Educational Psychology Review (2021) 33:693–712
https://doi.org/10.1007/s10648-020-09556-8

INTERVENTION STUDY

Do Judgments of Learning Directly Enhance Learning of Educational Materials?

Robert Ariel[1] • Jeffrey D. Karpicke[2] • Amber E. Witherby[3] • Sarah K. Tauber[4]

Published online: 30 July 2020

Abstract

When people make judgments of learning (JOLs) after studying paired associates, the process they engage in to monitor their learning can directly enhance learning for some types of material (Soderstrom et al. 2015). The current experiments investigated whether JOLs directly enhance learning educationally relevant texts. Across 5 experiments (N = 703), people read several sections of an educational textbook with or without JOLs embedded between each section. We manipulated whether JOLs queried one's understanding of the text at the aggregate level (Experiment 1) or for specific concepts in the text (Experiment 2a, 2b, 3, and 4). We also manipulated whether JOLs were framed to afford covert retrieval practice by prompting judgments with either the target information present or absent (Experiment 3). In most cases, instructing students to make JOLs did not improve comprehension above and beyond just reading the text. However, when people were instructed to retrieve information prior to making JOLs (Experiment 4), large learning gains occurred. These results indicate that JOLs in their standard form are unlikely to produce educational benefits to text comprehension in part because learners do not spontaneously retrieve criterial information when making metacomprehension judgments.

Keywords Judgments of learning · Judgment reactivity · Metacomprehension · Metacognition

Metacognitive models of learning assume that monitoring exerts an indirect effect on learning outcomes (Ariel et al. 2009; Butler and Winne 1995; Metcalfe 2009; Thiede and Dunlosky 1999; Winne and Hadwin 1998). People monitor—or judge—their learning to assess whether

✉ Robert Ariel
rariel@vwu.edu

[1] Department of Psychology, Virginia Wesleyan University, Virginia Beach, VA 23455, USA

[2] Department of Psychological Sciences, Purdue University, West Lafayette, IN 47907, USA

[3] Department of Psychology, Iowa State University, Ames, IA 50011, USA

[4] Department of Psychology, Texas Christian University, Fort Worth, TX 76129, USA

information has been sufficiently encoded and use feedback from monitoring to inform control processes they utilize to learn material. Monitoring in this case affects the quality of control decisions during study (e.g., study-time allocation, strategy use) and control decisions in turn affect learning (Metcalfe and Finn 2008; Thiede et al. 2003). Recent evidence, however, indicates that monitoring learning by making judgments of learning (JOLs) can also directly enhance learning for certain material (sometimes called "judgment reactivity"; Janes et al. 2018; Mitchum et al. 2016; Soderstrom et al. 2015; Witherby and Tauber 2017). These recent findings suggest that simply instructing students to make JOLs when studying course material could improve academic achievement.

Although recent evidence for direct effects of JOLs on learning have led to much enthusiasm and speculation about the educational benefits of JOLs, positive effects of making JOLs on learning are not ubiquitous (Double et al. 2018; Janes et al. 2018; Mitchum et al. 2016), and no prior research has examined whether making JOLs produces direct benefits for learning educationally relevant material. In the current experiments, we investigated whether making JOLs after reading sections of an educational textbook directly improves text comprehension. These experiments are the first to investigate potential reactive effects of JOLs on text learning. Before describing the current experiments in more detail, we briefly review evidence that JOLs can directly improve learning and describe a potential mechanism that could contribute to the direct benefits of making JOLs on text comprehension.

The Direct Effects of JOLs on Learning

The procedure for making JOLs typically involves instructing learners to rate their confidence on a 0 to 100% scale in their ability to later recall or answer questions about some recently studied material (e.g., paired associate items, category exemplars, or sections of texts). A growing body of evidence indicates that the act of making JOLs in itself can improve retention of certain types of information (Arbuckle and Cuddy 1969; Double et al. 2018; Janes et al. 2018; Mitchum et al. 2016; Soderstrom et al. 2015; Sommer et al. 1995; Tauber and Witherby 2019; Witherby and Tauber 2017; Zechmeister and Shaughnessy 1980). For example, when college students study semantically related word pairs (e.g., *king—crown*), cued recall is improved when they make JOLs after studying each item relative to when they do not make JOLs (Janes et al. 2018; Mitchum et al. 2016; Soderstrom et al. 2015). Memory improvements due to making JOLs occur across both short and long retention intervals (Witherby and Tauber 2017), and benefits have been observed with free recall of word lists (Zechmeister and Shaughnessy 1980) and recognition memory for faces (Sommer et al. 1995). These results indicate that instructing students to make JOLs may be a simple educational intervention that provides learning benefits for certain materials that transfer across time and assessment formats (see Barnett and Ceci 2002).

Although the current evidence is promising, there are conditions in which instructing learners to make JOLs produces no direct benefits to learning or may even harm it. One condition is when the to-be-learned items are semantically unrelated word pairs (Janes et al. 2018; Mitchum et al. 2016; Soderstrom et al. 2015; see Double et al. 2018 for a review). Memory is enhanced when people make JOLs for related word pairs (e.g., *king—crown*) but is unaffected (Kelemen and Weaver III 1997; Soderstrom et al. 2015) or sometimes impaired (Mitchum et al. 2016) when learners make JOLs for unrelated word pairs (e.g., *dog—spoon*). These findings have led some authors to propose that the benefits of making JOLs occur only when learners are strengthening existing associations (Soderstrom et al. 2015). If making JOLs only enhances retention of existing associations, then this procedure may provide limited benefits for learning educationally relevant material that requires forming new associations among concepts.

Other evidence indicates there may be individual differences in who directly benefits from making metacognitive judgments. For example, JOLs do not directly enhance older adults' learning of semantically related word pairs (Tauber and Witherby 2019), and individual differences in self-confidence moderate the effects of making confidence judgments on reasoning (Double and Birney 2017). Specifically, when people are required to make confidence judgments while completing Raven's Progressive Matrices problems, people who have high initial self-confidence display enhanced reasoning but people with low initial self-confidence display impaired reasoning (Double and Birney 2017). The type of assessment format could also influence the effects of JOLs on learning. Myers et al. (2020) recently reported that JOLs enhance learning for semantically related word pairs on cued recall tests but not for free recall tests.

A final condition where no direct effects of JOLs occur is when learners are instructed to either generate or retrieve target information in addition to making JOLs (Dougherty et al. 2018; Soderstrom et al. 2015). Making JOLs may not provide learning benefits under these conditions because JOLs themselves may enhance learning by encouraging covert retrieval of target information (Jönsson et al. 2012; Kelemen and Weaver III 1997; Kimball & Metcalfe, 2003; Spellman and Bjork 1992; Son and Metcalfe 2005). To illustrate, consider the modal method for prompting JOLs. Learners typically study a list of paired associates (e.g., *king—crown*) and make JOLs sometime after study with the target word absent (e.g., *king*—?). Prompting JOLs using this target-absent format affords an opportunity to practice retrieval of the target item which subsequently enhances memory for it (Spellman and Bjork 1992).

