Educ Psychol Rev
DOI 10.1007/s10648-015-9311-9

INTERVENTION STUDY

A Classroom Study on the Relationship Between Student Achievement and Retrieval-Enhanced Learning

Shana K. Carpenter[1] · **Terry J. S. Lund**[1] ·
Clark R. Coffman[1] · **Patrick I. Armstrong**[1] ·
Monica H. Lamm[1] · **Robert D. Reason**[1]

Abstract Retrieval practice has been shown to produce powerful learning gains in laboratory experiments but has seldom been explored in classrooms as a means of enhancing students' learning of their course-relevant material. Furthermore, research is lacking concerning the role of individual differences in learning from retrieval. The current study explored the effects of retrieval in a large undergraduate introductory biology course as a function of individual differences in student achievement. Students completed in-class exercises that required them to retrieve course information (e.g., recalling definitions for terms and labeling diagrams) followed by feedback or to simply copy the information without retrieving it. A later quiz over the information showed that high-performing students benefited more from retrieving than copying, whereas middle- and low-performing students benefited more from copying than retrieving. When asked to predict their quiz scores following the in-class exercises, high-performers demonstrated better overall metacognitive calibration compared to middle- or low-performers. These results highlight the importance of individual differences in learning from retrieval and encourage future research using course-relevant material to consider the role of student achievement in classroom-based interventions.

Keywords Retrieval-enhanced learning · Student achievement · Individual differences · Metacognition

Many studies have shown that retrieval enhances learning. When students are given a test or quiz over information they have studied, the act of retrieving information leads to significant improvements in long-term memory, even when compared to additional opportunities to restudy the information. For example, Kang et al. (2007) found that undergraduate students' learning of scientific articles was significantly enhanced by reading the articles once and then

✉ Shana K. Carpenter
shacarp@iastate.edu

[1] Department of Psychology, Iowa State University, W112 Lagomarcino Hall, Ames, IA 50011, USA

Published online: 16 May 2015

completing short-answer quizzes over the articles, relative to simply reading them twice. This benefit of retrieval over restudying is typically referred to as the *testing effect* or *retrieval-enhanced learning* (for recent reviews, see Rawson and Dunlosky 2011; Roediger and Butler 2011).

Retrieval-enhanced learning is widely studied, having been demonstrated in over 100 studies in the last several years alone. Retrieval has been known to produce significant learning gains for a wide variety of materials, including word lists (e.g., Carpenter 2009, 2011; Halamish and Bjork 2011; Karpicke and Zaromb 2010; Kornell et al. 2011; Kuo and Hirshman 1997; Peterson and Mulligan 2013; Zaromb and Roediger 2010), foreign language vocabulary (Carpenter et al. 2008; Coppens et al. 2011; Finn and Roediger 2011; Kang and Pashler 2014; Karpicke and Roediger 2008; Pyc and Rawson 2010; Toppino and Cohen 2009; Vaughn and Rawson 2011; Vaughn et al. 2013), text passages (Agarwal et al. 2008; Butler 2010; Clark and Svinicki 2014; Hinze and Wiley 2011; Kubik et al. 2014; Roediger and Karpicke 2006), and video-recorded lectures (Butler and Roediger 2007). Given the consistency of these findings in laboratory studies, researchers have advocated more frequent use of retrieval-based approaches as means of promoting students' learning in authentic educational environments (e.g., Dunlosky et al. 2013; Pashler et al. 2007; Roediger and Pyc 2012).

Confidence in the power of retrieval to improve educational outcomes is tempered by the lack of classroom-based studies on this topic, however. Compared to hundreds of laboratory-based demonstrations of retrieval-enhanced learning under highly controlled conditions, we know much less about the promises—and potential pitfalls—of using retrieval to promote learning in classrooms. Particularly important—but especially lacking in the current literature—are studies comparing the effects of retrieval versus restudying on students' learning of their course material. Determining whether benefits of retrieval hold up in these authentic contexts is critical to establishing the generality of the effect, and the degree to which the use of retrieval to promote student learning should be advocated, and under what conditions.

Retrieval-Enhanced Learning in the Classroom

To date, only a handful of studies has explored the effectiveness of retrieval vs restudying on students' learning of course-relevant material. Some studies have observed significant benefits of retrieval over restudying in middle-school classrooms (e.g., Carpenter et al. 2009; Roediger et al. 2011) and in college-level online courses (e.g., McDaniel et al. 2007; McDaniel et al. 2012). However, others have observed no significant benefits of retrieval over alternative, nonretrieval-based activities (e.g., creating concept maps) on elementary school children's learning of science concepts (Karpicke et al. 2014, experiment 1).

These studies provide some support for the use of retrieval as a learning tool in classrooms but suggest that there may be some limitations to its effectiveness. Factors that drive the effectiveness of retrieval could be linked with a variety of individual and situational factors that are present in classrooms. Unlike a laboratory setting that strives to control such factors, a classroom includes a wider range of individual differences in student achievement, prior knowledge of the material, interest, and motivation. It is presently unknown how such factors might interact with the effects of retrieval, as previous studies addressing the role of individual differences in learning from retrieval—particularly in classroom settings—are lacking. Given the diversity among students in classrooms and the need for more classroom-based research on retrieval-enhanced learning, the primary goal of the current study was to provide data on the role of individual differences in learning from retrieval.

In the current study, undergraduate students in an introductory biology course learned information from their course pertaining to the topic of reproduction. Students completed different in-class exercises, some of which required retrieval (e.g., recalling definitions for terms such as *primary oocyte* and constructing a diagram of the process of *oogenesis*), and some that provided exposure to the information without requiring retrieval (e.g., copying the definitions rather than recalling them, or labeling a diagram that was already provided). All students completed a quiz 5 days later assessing knowledge of the information learned from the exercises.

The Importance of Student Achievement

To examine retrieval-enhanced learning as a function of individual differences, we focused on a measure that varies considerably across students and is known to influence learning—student achievement. Previous research has shown that the level of knowledge students have achieved on a given topic can be a powerful predictor of further learning. Specifically, students who have learned more on a given topic acquire new knowledge on that topic more readily than students who have not learned as much. In the literature on text comprehension, for example, students who score higher on a pretest measuring knowledge of a particular topic (e.g., the formation of stars), compared to those who score lower, demonstrate better learning of a never-before-seen passage relating to that topic. This effect has been shown for high school students learning about concepts in physical science (Boscolo and Mason 2003) and biology (McNamara et al. 1996), undergraduate students learning about concepts in biology (McNamara 2001), and both undergraduate and graduate students learning about concepts in physics (e.g., Alexander et al. 1994).

This association between prior knowledge and new learning can be influenced—and sometimes *reversed*—by other factors. For example, high-knowledge learners show the usual advantage over low-knowledge learners when the coherency of a text passage is reduced (e.g., by eliminating connective links between paragraphs or replacing nouns with pronouns), but this advantage for high-knowledge learners can be reduced or eliminated when the coherency of the text is increased (e.g., McNamara 2001; McNamara et al. 1996; see also Kalyuga et al. 2013). One explanation for this effect is that low-coherence text requires effortful, active processing to fill in the gaps, and high-knowledge learners are better equipped to do this than are low-knowledge learners, who lack the appropriate background knowledge to make these inferences. High-coherence text is less dependent on inferences and may even produce inferior learning among high-knowledge students due to its tendency to discourage active processing (e.g., see McNamara 2001).

In a related body of research on the *expertise reversal effect*, certain instructional methods have been shown to affect learning in different ways depending upon the expertise of the learner. For example, Lee et al. (2006) found that middle-school students' learning of chemistry concepts via a computerized simulation depended upon how the information was represented, as well as students' prior knowledge of science. Components of the computer simulation were either represented by a verbal label accompanied by a visual icon (e.g., "temperature" with an image of a burner), or only a verbal label (e.g., "temperature"). Whereas low-knowledge students learned better when the components were represented by the easier-to-understand visual icon with a verbal label compared to the verbal label alone, high-knowledge learners demonstrated the opposite pattern. In a study by Leppink et al. (2012), undergraduate students learned statistics concepts by either providing arguments to support their answer to a true/false statement (e.g., the sample mean is 10), or by reading arguments that had already been provided for them. Students who scored high on a pretest of statistical reasoning learned better from coming up with their own arguments than from reading the examples, whereas students who scored low on the same pretest learned better from reading the examples than from trying to come up with their own arguments. In another study, Cooper et al. (2001, experiment 4) taught high school students how to use a computerized spreadsheet program to make different types of calculations. After completing a tutorial on the program, students were given new problems and were asked to either imagine performing the steps in the calculations or to refer to the on-screen instructions that walked them through each step. Students who performed higher in their mathematics classes learned the calculations better through imagining than through following examples, whereas students who performed lower in their mathematics classes learned better through following examples than through imagining.

