M. J., & Bodner, G. E. (2014). All varieties of encoding variability are not created equal: Separating variable processing from variable tasks. *Journal of Memory and Language, 73*, 43–58. https://doi.org/10.1016/j.jml.2014.02.004

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 3–25). Oxford: Oxford University Press. https://doi.org/10.1093/acprof:oso/9780195169669.003.0001

Hunt, R. R. (2012). The co-action of similarity and difference in memory. In B. H. Ross (Ed.), *Psychology of learning and motivation* (Vol. 56, pp. 1–46). San Diego, CA: Academic Press. https://doi.org/10.1016/B978-0-12-394393-4.00001-7

Hunt, R. R., & Einstein, G. O. (1981). Relational and item-specific information in memory. *Journal of Verbal Learning and Verbal Behavior, 20*(5), 497–514. https://doi.org/10.1016/S0022-5371(81)90138-9

Hunt, R. R., & McDaniel, M. A. (1993). The enigma of organization and distinctiveness. *Journal of Memory and Language*, *32*(4), 421–445. https://doi.org/10.1006/jmla.1993.1023

Hunt, R. R., & Seta, C. E. (1984). Category size effects in recall: The roles of relational and individual item information. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *10*(3), 454–464. https://doi.org/10.1037/0278-7393.10.3.454

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

Jarosz, A. F., & Wiley, J. (2014). What are the odds? A practical guide to computing and reporting Bayes factors. *The Journal of Problem Solving*, *7*(1), 2–9. https://doi.org/10.7771/1932-6246.1167

Jönsson, F. U., Hedner, M., & Olsson, M. J. (2012). The testing effect as a function of explicit testing instructions and judgments of learning. *Experimental Psychology, 59*, 251–257. https://doi.org/10.1027/1618-3169/a000150

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

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

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: Overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*(4), 449–468. https://doi.org/10.1037/a0017350

Lakens, D., & Caldwell, A. R. (2019). *Simulation-based power-analysis for factorial ANOVA designs.* PsyArXiv. https://doi.org/10.31234/osf.io/baxsf

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychological Science, 18*(3), 159–163. https://doi.org/10.1111/j.1467-8721.2009.01628.x

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

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

Mitchum, A. L., Kelley, C. M., & Fox, M. C. (2016). When asking the question changes the ultimate answer: Metamemory judgments change memory. *Journal of Experimental Psychology: General, 145*(2), 200–219. https://doi.org/10.1037/a0039923

Mulligan, N. W. (2000). Perceptual interference at encoding enhances item-specific encoding and disrupts relational encoding: Evidence from multiple recall tests. *Memory & Cognition, 28*, 539–546. https://doi.org/10.3758/BF03201244

Mulligan, N. W. (2002). The emergence of item-specific encoding effects in between-subjects designs: Perceptual interference and multiple recall tests. *Psychonomic Bulletin & Review, 9*, 375–382. https://doi.org/10.3758/BF03196296

Murray, D. J. (1983). *A history of Western psychology*. Englewood Cliffs, NY: Prentice Hall.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at prediction subsequent recall: The delayed-JOL effect. *Psychological Science, 2*(4), 267–270. https://doi.org/10.1111/j.1467-9280.1991.tb00147.x

Nelson, T. O., & Narens, L. (1994). Why investigate metacognition? In J. Metcalfe, & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 1–25). Cambridge, MA: MIT Press.

Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment recall and monitoring (PRAM). *Psychological Methods*, *9*, 53–69. https://doi.org/10.1037/1082-989X.9.1.53

Otani, H., & Hodge, M. (1991). Does Hypermnesia occur in recognition and cued-recall? *The American Journal of Psychology, 104*(1), 101–116. https://doi.org/10.2307/1422853

Otani, H., Schwartz, B. L., & Knoll, A. R. (2019). History of methods in memory science: From Ebbinghaus to fMRI. In H. Otani & B. L. Schwartz (Eds.), *Handbook of research methods in human memory* (pp. 1–18). New York: Routledge.

Otani, H., Von Glahn, N. R., Libkuman, T. M., Goernert, P. N., & Kato, K. (2014). Emotional salience and the isolation effect. *The Journal of General Psychology*, *141*(1), 35–46. https://doi.org/10.1080/00221309.2013.848180

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

Schmidt, S. R., & Schmidt, C. R. (2017). Revisiting von Restorff's early isolation effect. *Memory & Cognition*, *45*, 194–207. https://doi.org/10.3758/s13421-016-0651-6

Schwartz, B. L., & Metcalfe, J. (1994). Methodological problems and pitfalls in the study of human metacognition. In Metcalfe, J. & Shimamura, A. P. (Eds.) *Metacognition: Knowing about knowing* (pp. 93–113). Cambridge, MA: MIT Press.

Schwartz, B. L., & Efklides, A. (2012). Metamemory and memory efficiency: Implications for student learning. *Journal of Applied Research in Memory and Cognition*, *1*(3), 145–151. https://doi.org/10.1016/j.jarmac.2012.06.002

Soderstrom, N. C., Clark, C. T., Halamish, V., & Bjork, E. L. (2015). Judgments of learning as memory modifiers. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(2), 553–558. https://doi.org/10.1037/a0038388

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: Evidence for a two-stage process. *Memory & Cognition, 33*, 1116–1129. https://doi.org/10.3758/BF03193217

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), 315–316. https://doi.org/10.1111/j.1467-9280.1992.tb00680.x

Tauber, S. K., Dunlosky, J., & Rawson, K. A. (2015). The influence of retrieval practice versus delayed judgments of learning on memory: Resolving a memory-metamemory paradox. *Experimental*

*Psychology, 62*, 254–263. https://doi.org/10.1027/1618-3169/a000296

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

Thiede, K. W., & Dunlosky, J. (1999). Toward a general model of self-regulated study: An analysis of selection of items for study and self-paced study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *25*(4), 1024–1037. https://doi.org/10.1037/0278-7393.25.4.1024

Van Overschelde, J. P., & Nelson, T. O. (2006). Delayed judgments of learning cause both a decrease in absolute accuracy (calibration) and an increase in relative accuracy (resolution). *Memory & Cognition, 34*, 1527–1538. https://doi.org/10.3758/BF03195916

Van Overschelde, J. P., Rawson, K. A., & Dunlosky, J. (2004). Category norms: An updated and expanded version of the Battig and Montague (1969) norms. *Journal of Memory and Language*, *50*(3), 289–335. https://doi.org/10.1016/j.jml.2003.10.003

Witherby, A. E., & Tauber, S. K. (2017). The influence of judgments of learning on long-term learning and short-term performance. *Journal of Applied Research in Memory and Cognition, 6*(4), 496–503. https://doi.org/10.1016/j.jarmac.2017.08.004

Yang, H., Cai, Y., Liu, Q., Zhao, X., Wang, Q., Chen, C., & Xue, G. (2015). Differential neural correlates underlie judgment of learning and subsequent memory performance. *Frontiers in Psychology, 6*, 12. https://doi.org/10.3389/fpsyg.2015.01699

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