Retrieval practice is a highly effective learning strategy (for a review, see Karpicke 2017) that is likely to improve students' text comprehension when used in the service of making a JOL. However, it is not certain that metacomprehension judgments encourage retrieval practice in the same manner as JOLs for paired associates. Morris (1990) argued that learners base their JOLs for texts on the momentary accessibility of text content (i.e., retrieval), and students report using retrieval in the form of self-testing to identify how well they understand course content that they expect to appear on exams (Hartwig and Dunlosky 2012; Kornell and Son 2009; Morehead et al. 2016). However, when learners are instructed to retrieve prior to making JOLs for key term definitions (e.g., *What is the availability heuristic?*), their metacomprehension accuracy improves compared with when they are not given retrieval instructions (Dunlosky et al. 2005). These results suggest that learners may not spontaneously retrieve prior to making JOLs without explicit instructions to do so. A major goal of the current experiments was to evaluate if retrieval practice plays a role in producing direct effects of JOLs on text comprehension.

Overview of the Present Experiments

Across five experiments, we examined whether instructing learners to make JOLs would improve their comprehension of educationally relevant texts. The present experiments are the first to investigate whether JOLs directly enhance learning for educationally relevant material as speculated by Soderstrom et al. (2015) and many others. In Experiment 1, learners read several sections of a textbook chapter about minerals and some were instructed to make aggregate JOLs immediately after reading each section. Experiments 2a, 2b, 3, and 4 used a similar procedure but required learners to make several term-specific JOLs, which probe learners for their understanding of specific concepts within a section of text rather than probing their general understanding of the text as a whole. An example of a term-specific JOL is to ask learners to rate their confidence in how well they understand how minerals are made. During a later test, learners would be asked to answer a question that corresponded to that key term (*How are minerals made?*). This procedure mirrors the target-absent format for JOL prompts that is widely adopted in metacognitive experiments examining paired associate learning. If making JOLs enhances learning by inducing retrieval practice, then one might expect this procedure to enhance text comprehension more than alternative methods of prompting JOLs that do not afford a retrieval practice opportunity.

Experiment 1

In experiment 1, subjects studied an educational text that was divided into five sections. Some subjects were instructed to make aggregate JOLs immediately after reading each section, and others read the text without making JOLs. Aggregate JOLs require subjects to rate their overall confidence in their comprehension of the entire section. The aggregate JOL procedure was first developed by Maki and Berry (1984), and for over three decades, it has served as the standard method for examining metacomprehension (for reviews, see Dunlosky and Lipko (2007) and Thiede and de Bruin (2017)). To date, no research using this standard method has included a control group that did not make JOLs. Thus, it is unclear whether aggregate JOLs have reactive effects on learning. If aggregate JOLs encourage subjects to engage in covert retrieval when they assess their own learning, then making JOLs would be expected to enhance performance on a final comprehension test relative to only reading the text and not making JOLs.

Method

Subjects and Design

Eighty subjects ($M_{\text{age}} = 36$, $SE = 1.30$; 40% female) were recruited online through a Human Intelligence Task (HIT) posted on Amazon Mechanical Turk. All subjects were high school graduates. Most subjects had some college experience (89%) and nearly half had earned a bachelor’s degree or higher (48%). The experiment lasted approximately 10 min and subjects were paid $1.00 for their participation. To avoid low effort responding, we restricted recruitment in all experiments reported here to subjects who had completed 1000 or more HITs with an acceptance rate greater than 95% (for rationale, see Peer et al. 2014). Recruitment was always restricted to geolocations within the USA. Subjects were randomly assigned to the JOL group ($n = 41$) or no JOL group ($n = 39$).

Materials and Procedure

A science text on *Minerals* was adapted from Freudenrich et al. (2009). The text was 522 words and had a Flesch reading ease score of 49.1 and a Flesch-Kincaid grade level of 10. The text was divided into five sections with roughly 100 words per section. Each of the five sections described a characteristic of minerals (e.g., minerals are made by geological processes; minerals are inorganic substances), and each section had a descriptive heading (*Geological Processes, Inorganic Substances, Crystalline Solids, Elements, and Compounds*). Subjects' reading time was self-paced, with the constraint that they had to spend at least 30 s studying each section. When subjects were finished studying a section, they were required to click a button to proceed to the next task. Sections were presented individually, and presentation order was fixed. The complete text and test questions can be accessed at https://osf.io/p4rf8/.

Immediately after studying each section, subjects in the JOL group made an aggregate JOL without the text present. Specifically, subjects rated how confident they were that they understood the section that they had just read, and they responded using a sliding scale from 0 (not very confident) to 100 (very confident). JOLs were self-paced. After making their JOL, subjects were presented with the next section. Subjects in the no JOL group did not make an aggregate JOL following each section. Instead, when they were done studying a section, they clicked a button to begin studying the next section.

Following the study phase, subjects took a 12-item short-answer test. The test consisted of factual questions drawn directly from each section of text (e.g., *How are minerals made?*; *What are inorganic substances?*). Six of the questions were about content from three sections (i.e., two questions per section), and six questions were about content from the remaining two sections (i.e., three questions per section). Questions were presented individually, and presentation order was randomized. The test was self-paced with the constraint that after 10 s a button appeared that allowed subjects to advance to the next question, thereby requiring them to spend at least 10 s answering each question. The questions did not require subjects to make inferences about what they read. Instead, all questions could be answered using one's memory for content directly stated in the text (cf. Wiley et al. 2005).

Finally, after completing the test, subjects rated how familiar they were with the content prior to reading the text. Subjects responded using a sliding scale from 0 (not very familiar) to 100 (very familiar). The familiarity rating was self-paced. Familiarity was measured after completing the test to minimize potential reactive effects of these judgments on learning. One implication of this decision is that familiarity ratings may be influenced by experience with the task as well as domain familiarity. At the end of the experiment, the subjects were thanked for their participation and shown a debriefing screen that explained the purpose of the experiment.

Results and Discussion

Responses to the short answers questions on the criterion test were scored independently by two coders using a predetermined rubric. The same coders scored data for every experiment and were blind to the experimental groups. All disagreements were resolved by the first author. Analyses and conclusions were based on 95% confidence intervals for effect sizes and estimated Bayes factors (i.e., BF_{10}) for contrasts. Bayes factors less than 1 indicate increasing evidence for a model favoring a null effect and Bayes factors greater than 1 reflect increasing evidence that there is a non-zero difference between groups. Data for all experiments are accessible at https://osf.io/terdu.

Figure 1 shows the mean percentage of questions answered correctly for the JOL and no JOL groups. Instructing learners to make aggregate JOLs after reading each section of text did not improve their comprehension relative to just reading the text (48.27% vs. 45.87%), $t(78) = 0.47$, $d = 0.11$ [−0.33, 0.54]. The estimated Bayes factor indicated the data provide moderate

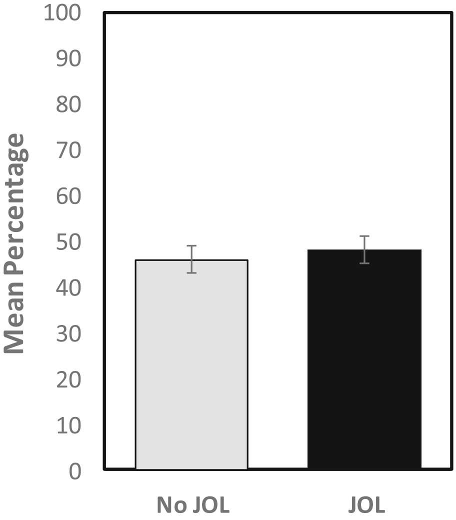

Fig. 1 Mean percentage of correct responses on the criterion test in Experiment 1 for the no JOL group and the JOL group which made aggregate JOLs for each section of text. Error bars represent standard error of the mean

support for a model favoring a null effect, $BF_{10} = 0.26$. There was no evidence that this null effect is due to baseline differences in domain familiarity because there were no differences in subjects' ratings of their pre-experiment familiarity with the text content between learners who made JOLs ($M = 41.51$; $SE = 5.00$) and learners who did not make JOLs ($M = 39.79$; $SE = 4.44$), $t(78) = 0.26$, $d = 0.06$ [− 0.38, 0.50], $BF_{10} = 0.24$. This conclusion was further supported by the results from a Bayesian ANCOVA that examined the effects of JOLs on test performance while controlling for familiarity ratings, $BF_{10} = 0.25$.

