**Hildenbrand**, L., & Wiley, J. (2021). Can closed-ended practice tests promote understanding from text? In *Proceedings of the 43rd Annual Conference of the Cognitive Science Society*, 327-333. Available at: https://cognitivesciencesociety.org/program/

# Can Closed-ended Practice Tests Promote Understanding from Text?

**Lena Hildenbrand (lhilde3@uic.edu)**
Department of Psychology, University of Illinois at Chicago
Chicago, IL 60607 USA

**Jennifer Wiley (jwiley@uic.edu)**
Department of Psychology, University of Illinois at Chicago
Chicago, IL 60607 USA

**Abstract**

Many studies have demonstrated that testing students on to-be-learned materials can be an effective learning activity. However, past studies have also shown that some practice test formats are more effective than others. Open-ended recall or short answer practice tests may be effective because the questions prompt deeper processing as students must generate an answer. With closed-ended testing formats such as multiple-choice or true-false tests, there are concerns that they may prompt only superficial processing, and that any benefits will not extend to non-practiced information or over time. They also may not be effective for improving comprehension from text as measured by how-and-why questions. The present study explored the utility of practice tests with closed-ended questions to improve learning from text. Results showed closed-ended practice testing can lead to benefits even when the learning outcome was comprehension of text.

**Keywords:** learning; comprehension; metacomprehension

## Can Closed-ended Practice Tests Promote Understanding from Text?

Testing students on to-be-learned materials can be an effective learning activity that enhances retention relative to comparison conditions such as re-reading (see Roediger & Karpicke, 2006 for a review). Widely replicated in the literature, this phenomenon has been termed the 'testing effect'. Interestingly, testing has been found to be a powerful tool to improve learning even when initial performance on practice tests is low and no feedback or opportunities for restudy are provided, although the effects may be further enhanced when they are (Butler & Roediger, 2008; Little et al., 2012; Metcalfe, 2017). Different accounts have been offered for *why* testing is so effective for retention. Some argue that retrieval practice serves to enhance the accessibility of an item stored in memory (Karpicke & Smith, 2012). Others have suggested that retrieval prompts learners to reorganize and supplement information stored in memory, and to broaden rather than focus the associated semantic network (Carpenter, 2011). More recently, support for the latter account has been growing, especially under circumstances where the goal is learning from text rather than just memory for word lists or facts.

Both research on word lists and text passages suggests that some practice test formats are more effective than others. In work exploring learning from word lists (Carpenter & DeLosh, 2006) and foreign language vocabulary (Carrier & Pashler, 1992), the effect of practice testing on memory seems to increase along with the degree of elaboration that the practice tests encourage. In work exploring more complex learning from lecture and text materials, open-ended recall or short answer practice tests have been found to be more effective than less demanding options such as cued-recall or multiple-choice tests (Butler & Roediger, 2007; Hinze & Wiley, 2011; McDaniel, Anderson, Derbish, & Morrisette, 2007). *How* practice test questions engage the learner is what really seems to matter. Using open-ended practice essay tests, Hinze, Wiley, and Pellegrino (2013) found that when students engaged in more constructive processing and attempted to integrate information into a coherent situation model or mental model for the content, these activities supported better comprehension outcomes. Open-ended practice tests may be effective because the questions prompt the reader to generate a response and engage in deeper processing of the to-be-learned materials.

In contrast to open-ended questions for which learners must generate the appropriate answer, closed-ended tests such as multiple-choice and true-false tests are often regarded as rather superficial in the sense that they are thought to primarily rely on recognition processes (Brabec et al., 2020; Little et al., 2012). There are concerns that any benefits of practice will be limited to the practiced information or to immediate memory and will not be robust over time or transfer to non-practiced information. However, such charges may be premature. In a study exploring memory for information presented in text passages, Little et al. (2012) demonstrated that multiple-choice tests can be constructed in such a way that they boost recall of not just practiced but also related unpracticed information. In this study, participants were asked to study two text passages. Each participant completed either a multiple-choice or cued-recall practice test for one passage, and no test for the other. The passage that was not tested served as the baseline memory comparison condition. Following a 5-minute distractor task, the final test was administered in cued-recall format and included items for practiced content, unpracticed content from the tested passage, and also content from the second, untested passage. On this final memory test, both multiple-choice practice and cued-recall practice resulted in better performance on content that was explicitly tested compared to content of the untested passage. Additionally, for multiple-choice practice, a transfer effect could be observed for unpracticed content of the tested passage. Final test performance was improved for this content