These studies suggest that a higher degree of baseline knowledge benefits learning on tasks that require learners to supply information that is not currently present. This appears to be the case for making inferences while reading a text (e.g., McNamara et al. 1996), and for coming up with solutions to problems rather than reading worked examples (Cooper et al. 2001; Leppink et al. 2012). When the task requires processing of information that is already present (e.g., following a worked example), having a high degree of knowledge may not help—and may even *hurt*—learning because the task encourages processing that is redundant with current knowledge and may lead to disengagement or distraction (e.g., see Lee and Kalyuga 2014). Indeed, in the study by Leppink et al., after learning the material through worked examples, a later test over the material showed that high-knowledge students actually performed slightly *worse* than low-knowledge students.

Applied to the current study, this suggests that retrieval practice may be more effective for high-knowledge students than for low-knowledge students. Retrieval requires an active search of memory, and high-knowledge students may be better equipped to engage in this processing because they have acquired knowledge of the topic that will facilitate successful retrieval. If low-knowledge students have acquired very little knowledge of the topic, there may be less information in memory to retrieve, and therefore, retrieval may be ineffective for learning, or even counter-effective, compared to simply reading the material. Thus, in the current study, retrieval-enhanced learning may be expected to be more pronounced for students who have achieved a higher degree of knowledge over the course material than for students who have achieved a lower degree of knowledge.

Metacognition

Prior knowledge has been shown to predict not only students' learning, but also their *perceptions* of their learning. Research on metacognition has shown that students can be poor predictors of their own learning, often giving estimates of their knowledge that exceed actual performance as measured by a later test (e.g., Carpenter et al. 2013; Castel et al. 2007; Dunlosky and Lipko 2009; Finn and Metcalfe 2007; Kornell and Bjork 2009). Furthermore, low performers show a greater tendency than high performers to be overconfident. For example, immediately before the exam in a college-level psychology course, Miller and Geraci (2011) asked students to predict their score. Across a series of experiments, low-performing students (those who scored within the lower quartile of the class) *overpredicted* their scores by as much as 18 %, whereas high-performing students (those who scored within the upper quartile of the class) consistently *underpredicted* their scores. Similar results were observed by Bol et al. (2005), who asked students in an undergraduate education course to predict their final exam scores, and found that lower-performing students overpredicted their scores by about twice as much as higher-performing students (12 vs 6 %, respectively).

Other studies show that students often fail to appreciate the benefits of retrieval, predicting that they will recall the information better after having read or restudied it compared to having retrieved it (Agarwal et al. 2008). Presently, however, it is unknown whether this effect occurs in a classroom setting for students learning course material, and whether it varies according to student achievement level. To explore this, after students completed different types of retrieval exercises in the current study, we asked them to estimate how well they would score on a later quiz over the information. Comparing students' predicted scores with their actual scores allows an assessment of metacognitive awareness as a function of student achievement and the type of exercise (retrieval-based, or nonretrieval-based) that they used to learn the material.

Method

Participants

A total of 311 students from an introductory biology course at a large Midwestern University were invited to participate in the study in exchange for class participation credit. Thirty-six students either did not complete the in-class exercises or were absent on the day the follow-up quiz was administered, resulting in 275 students who completed all phases of the study.

Materials and Design

The study materials consisted of five term definition pairs and a diagram depicting the process of oogenesis (see Appendix A). These materials were part of the regular course curriculum. On the day the study was conducted, this information had not yet been covered by the instructor in class but had been included on a study guide that students received outside of class, and also in the assigned readings that students received the previous week.

Students completed in-class exercises that required them to engage with this material in one of four different ways, modeled after common instructional methods that have been used to learn this type of material. In the *Copy Definitions* + *Label Diagram* condition, students were provided with a sheet of paper that provided the terms and definitions, along with an unlabeled diagram. Students were asked to copy the definitions and label each of the terms within the diagram. In the *Copy Definitions* + *Draw Diagram* condition, students were also provided with the terms and definitions and were asked to copy the definitions, but this time, no diagram was provided, and students were asked to draw the diagram and label each of the corresponding terms. In the *Recall Definitions* + *Label Diagram* condition, students were given the terms (but not the definitions) and an unlabeled diagram. They were asked to recall the definition for each term and label each term within the diagram. Finally, in the *Recall Definitions* + *Draw Diagram* condition, students were provided with only the terms and asked to recall the definitions, draw the diagram, and label each of the terms within the diagram. Thus, the first two conditions did not require students to retrieve the definitions, whereas the last two conditions did.

Each student was assigned to one of these four conditions based on the seating arrangement of the class, which was randomly determined at the beginning of the semester. The classroom was organized into eight seating "zones," each consisting of two rows of seats that shared a tier within the classroom. The conditions were distributed in alternating fashion by zone, such that students in zones 1 and 5 received the Recall Definitions + Draw Diagram condition (n=82), students in zones 2 and 6 received the Copy Definitions + Draw Diagram condition (n=66), students in zones 3 and 7 received the Recall Definitions + Label Diagram condition (n=73), and students in zones 4 and 8 received the Copy Definitions + Label Diagram condition (n= 54). During the next class period, 5 days later, all students were given an unannounced quiz over the information from the in-class exercises.

Procedure

The study was conducted at the beginning of class. Following some introductory announcements by the instructor, students were informed that they would be working on some in-class exercises related to things they were learning in the course. They were asked to work on the exercises individually, without help from notes, books, or other resources, and to raise their hands when finished. Students were given an opportunity to ask questions, and then, the exercises were distributed according to the system described above.

Students completed the exercises at their own pace. Upon completion, each student raised his/her hand. At that point, the worksheet was collected, and the student was then given an answer sheet containing the terms and definitions, as well as the complete labeled diagram (see Appendix A). Students were asked to use this information to reflect on their accuracy on the exercise they had just completed. After reviewing the answer sheet at their own pace, each student was asked to make a judgment of learning (JOL) regarding how well they believed they would score on a multiple-choice quiz over this information. Students were asked to make two JOLs, one estimating their score (expressed as percent correct) on a multiple-choice quiz if it were given immediately, and the other estimating their score on the same quiz if it were given during the following week. Asking students to make both JOLs provided an opportunity to assess their predictions of their own performance both immediately and after a delay. Students wrote down these two JOL values on their answer keys, and then returned the sheets to the researchers.

After finishing the exercise on oogenesis, each student was given a similar exercise pertaining to the process of spermatogenesis, another topic that would soon be covered in the course. Anticipating that students would complete the oogenesis exercises at different times (e.g., the Recall Definitions + Draw Diagram condition takes longer than the Copy Definitions + Label Diagram), the purpose of this second exercise was to provide some additional course-relevant activities for students to work on (to keep them engaged) while other students completed the oogenesis exercises. Performance on the second exercise was not of primary interest, however, and will not be discussed. Some of the content relating to spermatogenesis is similar (but not identical) to that of oogenesis, raising the possibility that portions of the filler exercise could have provided some exposure to concepts that were encountered on the target exercise over oogenesis. Only performance on the oogenesis exercise was analyzed, with the understanding that the filler exercise could have provided additional feedback over some of the concepts. All students completed both exercises (lasting approximately 20 min altogether), and after all materials were collected from students, the instructor commenced with regular class activities.

During the next class period (5 days later), students were given an unannounced quiz over the information from the in-class exercises. All students received the same double-sided sheet of paper containing 20 quiz questions (10 questions first pertaining to oogenesis, and then 10 to spermatogenesis). In consultation with the revised Bloom's taxonomy (Anderson et al. 2001), the first five questions from each topic were designed to assess knowledge and the last five to assess comprehension (see Appendix B). Students were asked to complete the quiz individually and were encouraged to do their best but were informed that their score would not count toward their grade in the course. As with the in-class exercises, participation credit was granted for completion of the quiz, regardless of students' scores. At the bottom of the second page of the quiz, students were asked to indicate how much, over the last 5 days, they had studied any of the information contained on the quiz. Students indicated their response using a 0–5 scale (0 representing "not at all" and 5 representing "a lot"). After all students handed in their quizzes, they were given a debriefing concerning the nature of the study, were encouraged to ask questions, and were provided with the contact information of the primary researcher.