For completeness, the JOL group's mean JOL magnitude and relative accuracy are presented in Table 1 for all experiments. Relative accuracy was computed by calculating gamma correlations between JOLs and test performance for each subject (for rationale, see Gonzalez and Nelson (1996) and Nelson (1984)). These gamma correlations reflect the degree to which JOLs discriminate at the item level between material subsequently recalled on the final test and material not recalled. Relative accuracy for Experiment 1 should be interpreted

Table 1 Mean judgment of learning (JOL) magnitude and mean relative accuracy for JOLs (mean gamma correlation between JOLs and test performance) in each experiment

Experiment and group	Judgment format	JOL	Relative accuracy
Experiment 1			
JOL group	Aggregate	71.12 (2.89)	0.08 (0.09)
Experiment 2a			
JOL group	Target-absent	70.14 (2.36)	0.08 (0.06)
Experiment 2b			
JOL group	Target-absent	79.13 (1.98)	0.22 (0.07)
Experiment 3			
Target-absent JOL group	Target-absent	71.20 (2.52)	− 0.06 (0.06)
Target-present JOL group	Target-present	82.21 (2.06)	0.17 (0.07)
Experiment 4			
JOL group	Target-absent	71.68 (2.17)	0.01 (0.06)
Retrieval practice + JOL group	Target-absent	76.71 (2.25)	0.25 (0.05)

Standard error of the mean are in parenthesis cautiously because only 5 data points per subject were available to compute relative accuracy. In subsequent experiments, this computation was based on at least 12 data points per subject.

The results of Experiment 1 failed to support the hypothesis that aggregate JOLs directly enhance learning for educational material. Aggregate JOLs may have failed to enhance learning of texts because the scope of information probed by aggregate JOLs was too general. Learners may not retrieve the specific concepts from the text that align with the tested material when monitoring at the aggregate level.

Retrieving text content for aggregate JOLs is also much more effortful and time consuming than is retrieving single words in paired associate learning tasks. Learners might not even attempt to retrieve text content when making aggregate JOLs. Consider two metacomprehension interventions that are effective for increasing the relative accuracy of aggregate JOLs during reading—the delayed summary effect (Anderson and Thiede 2008; Thiede and Anderson 2003) and the delayed key word effect (Thiede et al. 2003; Thiede et al. 2005). Both these interventions involve providing learners explicit instructions that encourage them to retrieve criterial information from the texts for purpose of summarizing or creating key words that capture the essence of the text. The efficacy of these interventions for improving the relative accuracy of JOLs during reading suggests aggregate JOLs do not spontaneously elicit retrieval when learners are not given explicit instructions to use a strategy that encourages them to retrieve. In subsequent experiments, we adopted a JOL procedure that aligns closer to the procedure used for paired associate learning to increase the likelihood that learners would attempt to retrieve criterial information during monitoring.

Experiment 2a

Experiments 2a and 2b adopted a different JOL procedure (relative to Experiment 1) to evaluate potential direct effects of JOLs on learning. Instead of making aggregate JOLs, which require learners to monitor learning for an entire section of a text, learners were instructed to make a series of term-specific JOLs after reading each section (Dunlosky et al. 2002). Each term-specific JOL required subjects to judge their learning for a specific piece of key content from the section of text they just read. For example, after reading the first section of the text on minerals in Experiment 2a, subjects were asked to make two separate JOLs. The first JOL required them to rate how confident they were that they understood how minerals are made. The second JOL required them to rate how confident they were that they understood the different types of geological processes they read about. These two JOLs correspond to two of the later criterial test questions subjects were required to answer which were (1) How are minerals made? and (2) What is an example of a geological process?

As discussed previously, the format for term-specific JOLs is similar to the target-absent JOL procedure used in paired associate learning paradigms that can produce direct benefits to learning. In the paired associate procedure, the JOL prompt can serve as a cue to retrieve the target item (Kimball & Metcalfe, 2003; Jönsson et al. 2012; Spellman and Bjork 1992; Son and Metcalfe 2005). Term-specific JOLs might also afford retrieval practice because the target information is absent when subjects make their judgments. If subjects attempt to retrieve information when monitoring learning for the key concepts, then monitoring should produce direct benefits to learning by virtue of retrieval practice.

Method

Subjects and Design

Ninety subjects were recruited from Amazon Mechanical Turk following the methods used in Experiment 1 ($M_{\text{age}} = 36$, $SE = 1.18$; 55% female). All subjects were high school graduates. Most subjects had some college experience (84%) and nearly half had earned a bachelor's degree or higher (47%). The experiment lasted approximately 15 min and subjects were paid $1.50 for their participation. Subjects were randomly assigned to the JOL group ($n = 45$) or to the no JOL group ($n = 45$).

Materials and Procedure

The materials were identical to those used in Experiment 1. The procedure was nearly identical to Experiment 1 with the only deviation being that subjects in the JOL group made term-specific JOLs after reading each section. The term-specific JOLs were phrased such that they matched the content of the short-answer questions that would appear on the test. For example, for the test question "How are minerals made?", the corresponding term-specific JOL was "How confident are you that you understand how minerals are made?". Thus, subjects made a total of 12 term-specific JOLs (two or three per section) that corresponded to the 12 short-answer questions on the test. JOLs were self-paced and subjects made them using the same sliding scale used in Experiment 1. Term-specific JOLs were presented individually in a fixed order that matched the order presented in the text. After the study phase, all subjects took the short-answer test and made a familiarity rating as in Experiment 1.

Results and Discussion

The mean percentage of questions answered correctly on the final test is presented in the left panel of Fig. 2. Figure 2 illustrates that learners who generated term-specific JOLs after reading did not perform substantially better on the criterion test compared with learners who simply read the text (49.29% vs. 44.44%), $t(88) = 1.06$, $d = 0.22$ [− 0.19, 0.64], $BF_{10} = 0.34$. There were no differences in pre-experiment domain familiarity between learners who made JOLs ($M = 46.33$; $SE = 3.80$) and learners who did not make JOLs ($M = 41.22$; $SE = 4.76$), $t(88) = 0.84$, $d = 0.18$ [− 0.24, 0.59], $BF_{10} = 0.30$. A Bayesian ANCOVA indicated that JOLs also do not improve test performance when controlling for domain familiarity, $BF_{10} = 0.31$. Mean JOLs and relative accuracy for the JOL group is presented in Table 1 for interested readers. Overall, these results indicate that the proposed educational benefits of making JOLs on learning may not transfer well to educational texts.