compared to content of the untested passage. For cued recall practice, no such transfer could be observed. These results show that multiple-choice practice testing can be a useful tool to improve memory for information presented in text passages and may sometimes even be more effective than a testing format that requires some generation (cued recall). However, these enhanced benefits from multiple-choice practice tests may be due to a specific feature in their design. The multiple-choice items were constructed such that there were pairs of questions that incorporated the same competitive alternatives. Each item that appeared on the practice test included an alternative answer option that would seem plausible but was incorrect, with the goal that participants would be prompted to consider *why* it was incorrect. In the paired question from the "unpracticed" test, this alternative option was then the correct answer, and the other option was now the plausible but incorrect distractor. When one of these associated items was tested during practice, it led to better performance on the other unpracticed item.

While the Little et al. (2012) study focused on multiple-choice testing, Brabec et al. (2020) used the same text materials to test if it might be possible to see the same benefits from true-false testing. They constructed true-false questions that included alternatives (false statements) which directly competed with the true statements. In this study, a no-test condition was compared to true-false testing only (no cued recall). Participants completed a true-false practice test for one passage and no test for the other. The final test was again a cued-recall test for both passages administered following a 5-minute distractor task. In addition, another set of participants were tested after a 48-hour delay. Again, the final test assessed two categories of content from the tested passage: practiced and not practiced. Both were compared to memory for the content of the untested passage. Results showed that the true-false practice test improved final test performance for both practiced and not practiced content from the tested passage compared to content from the untested passage. These same effects were found both after a 5-minute delay and after 48 hours. Thus, these true-false testing benefits were found to be robust over time, which stands in stark contrast to the suggestion that this format might only produce immediate gains.

In sum, these studies show that both multiple-choice and true-false practice testing can be used as a tool to promote memory for information presented in text passages. Yet, because both studies used a no-test comparison rather than a re-read comparison, it is unclear whether any benefits are the result of testing versus more simply from re-exposure to the information. When Brabec et al. (2020) compared true-false practice testing to re-study, practice testing did not add a significant benefit over re-exposure. Further, while benefits for "unpracticed" content could be observed in the above studies, this was on items that were designed to be directly associated with items that were explicitly tested. The items in each pair included the same competitive alternatives, which suggests this result may not have required much transfer. Finally, it is important to note that these previous studies have primarily focused on learning from text as measured with retention measures (memory for information) rather than comprehension outcomes (ability to use information to answer how-and-why questions).

To remember what a text said is not necessarily the same as having understood it. As is highlighted in Kintsch's (1994) comprehension framework, readers process text at multiple levels of representation. The surface level encodes the exact words that are used. The textbase level represents the propositional contents of a text. Meanwhile, at the situation-model or mental-model level, the reader attempts to represent what a text is about by making connections between ideas from the text and with prior knowledge. It is this level of representation that best predicts performance on tests of comprehension where the reader is asked to go beyond the explicitly stated information (Kintsch, 1994; Mayer, 1989; Otero, Leon, & Graesser, 2002; Wiley, Griffin, & Thiede, 2005). Considering the difference between outcome measures that test for retention of information and comprehension of information, it was unclear whether the benefits of closed-ended practice testing would extend to contexts where students must not only remember facts but use the information from the text to make new connections. Further, because multiple-choice and true-false practice tests have only been examined independently of each other with no direct comparisons made between them, it was of interest to test whether these two formats might promote comprehension to the same extent.

## Differences Between True-false and Multiple-choice Tests

There are several differences between true-false and multiple-choice questions that may affect how individuals approach answering them. First, for multiple-choice questions it may be enough to simply recognize the best answer in the set of statements. In contrast, true-false questions may require a more thorough review of each statement that is provided.

Second, for multiple-choice questions it can be expected that among the answer choices provided there is some information that is both true and useful – the correct answer – while for true-false items it is possible that the information stated is either all true, all false, or partially true and partially false creating an overall invalid statement (Brabec, et al., 2020). For true-false tests, it has been shown that there is a tendency for students to mark statements as true more often than false, and that both reliability and validity of scores obtained for false items are greater than that obtained for those which are true (Cronbach, 1942). In the literature, this 'acquiescence bias' has been explained in terms of processing demands where initial acceptance of a statement is automatic but its rejection as false requires additional cognitive effort. Because true-false practice tests prompt students to at least consider rejecting each statement, this format may be associated with more effortful or deeper processing than multiple-choice practice tests.