Results

Student Achievement

Student achievement was operationalized as overall performance in the course based on credit earned from four mandatory exams (all of which were completed prior to the in-class experiment) and a number of daily in-class activities (excluding the in-class experiment) that were administered throughout the semester via individualized response systems, or "clickers." For the 275 students who completed the study, overall course performance ranged from 38 to 95 % (*M*=71.5 %, *SD*=11 %).

Students were partitioned into performance levels based on whether their overall course performance fell within the upper, middle, or lower one third of the students who completed the study. Table 1 displays mean course performance for these levels across the four experimental conditions. Across the four conditions, no significant differences in course performance were observed among high performers, $F(3, 86)=1.15$, $p=.33$, among middle performers, $F(3, 88)=1.35$, $p=.26$, or among low performers, $F(3, 89)=2.25$, $p=.09$.

Accuracy on the In-Class Exercises

Accuracy on the in-class exercises was scored by two independent raters who were knowledgeable about the content. Accuracy of the definitions was scored by awarding two points for a fully correct answer, one point for a partially correct answer and zero point for an incorrect answer. Accuracy of the diagrams was scored by awarding one point for each component of the diagram that was correctly drawn, plus one additional point if the correct label was included. For the two conditions requiring students to label the diagram but not draw it (the Copy Definitions + Label Diagram condition and the Recall Definitions + Label Diagram condition), one point was automatically included for the presence of the component in the diagram, and accuracy of the labeling was scored by awarding one point for each component

Table 1 Mean course performance for high, middle, and low performers across the four experimental conditions

	Conditions			
	Copy definition + Label diagram	Copy definition + Draw diagram	Recall definition + Label diagram	Recall definition + Draw diagram
High performers	.84 (.06)	.82 (.05)	.84 (.05)	.84 (.05)
	$n=13$	$n=19$	$n=23$	$n=35$
Middle performers	.70 (.03)	.71 (.03)	.72 (.03)	.72 (.03)
	$n=18$	$n=19$	$n=27$	$n=28$
Low performers	.58 (.07)	.61 (.04)	.62 (.04)	.60 (.05)
	$n=23$	$n=28$	$n=23$	$n=19$

Standard deviations are given in parentheses. High performers earned course scores that fell within the upper 1/3 of the sample, middle performers earned course scores that fell within the middle 1/3, and low performers earned course scores that fell within the lower 1/3

that was correctly labeled. Across all conditions, inter-rater correlations for the accuracy of the definitions and diagrams (excluding the two conditions that involved merely copying the definitions) ranged from .72 to .91, ps<.001. Accuracy across all conditions was computed by averaging the two raters' scores.

Table 2 displays accuracy on the experimental exercises as a function of student course performance and condition. For the conditions that required retrieval of the definitions (Recall Definitions + Label Diagram and Recall Definitions + Draw Diagram), high performers retrieved the definitions better than middle performers [ts>2.38, ps<.03, ds>.65] or low-performers [ts>3.86, ps<.001, ds>1.16]. Though middle performers retrieved the definitions better than low performers in the Recall Definition + Label Diagram condition [$t(48)=3.07$, $p=.003$, $d=.88$], this was not true for the Recall Definition + Draw Diagram condition [$t(45)=.49$, $p=.62$]. In the Recall Definitions + Draw Diagram condition, high performers drew and labeled the diagram better than middle performers [$t(61)=2.03$, $p=.047$, $d=.52$] or low performers [$t(52)=3.89$, $p<.001$, $d=1.22$], and middle performers drew and labeled the diagram better than low performers [$t(45)=2.11$, $p=.04$, $d=.67$]. In the Copy Definitions +

Table 2 Accuracy on the in-class exercises for high, middle, and low performers across the four experimental conditions

	Conditions							
	Copy definition + Label diagram		Copy definition + Draw diagram		Recall definition + Label diagram		Recall definition + Draw diagram	
	Definition	Diagram	Definition	Diagram	Definition	Diagram	Definition	Diagram
High performers	1.00 (.00)	.92 (.02)	1.00 (.00)	.32 (.06)	.32 (.05)	.87 (.03)	.37 (.04)	.35 (.05)
Middle performers	1.00 (.00)	.88 (.03)	1.00 (.00)	.52 (.06)	.19 (.02)	.73 (.03)	.18 (.03)	.20 (.05)
Low performers	1.00 (.00)	.86 (.02)	1.00 (.00)	.30 (.05)	.09 (.02)	.75 (.04)	.15 (.04)	.06 (.03)

Standard errors are given in parentheses. The 100 % accuracy for the definitions in the first two conditions indicates that students copied the definitions verbatim without making any errors

Draw Diagram condition, middle performers drew and labeled the diagram better than high performers [*t*(36)=2.49, *p*=.018, *d*=.81] or low performers [*t*(45)=2.74, *p*=.009, *d*=.82], with no significant difference between the latter two, *t*(45)=.25.

For the conditions requiring students to label the diagram that was provided, high performers labeled the diagram better than middle or low performers. In the Recall Definitions + Label Diagram condition, high performers labeled the diagram better than middle performers [*t*(48)=3.48, *p*=.001, *d*=1.00] or low performers [*t*(44)=2.5, *p*=.016, *d*=.74], with no significant difference between the latter two, *t*(48)=.29, *p*=.77. Though the same pattern occurred for the Copy Definition + Label Diagram condition, accuracy was fairly high for all students. A marginally significant advantage occurred for high performers over low performers [*t*(34)= 1.77, *p*=.08, *d*=.64], but not between high and middle performers [*t*(29)=1.13, *p*=.27], or between middle and low performers [*t*(39)=.61, *p*=.55].

Quiz Scores

Quiz scores were examined as a function of experimental condition, type of quiz question (knowledge vs comprehension), and student course performance. Table 3 displays the mean quiz scores as a function of experimental condition and course performance for knowledge questions (upper half) and comprehension questions (lower half). For knowledge questions, high performers appeared to benefit more from exercises requiring retrieval of the definitions compared to copying of the definitions, whereas middle performers and low performers appeared to show the opposite pattern. Within each of the three course performance levels, no significant differences were observed between the two conditions requiring retrieval of the definitions [*t*s<.34, *p*s>.70], or between the two conditions requiring copy of the definitions [*t*s<.95, *p*s>.35]. Thus, to increase the sample sizes within each group for the comparisons of interest—retrieval vs copying—we combined the two conditions requiring retrieval of the definitions (Recall Definitions + Draw Diagram and Recall Definitions + Label Diagram), and the two conditions requiring copy of the definitions (Copy Definitions + Label Diagram and Copy Definitions + Draw Diagram). A 3×2 (performance×retrieval) between-subject analysis of variance (ANOVA) revealed a significant interaction, $F(2, 269)=5.04$, $p=.007$, $\eta^2=.04$, in

Table 3 Mean quiz scores for knowledge and comprehension questions across the four experimental conditions as a function of student course performance

	Conditions			
	Copy definition + Label diagram	Copy definition + Draw diagram	Recall definition + Label diagram	Recall definition + Draw diagram
Knowledge questions				
High performers	.72 (.09)	.65 (.05)	.79 (.04)	.80 (.03)
Middle performers	.73 (.06)	.65 (.06)	.67 (.05)	.65 (.05)
Low performers	.63 (.05)	.56 (.05)	.47 (.05)	.47 (.05)
Comprehension questions				
High performers	.65 (.05)	.55 (.05)	.62 (.06)	.65 (.04)
Middle performers	.52 (.05)	.59 (.05)	.51 (.04)	.47 (.04)
Low performers	.50 (.06)	.39 (.04)	.46 (.05)	.41 (.05)

Standard errors are given in parentheses that high-performing students benefitted more from exercises that required retrieval vs copying, whereas middle- and low-performing students did not.