Experiment 2b

Just as making aggregate JOLs did not improve learning in Experiment 1, making term-specific JOLs did not improve learning in Experiment 2a. It is possible that these null results were specific to the educational materials used in in the previous experiments because to-be-learned materials can moderate the effects of JOLs on learning (see Double et al. 2018). To ensure that the texts we selected for Experiments 1 and 2a were not limiting the direct benefits

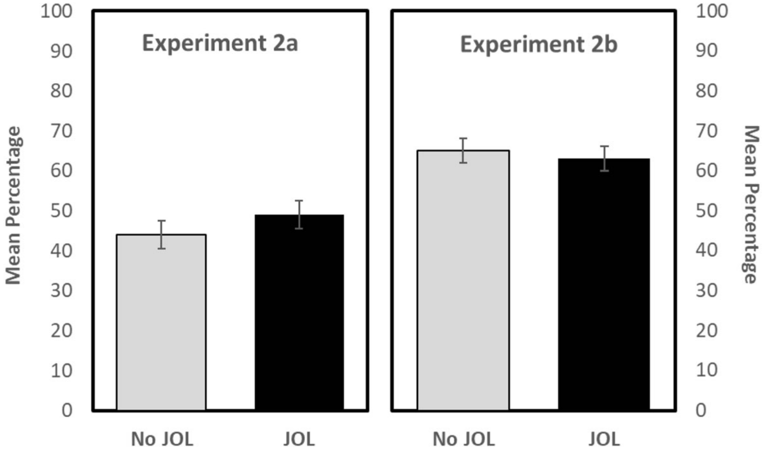

Fig. 2 Mean percentage of correct responses on the criterion test in Experiment 2a (left panel) and Experiment 2b (right panel). Subjects in the JOL groups made target-absent JOLs in both experiments. Error bars represent standard error of the mean

of making JOLs on learning, we used a new set of educational texts in Experiment 2b. If making term-specific JOLs improves learning and the previous null results are due to the types of material we selected, then subjects who made JOLs after reading these new texts should perform better on the criterion test than subjects who read without making JOLs.

Method

Subjects and Design

Ninety-nine subjects were recruited from Amazon Mechanical Turk ($M_{\text{age}} = 33$, $SE = 0.92$; 39% female). All subjects were high school graduates. Most subjects had some college experience (83%) and nearly half had earned a bachelor's degree or higher (45%). The experiment lasted approximately 15 min and subjects were paid $1.50 for their participation. Subjects were randomly assigned to the JOL group ($n = 50$) or to the no JOL group ($n = 49$).

Materials and Procedure

The primary change in Experiment 2b was the to-be-learned materials. Specifically, subjects studied four texts that were selected from a test preparation book for the Test of English as a Foreign Language (TOEFL; Rogers 2001). Each text covered a different topic (Jupiter, Tumbleweeds, Diving, and Killdeer). The texts were 199, 206, 222, and 221 words in length, and the Flesch reading ease scores were 58.5, 64.3, 54.6, and 62.2, respectively. The change in materials resulted in minor changes to the procedure relative to Experiment 2a. Subjects studied each text, and in the JOL group, subjects made four term-specific JOLs per text (i.e., 16 in total). These 16 term-specific JOLs corresponded to the 16 questions on the short-answer test. Subjects in the no JOL group did not make JOLs. Other than the change in materials and the number of JOLs and test questions, the rest of the procedure was identical to Experiment 2a.

Results and Discussion

The right panel of Fig. 2 shows the mean percentage of questions answered correctly on the final test for the JOL and no JOL groups. Making term-specific JOLs did not produce learning benefits relative to just reading (63.13% vs, 65.36%), $t(97) = -0.52$, $d = -0.10$ [− 0.50, 0.29], and the estimated Bayes factor provided moderate support for this null effect, $BF_{10} = 0.24$. Mean JOLs and mean relative accuracy for the JOL group is presented in Table 1 for interested readers. There were no differences in pre-experiment domain familiarity for learners who made JOLs ($M = 29.04$; $SE = 3.54$) and learners who did not ($M = 32.24$; $SE = 3.67$), $t(88) = 0.63$, $d = 0.13$ [− 0.27, 0.52], $BF_{10} = 0.25$, and making JOLs did not improve final test performance even when controlling for domain familiarity, $BF_{10} = 0.10$.

Previous research indicates that JOLs may selectively enhance learning for only certain types of material (for a review, see Double et al. 2018). Experiment 2a examined learning passages of an earth science textbook and Experiment 2b examined learning in several different scientific, historical, and vocational domains. In both experiments, Bayesian analyses indicated the data support a null model for the effects of making JOLs on learning. Thus, the null results we observed are unlikely to be material-specific and instead must be due to the evaluative process learners engage in when making their JOLs.

Experiment 3

Experiments 1, 2a, and 2b failed to support the hypothesis that making JOLs would enhance learning of educationally relevant texts. Given that retrieval practice typically improves learning outcomes more than restudy (for a review, see Rowland 2014), JOLs that afford retrieval practice should increase learning more than JOLs that afford an additional study opportunity. Experiment 3 directly evaluated this hypothesis using a new methodological approach for examining metacomprehension, which involved querying term-specific JOLs with the content subjects were required to know when answering the upcoming test questions (henceforth referred to as target-present JOLs). These new target-present JOLs were contrasted with the target-absent JOLs used in Experiment 2a and compared with performance of a no JOL group.

Target-absent JOLs emphasize the test question and afford opportunities for retrieval practice during monitoring (e.g., How confident are you that you understand how minerals are made?), whereas target-present JOLs emphasize the answer to each question and provide a restudy opportunity (e.g., How confident are you that you understand that minerals are made by geological processes?). If subjects engage in retrieval practice when making target-absent JOLs in the current experiment, they should display higher test performance compared with subjects who make target-present JOLs or no JOLs after reading.

Although the current experiment is the first to utilize target-present JOLs for text material, several experiments have used an analogous procedure to examine JOL accuracy for paired associates (Connor et al. 1997; Dunlosky and Nelson 1992). A classic finding from this research is that delayed JOLs are more accurate at predicting future memory performance when JOL queries encourage retrieval (e.g., How confident are you that you will remember the second word of this pair: *dog*—?) than when JOL queries present a restudy opportunity (e.g., How confident are you that you will remember the second word of this pair: *dog*—*spoon*). The same pattern should be present in the current experiment because target-absent JOLs should encourage subjects to attend to diagnostic retrieval cues during monitoring (e.g., retrieval fluency and target accessibility) that are not available for target-present JOLs.

Method

Subjects and Design

One hundred seventy-four subjects were recruited from Amazon Mechanical Turk ($M_{\text{age}} = 34$, $SE = 0.72$; 37% female). Only one subject was not a high school graduate. Most subjects had some college experience (82%) and 43% earned a bachelor's degree or higher. The experiment lasted approximately 15 min and subjects were paid $1.50 for their participation. Subjects were randomly assigned to one of three groups: target-absent JOL group ($n = 59$), target-present JOL group ($n = 58$), or no JOL group ($n = 57$).

Materials and Procedure

The materials used in Experiment 3 were identical to those used in Experiments 1 and 2a. The procedure for the no JOL group was identical to that of Experiments 1 and 2a. The procedure for the target-absent JOL group was identical to the procedure for the JOL group in Experiment 2a. Subjects in the target-present JOL group were treated similarly to subjects in the target-absent JOL group. The primary difference was that JOL prompts for the target-present JOL group included the target information that subjects would attempt to retrieve on the test, whereas the JOL prompts for the target-absent JOL group did not. For example, the target-absent JOL group received the prompt, "How confident are you that you understand how minerals are made?", whereas the target-present JOL group received the prompt, "How confident are you that you understand that minerals are made by geological processes?". In all other respects, the procedure was identical to the one used in previous experiments.