Third, responses on multiple-choice tests reveal what students think is the best answer. Yet, even if the correct answer choice is selected, students may still hold incorrect beliefs regarding the other possible answer choices. When taking a true-false test, students must evaluate each statement separately which reveals more clearly what they really do and do not know.

These observations suggest that multiple-choice and true-false testing formats may differ in how students evaluate the response options, and in how deeply information is processed. If true-false questions prompt students to evaluate all statements more carefully, then a true-false format for practice tests should be better than a multiple-choice format as an opportunity to promote understanding from text.

## Promoting Metacomprehension with Closed-ended Practice Tests

Practice testing can also have indirect benefits such that it can help learners to diagnose what they do and do not know. In the literature, this awareness of one's own level of retention or understanding of information is respectively known as metamemory or metacomprehension. The focus for this study is on metacomprehension, which is typically assessed by asking individuals to make judgments of learning (JOLs) or comprehension (JOCs), and those judgments are then compared to measures of actual performance on inference or comprehension tests (Wiley, Griffin, & Thiede, 2005). The accuracy of these metacognitive judgments depends on the utility of the cues that individuals use to make such judgments. Some cues are more predictive than others of actual performance (Koriat, 1997; Thiede et al., 2010). For example, it has been demonstrated that JOLs are sensitive to manipulations of superficial cues such as audio volume or font size of materials even when the manipulations do not impact actual learning (Rhodes & Castel, 2009).

The postdiction effect is one of the most consistent effects in the metacomprehension literature. Taking a practice or initial test improves the accuracy of predicting performance on a later test (Griffin, Jee, & Wiley, 2009; Little & McDaniel, 2015; Maki & Serra, 1992; Thiede & Anderson, 2003; Thiede et al., 2009). When a testing activity precedes making JOLs or JOCs, this allows the learner to use that prior test experience as a basis for making their judgments. In other words, what is supposed to be a prediction of performance based on monitoring during learning is turned into a postdiction where test experience is now a salient, concrete cue available to the learner (Griffin et al., 2009). Even when explicitly told to predict future performance, it appears that learners will resort to such past experience cues if they are available (Finn & Metcalfe, 2007).

Still, even if accuracy improves for postdictions after taking any practice test, some practice test opportunities may improve metacomprehension more than others. A recent review concluded that the conditions that are most likely to improve metacomprehension accuracy are those that actively engage learners and elicit deeper processing (Griffin, Mielicki, & Wiley, 2019). Some of the highest levels of predictive metacomprehension accuracy have been seen when students engage in generative activities such as self-explanation or drawing before making their JOCs. These activities encourage efforts to integrate ideas and construct a situation model and as such make predictive cues more accessible (Griffin, Wiley, & Thiede, 2008; Wiley, 2019). If multiple-choice and true-false tests differ in the extent to which they prompt learners to engage in deeper processing and change the likelihood of accessing their situation model during testing, these different formats may also have different effects on metacomprehension accuracy.

## The Present Study

The present study examined three questions. The first question was if closed-ended practice tests would be seen to promote better understanding from text, as measured by a final open-ended comprehension question. Performance of students who engaged in closed-ended practice testing was compared to students in a re-reading condition. The second question was whether differences would be seen between true-false and multiple-choice practice test formats in improving understanding from text. The third question was if the two practice test formats would have different effects on metacomprehension monitoring accuracy. If true-false practice tests prompt students to engage in more in-depth processing of each response option and to verify individual propositions using their situation model of the text, then they may be more likely to yield better understanding than multiple-choice practice tests. Further, if true-false questions prompt more reasoning from the situation model, then experience-based cues that result from accessing the situation model during the practice tests would be expected to also increase metacognitive accuracy.

## Method

**Participants** Participants were 115 undergraduates (69% female, mean age = 18.7 years) at a large urban university who received course credit as part of an introductory psychology subject pool. Students identified as 16% Asian, 10% Black/African American, 27% Hispanic/Latinx, 13% Indian, and 26% White/Caucasian. All participants were fluent English speakers but 56% indicated they were bilingual.