Figure 1 (left panel) displays this interaction. High performers' scores were significantly higher following retrieval than copying of the definitions, $t(88)=2.45$, $p=.016$, $d=.50$, middle performers' scores were similar following retrieval vs copying, $t(90)=.53$, $p=.60$, and low performers' scores were actually *lower* following retrieval than copying, $t(91)=2.31$, $p=.02$, $d=.48$. The same 3×2 ANOVA revealed no overall main effect of retrieval vs copying, $F(1, 269)=.12$, $p=.73$, but a main effect of performance level, $F(2, 269)=17.31$, $p<.001$, $\eta^2=.11$, in that high performers achieved higher quiz scores overall compared to middle performers or low performers.

As a supplemental analysis, we computed correlations between student course performance and later quiz scores for the groups that learned the definitions through retrieval vs copying. Consistent with the interaction described above, this correlation was stronger for those students who learned the definitions through retrieval [$r(155)=.44$, $p<.001$] than through copying [$r(120)=.21$, $p=.02$], *Fisher's r-to-z transformation*=2.11, $p=.03$.

For comprehension questions (Fig. 1, right panel), the same 3×2 ANOVA revealed only a main effect of performance level, $F(2, 269)=12.13$, $p<.001$, $\eta^2=.08$. There was no main effect of retrieval vs copying, and no interaction, Fs<1.29, ps>.27. Correlations between student course performance and later quiz scores were similar for those who learned the definitions through retrieval [$r(155)=.39$, $p<.001$] vs copying [$r(120)=.34$, $p<.001$], *Fisher's r-to-z-transformation*=.47, $p=.64$, indicating a positive relationship between student course performance and later quiz scores that did not depend upon whether students learned definitions through retrieval vs copying.

Metacognition

Students' metacognitive calibration was assessed by comparing predicted quiz scores with actual quiz scores. Although students made two predictions—one pertaining to an immediate

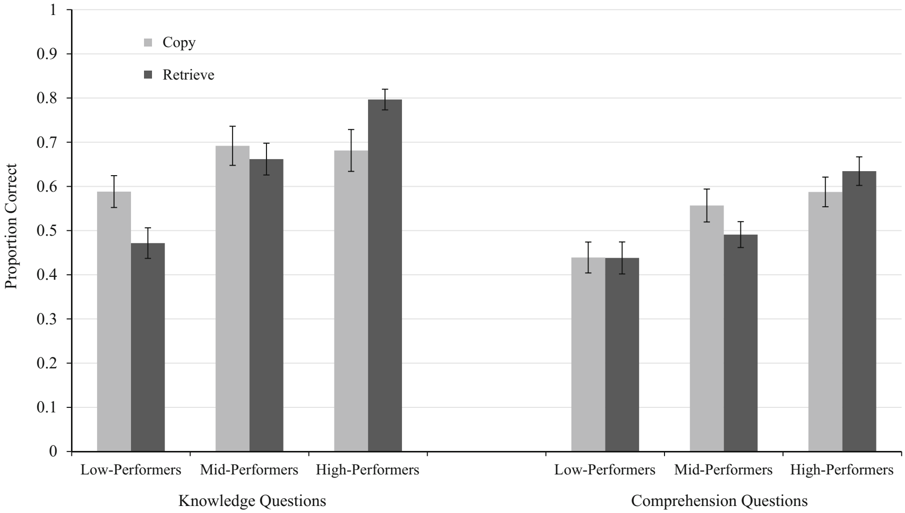


Fig. 1 Mean quiz scores for knowledge questions (*left panel*) and comprehension questions (*right panel*) as a function of student course performance and whether definitions were learned through retrieval vs copying. *Error bars* represent standard errors

quiz and one pertaining to a quiz "next week"—only the latter is relevant in computing calibration, as the quiz itself was administered the week after the in-class exercises. Calibration scores were based on knowledge questions (and not comprehension questions), as the knowledge questions were identical to those that students encountered on the in-class exercises and would thus represent the information that was available when students made their predictions.

Table 4 shows students' predicted vs actual quiz scores on knowledge questions for high performers, middle performers, and low performers. Individual calibration scores (computed by subtracting the actual quiz score from the predicted quiz score) were computed for each student. Across performance levels, no significant differences in calibration scores were observed between the two conditions requiring retrieval of the definitions [*t*s<.50, *p*s>.60], or between the two conditions requiring copying of the definitions [*t*s<.40, *p*s>.70]. Thus, like before, we combined scores from the two conditions requiring retrieval of the definitions (Recall Definitions + Label Diagram and Recall Definitions + Draw Diagram) and the two conditions requiring copying of the definitions (Copy Definitions + Label Diagram and Copy Definitions + Draw Diagram). A 3×2 (performance×retrieval) between-subject ANOVA on the calibration scores revealed a significant interaction, $F(2, 264)=3.21$, $p=.042$, $\eta^2=.02$.

This interaction is displayed in Fig. 2. High performers demonstrated better calibration following retrieval than copying, $t(87)=2.11$, $p=.037$, $d=.44$, middle performers demonstrated a similar but nonsignificant pattern, $t(89)=.89$, $p=.37$, and low performers demonstrated slightly *worse* calibration following retrieval than copying, $t(88)=1.45$, $p=.15$. The ANOVA revealed no overall main effect of retrieval vs copying, $F(1, 264)=.83$, $p=.36$, but did reveal a significant main effect of performance level, $F(2, 264)=10.74$, $p<.001$, $\eta^2=.07$, in that overall calibration was better for high performers than for middle performers or low performers. The overall calibration score of high performers ($M=4.45$ %) did not differ significantly from zero, $t(88)=1.47$, $p=.15$, indicating a close match between students' predicted scores and their actual scores. However, the calibration score of

Table 4 Mean predicted and actual quiz scores across the four experimental conditions as a function of student course performance

	Conditions			
	Copy definition + Label diagram	Copy definition + Draw diagram	Recall definition + Label diagram	Recall definition + Draw diagram
Predicted scores				
High performers	.84 (.04)	.79 (.04)	.78 (.05)	.80 (.03)
Middle performers	.89 (.03)	.81 (.06)	.75 (.04)	.76 (.04)
Low performers	.84 (.03)	.77 (.03)	.80 (.04)	.76 (.06)
Actual scores				
High performers	.72 (.09)	.65 (.05)	.79 (.04)	.79 (.03)
Middle performers	.73 (.06)	.67 (.06)	.67 (.05)	.65 (.05)
Low performers	.61 (.05)	.57 (.05)	.47 (.05)	.48 (.05)
Calibration scores (predicted score—actual score)				
High performers	.12 (.11)	.14 (.07)	−.01 (.06)	.01 (.03)
Middle performers	.16 (.07)	.14 (.07)	.08 (.06)	.11 (.06)
Low performers	.23 (.06)	.20 (.05)	.33 (.06)	.28 (.07)

Standard errors are given in parentheses. Values in the table exclude data from five students who did not provide a JOL

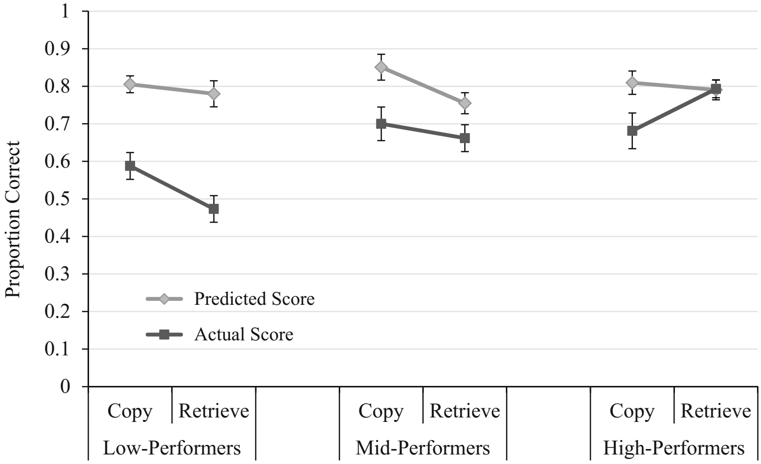


Fig. 2 Mean predicted and actual quiz scores for knowledge questions as a function of student course performance and whether definitions were learned through retrieval vs copying. Means exclude data from five students who did not provide a judgment of learning. *Error bars* represent standard errors

middle performers (M=11.59 %) was significantly greater than zero, $t(90)=3.67$, $p<.001$, $d=.39$, as was the calibration score of low performers (M=25.82 %), $t(89)=8.39$, $p<.001$, $d=.88$, indicating overconfidence. Correlation analysis between student course performance and calibration scores confirmed the same pattern by revealing a negative relationship, $r(270)=-.27$, $p<.001$, indicating that overconfidence increased as student course performance decreased.