Results and Discussion

Figure 3 presents the mean percentage of test questions answered correctly for each group. Consistent with the findings from Experiment 2a and 2b, making target-absent JOLs did not enhance learning more than just reading the text (40.31% vs. 42.37%), $t(114) = -0.41$, $d = 0.08$ [-0.44, 0.29], $BF_{10} = 0.21$. Target-present JOLs also did not significantly enhance learning more than just reading (50.09% vs. 42.37%), $t(113) = 1.65$, $d = 0.31$ [-0.06, 0.67], $BF_{10} = 0.67$. However, target-present JOLs did produce better performance than did target-absent JOLs (50.09% vs. 40.31%), $t(115) = 2.01$, $d = 0.37$ [0.01, 0.74], but the Bayes factor suggests evidence for this difference is weak, $BF_{10} = 1.19$. There were no differences in domain familiarity among the target-absent JOL group ($M = 41.02$, $SE = 4.05$), target-present JOL group ($M = 43.05$, $SE = 4.06$), and the no JOL group ($M = 37.35$, $SE = 4.41$), $F(2, 171) = 0.48$, $MSE = 479.34$, $p = 0.62$, $\eta_p^2 = 0.01$, $BF_{10} = 0.09$. A Bayesian ANCOVA examining the effects of JOL group on final test performance controlling for familiarity supported the null effect, $BF_{10} = 0.40$.

Mean JOL magnitude and mean relative accuracy for JOLs are presented in Table 1. The mean magnitude of JOLs was significantly higher for the target-present JOL group than for the target-absent JOL group (82.21 vs. 71.20), $t(115) = 3.38$, $d = 0.62$ [0.25, 0.99], $BF_{10} = 0.21$. The effect of JOL format on the relative accuracy of JOLs was examined by computing gamma correlations between JOLs and test performance. Mean gamma correlations were significantly higher for the target-present JOL group ($M = 0.17$; $SE = 0.07$) than for the target-absent JOL group ($M = -0.06$; $SE = 0.06$), $t(115) = 2.46$, $d = 0.45$ [0.09, 0.82], $BF_{10} =$

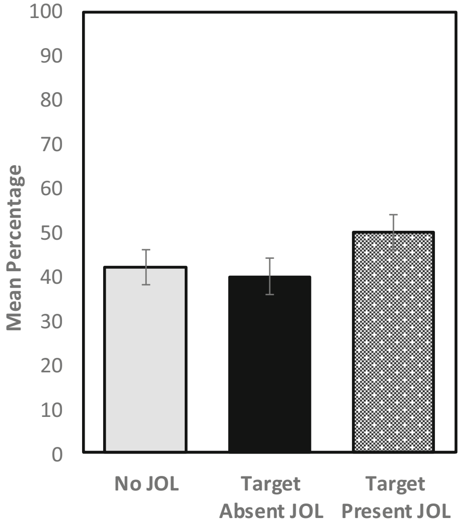

Fig. 3 Mean percentage of correct responses on the criterion test in Experiment 3. All subjects in the JOL groups made term-specific JOLs. Error bars represent standard error of the mean

2.87. This outcome is surprising because it suggests that the typical recommendations about how to query JOLs to maximize their accuracy may not generalize to complex educational materials like text. Perhaps target accessibility is less diagnostic of text comprehension than of paired associate memory performance. Alternatively, target accessibility may not be a cue that learners use at all when making JOLs for texts because JOLs do not effectively elicit retrieval of criterial information (see Dunlosky et al. 2005). Both the null effects for JOLs on learning and the relative accuracy advantage for target-present JOL format over target-absent JOL format are consistent with this latter hypothesis. If JOLs do not elicit retrieval of the target, then it would be advantageous for students to make JOLs with the target-present during monitoring to draw their attention to the important concepts they need to understand. In Experiment 4, we investigated the retrieval dynamics involved in making JOLs and directly evaluated the effects of requiring covert retrieval before making JOLs on text comprehension.

Experiment 4

The results of the previous experiments are puzzling. Why does the act of making JOLs fail to directly improve learning for educational text materials when this activity does benefit learning of other types of material? One answer to this question is that retrieval processes elicited for JOLs are qualitatively different than the retrieval processes utilized when learners are instructed to use retrieval practice (Tauber et al. 2015). A key difference between JOLs and retrieval practice is that JOLs are assumed to afford a covert retrieval opportunity, whereas retrieval practice typically involves overt retrieval. Covert retrieval can produce similar learning benefits as overt retrieval for paired associate items and word lists (Putnam and Roediger 2013; Smith et al. 2013) but recent evidence suggests that covert retrieval may be less effective for learning more complex material like key term definitions (Tauber et al. 2018). A second difference between the retrieval processes for JOLs and retrieval practice is that JOLs may elicit less effortful search and premature termination compared with explicit instructions to retrieve (Son and Metcalfe 2005). Both differences could explain why JOLs do not provide direct benefits to learning for educational texts.

Experiment 4 tested whether the retrieval dynamics differ between JOLs and retrieval practice by contrasting the effects of making JOLs on learning to the benefits of retrieval practice using a 2 (JOL vs. no JOL) × 2 (retrieval practice vs. no retrieval practice) factorial design. Subjects read several sections of the same educational text used in Experiment 3 and some made target-absent JOLs in the same manner as in previous experiments. Importantly, subjects in the retrieval practice group and the retrieval practice + JOL group were asked to overtly recall target information for each key term. For the retrieval practice + JOL group, these retrieval practice trials occurred immediately before subjects made JOLs for each key term. This prejudgment recall and monitoring procedure (PRAM; Nelson et al. 2004) has been adopted in several experiments to examine how retrieval influences metacognitive monitoring including with key term definitions (Dunlosky et al. 2005).

If the retrieval processes for JOLs and explicit retrieval instructions are qualitatively different, then retrieval practice should enhance learning more than making JOLs. Moreover, these differences should be attenuated when subjects are instructed to use overt retrieval before making their JOLs.

Subjects and Design

Two hundred sixty subjects were recruited from Amazon Mechanical Turk ($M_{\text{age}} = 34$, $SE = 0.59$; 45% female). All but two subjects were high school graduates. Most subjects had some college experience (82%) and had earned a bachelor's degree or higher (48%). The experiment lasted approximately 20 min and subjects were paid $1.80 for their participation. A 2 (JOL vs. no JOL) × 2 (retrieval practice vs. no retrieval practice) factorial design was used for this experiment. Subjects were randomly assigned to one of four groups: JOL ($n = 64$), no JOL ($n = 65$), retrieval practice ($n = 66$), and retrieval practice + JOL ($n = 65$).

Materials and Procedure

The materials were identical to those used in Experiments 1, 2a, and 3. The procedure for the JOL group was identical to the JOL group in Experiment 2a (and the target-absent JOL group in Experiment 3) and the procedure for the no JOL group was identical to the procedure for the no JOL groups in previous experiments. The procedure for the retrieval practice group and the retrieval practice + JOL groups differed from the JOL group and the no JOL group only in the study phase of the task. After studying each section of the text, subjects in the retrieval practice group were prompted to practice retrieval by answering two or three short-answer questions about the section of text they just read. The short-answer questions were the same as those used on the short-answer test in Experiments 1, 2a, and 3. Question order was fixed and appeared in the same order that the information appeared in the text. Subjects were given an unlimited amount of time to answer the questions, with the constraint that they had to spend at least 10 s retrieving each answer. After 10 s, a button appeared and subjects could advance when they were ready. Subjects were given feedback about the correct answer after responding to each question. Feedback was self-paced, with the constraint that subjects had to spend 5 s reading it before a button appeared to move on. After answering all questions for a section, subjects studied the next section. The procedure for the retrieval practice + JOL group was identical to the retrieval practice group, with the exception that subjects in this group also made a target-absent JOL following feedback for each short-answer question during the study phase.