**Design** The design of the study was between-subjects with three conditions: true-false practice test questions, multiple-choice practice test questions, and a re-read (no practice test) comparison condition. A between-participants design was selected to avoid carryover effects between conditions and more specifically so that it would be possible to better isolate the effects of the different practice tests on final test performance. Participants were randomly assigned to condition, and the conditions did not differ on ACT scores (a standardized assessment of college preparedness), Gates McGinitie vocabulary scores, or self-reported prior knowledge ratings (see Table 1; $F$s < 1.08).

Table 1: Sample Descriptives.

| | TF practice | | MC practice | | Re-read | |
|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| ACT | 25.04 | 4.56 | 25.63 | 4.75 | 25.13 | 4.52 |
| GMRT | 0.61 | 0.71 | 0.63 | 0.17 | 0.65 | 0.17 |
| PK: V | 3.08 | 1.48 | 3.05 | 1.78 | 3.55 | 2.04 |
| PK: E | 5.90 | 2.62 | 5.95 | 2.21 | 5.60 | 2.34 |
| PK: I | 2.79 | 1.75 | 3.27 | 2.02 | 3.16 | 1.97 |

Note: GMRT: Gates McGinitie vocabulary scores (reported in proportion correct out of total); PK: V prior knowledge ratings for volcanoes, PK: E evolution, and PK: I ice ages.

**Materials** Materials included three expository science texts, and two versions of practice tests (a true-false and a multiple-choice format). Three expository science texts were adapted from Griffin, Wiley, and Thiede (2008, 2019) and described complex phenomena in the natural sciences, namely volcanic eruptions, evolution, and ice ages. The texts ranged between 829 to 1080 words in length, had Flesch-Kincaid grade levels of 11–12, and reading ease scores in the difficult range of 31–49. The practice tests were adapted from the 5 multiple-choice inference questions used for each topic in Griffin, Wiley, and Thiede (2019). Because the current study focused on comprehension and not just memory for text, inference questions were used as stimuli. Inference questions test connections between ideas and ideas that go beyond that which is explicitly stated in the text. They require access to one's situation model for the text rather than just verbatim memory. To keep the content of the test questions as similar as possible, true-false inference verification items (20 total for each text) were constructed to map onto the 4 response options for each of the 5 multiple-choice questions. The wording of some statements was slightly modified so that the number of true vs false items was balanced (10 each) in the true-false format.

**Procedure** The experiment was fully computerized and administered online in two parts through Qualtrics.

Instructions and procedure were based primarily on Experiment 3 by Hinze, Wiley, and Pellegrino (2013). For Part 1, all participants were told they would read 3 short texts about science topics and that they would be tested on their understanding after 48 hours. Also, all were told they would perform a short activity to help them prepare for the final test and that for many of these activities they could not revisit the original text later.

All participants read the 3 texts self-paced and in the same topic order (Volcanoes, Evolution, Ice Ages). After reading, they engaged in their assigned practice test or re-reading, with type of activity held constant for all texts. Practice activities were presented in the same topic order as initial reading.

After completing these practice activities, participants were then asked to make a JOC for each topic. They were asked to judge how well they thought they understood each of the texts on a scale of 0, very poorly to 5, very well (Glenberg, Wilkinson, & Epstein, 1982).

Following Little and McDaniel (2015), participants completed JOCs immediately following all three practice tests or re-reading, and the final test was delayed until 48 hours after the initial session. A link to Part 2 was emailed 48 hours after Part 1 was completed. For Part 2, participants were told they would be tested on their understanding of the 3 texts. The final comprehension test for each topic consisted of an essay task (adapted from Sanchez & Wiley, 2006), asking the reader to explain in a minimum of 5 sentences 'how and why' each of the scientific phenomena occur. Previous work has shown that performance on how-and-why essay questions reliably correlates with performance on inference questions such as those used as a practice test in the current experiment (Hinze, Wiley, & Pellegrino, 2013; Sanchez & Wiley, 2006; Wiley et al., 2009).

The essay task was the same across all conditions and participants had to spend a minimum of 2 minutes answering each of the three questions. The final tests on the three topics were again presented in the same order in which participants had initially read them.

Finally, information was collected on prior knowledge for these science topics as part of an exit survey using 0-10 rating scales. These same questions were also asked in a pre-screening survey at the start of the semester. Students also indicated their scores on the ACT (a standardized assessment of college preparedness) and completed the Gates McGinitie Vocabulary Test (Version 10-12) as part of pre-screening.