It is informative as well to compare students' predictions of their scores on an immediate quiz vs a delayed quiz. Are students aware that forgetting occurs over time, such that a delayed quiz would likely yield lower scores than an immediate quiz? To the contrary, students' predictions for the delayed quiz were actually higher than for the immediate quiz. This was true for high performers (77 vs 80 %, respectively), middle performers (64 vs 79 %, respectively), and low performers (63 vs 79 %, respectively). A possible explanation for this pattern is that while thinking about a future quiz and how they would score, students may have assumed that they would have the opportunity to study for that quiz. As a result, they would expect their score on a future quiz to be higher than on a quiz given right at that moment. The fact that students were asked to report how much they studied prior to taking the quiz provides an opportunity to explore calibration under conditions in which this assumption was met—i.e., when students actually did study the information. Table 5 reports predicted quiz scores vs actual quiz scores for high, middle, and low performers as a function of how much students reported studying the information prior to the quiz. Table 5 indicates that even when students studied the information, they still overpredicted their own scores, and this overconfidence was greatest for lower performing students.

Explaining Different Learning Patterns as a function of Student Course Performance

Why are high performers more likely than middle or low performers to benefit from retrieval? There are at least two possibilities. First, quiz scores may reflect different degrees of

Table 5 Mean predicted and actual quiz scores as a function of student course performance and the amount of studying that students engaged in prior to the quiz

	Amount of studying (0="not at all," 5="a lot")				
	0	>0	>1	>2	>3
High performers	(*n*=62)	(*n*=26)	(*n*=11)	(*n*=4)	(*n*=1)
Predicted score	79 %	82 %	82 %	68 %	100 %
Actual score	76 %	75 %	67 %	65 %	80 %
Middle performers	(*n*=57)	(*n*=32)	(*n*=20)	(*n*=6)	(*n*=1)
Predicted score	76 %	85 %	87 %	90 %	60 %
Actual score	67 %	71 %	74 %	77 %	20 %
Low performers	(*n*=48)	(*n*=40)	(*n*=19)	(*n*=7)	(*n*=1)
Predicted score	75 %	85 %	88 %	90 %	70 %
Actual score	52 %	55 %	58 %	51 %	40 %

These values are reported for descriptive purposes, as very few students reported engaging in high amounts of studying (i.e., only one student in each of the three groups indicated a value greater than 3). Values in the table exclude five students who did not provide a judgment of learning, and four additional students who did not indicate whether they studied

postretrieval exposure to the material. Given that the quiz was administered 5 days after the in-class exercises, it is possible that students may have studied some of the material in-between completing the exercises and taking the quiz, and if so, performance could reflect the effects of studying (which may be more likely among high performers) and not the effects of the retrieval exercises per se. However, this possibility seems unlikely, as the amount that students reported studying was not correlated with later quiz scores, $r(270)=-.014$, $p=.82$.

The other possibility is that retrieval-enhanced learning is linked with the rate of success at recalling the definitions during the in-class exercises. Analysis of performance on the in-class exercises supports this idea (Table 2). In the conditions requiring retrieval of the definitions, high performers recalled 35 % of the definitions correctly ($SD=24$ %), middle performers recalled 18 % correctly ($SD=15$ %), and low performers recalled 11 % correctly ($SD=13$ %). Independent sample t tests confirmed that high performers recalled significantly more than middle performers [$t(111)=4.52$, $p<.001$, $d=.86$] and low performers [$t(98)=5.85$, $p<.001$, $d=1.24$], and middle performers recalled significantly more than low performers [$t(95)=2.22$, $p=.03$, $d=.46$]. Furthermore, across all performance levels, the success rate of initial retrieval of the definitions was significantly correlated with later scores on the knowledge quiz questions, $r(155)=.38$, $p<.001$.

Thus, learning from retrieval appears to be linked with the amount of information that students initially retrieve. Students who recall more of the information—i.e., high performers—may benefit from retrieval because they gain additional exposure to this information through recall, and then gain further exposure through studying the answers during feedback. If students do not know enough to successfully retrieve the information—or if they are seeing it for the first time—retrieval may be ineffective because the information has not been effectively encoded, in which case learning would take place primarily by studying the answer sheet. Thus, an important contributor to the effectiveness of retrieval in classroom situations could be the degree to which a student has encoded and can successfully retrieve the information.

Discussion

In a large introductory biology course, classroom-based retrieval exercises were more effective for high performers than for middle or low performers. Whereas high performers learned biology definitions better when they had to retrieve the definitions, low performers learned the definitions better when they had to copy them. This finding is consistent with studies on the expertise reversal effect (e.g., Kalyuga 2007; Kalyuga et al. 2003; Lee and Kalyuga 2014), showing that high performers typically benefit from methods that require them to fill in or elaborate on information that is not currently present, whereas low performers benefit from additional processing of information that is provided for them (e.g., Cooper et al. 2001; McNamara et al. 1996).

In the current study, a likely possibility for this interaction is that high performers have a greater degree of baseline knowledge on the topic that permits successful retrieval on the initial test. Initial retrieval accuracy of the definitions was significantly correlated with later quiz scores. This is somewhat different from laboratory studies showing that initial retrieval success is not always necessary for benefits of retrieval to occur, as long as corrective feedback is provided. Some laboratory studies have shown, for example, that students learn better from failed retrieval attempts than from merely reading, as long as the correct answer is provided after the retrieval attempt (e.g., Kornell et al. 2009; Pashler et al. 2005). In the current study, students were always provided with the correct answers after trying to retrieve them, but they still benefited more from successful retrieval than from failed retrieval.

There may be important differences between laboratory- and classroom-based studies that influence the effectiveness of learning from feedback, however. Typical laboratory studies involve relatively simple stimuli that are learned under highly controlled conditions, with the goal of controlling factors such as the level of prior knowledge of the to-be-learned material (keeping it at minimal-to-none) and—unless it is directly manipulated—the learner's motivation. In an environment that minimizes distractions and encourages task engagement, failing to retrieve an item in a laboratory study (assuming feedback is provided) may not be accompanied by any harmful consequences to learning. Classrooms, on the other hand, are less controlled environments that contain students with a broad range of background knowledge, interests, and motivation. Compared to a laboratory study, failing to retrieve curriculum-based information in a classroom study could more likely reflect lapses in prior knowledge, interest, or motivation that could perpetuate suboptimal learning. It may be the case, therefore, that successful initial retrieval is more important to retrieval-enhanced learning in the classroom than in the laboratory.

Indeed, at least one recent study has shown that positive effects of retrieval practice do not always occur in classroom environments, even when feedback is provided. Karpicke et al. (2014, experiment 1) gave elementary-school children a series of retrieval exercises over science concepts and found that children only retrieved about 10 % correct initially. On a later test, retrieval was *not* more effective than alternative, nonretrieval-based activities that required children to interact with the material, and this was true even though children received feedback after their initial retrieval attempts. In a follow-up experiment in which the retrieval conditions were more likely to facilitate success (i.e., by reducing the amount of information to retrieve and improving its organizational structure), retrieval demonstrated its usual advantage over simply reading the material (Karpicke et al., experiment 3). Thus, like the current results and unlike results of some laboratory-based experiments, failed retrieval attempts followed by feedback may not always benefit learning in authentic educational environments, particularly when the level of initial retrieval is low.

Consistent with this finding, previous studies in classroom environments have shown significant overall advantages of retrieval vs rereading under conditions that increase the success of initial retrieval—for example, by administering a quiz immediately after a lesson (which resulted in 89 % correct on the quiz and a later advantage of retrieval (91 %) over rereading (83 %) in the study by Roediger et al. 2011), or by allowing students multiple retrieval opportunities with feedback (which resulted in over 95 % correct on the quizzes and a later advantage of retrieval (87 %) over rereading (75 %) in the study by McDaniel et al. 2012). These results highlight the importance of providing guidance or scaffolding to facilitate retrieval in classroom environments, and the current results suggest that this may be especially important for lower performing students.

