aps ASSOCIATION FOR PSYCHOLOGICAL SCIENCE

Psychological Science in the Public Interest
14(1) 4–58
© The Author(s) 2013
Reprints and permission:
sagepub.com/journalsPermissions.nav
DOI: 10.1177/1529100612453266
http://pspi.sagepub.com
SAGE

Improving Students' Learning With Effective Learning Techniques: Promising Directions From Cognitive and Educational Psychology

John Dunlosky[1], Katherine A. Rawson[1], Elizabeth J. Marsh[2], Mitchell J. Nathan[3], and Daniel T. Willingham[4]

[1]Department of Psychology, Kent State University; [2]Department of Psychology and Neuroscience, Duke University; [3]Department of Educational Psychology, Department of Curriculum & Instruction, and Department of Psychology, University of Wisconsin–Madison; and [4]Department of Psychology, University of Virginia

Summary

Many students are being left behind by an educational system that some people believe is in crisis. Improving educational outcomes will require efforts on many fronts, but a central premise of this monograph is that one part of a solution involves helping students to better regulate their learning through the use of effective learning techniques. Fortunately, cognitive and educational psychologists have been developing and evaluating easy-to-use learning techniques that could help students achieve their learning goals. In this monograph, we discuss 10 learning techniques in detail and offer recommendations about their relative utility. We selected techniques that were expected to be relatively easy to use and hence could be adopted by many students. Also, some techniques (e.g., highlighting and rereading) were selected because students report relying heavily on them, which makes it especially important to examine how well they work. The techniques include elaborative interrogation, self-explanation, summarization, highlighting (or underlining), the keyword mnemonic, imagery use for text learning, rereading, practice testing, distributed practice, and interleaved practice.

To offer recommendations about the relative utility of these techniques, we evaluated whether their benefits generalize across four categories of variables: learning conditions, student characteristics, materials, and criterion tasks. Learning conditions include aspects of the learning environment in which the technique is implemented, such as whether a student studies alone or with a group. Student characteristics include variables such as age, ability, and level of prior knowledge. Materials vary from simple concepts to mathematical problems to complicated science texts. Criterion tasks include different outcome measures that are relevant to student achievement, such as those tapping memory, problem solving, and comprehension.

We attempted to provide thorough reviews for each technique, so this monograph is rather lengthy. However, we also wrote the monograph in a modular fashion, so it is easy to use. In particular, each review is divided into the following sections:

1. General description of the technique and why it should work
2. How general are the effects of this technique?
   2a. Learning conditions
   2b. Student characteristics
   2c. Materials
   2d. Criterion tasks
3. Effects in representative educational contexts
4. Issues for implementation
5. Overall assessment

Corresponding Author:
John Dunlosky, Psychology, Kent State University, Kent, OH 44242
E-mail: jdunlosk@kent.edu

The review for each technique can be read independently of the others, and particular variables of interest can be easily compared across techniques.

To foreshadow our final recommendations, the techniques vary widely with respect to their generalizability and promise for improving student learning. Practice testing and distributed practice received high utility assessments because they benefit learners of different ages and abilities and have been shown to boost students' performance across many criterion tasks and even in educational contexts. Elaborative interrogation, self-explanation, and interleaved practice received moderate utility assessments. The benefits of these techniques do generalize across some variables, yet despite their promise, they fell short of a high utility assessment because the evidence for their efficacy is limited. For instance, elaborative interrogation and self-explanation have not been adequately evaluated in educational contexts, and the benefits of interleaving have just begun to be systematically explored, so the ultimate effectiveness of these techniques is currently unknown. Nevertheless, the techniques that received moderate-utility ratings show enough promise for us to recommend their use in appropriate situations, which we describe in detail within the review of each technique.

Five techniques received a low utility assessment: summarization, highlighting, the keyword mnemonic, imagery use for text learning, and rereading. These techniques were rated as low utility for numerous reasons. Summarization and imagery use for text learning have been shown to help some students on some criterion tasks, yet the conditions under which these techniques produce benefits are limited, and much research is still needed to fully explore their overall effectiveness. The keyword mnemonic is difficult to implement in some contexts, and it appears to benefit students for a limited number of materials and for short retention intervals. Most students report rereading and highlighting, yet these techniques do not consistently boost students' performance, so other techniques should be used in their place (e.g., practice testing instead of rereading).

Our hope is that this monograph will foster improvements in student learning, not only by showcasing which learning techniques are likely to have the most generalizable effects but also by encouraging researchers to continue investigating the most promising techniques. Accordingly, in our closing remarks, we discuss some issues for how these techniques could be implemented by teachers and students, and we highlight directions for future research.

Introduction

If simple techniques were available that teachers and students could use to improve student learning and achievement, would you be surprised if teachers were not being told about these techniques and if many students were not using them? What if students were instead adopting ineffective learning techniques that undermined their achievement, or at least did not improve it? Shouldn't they stop using these techniques and begin using ones that are effective? Psychologists have been developing and evaluating the efficacy of techniques for study and instruction for more than 100 years. Nevertheless, some effective techniques are underutilized—many teachers do not learn about them, and hence many students do not use them, despite evidence suggesting that the techniques could benefit student achievement with little added effort. Also, some learning techniques that are popular and often used by students are relatively ineffective. One potential reason for the disconnect between research on the efficacy of learning techniques and their use in educational practice is that because so many techniques are available, it would be challenging for educators to sift through the relevant research to decide which ones show promise of efficacy and could feasibly be implemented by students (Pressley, Goodchild, Fleet, Zajchowski, & Evans, 1989).

Toward meeting this challenge, we explored the efficacy of 10 learning techniques (listed in Table 1) that students could use to improve their success across a wide variety of content domains.[1] The learning techniques we consider here were chosen on the basis of the following criteria. We chose some techniques (e.g., self-testing, distributed practice) because an initial survey of the literature indicated that they could improve student success across a wide range of conditions. Other techniques (e.g., rereading and highlighting) were included because students report using them frequently. Moreover, students are responsible for regulating an increasing amount of their learning as they progress from elementary grades through middle school and high school to college. Lifelong learners also need to continue regulating their own learning, whether it takes place in the context of postgraduate education, the workplace, the development of new hobbies, or recreational activities.

Thus, we limited our choices to techniques that could be implemented by students without assistance (e.g., without requiring advanced technologies or extensive materials that would have to be prepared by a teacher). Some training may be required for students to learn how to use a technique with fidelity, but in principle, students should be able to use the techniques without supervision. We also chose techniques for which a sufficient amount of empirical evidence was available to support at least a preliminary assessment of potential efficacy. Of course, we could not review all the techniques that meet these criteria, given the in-depth nature of our reviews, and these criteria excluded some techniques that show much promise, such as techniques that are driven by advanced technologies.

Because teachers are most likely to learn about these techniques in educational psychology classes, we examined how some educational-psychology textbooks covered them (Ormrod, 2008; Santrock, 2008; Slavin, 2009; Snowman,

Table 1. Learning Techniques

Technique	Description
1. Elaborative interrogation	Generating an explanation for why an explicitly stated fact or concept is true
2. Self-explanation	Explaining how new information is related to known information, or explaining steps taken during problem solving
3. Summarization	Writing summaries (of various lengths) of to-be-learned texts
4. Highlighting/underlining	Marking potentially important portions of to-be-learned materials while reading
5. Keyword mnemonic	Using keywords and mental imagery to associate verbal materials
6. Imagery for text	Attempting to form mental images of text materials while reading or listening
7. Rereading	Restudying text material again after an initial reading
8. Practice testing	Self-testing or taking practice tests over to-be-learned material
9. Distributed practice	Implementing a schedule of practice that spreads out study activities over time
10. Interleaved practice	Implementing a schedule of practice that mixes different kinds of problems, or a schedule of study that mixes different kinds of material, within a single study session

Note. See text for a detailed description of each learning technique and relevant examples of their use.