**Coding and scoring** Responses on the essay task were scored using rubrics based on prior work on these texts (Sanchez & Wiley, 2006; Wiley et al., 2009). To create the rubrics, each explanatory text was analyzed for its underling causal model. Each response was scored for the presence or absence of 5 correct causal concepts, and a proportion score was computed out of the possible total of 5 concepts. This served as the comprehension outcome measure. This scoring was corroborated using latent semantic analysis (LSA, Landauer, Foltz, & Laham, 1998) where the semantic overlap between a constructed model response and each participant response was computed, with numbers closer to 1 representing a greater degree of semantic overlap. Responses were edited to correct misspellings, and to expand contractions and abbreviations. Semantic overlap was computed using a one-to-many, document-to-document analysis using the general reading up to first year college LSA space with maximal factors included. The correlation between the proportion scores and LSA overlap with model responses was positive and significant, $r = .43$, $p < .001$.

For metacomprehension outcomes, two different measures were examined. One was confidence bias, which is the signed difference between a learner's perceived comprehension (JOC) and their actual performance on the final comprehension test (the essay task). This difference reflects over- or under-confidence in learning. The closer this measure is to 0, the better a learner is calibrated to estimate their own performance. The other measure of interest was relative accuracy, which is the intra-individual correlation

between JOCs on each topic and performance on each topic. In other words, relative accuracy describes a learner's ability to differentiate between the texts they understood better from the texts they understood less well – the higher the correlation between judgments and actual performance the better a learner is aware of what they do and do not know. When correlations are not significantly different from zero, it suggests a lack of any ability to discriminate among the topics. A significant positive correlation suggests some ability to accurately detect relative levels of understanding. Test scores and confidence bias are displayed as proportions; relative accuracy is displayed as correlations.

## Results

**Comprehension outcomes** A one-way analysis of variance (ANOVA) was conducted to compare final test performance between the different types of practice activities that participants engaged in. As shown in the leftmost set of bars in Figure 1, the effect of practice condition was significant, $F(2, 113) = 3.22$, $MSE = .02$, $p = .04$, $\eta_p^2 = .05$. Planned comparisons indicated that participants assigned to the true-false testing condition performed significantly better than did those in the re-read comparison condition. Those assigned to the multiple-choice practice condition also performed significantly better on the final test than did those in the re-read control condition. There was no difference in overall performance on the final test between the two practice testing conditions (true-false vs multiple-choice).

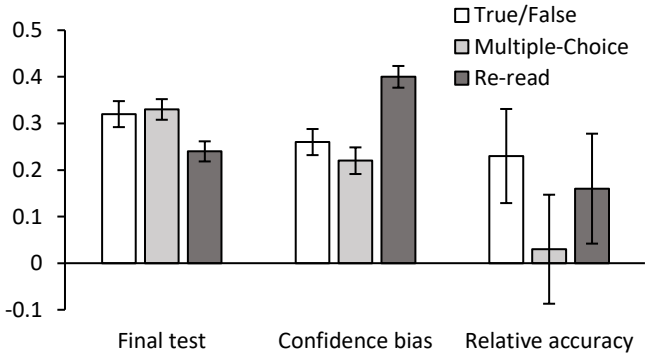


Figure 1: Means (and *SEs*) for Proportion Correct on Final Test, Confidence Bias, and Relative Accuracy Correlations

**Metacomprehension outcomes** A one-way analysis of variance (ANOVA) was conducted to examine effects of practice condition on confidence bias for JOCs provided before the final test. As shown in the middle set of bars in Figure 1, the effect of condition was significant, $F(2, 113) = 12.21$, $MSE = .27$, $p < .001$, $\eta_p^2 = .18$. Planned comparisons revealed that participants assigned to the re-read comparison condition were significantly more overconfident than those assigned to both the true-false practice testing condition and the multiple-choice practice testing condition. There was no significant difference in confidence bias between the two practice testing conditions.

The levels of relative accuracy are shown in the rightmost set of bars in Figure 1. For relative accuracy, the effect of condition did not reach significance, $F(2, 113) = .88$ $MSE = .43$, $p = .42$, $\eta_p^2 = .07$. However, relative accuracy was significantly above zero when participants took true-false practice tests, $t(38) = 2.30$, $p = .03$. Relative accuracy was not significantly above zero in the other two conditions: multiple choice, $t(36) = 0.20$, $p = .84$; re-read, $t(38) = 1.34$, $p = .19$.