Retrieval success per se may not be the only factor contributing to the current results. The degree of success while retrieving information during a class activity could reflect individual student differences in motivation, interest, cognitive abilities, propensity to learn from feedback, or other factors. High performers may embody these characteristics more so than middle or low performers, such that retrieval success may be only a partial contributor—or even a by-product—of other individual characteristics. Though the influence of such additional factors cannot be ascertained from the current study, one previous study showed that the frequency of students' reported use of retrieval practice as a study strategy was positively related to student achievement (Hartwig and Dunlosky 2012). An interesting possibility, therefore, is that higher performing students may be more familiar with the use of retrieval practice, increasing the likelihood that they use it effectively to learn course material.

Just as the role of a given construct (i.e., student achievement) may relate to retrieval-enhanced learning for a variety of reasons, retrieval-enhanced learning could reflect a number of additional constructs. A small number of studies has begun to explore these possibilities, so far reporting that individual differences in retrieval-enhanced learning do not appear to be linked with working memory (Brewer and Unsworth 2012), but one study reporting that these differences were accounted for by the interactive effects of working memory capacity and trait test anxiety (Tse and Pu 2012). Along similar lines, the positive effects of retrieval have been attenuated by the application of performance pressure at the time of retrieval (Hinze and Rapp 2014). These initial studies suggest that there may be important individual and situational factors underlying retrieval-based learning—particularly as they relate to performance, or one's *perception* of their own performance—that have much potential for further exploration.

In the current study, effects of retrieval were only observed for quiz questions that tested memory of the same definitions that appeared on the retrieval exercises (i.e., knowledge questions) and did not occur for never-before-seen questions that tested a higher degree of understanding (i.e., comprehension questions). This is consistent with the results of some recent studies showing that retrieving the answer to a particular question does not always facilitate retrieval of the answer to a related but never-before-seen question (e.g., Hinze and Wiley 2011; Wooldridge et al. 2014). The degree of retrieval-enhanced facilitation to new questions may depend, in part, on how the retrieved information relates to the never-before-seen question. In the study by Wooldridge et al., final test questions were drawn from the same textbook chapter, but may not have tapped the same concepts, as the information that was originally retrieved. In Hinze and Wiley's study (experiments 1–2), final test questions were drawn from the same paragraph of text as the retrieved information, but it is possible that students did not draw a connection between the retrieved information (e.g., the fact that daughter cells are created from parent cells in mitosis) and the nonretrieved information (e.g., the fact that daughter cells are genetically identical) that would facilitate later memory for the nonretrieved information. Studies demonstrating retrieval-induced transfer have often involved initial testing conditions in which the retrieved information bears a strong link with—and may even prompt explicit recall of—the nonretrieved information (e.g., Butler 2010; Carpenter and Kelly 2012; Chan et al. 2006). Indeed, in a third experiment, Hinze and Wiley found that free recall of entire paragraphs—which is more likely than short-answer questions to activate knowledge of the entire passage—resulted in better performance (compared to rereading the passage) on later, never-before-seen multiple-choice questions. In another recent study, Bjork et al. (2014) observed positive effects of retrieval on the concepts that served as incorrect lures on a multiple-choice test. To the extent that students processed each of the lures as potential answers to the question, they may have activated or retrieved information associated with each lure while answering the question, increasing the chances that a later question tapping knowledge of the lure would be answered correctly. Along similar lines, McDaniel et al. (2012) found that answering a quiz question in an undergraduate Brain and Behavior course (e.g., "Information coming INTO a structure (arriving) is called:" answer: "afferent") facilitated later performance on a conceptually related but nonidentical question (e.g., "Information leaving a nervous system structure is called:" answer: "efferent").

This provides some insight into why recalling answers to knowledge-based questions did not facilitate later performance on comprehension-based questions in the current study. Even though the knowledge and comprehension questions related to the same content, it seems unlikely that students would have needed to activate comprehension-based information in order to answer the knowledge-based questions. On the other hand, answering comprehension-based questions would seem to require knowledge-level representations; so, practice at retrieving comprehension-based information may be more likely to transfer to knowledge-based information than the other way around (e.g., see McDaniel et al. 2013). Indeed, one recent study reported that high-level quiz questions (those requiring application, evaluation, and analysis of information) were more effective than low-level questions (those requiring mere recall of information) at promoting later exam performance on both high-level *and* low-level questions (Jensen et al. 2014). Similarly, Hinze et al. (2013) found that students who read a science passage while expecting a future test containing higher-order inference-based questions scored higher than students who read the same passage while expecting a future test containing detail-based questions that were explicitly stated in the passage. Furthermore, students who expected the inference-based test performed better on inference-based questions *and* detail-based questions, relative to students who expected the detail-based test.

Thus, the degree of transfer resulting from retrieval may depend on how students approach, and engage with, the retrieval task. In the absence of conditions that promote connectivity between the retrieved information and later transfer questions (as in the current study), retrieval-enhanced learning may be relatively specific to the information that was practiced. Factors that promote this connectivity, however—through using higher-order questions during practice (Jensen et al. 2014), test expectancy instructions (Hinze et al. 2013, experiment 2), or the construction of explanations during retrieval of complex text materials (Hinze et al. 2013, experiment 3)—may increase the flexibility of retrieval-enhanced learning. Though much of the literature on retrieval practice has focused on measuring direct retention of relatively specific types of knowledge, a timely and worthwhile goal for future studies is to develop and apply retrieval-based methods for promoting the types of higher-order comprehension and application skills that align with educational goals (e.g., Carpenter 2012; Pellegrino 2012).

Finally, we found that metacognitive calibration was better for high performers than for middle or low performers. This is consistent with prior studies conducted in university classrooms showing that low-performing students tend to overpredict exam scores more than high-performing students (e.g., Bol et al. 2005; Miller and Geraci 2011). The relationship between achievement and metacognitive awareness exists even when students do not make specific performance predictions. For example, Schraw and Dennison (1994) administered a 52-item survey measuring everyday metacognitive monitoring behavior among students (e.g., "I ask myself periodically if I am meeting my goals" and "I find myself pausing regularly to check my comprehension") and found that greater monitoring was associated with higher scores on a subsequent test of reading comprehension. Thus, students who show higher academic achievement also tend to show a higher degree of metacognitive monitoring.

Consistent with previous research, we also found that students displayed improved metacognitive calibration following retrieval (Agarwal et al. 2008; Little and McDaniel 2014; Tullis et al. 2013); however, this was only apparent for high performers and not for low performers. This could have been driven, in part, by the tendency for the conditions involving copying to increase the perceived ease of processing of the material, which has been known to inflate judgments of learning (e.g., Benjamin et al. 1998; Carpenter and Olson 2012; Diemand-Yauman et al. 2011; Rhodes and Castel 2008; Serra and Dunlosky 2010). The fact that this occurred for high performers suggests that high achievement may not inoculate students from using a (sometimes faulty) ease-of-processing heuristic while making judgments of learning.

Such findings are consistent as well with a recent study by Szpunar et al. (2014), who showed that students' learning of statistics concepts from a video-taped lecture increased, and metacognitive calibration improved, by inserting quizzes periodically throughout the lecture. In addition, at least one classroom-based study has documented an improvement in students' predictions of their own exam scores over the course of a semester (with overconfidence initially very high but then decreasing with each subsequent exam), and that high performers were more likely than low performers to show this improvement (Hacker et al. 2000). Thus, whereas high performers appear to get better at aligning their predictions with performance as a result of practice, middle and low performers may be in need of additional metacognitive training to improve calibration.

In conclusion, the current classroom-based study highlights the important role that individual student achievement can play in the effectiveness of retrieval-based learning. High performers were more likely than middle or low performers to benefit from retrieval and were more likely to accurately estimate their own performance on a later quiz. Given the wide range of student achievement that is present in many classrooms, these results encourage researchers to consider individual differences in student achievement when evaluating the effectiveness of educational interventions.