Table 2. Examples of the Four Categories of Variables for Generalizability

Materials	Learning conditions	Student characteristics[a]	Criterion tasks
Vocabulary	Amount of practice (dosage)	Age	Cued recall
Translation equivalents	Open- vs. closed-book practice	Prior domain knowledge	Free recall
Lecture content	Reading vs. listening	Working memory capacity	Recognition
Science definitions	Incidental vs. intentional learning	Verbal ability	Problem solving
Narrative texts	Direct instruction	Interests	Argument development
Expository texts	Discovery learning	Fluid intelligence	Essay writing
Mathematical concepts	Rereading lags[b]	Motivation	Creation of portfolios
Maps	Kind of practice tests[c]	Prior achievement	Achievement tests
Diagrams	Group vs. individual learning	Self-efficacy	Classroom quizzes

[a]Some of these characteristics are more state based (e.g., motivation) and some are more trait based (e.g., fluid intelligence); this distinction is relevant to the malleability of each characteristic, but a discussion of this dimension is beyond the scope of this article.
[b]Learning condition is specific to rereading.
[c]Learning condition is specific to practice testing.

McCown, & Biehler, 2009; Sternberg & Williams, 2010; Woolfolk, 2007). Despite the promise of some of the techniques, many of these textbooks did not provide sufficient coverage, which would include up-to-date reviews of their efficacy and analyses of their generalizability and potential limitations. Accordingly, for all of the learning techniques listed in Table 1, we reviewed the literature to identify the generalizability of their benefits across four categories of variables—materials, learning conditions, student characteristics, and criterion tasks. The choice of these categories was inspired by Jenkins' (1979) model (for an example of its use in educational contexts, see Marsh & Butler, in press), and examples of each category are presented in Table 2. *Materials* pertain to the specific content that students are expected to learn, remember, or comprehend. *Learning conditions* pertain to aspects of the context in which students are interacting with the to-be-learned materials. These conditions include aspects of the learning environment itself (e.g., noisiness vs. quietness in a classroom), but they largely pertain to the way in which a learning technique is implemented. For instance, a technique could be used only once or many times (a variable referred to as *dosage*) when students are studying, or a technique could be used when students are either reading or listening to the to-be-learned materials.

Any number of *student characteristics* could also influence the effectiveness of a given learning technique. For example, in comparison to more advanced students, younger students in early grades may not benefit from a technique. Students' basic cognitive abilities, such as working memory capacity or general fluid intelligence, may also influence the efficacy of a given technique. In an educational context, domain knowledge refers to the valid, relevant knowledge a student brings to a lesson. Domain knowledge may be required for students to use some of the learning techniques listed in Table 1. For instance,

the use of imagery while reading texts requires that students know the objects and ideas that the words refer to so that they can produce internal images of them. Students with some domain knowledge about a topic may also find it easier to use self-explanation and elaborative interrogation, which are two techniques that involve answering "why" questions about a particular concept (e.g., "Why would particles of ice rise up within a cloud?"). Domain knowledge may enhance the benefits of summarization and highlighting as well. Nevertheless, although some domain knowledge will benefit students as they begin learning new content within a given domain, it is not a prerequisite for using most of the learning techniques.

The degree to which the efficacy of each learning technique obtains across long retention intervals and generalizes across different *criterion tasks* is of critical importance. Our reviews and recommendations are based on evidence, which typically pertains to students' objective performance on any number of criterion tasks. Criterion tasks (Table 2, rightmost column) vary with respect to the specific kinds of knowledge that they tap. Some tasks are meant to tap students' memory for information (e.g., "What is operant conditioning?"), others are largely meant to tap students' comprehension (e.g., "Explain the difference between classical conditioning and operant conditioning"), and still others are meant to tap students' application of knowledge (e.g., "How would you apply operant conditioning to train a dog to sit down?"). Indeed, Bloom and colleagues divided learning objectives into six categories, from memory (or knowledge) and comprehension of facts to their application, analysis, synthesis, and evaluation (B. S. Bloom, Engelhart, Furst, Hill, & Krathwohl, 1956; for an updated taxonomy, see L. W. Anderson & Krathwohl, 2001).

In discussing how the techniques influence criterion performance, we emphasize investigations that have gone beyond demonstrating improved memory for target material by measuring students' comprehension, application, and transfer of knowledge. Note, however, that although gaining factual knowledge is not considered the only or ultimate objective of schooling, we unabashedly consider efforts to improve student retention of knowledge as essential for reaching other instructional objectives; if one does not remember core ideas, facts, or concepts, applying them may prove difficult, if not impossible. Students who have forgotten principles of algebra will be unable to apply them to solve problems or use them as a foundation for learning calculus (or physics, economics, or other related domains), and students who do not remember what operant conditioning is will likely have difficulties applying it to solve behavioral problems. We are not advocating that students spend their time robotically memorizing facts; instead, we are acknowledging the important interplay between memory for a concept on one hand and the ability to comprehend and apply it on the other.

An aim of this monograph is to encourage students to use the appropriate learning technique (or techniques) to accomplish a given instructional objective. Some learning techniques are largely focused on bolstering students' memory for facts (e.g., the keyword mnemonic), others are focused more on improving comprehension (e.g., self-explanation), and yet others may enhance both memory and comprehension (e.g., practice testing). Thus, our review of each learning technique describes how it can be used, its effectiveness for producing long-term retention and comprehension, and its breadth of efficacy across the categories of variables listed in Table 2.

Reviewing the Learning Techniques

In the following series of reviews, we consider the available evidence for the efficacy of each of the learning techniques. Each review begins with a brief description of the technique and a discussion about why it is expected to improve student learning. We then consider generalizability (with respect to learning conditions, materials, student characteristics, and criterion tasks), highlight any research on the technique that has been conducted in representative educational contexts, and address any identified issues for implementing the technique. Accordingly, the reviews are largely modular: Each of the 10 reviews is organized around these themes (with corresponding headers) so readers can easily identify the most relevant information without necessarily having to read the monograph in its entirety.

At the end of each review, we provide an overall assessment for each technique in terms of its relatively utility—low, moderate, or high. Students and teachers who are not already doing so should consider using techniques designated as *high utility*, because the effects of these techniques are robust and generalize widely. Techniques could have been designated as *low utility* or *moderate utility* for any number of reasons. For instance, a technique could have been designated as low utility because its effects are limited to a small subset of materials that students need to learn; the technique may be useful in some cases and adopted in appropriate contexts, but, relative to the other techniques, it would be considered low in utility because of its limited generalizability. A technique could also receive a low- or moderate-utility rating if it showed promise, yet insufficient evidence was available to support confidence in assigning a higher utility assessment. In such cases, we encourage researchers to further explore these techniques within educational settings, but students and teachers may want to use caution before adopting them widely. Most important, given that each utility assessment could have been assigned for a variety of reasons, we discuss the rationale for a given assessment at the end of each review.

Finally, our intent was to conduct exhaustive reviews of the literature on each learning technique. For techniques that have been reviewed extensively (e.g., distributed practice), however, we relied on previous reviews and supplemented them with any research that appeared after they had been published. For many of the learning techniques, too many articles have been published to cite them all; therefore, in our discussion of most of the techniques, we cite a subset of relevant articles.

1 Elaborative interrogation

Anyone who has spent time around young children knows that one of their most frequent utterances is "Why?" (perhaps coming in a close second behind "No!"). Humans are inquisitive creatures by nature, attuned to seeking explanations for states, actions, and events in the world around us. Fortunately, a sizable body of evidence suggests that the power of explanatory questioning can be harnessed to promote learning. Specifically, research on both elaborative interrogation and self-explanation has shown that prompting students to answer "Why?" questions can facilitate learning. These two literatures are highly related but have mostly developed independently of one another. Additionally, they have overlapping but nonidentical strengths and weaknesses. For these reasons, we consider the two literatures separately.

1.1 General description of elaborative interrogation and why it should work. In one of the earliest systematic studies of elaborative interrogation, Pressley, McDaniel, Turnure, Wood, and Ahmad (1987) presented undergraduate students with a list of sentences, each describing the action of a particular man (e.g., "The hungry man got into the car"). In the elaborative-interrogation group, for each sentence, participants were prompted to explain "Why did that particular man do that?" Another group of participants was instead provided with an explanation for each sentence (e.g., "The hungry man got into the car to go to the restaurant"), and a third group simply read each sentence. On a final test in which participants were cued to recall which man performed each action (e.g., "Who got in the car?"), the elaborative-interrogation group substantially outperformed the other two groups (collapsing across experiments, accuracy in this group was approximately 72%, compared with approximately 37% in each of the other two groups). From this and similar studies, Seifert (1993) reported average effect sizes ranging from 0.85 to 2.57.