## Discussion

The present findings demonstrated that closed-ended practice testing can promote understanding from text. Relative to re-reading, taking a multiple-choice or true-false practice test improved performance on a delayed test of comprehension. Since closed-ended testing was compared to a re-reading condition rather than to a no-test baseline in this study, it suggests the observed benefits of testing were not merely re-exposure effects.

Because closed-ended tests do not require the learner to actively generate an answer, there have been concerns that practice tests using closed-ended formats might only prompt superficial learning. Some preliminary work had suggested that closed-ended practice tests could improve memory for text, and this benefit might extend to unpracticed content (Brabec et al., 2020; Little et al., 2012). However, these benefits may have been the result of using pairs of questions that were closely related by design. What remained to be seen was whether closed-ended practice test formats could aid more complex learning from text including the application of understanding from text to answer comprehension questions.

Thus, an important contribution from the results of this study was showing that benefits from taking closed-ended practice tests can extend to contexts in which learning from text is assessed by measures of comprehension. Because this study used a how-and-why essay task as a final test, it was able to show that closed-ended practice testing could improve *understanding* and not just factual recall of the information. It also is important that in this study the benefits were seen after a 48-hour delay. Better performance on a delayed comprehension test helps to show that closed-ended practice tests are not just leading to short-lived advantages in memory and can translate to more durable learning gains.

Beyond showing improved performance on a delayed test of comprehension, this study also showed that taking a closed-ended practice test could influence the accuracy of metacomprehension judgments. Although no differences were seen due to which practice test format was used, taking either a multiple-choice or true-false practice test significantly decreased participants' overconfidence in their predicted final test performance compared to re-reading. This effect is consistent with the prior literature showing postdiction effects (Thiede et al., 2009).

Further, while no significant effects in relative accuracy were seen when comparing across the three conditions, it is potentially of interest that relative accuracy was significantly above zero only for those who completed the true-false practice tests. It will be important to investigate if this same

pattern suggesting a possible advantage for true-false practice tests may be seen in future studies that explore differences between these two closed-ended formats.

The most important next steps for future work will be adding feedback and the opportunity to re-study to this paradigm, which are conditions that have been argued to help students to maximize the benefits from practice testing (Butler & Roediger, 2008). The fact that closed-ended practice tests were shown to significantly improve comprehension and reduce confidence bias even in the absence of feedback and without there being an opportunity to restudy is encouraging. It is possible that in some way both of these closed-ended tests encourage the kind of 'deeper' processing known to promote learning from text. Although no differences between the two formats were significant in this sample, differences may become more obvious once feedback is added. Further, adding a restudy opportunity after the feedback will allow for the investigation of whether different closed-ended practice test formats might lead to differences in the kinds of restudy choices that students make, and whether they use the opportunity to correct misconceptions and address the gaps in their understanding, which could lead to even greater benefits such as from "errorful learning" (Metcalfe, 2017).

Some educators argue that testing in the classroom should be minimized, so that valuable time will not be taken away from classroom instruction (Roediger & Karpicke, 2006). However, practice tests have been documented to offer a number of benefits to students as a learning activity. Because closed-ended tests using multiple-choice or true-false formats are generally easier to administer and take up less time than do open-ended alternatives such as essay or short answer formats, an important goal for research is to ask whether (and under what conditions) closed-ended practice tests can offer the same benefits as open-ended alternatives. The present findings extend the literature to show that closed-ended practice testing can lead to benefits even when the learning goal is comprehension of text.

## Acknowledgments

The authors would like to thank Thomas D. Griffin and Tricia A. Guerrero for their thoughtful comments on this project as well as Meaghan Schmugge for help with coding. This research was partially supported by Grant R305A160008 from the Institute of Education Sciences to the second author, and by a University Fellowship from the University of Illinois at Chicago to the first author.

## References

Brabec, J. A., Pan, S. C., Bjork, E. L., & Bjork, R. A. (2020). True-false testing on trial: Guilty as charged or falsely accused? *Educational Psychology Review, 33*, 1-26.

Butler, A. C., & Roediger, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology, 19*(5), 514-527.

Butler, A. C., & Roediger, H. L. (2008). Feedback enhances the positive effects and reduces the negative effects of multiple-choice testing. *Memory & Cognition, 36*(3), 604-616.

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 37*(6), 1547-1552.

Carpenter, S. K., & DeLosh, E. L. (2006). Impoverished cue support enhances subsequent retention: Support for the elaborative retrieval explanation of the testing effect. *Memory & Cognition, 34*(2), 268-276.