Results and Discussion

Mean final test performance for each group is presented in Fig. 4. Both making JOLs and retrieving target information improved performance on the criterion test. Several planned comparisons were examined to evaluate the differential effects of making JOLs and retrieval practice on final performance. Comparisons of the JOL group and no JOL group revealed that making JOLs alone after reading did not substantially improve performance more than just reading the text, (47% vs. 41%), $t(127) = 1.45$, $d = 0.26$ [− 0.09, 0.60], $BF_{10} = 0.49$. In contrast, comparisons of the retrieval practice group to the no JOL group (58% vs. 41%), $t(127) = 4.07$, $d = 0.71$ [0.36, 1.06], $BF_{10} = 255.79$, and the retrieval practice + JOL group to the no JOL group (67% vs. 41%), $t(128) = 7.02$, $d = 1.23$ [0.86, 1.61], $BF_{10} = 7.212\text{e} +7$, both indicated that overtly retrieving target information after reading produced large learning gains compared with only reading. The retrieval practice + JOL group outperformed the standard JOL group (67% vs. 47%), $t(128) = 2.44$, $d = 0.43$ [0.08, 0.78], $BF_{10} = 9730$, and also outperformed the standard retrieval practice group (67% vs. 58%), $t(129) = 2.50$, $d = 0.44$ [0.09, 0.79], $BF_{10} =$ 3.06. The performance advantage for the retrieval practice + JOL group compared with the standard retrieval practice group was unexpected and suggests that there may be conditions in which JOLs enhance learning for texts (e.g., when JOLs occur following retrieval practice). What it is clear from these data is that JOLs alone are not an effective substitute for overt retrieval practice. The retrieval dynamics for JOLs and retrieval practice are qualitatively different.

As shown in Table 1, mean JOL magnitude did not significantly differ between the JOL group and the retrieval practice + JOL group, $t(127) = 1.61$, $p = 0.11$, $d = 0.28$ [− 0.06, 0.63], $BF_{10} = 0.61$. More critical for evaluating differences in retrieval dynamics for JOLs and overt retrieval practice, Table 1 also shows that requiring learners to retrieve target information before making JOLs improved their relative accuracy compared with querying JOLs without retrieval instructions (0.25 vs. 0.01), $t(123) = 3.03$, $p < 0.01$, $d = 0.54$ [0.18, 0.89], $BF_{10} = 0.97$. These results suggest that the utilization of retrieval-based cues increases during monitoring when learners are instructed to overtly retrieve target information before making JOLs.

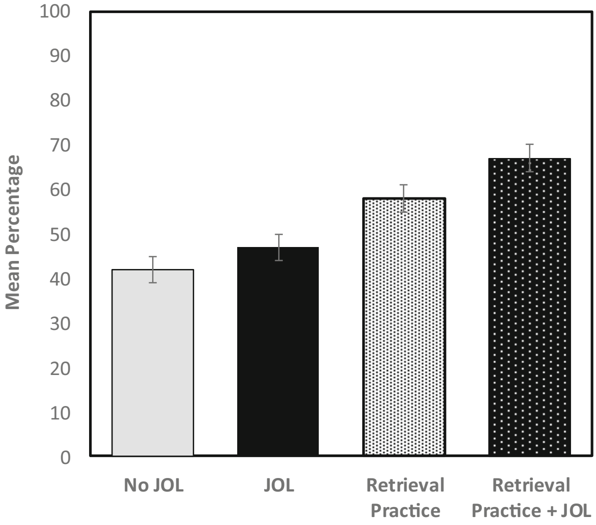

Fig. 4 Mean percentage of correct responses on the criterion test in experiment. All participants in the JOL groups made term-specific JOLs with the target-absent. Error bars represent standard error of the mean

Why is retrieval underutilized when learners make metacomprehension judgments without explicit retrieval instructions? One answer to this question may lie in research on students' self-regulated use of retrieval practice. Students self-report that they use retrieval in the form of self-testing to monitor their learning (Hartwig and Dunlosky 2012; Kornell and Son 2009). However, their decisions to self-test are also influenced by monitoring processes. When learners are allowed to choose whether to restudy, test, or drop material from learning, they prefer to restudy material they assign low JOLs, to self-test material they assign intermediate JOLs, and to drop material they assign high JOLs (Karpicke 2009). Students may prefer to use retrieval to monitor their learning only when they are unsure if they know the material. When they are confident they do not know it, they prefer to restudy it. When they are confident they do know it, they prefer to terminate practice for it. Perhaps when students monitor their comprehension, retrieval and subsequent accessibility-based cues are utilized only after other cues elicit insufficient evidence about learning. In such cases, learners may attempt to retrieve information from memory to test whether it is accessible.

Regardless of why students underutilize retrieval when making metacomprehension judgments, it is clear that they should be instructed to retrieve text content before making JOLs. Retrieval improves retention of material (Karpicke 2017) and can also improve metacomprehension accuracy (Dunlosky et al. 2005), especially when there is a delay between initial study and the retrieval activity (Anderson and Thiede 2008; Thiede et al. 2005).

General Discussion

The current experiments investigated the direct educational benefits of making JOLs on text comprehension. The pooled effect size for making JOLs on text comprehension across 5 experiments (excluding the retrieval practice + JOL group in Experiment 4) was small and not significantly different from zero, Pooled $d = 0.08$ [-0.03, 0.28]. On average, making JOLs produced a 3% increase in text learning compared with just reading. These data indicate that the large direct effects of making JOLs on learning semantically related paired associates (Janes et al. 2018; Soderstrom et al. 2015; Witherby and Tauber 2017) do not transfer well to educationally relevant text materials.

The present results highlight the need to be cautious about making educational recommendations based on findings from paired associate research alone. Two robust outcomes from metacognitive research on paired associate learning did not generalize to more complex educationally relevant material. First, as mentioned above, making JOLs produced minimal direct benefits to learning educational texts. Second, and perhaps more surprising, long held recommendations about the best way to elicit JOLs to maximize monitoring accuracy did not appear to hold true for learning from texts. That is, the conventional wisdom that JOLs are more accurate when elicited with the target information absent than when elicited with the target information present was not supported by our findings. In fact, the opposite outcome occurred. Relative accuracy was better when JOLs were queried with the target-present than when queried with the target-absent (see Table 1), a novel finding we have since replicated (Ariel and Karpicke 2020). These data suggest that researchers and educators may need to rethink how they elicit JOLs when assessing students' comprehension.

Metacomprehension, like comprehension, can be evaluated at multiple levels (Kintsch 1988). The current research focused on monitoring one's understanding of factual information in a text but metacomprehension research often focuses on monitoring comprehension at a deeper level like one's situation model level (Wiley et al. 2005). The situation model level focuses on understanding the gist of the text (Kintsch 1988). Typically, research focused on monitoring one's situation model uses aggregate JOLs like the JOLs we used in Experiment 1. However, research examining one's situation model usually incorporates tests that require learners to apply knowledge and make inferences about the text they read.