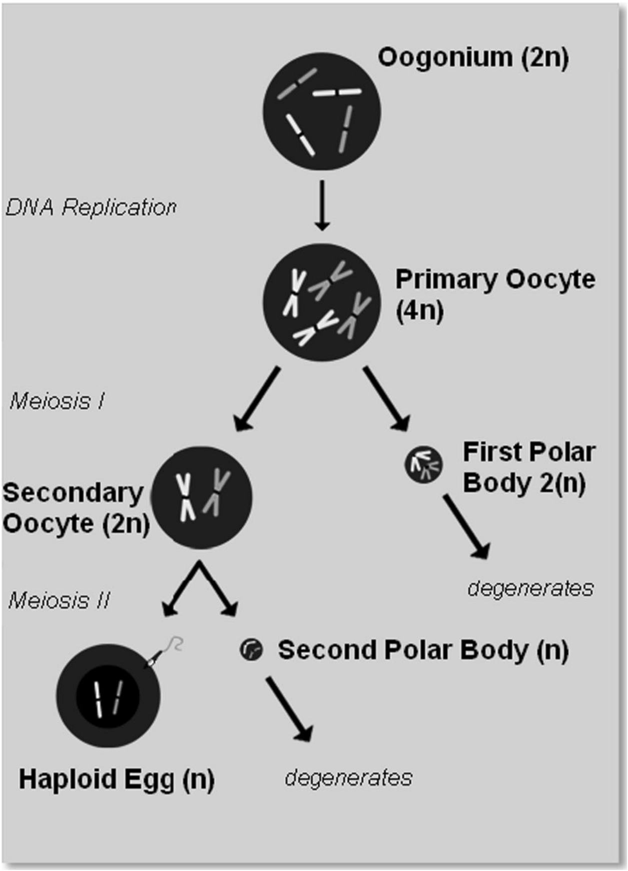
Author Note Shana K. Carpenter, Terry J. S. Lund, and Patrick I. Armstrong, Department of Psychology, Iowa State University; Clark R. Coffman, Department of Genetics, Development, and Cell Biology, Iowa State University; Monica H. Lamm, Department of Chemical and Biological Engineering, Iowa State University; Robert D. Reason, School of Education, Iowa State University.

Portions of this study were presented at the annual meeting of the Society for the Advancement of Biology Education Research (Minneapolis, MN, July 2014). We thank Tyler McGrath for creating the images for the in-class exercises, and Chad Fernandez, Shuhebur Rahman, Andrew Woods, and Kristi Harris for their assistance with scoring the quizzes and in-class exercises.

Appendix A

Materials used in the current study.

1. **Polar body**: a cell produced by asymmetric cell divisions during meiosis.
2. **Oogonium**: the germ cell progenitor found in females.
3. **Primary oocyte**: an oogonium that has undergone some of the early stages of meiosis and replicated its DNA.
4. **Secondary oocyte**: an oocyte that has undergone meiosis I and is now 2n.
5. **Haploid egg**: the mature female reproductive cell that contains one copy of the genome.



Appendix B

Quiz questions used in the current study (with correct answers bolded). Questions 1–5 were designed to measure knowledge of the term definitions, and questions 6–10 were designed to measure comprehension.

1. Which of the following has undergone some of the early stages of meiosis and replicated its DNA?
   A. haploid egg
   B. **primary oocyte**
   C. polar body
   D. oogonium
   E. I do not know

2. Which describes a cell produced by asymmetric cell divisions during meiosis?
   A. **polar body**
   B. haploid egg
   C. oocyte
   D. Barr body
   E. I do not know

3. What is the germ cell progenitor found in females?
   A. oocyte
   B. gamete
   C. embryo
   D. **oogonium**
   E. I do not know

4. What is the female reproductive cell that contains one copy of the genome?
   A. oogonium
   B. polar body
   C. **haploid egg**
   D. zygote
   E. I do not know

5. Which of the following has completed meiosis I and is now 2n?
   A. primary oocyte
   B. **secondary oocyte**
   C. tertiary oocyte
   D. quaternary oocyte
   E. I do not know

6. A medication is developed that causes the oocyte to perpetually undergo uneven division. What is the result?
   A. many polar bodies and a fertile person
   B. **many polar bodies and an infertile person**
   C. few polar bodies and a fertile person

D. few polar bodies and an infertile person
E. I do not know

7. A patient suffers from a disease which arrests the development of the secondary oocyte. Which of the following CANNOT occur?
A. The first part of meiosis I
B. **The completion of meiosis I**
C. DNA replication
D. Formation of polar bodies
E. I do not know

8. A patient suffers from a condition that prevents her germ cells' DNA from replicating. What effect does this have?
A. Polar bodies will degenerate
B. Meiosis I, but not meiosis II will occur
C. **A secondary oocyte will not form**
D. A 2n egg will form
E. I do not know

9. Medication has been developed to prevent the early stages of meiosis. Which of the following stages is arrested?
A. Haploid egg
B. Secondary oocyte
C. **Primary oocyte**
D. Secondary polar body
E. I do not know

10. If a patient's germ cell development stops after meiosis II, is she fertile?
A. No, the DNA has not replicated
B. Yes, the polar body has degenerated
C. No, the haploid egg has not formed
D. **Yes, this is normal development**
E. I do not know

References

Agarwal, P. K., Karpicke, J. D., Kang, S. H. K., Roediger, H. L. III, & McDermott, K. B. (2008). Examining the testing effect with open- and closed-book tests. *Applied Cognitive Psychology, 22*, 861–876.

Alexander, P. A., Kulikowich, J. M., & Schulze, S. K. (1994). How subject-matter knowledge affects recall and interest. *American Educational Research Journal, 31*, 313–337.

Anderson, L. W., Krathwohl, D. R., Airasian, P. W., Cruikshank, K. A., Mayer, R. E., Pintrich, P. R., Raths, J., & Wittrock, M. C. (2001). *A taxonomy for learning, teaching, and assessing: a revision of bloom's taxonomy of educational objectives*. New York: Longman.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1998). The mismeasure of memory: when retrieval fluency is misleading as a metacognitive index. *Journal of Experimental Psychology: General, 127*, 55–68.

Bjork, E. L., Little, J. L., & Storm, B. C. (2014). Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory & Cognition, 3*, 1–6.

Bol, L., Hacker, D. J., O'Shea, P., & Allen, D. (2005). The influence of overt practice, achievement level, and explanatory style on calibration accuracy and performance. *The Journal of Experimental Education, 73*, 269–290.

Boscolo, P., & Mason, L. (2003). Topic knowledge, text coherence, and interest: how they interact in learning from instructional texts. *Journal of Experimental Education, 71*, 126–148.

Brewer, G. A., & Unsworth, N. (2012). Individual differences in the effects of retrieval from long-term memory. *Journal of Memory & Language, 66*, 407–415.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 36*, 1118–1133.

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

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35*, 1563–1569.

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*, 1547–1552.

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21*, 279–283.

Carpenter, S. K., & Kelly, J. W. (2012). Tests enhance retention and transfer of spatial learning. *Psychonomic Bulletin & Review, 19*, 443–448.

Carpenter, S. K., & Olson, K. M. (2012). Are pictures good for learning new vocabulary in a foreign language? Only if you think they are not. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 38*, 92–101.

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*, 438–448.

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U. S. history facts. *Applied Cognitive Psychology, 23*, 760–771.

Carpenter, S. K., Wilford, M., Kornell, N., & Mullaney, K. M. (2013). Appearances can be deceiving: instructor fluency increases perceptions of learning without increasing actual learning. *Psychonomic Bulletin & Review, 20*, 1350–1356.

Castel, A. D., McCabe, D. P., & Roediger, H. L., III. (2007). Illusions of competence and overestimation of associative memory for identical items: evidence from judgments of learning. *Psychonomic Bulletin & Review, 14*, 107–111.

Chan, J. C. K., McDermott, K. B., & Roediger, H. L., III. (2006). Retrieval-induced facilitation: initially non-tested material can benefit from prior testing of related material. *Journal of Experimental Psychology: General, 135*, 553–571.

Clark, D.A., & Svinicki, M. (2014). The effect of retrieval on post-task enjoyment of studying. *Educational Psychology Review, 27,* 51-67.

Cooper, G., Tindall-Ford, S., Chandler, P., & Sweller, J. (2001). Learning by imagining. *Journal of Experimental Psychology: Applied, 7*, 68–82.

Coppens, L. C., Verkoeijen, P. P. J. L., & Rikers, R. M. J. P. (2011). Learning Adinkra symbols: the effect of testing. *Journal of Cognitive Psychology, 3*, 351–357.