As illustrated above, the key to elaborative interrogation involves prompting learners to generate an explanation for an explicitly stated fact. The particular form of the explanatory prompt has differed somewhat across studies—examples include "Why does it make sense that…?", "Why is this true?", and simply "Why?" However, the majority of studies have used prompts following the general format, "Why would this fact be true of this [X] and not some other [X]?"

The prevailing theoretical account of elaborative-interrogation effects is that elaborative interrogation enhances learning by supporting the integration of new information with existing prior knowledge. During elaborative interrogation, learners presumably "activate schemata . . . These schemata, in turn, help to organize new information which facilitates retrieval" (Willoughby & Wood, 1994, p. 140). Although the integration of new facts with prior knowledge may facilitate the organization (Hunt, 2006) of that information, organization alone is not sufficient—students must also be able to discriminate among related facts to be accurate when identifying or using the learned information (Hunt, 2006). Consistent with this account, note that most elaborative-interrogation prompts explicitly or implicitly invite processing of both similarities and differences between related entities (e.g., why a fact would be true of one province versus other provinces). As we highlight below, processing of similarities and differences among to-be-learned facts also accounts for findings that elaborative-interrogation effects are often larger when elaborations are precise rather than imprecise, when prior knowledge is higher rather than lower (consistent with research showing that preexisting knowledge enhances memory by facilitating distinctive processing; e.g., Rawson & Van Overschelde, 2008), and when elaborations are self-generated rather than provided (a finding consistent with research showing that distinctiveness effects depend on self-generating item-specific cues; Hunt & Smith, 1996).

1.2 How general are the effects of elaborative interrogation?

1.2a Learning conditions. The seminal work by Pressley et al. (1987; see also B. S. Stein & Bransford, 1979) spawned a flurry of research in the following decade that was primarily directed at assessing the generalizability of elaborative-interrogation effects. Some of this work focused on investigating elaborative-interrogation effects under various learning conditions. Elaborative-interrogation effects have been consistently shown using either incidental or intentional learning instructions (although two studies have suggested stronger effects for incidental learning: Pressley et al., 1987; Woloshyn, Willoughby, Wood, & Pressley, 1990). Although most studies have involved individual learning, elaborative-interrogation effects have also been shown among students working in dyads or small groups (Kahl & Woloshyn, 1994; Woloshyn & Stockley, 1995).

1.2b Student characteristics. Elaborative-interrogation effects also appear to be relatively robust across different kinds of learners. Although a considerable amount of work has involved undergraduate students, an impressive number of studies have shown elaborative-interrogation effects with younger learners as well. Elaborative interrogation has been shown to improve learning for high school students, middle school students, and upper elementary school students (fourth through sixth graders). The extent to which elaborative interrogation benefits younger learners is less clear. Miller and Pressley (1989) did not find effects for kindergartners or first graders, and Wood, Miller, Symons, Canough, and Yedlicka (1993) reported mixed results for preschoolers. Nonetheless, elaborative interrogation does appear to benefit learners across a relatively wide age range. Furthermore, several of the studies involving younger students have also established elaborative-interrogation effects for learners of varying ability levels, including fourth through twelfth graders with learning disabilities (C. Greene, Symons, & Richards, 1996; Scruggs, Mastropieri, & Sullivan, 1994) and sixth through eighth graders with mild

cognitive disabilities (Scruggs, Mastropieri, Sullivan, & Hesser, 1993), although Wood, Willoughby, Bolger, Younger, and Kaspar (1993) did not find effects with a sample of low-achieving students. On the other end of the continuum, elaborative-interrogation effects have been shown for high-achieving fifth and sixth graders (Wood & Hewitt, 1993; Wood, Willoughby, et al., 1993).

Another key dimension along which learners differ is level of prior knowledge, a factor that has been extensively investigated within the literature on elaborative interrogation. Both correlational and experimental evidence suggest that prior knowledge is an important moderator of elaborative-interrogation effects, such that effects generally increase as prior knowledge increases. For example, Woloshyn, Pressley, and Schneider (1992) presented Canadian and German students with facts about Canadian provinces and German states. Thus, both groups of students had more domain knowledge for one set of facts and less domain knowledge for the other set. As shown in Figure 1, students showed larger effects of elaborative interrogation in their high-knowledge domain (a 24% increase) than in their low-knowledge domain (a 12% increase). Other studies manipulating the familiarity of to-be-learned materials have reported similar patterns, with significant effects for new facts about familiar items but weaker or nonexistent effects for facts about unfamiliar items. Despite some exceptions (e.g., Ozgungor & Guthrie, 2004), the overall conclusion that emerges from the literature is that high-knowledge learners will generally be best equipped to profit from the elaborative-interrogation technique. The benefit for lower-knowledge learners is less certain.

One intuitive explanation for why prior knowledge moderates the effects of elaborative interrogation is that higher knowledge permits the generation of more appropriate explanations for why a fact is true. If so, one might expect final-test performance to vary as a function of the quality of the explanations generated during study. However, the evidence is mixed. Whereas some studies have found that test performance is better following adequate elaborative-interrogation responses (i.e., those that include a precise, plausible, or accurate explanation for a fact) than for inadequate responses, the differences have often been small, and other studies have failed to find differences (although the numerical trends are usually in the anticipated direction). A somewhat more consistent finding is that performance is better following an adequate response than no response, although in this case, too, the results are somewhat mixed. More generally, the available evidence should be interpreted with caution, given that outcomes are based on conditional post hoc analyses that likely reflect item-selection effects. Thus, the extent to which elaborative-interrogation effects depend on the quality of the elaborations generated is still an open question.

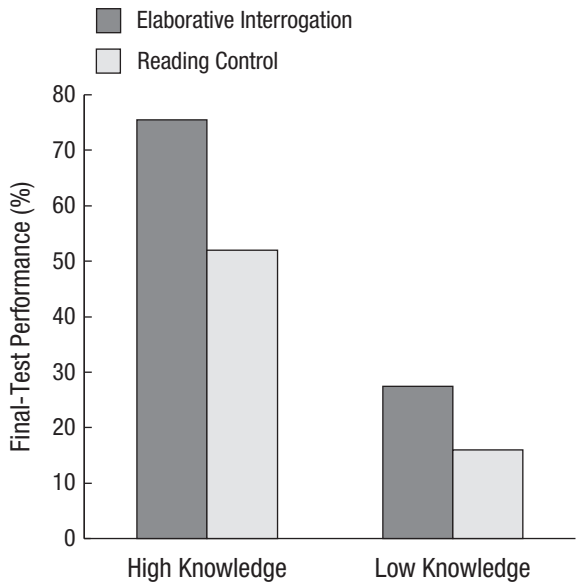


Fig. 1. Mean percentage of correct responses on a final test for learners with high or low domain knowledge who engaged in elaborative interrogation or in reading only during learning (in Woloshyn, Pressley, & Schneider, 1992). Standard errors are not available.

1.2c Materials. Although several studies have replicated elaborative-interrogation effects using the relatively artificial "man sentences" used by Pressley et al. (1987), the majority of subsequent research has extended these effects using materials that better represent what students are actually expected to learn. The most commonly used materials involved sets of facts about various familiar and unfamiliar animals (e.g., "The Western Spotted Skunk's hole is usually found on a sandy piece of farmland near crops"), usually with an elaborative-interrogation prompt following the presentation of each fact. Other studies have extended elaborative-interrogation effects to fact lists from other content domains, including facts about U.S. states, German states, Canadian provinces, and universities; possible reasons for dinosaur extinction; and gender-specific facts about men and women. Other studies have shown elaborative-interrogation effects for factual statements about various topics (e.g., the solar system) that are normatively consistent or inconsistent with learners' prior beliefs (e.g., Woloshyn, Paivio, & Pressley, 1994). Effects have also been shown for facts contained in longer connected discourse, including expository texts on animals (e.g., Seifert, 1994); human digestion (B. L. Smith, Holliday, & Austin, 2010); the neuropsychology of phantom pain (Ozgungor & Guthrie, 2004); retail, merchandising, and accounting (Dornisch & Sperling, 2006); and various science concepts (McDaniel & Donnelly, 1996). Thus, elaborative-interrogation effects are relatively robust across factual material of different kinds and with different contents. However, it is important to note that elaborative interrogation has been applied (and may be applicable) only to discrete units of factual information.