Given that aggregate JOLs failed to enhance learning for factual information that could be answered using verbatim recall of information from the text, it is unlikely that making aggregate JOLs would improve deeper comprehension. However, there is evidence that the types of assessment test that learners expect can change the cues they attend to during monitoring. When learners are exposed to inference questions prior to study, they are more likely to attend to cues relevant to their situation model than if they expect to be tested on their memory for what they read (Thiede et al. 2011). It is possible that interventions that draw attention to cues relevant to one's situation model during monitoring would alter the effects of JOLs on comprehension, especially if these interventions encourage relational processing.

The semantic relationship among items seems to play an important role in the direct effects of making JOLs on learning. Making JOLs produces robust positive effects on learning when material is semantically related but can have null or even negative effects on learning when the material is semantically unrelated (Janes et al. 2018; Mitchum et al. 2016; Soderstrom et al. 2015; Witherby and Tauber 2017). Learning from educational texts requires forming new associations among items, which may explain why JOLs were not effective for improving text comprehension in the current experiments. Perhaps if learners were experts in the domain they were reading and could leverage their knowledge of preexisting associations among information in the text, JOLs would have improved comprehension. If so, one might expect domain familiarity would influence the effects of JOLs on learning in the current experiments, but we found no evidence to support this claim. However, it is possible that our domain familiarity judgments were influenced by task experience because familiarity was assessed after the task was completed. Thus, we cannot conclusively rule out this hypothesis.

A major assumption of the current experiments was that any direct effects of making JOLs on comprehension would likely stem from covert retrieval practice. This assumption was motivated by evidence that learners attempt to retrieve target information when making JOLs and base their metacognitive judgments on how much and how quickly information comes to mind (Benjamin and Bjork 1996; Baker and Dunlosky 2006; Koriat 1997; Morris 1990; Schwartz 1994). Contrary to this assumption, our results indicate that if target accessibility is influencing metacomprehension judgments, the retrieval processes affecting judgments are substantially different from the retrieval processes utilized when learners are instructed to engage in retrieval practice. Learners who overtly retrieved prior to making JOLs displayed improved relative accuracy (Table 1) and increased comprehension in line with the expected benefits of retrieval practice when compared with a standard JOL condition (Fig. 4). One explanation for these outcomes is that learners terminated their retrieval too soon when making their JOLs in the standard condition, perhaps basing their judgments on initial ease of access alone (see Tauber et al. 2015). Alternatively, they may have avoided retrieving text content at all and instead based their judgments on other factors such as domain familiarity (Griffin et al. 2009; Maki and Serra 1992) and ease of reading the text (Rawson and Dunlosky 2002).

Differences in retrieval dynamics for target-absent and target-present JOLs could also explain why target-present JOLs produced better relative accuracy for texts than target-absent JOLs in Experiment 3. Contrary to our expectations, the target-present JOL format may serve as a more precise retrieval cue for episodic content from the text than the target-absent format. For example, when learners are asked if they understand *that minerals are made* *by geological processes* (i.e., they make a target-present JOL), they may attempt to remember whether they read this specific fact in the text. Their memory for the event may be more diagnostic of the future memorability of the fact than the limited accessibility-based cues that learners attend to when they are just asked if they understand *how minerals are made* (i.e., they make a target-absent JOL).

Although standard JOLs have limited direct educational benefits, making JOLs may provide indirect educational benefits. Metacognitive models of learning assume that monitoring judgments drive students' decision about which information to restudy (Ariel et al. 2009; Butler and Winne 1995; Metcalfe 2009; Thiede and Dunlosky 1999; Winne and Hadwin 1998). An implicit assumption of these models is that students always spontaneously monitor their learning to strategically allocate their study time to material. However, learners' study choices are not always strategically driven by monitoring (Ariel et al. 2011; Dunlosky and Ariel 2011), and there is evidence for both individual and intraindividual differences in when and if people choose to monitor their online performance (for a review, see Jordano and Touron 2018). Students who fail to spontaneously monitor in the moment could benefit from interleaving JOLs in their course materials, especially when they have the opportunity to restudy material.

In summary, metacomprehension judgments in their current form are a poor substitute for retrieval practice and are unlikely to directly improve students' text comprehension. When educators recommend that their students make JOLs for their course content, they should instruct them to overtly retrieve information before making their judgments. Requiring overt retrieval produces large learning gains compared with standard JOL instructions and also improves metacomprehension accuracy.

Funding Information This research was supported in part by grants from the James S. McDonnell Foundation (220020483) and the Institute of Education Sciences in the U.S. Department of Education (R305A150546).

Compliance with Ethical Standards

Conflict of Interest The authors declare that they have no conflicts of interest.

Disclaimer The opinions expressed are those of the authors and do not represent the views of the James S. McDonnell Foundation, the Institute of Education Sciences, or the U.S. Department of Education.

References

Anderson, M. C., & Thiede, K. W. (2008). Why do delayed summaries improve metacomprehension accuracy? *Acta Psychologica, 128*(1), 110–118.

Arbuckle, T. Y., & Cuddy, L. L. (1969). Discriminations of item strength at time of presentation. *Journal of Experimental Psychology, 81*(1), 126–131.

Ariel, R. & Karpicke, J. D. (2020). *Prompting judgments of learning with criterial information improves metacomprehension accuracy*. Manuscript in preparation.

Ariel, R., Dunlosky, J., & Bailey, H. (2009). Agenda-based regulation of study-time allocation: when agendas override item-based monitoring. *Journal of Experimental Psychology: General, 138*(3), 432–447.

Ariel, R., Al-Harthy, I. S., Was, C. A., & Dunlosky, J. (2011). Habitual reading biases in the allocation of study time. *Psychonomic Bulletin & Review, 18*(5), 1015–1021.

Baker, J. M., & Dunlosky, J. (2006). Does momentary accessibility influence metacomprehension judgments? The influence of study-judgment lags on accessibility effects. *Psychonomic Bulletin & Review, 13*(1), 60–65.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn? A taxonomy for far transfer. *Psychological Bulletin, 128*(4), 612–637.

Benjamin, A. S., & Bjork, R. A. (1996). Retrieval fluency as a metacognitive index. In L. Reder (Ed.), *Implicit memory and metacognition* (pp. 309–338). Hillsdale, NJ: Erlbaum.

Butler, D. L., & Winne, P. H. (1995). Feedback and self-regulated learning: a theoretical synthesis. *Review of Educational Research, 65*(3), 245–281.

Connor, L. T., Dunlosky, J., & Hertzog, C. (1997). Age-related differences in absolute but not relative metamemory accuracy. *Psychology and Aging, 12*(1), 50–71.

Double, K. S., & Birney, D. P. (2017). Are you sure about that? Eliciting confidence ratings may influence performance on Raven’s progressive matrices. *Thinking & Reasoning, 23*(2), 190–206.

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.

Dougherty, M. R., Robey, A. M., & Buttaccio, D. (2018). Do metacognitive judgments alter memory performance beyond the benefits of retrieval practice? A comment on and replication attempt of Dougherty, Scheck, Nelson, and Narens (2005). *Memory & Cognition, 46*(4), 558–565.

Dunlosky, J., & Ariel, R. (2011). The influence of agenda-based and habitual processes on item selection during study. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*, 899–912.

Dunlosky, J., & Lipko, A. R. (2007). Metacomprehension: a brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*(4), 228–232.

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*(4), 374–380.

Dunlosky, J., Rawson, K. A., & McDonald, S. L. (2002). Influence of practice tests on the accuracy of predicting memory performance for paired associates, sentences, and text material. In T. J. Perfect & B. L. Schwartz (Eds.), *Applied metacognition* (pp. 68–92). New York, NY, US: Cambridge University Press.