Diemand-Yauman, C., Oppenheimer, D. M., & Vaughan, E. B. (2011). Fortune favors the bold (and the italicized): effects of disfluency on educational outcomes. *Cognition, 118*, 111–115.

Dunlosky, J., & Lipko, A. R. (2009). Metacomprehension: a brief history and how to improve its accuracy. *Current Directions in Psychological Science, 16*, 228–232.

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14*, 4–58.

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*, 238–244.

Finn, B., & Roediger, H. L., III. (2011). Enhancing retention through reconsolidation: negative emotional arousal following retrieval enhances later recall. *Psychological Science, 22*, 781–786.

Hacker, D. J., Bol, L., Horgan, D. D., & Rakow, E. A. (2000). Test prediction and performance in a classroom context. *Journal of Educational Psychology, 92*, 160–170.

Halamish, V., & Bjork, R. A. (2011). When does testing enhance retention? A distribution-based interpretation of retrieval as a memory modifier. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 37*, 801–812.

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

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology, 28*, 597–606.

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

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

Jensen, J. L., McDaniel, M. A., Woodard, S. M., & Kummer, T. A. (2014). Teaching to the test… or testing to teach? Exams requiring higher order thinking skills encourage greater conceptual understanding. *Educational Psychology Review, 26*, 307–329.

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*, 509–539.

Kalyuga, S., Ayres, P., Chandler, P., & Sweller, J. (2003). The expertise reversal effect. *Educational Psychologist, 38*, 23–31.

Kalyuga, S., Law, Y. K., & Lee, C. H. (2013). Expertise reversal effect in reading Chinese texts with added causal words. *Instructional Science, 41*, 481–497.

Kang, S. H. K., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation? *Journal of Applied Research in Memory & Cognition, 3*, 7–12.

Kang, S. H. K., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *European Journal of Cognitive Psychology, 19*, 528–558.

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science, 319*, 966–968.

Karpicke, J. D., & Zaromb, F. M. (2010). Retrieval mode distinguishes the testing effect from the generation effect. *Journal of Memory & Language, 62*, 227–239.

Karpicke, J.D., Blunt, J.R., Smith, M.A., & Karpicke, S.S. (2014). Retrieval-based learning: the need for guided retrieval in elementary-school children. *Journal of Applied Research in Memory & Cognition, 3,* 198-206.

Kornell, N., & Bjork, R. A. (2009). A stability bias in human memory: overestimating remembering and underestimating learning. *Journal of Experimental Psychology: General, 138*, 449–468.

Kornell, N., Hays, M. J., & Bjork, R. A. (2009). Unsuccessful retrieval attempts enhance subsequent learning. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35*, 989–998.

Kornell, N., Bjork, R. A., & Garcia, M. A. (2011). Why tests appear to prevent forgetting: a distribution-based bifurcation model. *Journal of Memory & Language, 65*, 85–97.

Kubik, V., Söderlund, H., Nilsson, L-G., & Jönsson, F.U. (2014). Individual and combined effects of enactment and testing on memory for action phrases. *Experimental Psychology, 61,* 347-355.

Kuo, T.-M., & Hirshman, E. (1997). The role of distinctive perceptual information in memory: studies of the testing effect. *Journal of Memory & Language, 36*, 188–201.

Lee, C. H., & Kalyuga, S. (2014). Expertise reversal effect and its instructional implications. In V. A. Benassi, C. E. Overson, & C. M. Hakala (Eds.), *Applying the science of learning in education: Infusing psychological science into the curriculum* (pp. 31–44). Retrieved from the Society for the Teaching of Psychology website: http://teachpsych.org/ebooks/asle2014/index.php.

Lee, H., Plass, J. L., & Homer, B. D. (2006). Optimizing cognitive load for learning from computer-based science simulations. *Journal of Educational Psychology, 98*, 902–913.

Leppink, J., Broers, N. J., Imbos, T., van der Vleuten, C. P. M., & Berger, M. P. F. (2012). Self-explanation in the domain of statistics: an expertise reversal effect. *Higher Education, 63*, 771–785.

Little, J. L., & McDaniel, M. A. (2014). Metamemory monitoring and control following retrieval practice for text. *Memory & Cognition, 43*, 85–98.

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*, 494–513.

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative assessment performance in a web-based class: an experimental study. *Journal of Applied Research in Memory & Cognition, 1*, 18–26.

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L., III. (2013). Quizzing in middle-school science: successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27*, 360–372.

McNamara, D. S. (2001). Reading both high-coherence and low-coherence texts: effects of text sequence and prior knowledge. *Canadian Journal of Experimental Psychology, 55*, 51–62.

McNamara, D. S., Kintsch, E., Songer, N. B., & Kintsch, W. (1996). Are good texts always better? Interactions of text coherence, background knowledge, and levels of understanding in learning from text. *Cognition & Instruction, 14*, 1–43.

Miller, T. M., & Geraci, L. (2011). Unskilled but aware: reinterpreting overconfidence in low-performing students. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 37*, 502–506.

Pashler, H., Cepeda, N. J., Wixted, J. T., & Rohrer, D. (2005). When does feedback facilitate learning of words ? *Journal of Experimental Psychology: Learning, Memory, & Cognition, 31*, 3–8.

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M. A., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning* (NCER 2007–2004). Washington, DC: U. S. Department of Education, National Center for Education Research, Institute of Education Sciences. Retrieved from http://ies.ed.gov/ncee/wwc/pdf/practiceguides/20072004.pdf.

Pellegrino, J. W. (2012). From cognitive principles to instructional practices: the devil is often in the details. *Journal of Applied Research in Memory & Cognition, 1*, 260–262.

Peterson, D. J., & Mulligan, N. W. (2013). The negative testing effect and multifactor account. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 39*, 1287–1293.

Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: mediator effectiveness hypothesis. *Science, 330*, 335.

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: how much is enough? *Journal of Experimental Psychology: General, 140*, 283–302.

Rhodes, M. G., & Castel, A. D. (2008). Memory predictions are influenced by perceptual information: evidence for metacognitive illusions. *Journal of Experimental Psychology: General, 137*, 615–625.

Roediger, H. L. III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15*, 20–27.

Roediger, H. L. III., & Karpicke, J. D. (2006). Test-enhanced learning: taking memory tests improves long-term retention. *Psychological Science, 17*, 249–255.

Roediger, H. L., III, & Pyc, M. A. (2012). Inexpensive techniques to improve education: applying cognitive psychology to enhance educational practice. *Journal of Applied Research in Memory & Cognition, 1*, 242–248.

Roediger, H. L. III, Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*, 382–395.

Schraw, G., & Dennison, R. S. (1994). Assessing metacognitive awareness. *Contemporary Educational Psychology, 19*, 460–475.

Serra, M. J., & Dunlosky, J. (2010). Metacomprehension judgments reflect the belief that diagrams improve learning from text. *Memory, 18*, 689–711.

Szpunar, K. K., Jing, H. G., & Schacter, D. L. (2014). Overcoming overconfidence in learning from video-recorded lectures: implications of interpolated testing for online education. *Journal of Applied Research in Memory & Cognition, 3,* 161-164.

Toppino, T. C., & Cohen, M. S. (2009). The testing effect and the retention interval: questions and answers. *Experimental Psychology, 56*, 252–257.

Tse, C.-S., & Pu, X. (2012). The effectiveness of test-enhanced learning depends on trait test anxiety and working-memory capacity. *Journal of Experimental Psychology: Applied, 18*, 253–264.

Tullis, J. G., Finley, J. R., & Benjamin, A. S. (2013). Metacognition of the testing effect: guiding learners to predict the benefits of retrieval. *Memory & Cognition, 41*, 429–442.

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion level effects on memory: what aspects of memory are enhanced by repeated retrieval? *Psychological Science, 22*, 1127–1131.

Vaughn, K. E., Rawson, K. A., & Pyc, M. A. (2013). Repeated retrieval practice and item difficulty: does criterion learning eliminate item difficulty effects? *Psychonomic Bulletin & Review, 20*, 1239–1245.

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: a cautionary note. *Journal of Applied Research in Memory & Cognition, 3*, 13–20.

Zaromb, F. M., & Roediger, H. L., III. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition, 38*, 995–1008.