1.2d Criterion tasks. Whereas elaborative-interrogation effects appear to be relatively robust across materials and learners, the extensions of elaborative-interrogation effects across measures that tap different kinds or levels of learning is somewhat more limited. With only a few exceptions, the majority of elaborative-interrogation studies have relied on the

following associative-memory measures: cued recall (generally involving the presentation of a fact to prompt recall of the entity for which the fact is true; e.g., "Which animal . . . ?") and matching (in which learners are presented with lists of facts and entities and must match each fact with the correct entity). Effects have also been shown on measures of fact recognition (B. L. Smith et al., 2010; Woloshyn et al., 1994; Woloshyn & Stockley, 1995). Concerning more generative measures, a few studies have also found elaborative-interrogation effects on free-recall tests (e.g., Woloshyn & Stockley, 1995; Woloshyn et al., 1994), but other studies have not (Dornisch & Sperling, 2006; McDaniel & Donnelly, 1996).

All of the aforementioned measures primarily reflect memory for explicitly stated information. Only three studies have used measures tapping comprehension or application of the factual information. All three studies reported elaborative-interrogation effects on either multiple-choice or verification tests that required inferences or higher-level integration (Dornisch & Sperling, 2006; McDaniel & Donnelly, 1996; Ozgungor & Guthrie, 2004). Ozgungor and Guthrie (2004) also found that elaborative interrogation improved performance on a concept-relatedness rating task (in brief, students rated the pairwise relatedness of the key concepts from a passage, and rating coherence was assessed via Pathfinder analyses); however, Dornisch and Sperling (2006) did not find significant elaborative-interrogation effects on a problem-solving test. In sum, whereas elaborative-interrogation effects on associative memory have been firmly established, the extent to which elaborative interrogation facilitates recall or comprehension is less certain.

Of even greater concern than the limited array of measures that have been used is the fact that few studies have examined performance after meaningful delays. Almost all prior studies have administered outcome measures either immediately or within a few minutes of the learning phase. Results from the few studies that have used longer retention intervals are promising. Elaborative-interrogation effects have been shown after delays of 1–2 weeks (Scruggs et al., 1994; Woloshyn et al., 1994), 1–2 months (Kahl & Woloshyn, 1994; Willoughby, Waller, Wood, & MacKinnon, 1993; Woloshyn & Stockley, 1995), and even 75 and 180 days (Woloshyn et al., 1994). In almost all of these studies, however, the delayed test was preceded by one or more criterion tests at shorter intervals, introducing the possibility that performance on the delayed test was contaminated by the practice provided by the preceding tests. Thus, further work is needed before any definitive conclusions can be drawn about the extent to which elaborative interrogation produces durable gains in learning.

1.3 Effects in representative educational contexts. Concerning the evidence that elaborative interrogation will enhance learning in representative educational contexts, few studies have been conducted outside the laboratory. However, outcomes from a recent study are suggestive (B. L. Smith et al., 2010). Participants were undergraduates enrolled in an introductory biology course, and the experiment was conducted during class meetings in the accompanying lab section. During one class meeting, students completed a measure of verbal ability and a prior-knowledge test over material that was related, but not identical, to the target material. In the following week, students were presented with a lengthy text on human digestion that was taken from a chapter in the course textbook. For half of the students, 21 elaborative interrogation prompts were interspersed throughout the text (roughly one prompt per 150 words), each consisting of a paraphrased statement from the text followed by "Why is this true?" The remaining students were simply instructed to study the text at their own pace, without any prompts. All students then completed 105 true/false questions about the material (none of which were the same as the elaborative-interrogation prompts). Performance was better for the elaborative-interrogation group than for the control group (76% versus 69%), even after controlling for prior knowledge and verbal ability.

1.4 Issues for implementation. One possible merit of elaborative interrogation is that it apparently requires minimal training. In the majority of studies reporting elaborative-interrogation effects, learners were given brief instructions and then practiced generating elaborations for 3 or 4 practice facts (sometimes, but not always, with feedback about the quality of the elaborations) before beginning the main task. In some studies, learners were not provided with any practice or illustrative examples prior to the main task. Additionally, elaborative interrogation appears to be relatively reasonable with respect to time demands. Almost all studies set reasonable limits on the amount of time allotted for reading a fact and for generating an elaboration (e.g., 15 seconds allotted for each fact). In one of the few studies permitting self-paced learning, the time-on-task difference between the elaborative-interrogation and reading-only groups was relatively minimal (32 minutes vs. 28 minutes; B. L. Smith et al., 2010). Finally, the consistency of the prompts used across studies allows for relatively straightforward recommendations to students about the nature of the questions they should use to elaborate on facts during study.

With that said, one limitation noted above concerns the potentially narrow applicability of elaborative interrogation to discrete factual statements. As Hamilton (1997) noted, "elaborative interrogation is fairly prescribed when focusing on a list of factual sentences. However, when focusing on more complex outcomes, it is not as clear to what one should direct the 'why' questions" (p. 308). For example, when learning about a complex causal process or system (e.g., the digestive system), the appropriate grain size for elaborative interrogation is an open question (e.g., should a prompt focus on an entire system or just a smaller part of it?). Furthermore, whereas the facts to be elaborated are clear when dealing with fact lists, elaborating on facts embedded in lengthier texts will require students to identify their own target facts. Thus, students may need some instruction about the kinds of content to which

elaborative interrogation may be fruitfully applied. Dosage is also of concern with lengthier text, with some evidence suggesting that elaborative-interrogation effects are substantially diluted (Callender & McDaniel, 2007) or even reversed (Ramsay, Sperling, & Dornisch, 2010) when elaborative-interrogation prompts are administered infrequently (e.g., one prompt every 1 or 2 pages).

1.5 Elaborative interrogation: Overall assessment. We rate elaborative interrogation as having moderate utility. Elaborative-interrogation effects have been shown across a relatively broad range of factual topics, although some concerns remain about the applicability of elaborative interrogation to material that is lengthier or more complex than fact lists. Concerning learner characteristics, effects of elaborative interrogation have been consistently documented for learners at least as young as upper elementary age, but some evidence suggests that the benefits of elaborative interrogation may be limited for learners with low levels of domain knowledge. Concerning criterion tasks, elaborative-interrogation effects have been firmly established on measures of associative memory administered after short delays, but firm conclusions about the extent to which elaborative interrogation benefits comprehension or the extent to which elaborative-interrogation effects persist across longer delays await further research. Further research demonstrating the efficacy of elaborative interrogation in representative educational contexts would also be useful. In sum, the need for further research to establish the generalizability of elaborative-interrogation effects is primarily why this technique did not receive a high-utility rating.

2 Self-explanation

2.1 General description of self-explanation and why it should work. In the seminal study on self-explanation, Berry (1983) explored its effects on logical reasoning using the Wason card-selection task. In this task, a student might see four cards labeled "A," "4," "D," and "3" and be asked to indicate which cards must be turned over to test the rule "if a card has A on one side, it has 3 on the other side" (an instantiation of the more general "if P, then Q" rule). Students were first asked to solve a concrete instantiation of the rule (e.g., flavor of jam on one side of a jar and the sale price on the other); accuracy was near zero. They then were provided with a minimal explanation about how to solve the "if P, then Q" rule and were given a set of concrete problems involving the use of this and other logical rules (e.g., "if P, then not Q"). For this set of concrete practice problems, one group of students was prompted to self-explain while solving each problem by stating the reasons for choosing or not choosing each card. Another group of students solved all problems in the set and only then were asked to explain how they had gone about solving the problems. Students in a control group were not prompted to self-explain at any point. Accuracy on the practice problems was 90% or better in all three groups. However, when the logical rules were instantiated in a set of abstract problems presented during a subsequent transfer test, the two self-explanation groups substantially outperformed the control group (see Fig. 2). In a second experiment, another control group was explicitly told about the logical connection between the concrete practice problems they had just solved and the forthcoming abstract problems, but they fared no better (28%).