Dunlosky, J., Rawson, K. A., & Middleton, E. L. (2005). What constrains the accuracy of metacomprehension judgments? Testing the transfer-appropriate-monitoring and accessibility hypotheses. *Journal of Memory and Language, 52*(4), 551–565.

Freudenrich, C., Benner, J., Bethal, D. Desonie, D., Karasov, C., Lusk, M., et al. (2009). *Earth science*. Retrieved from https://www.ck12.org/c/earth-science/minerals/

Gonzalez, R., & Nelson, T. O. (1996). Measuring ordinal association in situations that contain tied scores. *Psychological Bulletin, 119*(1), 159–165.

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition, 37*(7), 1001–1013.

Hartwig, M. K., & Dunlosky, J. (2012). Study strategies of college students: are self-testing and scheduling related to achievement? *Psychononimic Bulletin & Review, 19*(1), 126–134.

Janes, J. L., Rivers, M. L., & Dunlosky, J. (2018). The influence of making judgments of learning on memory performance: positive, negative, or both? *Psychonomic Bulletin & Review, 25*(6), 2356–2364.

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*(5), 251–257.

Jordano, M. L., & Touron, D. R. (2018). How often are thoughts metacognitive? Findings from research on self-regulated learning, think-aloud protocols, and mind-wandering. *Psychonomic Bulletin & Review, 25*(4), 1269–1286.

Karpicke, J. D. (2009). Metacognitive control and strategy selection: deciding to practice retrieval during learning. *Journal of Experimental Psychology: General, 138*(4), 469–486.

Karpicke, J. D. (2017). Retrieval-based learning: a decade of progress. In J. Wixted (Ed.), *Cognitive psychology of memory, Vol. 2 of Learning and memory: a comprehensive reference*. (J. H. Byrne, Series Ed.).

Kelemen, W. L., & Weaver III, C. A. (1997). Enhanced memory at delays: why do judgments of learning improve over time? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 23*(6), 1394–1409.

Kimball, D. R., & Metcalfe, J. (2003). Delaying judgments of learning affects memory, not metamemory. *Memory & Cognition, 31*(6), 918-929.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: a construction-integration model. *Psychological Review, 95*(2), 163–182.

Koriat, A. (1997). Monitoring one’s own knowledge during study: a cue-utilization approach to judgments of learning. *Journal of Experimental Psychology: General, 126*(4), 349–370.

Kornell, N., & Son, L. K. (2009). Learners’ choices and beliefs about self-testing. *Memory, 17*(5), 493–501.

Maki, R. H., & Berry, S. L. (1984). Metacomprehension of text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 10*(4), 663–679.

Maki, R. H., & Serra, M. (1992). The basis of test predictions for text material. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18*(1), 116–126.

Metcalfe, J. (2009). Metacognitive judgments and control of study. *Current Directions in Psychology Science, 18*(3), 159–163.

Metcalfe, J., & Finn, B. (2008). Evidence that judgments of learning are causally related to study choice. *Psychonomic Bulletin and Review, 15*(1), 174–179.

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.

Morehead, K., Rhodes, M. G., & DeLozier, S. (2016). Instructor and student knowledge of study strategies. *Memory, 24*(2), 257–271.

Morris, C. C. (1990). Retrieval processes underlying confidence in comprehension judgments. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(2), 223–232.

Myers, S. J., Rhodes, M. G., & Hausman, H. E. (2020). Judgments of learning (JOLs) selectively improve memory depending on the type of test. *Memory & Cognition*, 1–14.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin, 95*(1), 109–133.

Nelson, T. O., Narens, L., & Dunlosky, J. (2004). A revised methodology for research on metamemory: Pre-judgment Recall And Monitoring (PRAM). *Psychological Methods, 9*(1), 53–69.

Peer, E., Vosgerau, J., & Acquisti, A. (2014). Reputation as a sufficient condition for data quality on Amazon Mechanical Turk. *Behavior Research Methods, 46*(4), 1023–1031.

Putnam, A. L., & Roediger, M. A. (2013). The effects of response modality on retrieval. *Memory & Cognition, 41*(1), 36–48.

Rawson, K. A., & Dunlosky, J. (2002). Are performance predictions for text based on ease of processing? *Journal of Experimental Psychology: Learning, Memory, and Cognition, 28*(1), 69–80.

Rogers, B. (2001). *TOEFL CBT Success*. New Jersey: Peterson's.

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463.

Schwartz, B. L. (1994). Sources of information in metamemory: judgments of learning and feelings of knowing. *Psychonomic Bulletin & Review, 1*(3), 357–375.

Smith, M. A., Roediger III, H. L., & Karpicke, J. D. (2013). Covert retrieval practice benefits retention as much as overt retrieval practice. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 39*, 1712–1725.

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.

Sommer, W., Heinz, A., Leuthold, H., Matt, J., & Schweinberger, S. R. (1995). Metamemory, distinctiveness, and event-related potentials in recognition memory for faces. *Memory & Cognition, 23*(1), 1–11.

Son, L. K., & Metcalfe, J. (2005). Judgments of learning: evidence for a two-stage process. *Memory & Cognition, 33*(6), 1116–1129.

Spellman, B. A., & Bjork, R. A. (1992). When predictions create reality: judgments of learning may alter what they are intended to assess. *Psychological Science, 5*, 315–316.

Tauber, S. K., & Witherby, A. E. (2019). Do judgments of learning modify older adults' actual learning? *Psychology and Aging, 34*(6), 836–847.

Tauber, S. K., Dunlosky, J., & Rawson, K. A. (2015). The influence of retrieval practice versus delayed judgments of learning on memory: resolving a memory-metamemory paradox. *Experimental Psychology, 62*(4), 254–263.

Tauber, S. K., Witherby, A. E., Dunlosky, J., Rawson, K. A., Putnam, A. L., & Roediger III, H. L. (2018). Does covert retrieval benefit learning of key-term definitions? *Journal of Applied Research in Memory and Cognition, 7*(1), 106–115.

Thiede, K. W., & Anderson, M. C. (2003). Summarizing can improve metacomprehension accuracy. *Contemporary Educational Psychology, 28*(2), 129–160.

Thiede, K. W., & de Bruin, A. B. H. (2017). Self-regulated learning in reading. In D. H. Schunk & J. A. Greene (Eds.), *Handbook of self-regulation of learning and performance* (3rd ed., pp. 124–137). New York, NY: Routledge Press.

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*, 1024–1037.

Thiede, K. W., Anderson, M. C. M., & Therriault, D. (2003). Accuracy of metacognitive monitoring affects learning of texts. *Journal of Educational Psychology, 95*(1), 66–73.

Thiede, K. W., Dunlosky, J., Griffin, T. D., & Wiley, J. (2005). Understanding the delayed-keyword effect on metacomprehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 31*(6), 1267–1280.

Thiede, K. W., Wiley, J., & Griffin, T. D. (2011). Test expectancy affects metacomprehension accuracy. *British Journal of Educational Psychology, 81*(2), 264–273.

Wiley, J., Griffin, T. D., & Thiede, K. W. (2005). Putting the comprehension in metacomprehension. *The Journal of General Psychology, 132*(4), 408–428.

Winne, P. H., & Hadwin, A. F. (1998). Studying as self-regulated learning. In D. J. Hacker, J. Dunlosky, & A. C. Graesser (Eds.), *Metacognition in educational theory and practice* (pp. 277–304). Hillsdale, NJ: Erlbaum.

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.

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.