Carrier, M. L., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20*(6), 633-642.

Cronbach, L. J. (1942). Studies of acquiescence as a factor in the true-false test. *Journal of Educational Psychology, 33*(6), 401-415.

Finn, B., & Metcalfe, J. (2007). The role of memory for past test in the underconfidence with practice effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 33*(1), 238-244.

Glenberg, A. M., Wilkinson, A. C., & Epstein, W. (1982). The illusion of knowing: Failure in the self-assessment of comprehension. *Memory & Cognition, 10*(6), 597-602.

Griffin, T. D., Wiley, J., & Thiede, K. W. (2008). Individual differences, rereading, and self-explanation: Concurrent processing and cue validity as constraints on meta-comprehension accuracy. *Memory & Cognition, 36*(1), 93-103.

Griffin, T. D., Jee, B. D., & Wiley, J. (2009). The effects of domain knowledge on metacomprehension accuracy. *Memory & Cognition, 37*(7), 1001-13.

Griffin, T. D., Wiley, J., & Thiede, K. W. (2019). The effects of comprehension-test expectancies on metacomp-rehension accuracy. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 45*(6), 1066-1092.

Griffin, T. D., Mielicki, M. K., & Wiley, J. (2019). Improving students' metacomprehension accuracy. In J. Dunlosky & K. Rawson (Eds.), *Cambridge handbook of cognition and education* (pp. 619-646). New York, NY: Cambridge University Press.

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects with completion tests. *Memory, 19*(3), 290-304.

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language, 69*(2), 151-164.

Karpicke, J. D., & Smith, M. A. (2012). Separate mnemonic effects of retrieval practice and elaborative encoding. *Journal of Memory and Language, 67*(1), 17-29.

Kintsch, W. (1994). Text comprehension, memory, and learning. *American Psychologist, 49*(4), 294-303.

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.

Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse Processes, 25*(2-3), 259-284.

Little, J. L., Bjork, E. L., Bjork, R. A., & Angello, G. (2012). Multiple-choice tests exonerated, at least of some charges: fostering test-induced learning and avoiding test-induced forgetting. *Psychological Science, 23*(11), 1337-1344.

Little, J. L., & McDaniel, M. A. (2015). Metamemory monitoring and control following retrieval practice for text. *Memory & Cognition, 43*(1), 85-98.

Maki, R. H., & Serra, M. (1992). Role of practice tests in the accuracy of test predictions on text material. *Journal of Educational Psychology, 84*(2), 200-210.

Mayer, R. E. (1989). Models for understanding. *Review of Educational Research, 59*(1), 43-64.

McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology, 19*(4-5), 954-1446.

Metcalfe, J., (2017) Learning from errors. *Annual Review of Psychology, 68*, 465-489.

Otero, J., León, J. A., & Graesser, A. C. (Eds.). (2002). *The psychology of science text comprehension.* Lawrence Erlbaum Associates Publishers.

Rhodes, M. G., & Castel, A. D. (2009). Metacognitive illusions for auditory information: Effects on monitoring and control. *Psychonomic Bulletin & Review, 16*(3), 550-554.

Roediger, H. L. & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1*(3), 181-210.

Sanchez, C., & Wiley, J. (2006). An examination of the seductive details effect in terms of working memory capacity. *Memory & Cognition, 34*(2), 344-355.

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

Thiede, K. W., Griffin, T. D., Wiley, J., & Anderson, M. C. M. (2010). Poor metacomprehension accuracy as a result of inappropriate cue use. *Discourse Processes, 47*(4), 331-362.

Thiede, K. W., Griffin, T. D., Wiley, J., & Redford, J. (2009). Metacognitive monitoring during and after reading. In D. J. Hacker, J. Dunlosky & A. C. Graesser (Eds). *Handbook of metacognition in education, pp. 85-106*. Routledge.

Wiley, J. (2019). Picture this! Effects of photographs, diagrams, animations, and sketching on learning and beliefs about learning from a geoscience text. *Applied Cognitive Psychology, 33*(1), 9-19.

Wiley, J., Goldman, S., Graesser, A., Sanchez. C., Ash, I., & Hemmerich, J. (2009). Source evaluation, comprehension, and learning in internet science inquiry tasks. *American Educational Research Journal, 46*(4), 1060-1106.

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