As illustrated above, the core component of self-explanation involves having students explain some aspect of their processing during learning. Consistent with basic theoretical assumptions about the related technique of elaborative interrogation, self-explanation may enhance learning by supporting the integration of new information with existing prior knowledge. However, compared with the consistent prompts used in the elaborative-interrogation literature, the prompts used to elicit self-explanations have been much more variable across studies. Depending on the variation of the prompt used, the particular mechanisms underlying self-explanation effects may differ somewhat. The key continuum along which self-explanation prompts differ concerns the degree to which they are content-free versus content-specific. For example, many studies have used prompts that include no explicit mention of particular content from the to-be-learned materials (e.g., "Explain what the sentence means to you. That is, what new information does the sentence provide for you? And how does it relate to what you already know?"). On the other end of the continuum, many studies have used prompts that are much more content-specific, such that different prompts are used for

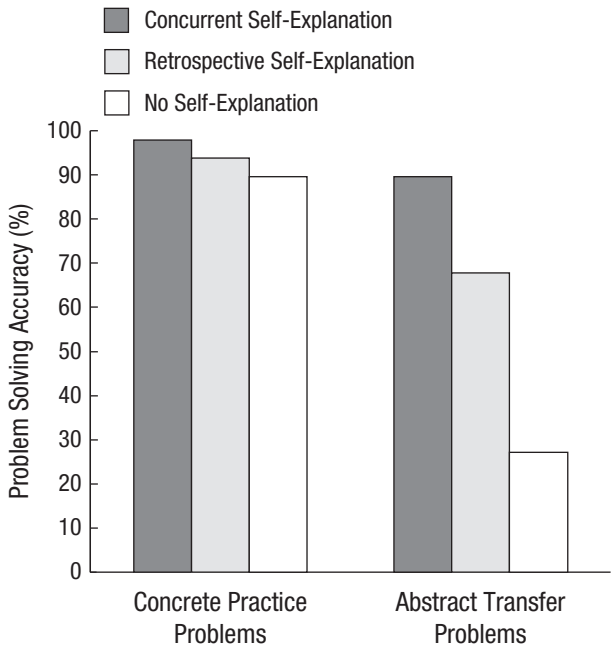


Fig. 2. Mean percentage of logical-reasoning problems answered correctly for concrete practice problems and subsequently administered abstract transfer problems in Berry (1983). During a practice phase, learners self-explained while solving each problem, self-explained after solving all problems, or were not prompted to engage in self-explanation. Standard errors are not available.

different items (e.g., "Why do you calculate the total acceptable outcomes by multiplying?" "Why is the numerator 14 and the denominator 7 in this step?"). For present purposes, we limit our review to studies that have used prompts that are relatively content-free. Although many of the content-specific prompts do elicit explanations, the relatively structured nature of these prompts would require teachers to construct sets of specific prompts to put into practice, rather than capturing a more general technique that students could be taught to use on their own. Furthermore, in some studies that have been situated in the self-explanation literature, the nature of the prompts is functionally more closely aligned with that of practice testing.

Even within the set of studies selected for review here, considerable variability remains in the self-explanation prompts that have been used. Furthermore, the range of tasks and measures that have been used to explore self-explanation is quite large. Although we view this range as a strength of the literature, the variability in self-explanation prompts, tasks, and measures does not easily support a general summative statement about the mechanisms that underlie self-explanation effects.

2.2 How general are the effects of self-explanation?

2.2a Learning conditions. Several studies have manipulated other aspects of learning conditions in addition to self-explanation. For example, Rittle-Johnson (2006) found that self-explanation was effective when accompanied by either direct instruction or discovery learning. Concerning potential moderating factors, Berry (1983) included a group who self-explained after the completion of each problem rather than during problem solving. Retrospective self-explanation did enhance performance relative to no self-explanation, but the effects were not as pronounced as with concurrent self-explanation. Another moderating factor may concern the extent to which provided explanations are made available to learners. Schworm and Renkl (2006) found that self-explanation effects were significantly diminished when learners could access explanations, presumably because learners made minimal attempts to answer the explanatory prompts before consulting the provided information (see also Aleven & Koedinger, 2002).

2.2b Student characteristics. Self-explanation effects have been shown with both younger and older learners. Indeed, self-explanation research has relied much less heavily on samples of college students than most other literatures have, with at least as many studies involving younger learners as involving undergraduates. Several studies have reported self-explanation effects with kindergartners, and other studies have shown effects for elementary school students, middle school students, and high school students.

In contrast to the breadth of age groups examined, the extent to which the effects of self-explanation generalize across different levels of prior knowledge or ability has not been sufficiently explored. Concerning knowledge level, several studies have used pretests to select participants with relatively low levels of knowledge or task experience, but no research has systematically examined self-explanation effects as a function of knowledge level. Concerning ability level, Chi, de Leeuw, Chiu, and LaVancher (1994) examined the effects of self-explanation on learning from an expository text about the circulatory system among participants in their sample who had received the highest and lowest scores on a measure of general aptitude and found gains of similar magnitude in each group. In contrast, Didierjean and Cauzinille-Marmèche (1997) examined algebra-problem solving in a sample of ninth graders with either low or intermediate algebra skills, and they found self-explanation effects only for lower-skill students. Further work is needed to establish the generality of self-explanation effects across these important idiographic dimensions.

2.2c Materials. One of the strengths of the self-explanation literature is that effects have been shown not only across different materials within a task domain but also across several different task domains. In addition to the logical-reasoning problems used by Berry (1983), self-explanation has been shown to support the solving of other kinds of logic puzzles. Self-explanation has also been shown to facilitate the solving of various kinds of math problems, including simple addition problems for kindergartners, mathematical-equivalence problems for elementary-age students, and algebraic formulas and geometric theorems for older learners. In addition to improving problem solving, self-explanation improved student teachers' evaluation of the goodness of practice problems for use in classroom instruction. Self-explanation has also helped younger learners overcome various kinds of misconceptions, improving children's understanding of false belief (i.e., that individuals can have a belief that is different from reality), number conservation (i.e., that the number of objects in an array does not change when the positions of those objects in the array change), and principles of balance (e.g., that not all objects balance on a fulcrum at their center point). Self-explanation has improved children's pattern learning and adults' learning of endgame strategies in chess. Although most of the research on self-explanation has involved procedural or problem-solving tasks, several studies have also shown self-explanation effects for learning from text, including both short narratives and lengthier expository texts. Thus, self-explanation appears to be broadly applicable.

2.2d Criterion tasks. Given the range of tasks and domains in which self-explanation has been investigated, it is perhaps not surprising that self-explanation effects have been shown on a wide range of criterion measures. Some studies have shown self-explanation effects on standard measures of memory, including free recall, cued recall, fill-in-the-blank tests, associative matching, and multiple-choice tests tapping explicitly stated information. Studies involving text learning have also shown effects on measures of comprehension, including diagram-drawing tasks, application-based questions, and tasks in which learners must make inferences on the basis of

information implied but not explicitly stated in a text. Across those studies involving some form of problem-solving task, virtually every study has shown self-explanation effects on near-transfer tests in which students are asked to solve problems that have the same structure as, but are nonidentical to, the practice problems. Additionally, self-explanation effects on far-transfer tests (in which students are asked to solve problems that differ from practice problems not only in their surface features but also in one or more structural aspects) have been shown for the solving of math problems and pattern learning. Thus, self-explanation facilitates an impressive range of learning outcomes.

In contrast, the durability of self-explanation effects is woefully underexplored. Almost every study to date has administered criterion tests within minutes of completion of the learning phase. Only five studies have used longer retention intervals. Self-explanation effects persisted across 1–2 day delays for playing chess endgames (de Bruin, Rikers, & Schmidt, 2007) and for retention of short narratives (Magliano, Trabasso, & Graesser, 1999). Self-explanation effects persisted across a 1-week delay for the learning of geometric theorems (although an additional study session intervened between initial learning and the final test; R. M. F. Wong, Lawson, & Keeves, 2002) and for learning from a text on the circulatory system (although the final test was an open-book test; Chi et al., 1994). Finally, Rittle-Johnson (2006) reported significant effects on performance in solving math problems after a 2-week delay; however, the participants in this study also completed an immediate test, thus introducing the possibility that testing effects influenced performance on the delayed test. Taken together, the outcomes of these few studies are promising, but considerably more research is needed before confident conclusions can be made about the longevity of self-explanation effects.

2.3 Effects in representative educational contexts. Concerning the strength of the evidence that self-explanation will enhance learning in educational contexts, outcomes from two studies in which participants were asked to learn course-relevant content are at least suggestive. In a study by Schworm and Renkl (2006), students in a teacher-education program learned how to develop example problems to use in their classrooms by studying samples of well-designed and poorly designed example problems in a computer program. On each trial, students in a self-explanation group were prompted to explain why one of two examples was more effective than the other, whereas students in a control group were not prompted to self-explain. Half of the participants in each group were also given the option to examine experimenter-provided explanations on each trial. On an immediate test in which participants selected and developed example problems, the self-explanation group outperformed the control group. However, this effect was limited to students who had not been able to view provided explanations, presumably because students made minimal attempts to self-explain before consulting the provided information.

R. M. F. Wong et al. (2002) presented ninth-grade students in a geometry class with a theorem from the course textbook that had not yet been studied in class. During the initial learning session, students were asked to think aloud while studying the relevant material (including the theorem, an illustration of its proof, and an example of an application of the theorem to a problem). Half of the students were specifically prompted to self-explain after every 1 or 2 lines of new information (e.g., "What parts of this page are new to me? What does the statement mean? Is there anything I still don't understand?"), whereas students in a control group received nonspecific instructions that simply prompted them to think aloud during study. The following week, all students received a basic review of the theorem and completed the final test the next day. Self-explanation did not improve performance on near-transfer questions but did improve performance on far-transfer questions.

2.4 Issues for implementation. As noted above, a particular strength of the self-explanation strategy is its broad applicability across a range of tasks and content domains. Furthermore, in almost all of the studies reporting significant effects of self-explanation, participants were provided with minimal instructions and little to no practice with self-explanation prior to completing the experimental task. Thus, most students apparently can profit from self-explanation with minimal training.

However, some students may require more instruction to successfully implement self-explanation. In a study by Didierjean and Cauzinille-Marmèche (1997), ninth graders with poor algebra skills received minimal training prior to engaging in self-explanation while solving algebra problems; analysis of think-aloud protocols revealed that students produced many more paraphrases than explanations. Several studies have reported positive correlations between final-test performance and both the quantity and quality of explanations generated by students during learning, further suggesting that the benefit of self-explanation might be enhanced by teaching students how to effectively implement the self-explanation technique (for examples of training methods, see Ainsworth & Burcham, 2007; R. M. F. Wong et al., 2002). However, in at least some of these studies, students who produced more or better-quality self-explanations may have had greater domain knowledge; if so, then further training with the technique may not have benefited the more poorly performing students. Investigating the contribution of these factors (skill at self-explanation vs. domain knowledge) to the efficacy of self-explanation will have important implications for how and when to use this technique.

An outstanding issue concerns the time demands associated with self-explanation and the extent to which self-explanation effects may have been due to increased time on task. Unfortunately, few studies equated time on task when comparing self-explanation conditions to control conditions involving other strategies or activities, and most studies involving self-paced practice did not report participants' time on task. In the few

studies reporting time on task, self-paced administration usually yielded nontrivial increases (30–100%) in the amount of time spent learning in the self-explanation condition relative to other conditions, a result that is perhaps not surprising, given the high dosage levels at which self-explanation was implemented. For example, Chi et al. (1994) prompted learners to self-explain after reading each sentence of an expository text, which doubled the amount of time the group spent studying the text relative to a rereading control group (125 vs. 66 minutes, respectively). With that said, Schworm and Renkl (2006) reported that time on task was not correlated with performance across groups, and Ainsworth and Burcham (2007) reported that controlling for study time did not eliminate effects of self-explanation.

Within the small number of studies in which time on task was equated, results were somewhat mixed. Three studies equating time on task reported significant effects of self-explanation (de Bruin et al., 2007; de Koning, Tabbers, Rikers, & Paas, 2011; O'Reilly, Symons, & MacLatchy-Gaudet, 1998). In contrast, Matthews and Rittle-Johnson (2009) had one group of third through fifth graders practice solving math problems with self-explanation and a control group solve twice as many practice problems without self-explanation; the two groups performed similarly on a final test. Clearly, further research is needed to establish the bang for the buck provided by self-explanation before strong prescriptive conclusions can be made.

2.5 Self-explanation: Overall assessment. We rate self-explanation as having moderate utility. A major strength of this technique is that its effects have been shown across different content materials within task domains as well as across several different task domains. Self-explanation effects have also been shown across an impressive age range, although further work is needed to explore the extent to which these effects depend on learners' knowledge or ability level. Self-explanation effects have also been shown across an impressive range of learning outcomes, including various measures of memory, comprehension, and transfer. In contrast, further research is needed to establish the durability of these effects across educationally relevant delays and to establish the efficacy of self-explanation in representative educational contexts. Although most research has shown effects of self-explanation with minimal training, some results have suggested that effects may be enhanced if students are taught how to effectively implement the self-explanation strategy. One final concern has to do with the nontrivial time demands associated with self-explanation, at least at the dosages examined in most of the research that has shown effects of this strategy.

3 Summarization

Students often have to learn large amounts of information, which requires them to identify what is important and how different ideas connect to one another. One popular technique for accomplishing these goals involves having students write summaries of to-be-learned texts. Successful summaries identify the main points of a text and capture the gist of it while excluding unimportant or repetitive material (A. L. Brown, Campione, & Day, 1981). Although learning to construct accurate summaries is often an instructional goal in its own right (e.g., Wade-Stein & Kintsch, 2004), our interest here concerns whether doing so will boost students' performance on later criterion tests that cover the target material.

3.1 General description of summarization and why it should work. As an introduction to the issues relevant to summarization, we begin with a description of a prototypical experiment. Bretzing and Kulhavy (1979) had high school juniors and seniors study a 2,000-word text about a fictitious tribe of people. Students were assigned to one of five learning conditions and given up to 30 minutes to study the text. After reading each page, students in a summarization group were instructed to write three lines of text that summarized the main points from that page. Students in a note-taking group received similar instructions, except that they were told to take up to three lines of notes on each page of text while reading. Students in a verbatim-copying group were instructed to locate and copy the three most important lines on each page. Students in a letter-search group copied all the capitalized words in the text, also filling up three lines. Finally, students in a control group simply read the text without recording anything. (A subset of students from the four conditions involving writing were allowed to review what they had written, but for present purposes we will focus on the students who did not get a chance to review before the final test.) Students were tested either shortly after learning or 1 week later, answering 25 questions that required them to connect information from across the text. On both the immediate and delayed tests, students in the summarization and note-taking groups performed best, followed by the students in the verbatim-copying and control groups, with the worst performance in the letter-search group (see Fig. 3).

Bretzing and Kulhavy's (1979) results fit nicely with the claim that summarization boosts learning and retention because it involves attending to and extracting the higher-level meaning and gist of the material. The conditions in the experiment were specifically designed to manipulate how much students processed the texts for meaning, with the letter-search condition involving shallow processing of the text that did not require learners to extract its meaning (Craik & Lockhart, 1972). Summarization was more beneficial than that shallow task and yielded benefits similar to those of note-taking, another task known to boost learning (e.g., Bretzing & Kulhavy, 1981; Crawford, 1925a, 1925b; Di Vesta & Gray, 1972). More than just facilitating the extraction of meaning, however, summarization should also boost organizational processing, given that extracting the gist of a text requires learners to connect disparate pieces of the text, as opposed to simply evaluating its individual components (similar to the way in which note-taking affords organizational processing; Einstein,

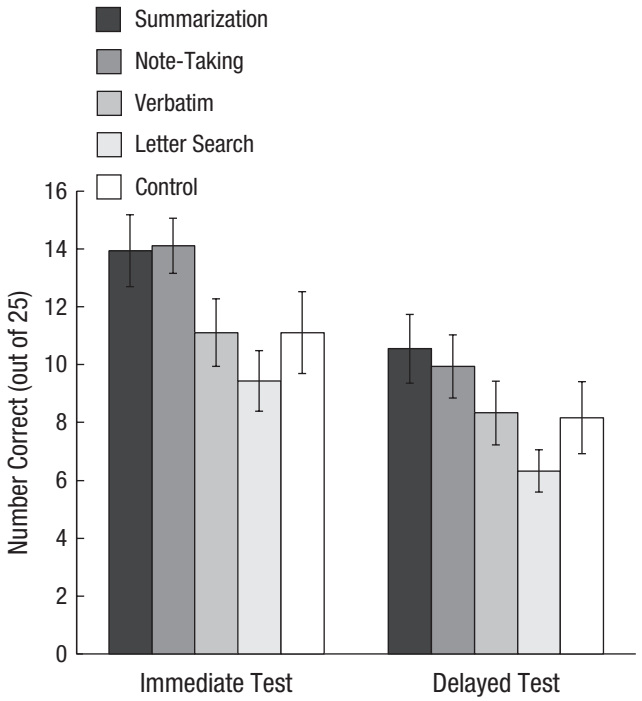


Fig. 3. Mean number of correct responses on a test occurring shortly after study as a function of test type (immediate or delayed) and learning condition in Bretzing and Kulhavy (1979). Error bars represent standard errors.

Morris, & Smith, 1985). One last point should be made about the results from Bretzing and Kulhavy (1979)—namely, that summarization and note-taking were both more beneficial than was verbatim copying. Students in the verbatim-copying group still had to locate the most important information in the text, but they did not synthesize it into a summary or rephrase it in their notes. Thus, writing about the important points in one's own words produced a benefit over and above that of selecting important information; students benefited from the more active processing involved in summarization and note-taking (see Wittrock, 1990, and Chi, 2009, for reviews of active/generative learning). These explanations all suggest that summarization helps students identify and organize the main ideas within a text.

So how strong is the evidence that summarization is a beneficial learning strategy? One reason this question is difficult to answer is that the summarization strategy has been implemented in many different ways across studies, making it difficult to draw general conclusions about its efficacy. Pressley and colleagues described the situation well when they noted that "summarization is not one strategy but a family of strategies" (Pressley, Johnson, Symons, McGoldrick, & Kurita, 1989, p. 5). Depending on the particular instructions given, students' summaries might consist of single words, sentences, or longer paragraphs; be limited in length or not; capture an entire text or only a portion of it; be written or spoken aloud; or be produced from memory or with the text present.

A lot of research has involved summarization in some form, yet whereas some evidence demonstrates that summarization works (e.g., L. W. Brooks, Dansereau, Holley, & Spurlin, 1983; Doctorow, Wittrock, & Marks, 1978), T. H. Anderson and Armbruster's (1984) conclusion that "research in support of summarizing as a studying activity is sparse indeed" (p. 670) is not outmoded. Instead of focusing on discovering when (and how) summarization works, by itself and without training, researchers have tended to explore how to train students to write better summaries (e.g., Friend, 2001; Hare & Borchardt, 1984) or to examine other benefits of training the skill of summarization. Still others have simply assumed that summarization works, including it as a component in larger interventions (e.g., Carr, Bigler, & Morningstar, 1991; Lee, Lim, & Grabowski, 2010; Palincsar & Brown, 1984; Spörer, Brunstein, & Kieschke, 2009). When collapsing across findings pertaining to all forms of summarization, summarization appears to benefit students, but the evidence for any one instantiation of the strategy is less compelling.

The focus on training students to summarize reflects the belief that the quality of summaries matters. If a summary does not emphasize the main points of a text, or if it includes incorrect information, why would it be expected to benefit learning and retention? Consider a study by Bednall and Kehoe (2011, Experiment 2), in which undergraduates studied six Web units that explained different logical fallacies and provided examples of each. Of interest for present purposes are two groups: a control group who simply read the units and a group in which students were asked to summarize the material as if they were explaining it to a friend. Both groups received the following tests: a multiple-choice quiz that tested information directly stated in the Web unit; a short-answer test in which, for each of a list of presented statements, students were required to name the specific fallacy that had been committed or write "not a fallacy" if one had not occurred; and, finally, an application test that required students to write explanations of logical fallacies in examples that had been studied (near transfer) as well as explanations of fallacies in novel examples (far transfer). Summarization did not benefit overall performance, but the researchers noticed that the summaries varied a lot in content; for one studied fallacy, only 64% of the summaries included the correct definition. Table 3 shows the relationships between summary content and later performance. Higher-quality summaries that contained more information and that were linked to prior knowledge were associated with better performance.

Several other studies have supported the claim that the quality of summaries has consequences for later performance. Most similar to the Bednall and Kehoe (2011) result is Ross and Di Vesta's (1976) finding that the length (in words) of an oral summary (a very rough indicator of quality) correlated with later performance on multiple-choice and short-answer questions. Similarly, Dyer, Riley, and Yekovich (1979) found that final-test questions were more likely to be answered correctly if the information needed to answer them had been included in an earlier summary. Garner (1982) used a different

Table 3. Correlations between Measures of Summary Quality and Later Test Performance (from Bednall & Kehoe, 2011, Experiment 2)

	Test		
Measure of summary quality	Multiple-choice test (factual knowledge)	Short-answer test (identification)	Application test
Number of correct definitions	.42*	.43*	.52*
Amount of extra information	.31*	.21*	.40*

Note. Asterisks indicate correlations significantly greater than 0. "Amount of extra information" refers to the number of summaries in which a student included information that had not been provided in the studied material (e.g., an extra example).

method to show that the quality of summaries matters: Undergraduates read a passage on Dutch elm disease and then wrote a summary at the bottom of the page. Five days later, the students took an old/new recognition test; critical items were new statements that captured the gist of the passage (as in Bransford & Franks, 1971). Students who wrote better summaries (i.e., summaries that captured more important information) were more likely to falsely recognize these gist statements, a pattern suggesting that the students had extracted a higher-level understanding of the main ideas of the text.

3.2 How general are the effects of summarization?

3.2a Learning conditions. As noted already, many different types of summaries can influence learning and retention; summarization can be simple, requiring the generation of only a heading (e.g., L. W. Brooks et al., 1983) or a single sentence per paragraph of a text (e.g., Doctorow et al., 1978), or it can be as complicated as an oral presentation on an entire set of studied material (e.g., Ross & Di Vesta, 1976). Whether it is better to summarize smaller pieces of a text (more frequent summarization) or to capture more of the text in a larger summary (less frequent summarization) has been debated (Foos, 1995; Spurlin, Dansereau, O'Donnell, & Brooks, 1988). The debate remains unresolved, perhaps because what constitutes the most effective summary for a text likely depends on many factors (including students' ability and the nature of the material).

One other open question involves whether studied material should be present during summarization. Hidi and Anderson (1986) pointed out that having the text present might help the reader to succeed at identifying its most important points as well as relating parts of the text to one another. However, summarizing a text without having it present involves retrieval, which is known to benefit memory (see the Practice Testing section of this monograph), and also prevents the learner from engaging in verbatim copying. The Dyer et al. (1979) study described earlier involved summarizing without the text present; in this study, no overall benefit from summarizing occurred, even though information that had been included in summaries was benefited (overall, this benefit was overshadowed by costs to the greater amount of information that had not been included in summaries). More generally, some studies have shown benefits from summarizing an absent text (e.g., Ross & Di Vesta, 1976), but some have not (e.g., M. C. M. Anderson & Thiede, 2008, and Thiede & Anderson, 2003, found no benefits of summarization on test performance). The answer to whether studied text should be present during summarization is most likely a complicated one, and it may depend on people's ability to summarize when the text is absent.

3.2b Student characteristics. Benefits of summarization have primarily been observed with undergraduates. Most of the research on individual differences has focused on the age of students, because the ability to summarize develops with age. Younger students struggle to identify main ideas and tend to write lower-quality summaries that retain more of the original wording and structure of a text (e.g., A. L. Brown & Day, 1983; A. L. Brown, Day, & Jones, 1983). However, younger students (e.g., middle school students) can benefit from summarization following extensive training (e.g., Armbruster, Anderson, & Ostertag, 1987; Bean & Steenwyk, 1984). For example, consider a successful program for sixth-grade students (Rinehart, Stahl, & Erickson, 1986). Teachers received 90 minutes of training so that they could implement summarization training in their classrooms; students then completed five 45- to 50-minute sessions of training. The training reflected principles of direct instruction, meaning that students were explicitly taught about the strategy, saw it modeled, practiced it and received feedback, and eventually learned to monitor and check their work. Students who had received the training recalled more major information from a textbook chapter (i.e., information identified by teachers as the most important for students to know) than did students who had not, and this benefit was linked to improvements in note-taking. Similar training programs have succeeded with middle school students who are learning disabled (e.g., Gajria & Salvia, 1992; Malone & Mastropieri, 1991), minority high school students (Hare & Borchardt, 1984), and underprepared college students (A. King, 1992).

Outcomes of two other studies have implications for the generality of the summarization strategy, as they involve individual differences in summarization skill (a prerequisite for

using the strategy). First, both general writing skill and interest in a topic have been linked to summarization ability in seventh graders (Head, Readence, & Buss, 1989). Writing skill was measured via performance on an unrelated essay, and interest in the topic (American history) was measured via a survey that asked students how much they would like to learn about each of 25 topics. Of course, interest may be confounded with knowledge about a topic, and knowledge may also contribute to summarization skill. Recht and Leslie (1988) showed that seventh- and eighth-grade students who knew a lot about baseball (as measured by a pretest) were better at summarizing a 625-word passage about a baseball game than were students who knew less about baseball. This finding needs to be replicated with different materials, but it seems plausible that students with more domain-relevant knowledge would be better able to identify the main points of a text and extract its gist. The question is whether domain experts would benefit from the summarization strategy or whether it would be redundant with the processing in which these students would spontaneously engage.

3.2c Materials. The majority of studies have used prose passages on such diverse topics as a fictitious primitive tribe, desert life, geology, the blue shark, an earthquake in Lisbon, the history of Switzerland, and fictional stories. These passages have ranged in length from a few hundred words to a few thousand words. Other materials have included Web modules and lectures. For the most part, characteristics of materials have not been systematically manipulated, which makes it difficult to draw strong conclusions about this factor, even though 15 years have passed since Hidi and Anderson (1986) made an argument for its probable importance. As discussed in Yu (2009), it makes sense that the length, readability, and organization of a text might all influence a reader's ability to summarize it, but these factors need to be investigated in studies that manipulate them while holding all other factors constant (as opposed to comparing texts that vary along multiple dimensions).

3.2d Criterion tasks. The majority of summarization studies have examined the effects of summarization on either retention of factual details or comprehension of a text (often requiring inferences) through performance on multiple-choice questions, cued recall questions, or free recall. Other benefits of summarization include enhanced metacognition (with text-absent summarization improving the extent to which readers can accurately evaluate what they do or do not know; M. C. M. Anderson & Thiede, 2008; Thiede & Anderson, 2003) and improved note-taking following training (A. King, 1992; Rinehart et al., 1986).

Whereas several studies have shown benefits of summarization (sometimes following training) on measures of application (e.g., B. Y. L. Wong, Wong, Perry, & Sawatsky, 1986), others have failed to find such benefits. For example, consider a study in which L. F. Annis (1985) had undergraduates read a passage on an earthquake and then examined the consequences of summarization for performance on questions designed to tap different categories of learning within Bloom et al.'s (1956) taxonomy. One week after learning, students who had summarized performed no differently than students in a control group who had only read the passages in answering questions that tapped a basic level of knowledge (fact and comprehension questions). Students benefited from summarization when the questions required the application or analysis of knowledge, but summarization led to *worse* performance on evaluation and synthesis questions. These results need to be replicated, but they highlight the need to assess the consequences of summarization on the performance of tasks that measure various levels of Bloom's taxonomy.

Across studies, results have also indicated that summarization helps later performance on generative measures (e.g., free recall, essays) more than it affects performance on multiple-choice or other measures that do not require the student to produce information (e.g., Bednall & Kehoe, 2011; L. W. Brooks et al., 1983; J. R. King, Biggs, & Lipsky, 1984). Because summarizing requires production, the processing involved is likely a better match to generative tests than to tests that depend on recognition.

Unfortunately, the one study we found that used a high-stakes test did not show a benefit from summarization training (Brozo, Stahl, & Gordon, 1985). Of interest for present purposes were two groups in the study, which was conducted with college students in a remedial reading course who received training either in summarization or in self-questioning (in the self-questioning condition, students learned to write multiple-choice comprehension questions). Training lasted for 4 weeks; each week, students received approximately 4 to 5 hours of instruction and practice that involved applying the techniques to 1-page news articles. Of interest was the students' performance on the Georgia State Regents' examination, which involves answering multiple-choice reading-comprehension questions about passages; passing this exam is a graduation requirement for many college students in the University System of Georgia (see http://www2.gsu.edu/~wwwrtp/). Students also took a practice test before taking the actual Regents' exam. Unfortunately, the mean scores for both groups were at or below passing, for both the practice and actual exams. However, the self-questioning group performed better than the summarization group on both the practice test and the actual Regents' examination. This study did not report pretraining scores and did not include a no-training control group, so some caution is warranted in interpreting the results. However, it emphasizes the need to establish that outcomes from basic laboratory work generalize to actual educational contexts and suggests that summarization may not have the same influence in both contexts.

Finally, concerning test delays, several studies have indicated that when summarization does boost performance, its effects are relatively robust over delays of days or weeks (e.g., Bretzing & Kulhavy, 1979; B. L. Stein & Kirby, 1992). Similarly, benefits of training programs have persisted several weeks after the end of training (e.g., Hare & Borchardt, 1984).

3.3 Effects in representative educational contexts. Several of the large summarization-training studies have been conducted in regular classrooms, indicating the feasibility of doing so. For example, the study by A. King (1992) took place in the context of a remedial study-skills course for undergraduates, and the study by Rinehart et al. (1986) took place in sixth-grade classrooms, with the instruction led by students' regular teachers. In these and other cases, students benefited from the classroom training. We suspect it may actually be more feasible to conduct these kinds of training studies in classrooms than in the laboratory, given the nature of the time commitment for students. Even some of the studies that did not involve training were conducted outside the laboratory; for example, in the Bednall and Kehoe (2011) study on learning about logical fallacies from Web modules (see data in Table 3), the modules were actually completed as a homework assignment. Overall, benefits can be observed in classroom settings; the real constraint is whether students have the skill to successfully summarize, not whether summarization occurs in the lab or the classroom.

3.4 Issues for implementation. Summarization would be feasible for undergraduates or other learners who already know how to summarize. For these students, summarization would constitute an easy-to-implement technique that would not take a lot of time to complete or understand. The only concern would be whether these students might be better served by some other strategy, but certainly summarization would be better than the study strategies students typically favor, such as highlighting and rereading (as we discuss in the sections on those strategies below). A trickier issue would concern implementing the strategy with students who are not skilled summarizers. Relatively intensive training programs are required for middle school students or learners with learning disabilities to benefit from summarization. Such efforts are not misplaced; training has been shown to benefit performance on a range of measures, although the training procedures do raise practical issues (e.g., Gajria & Salvia, 1992: 6.5–11 hours of training used for sixth through ninth graders with learning disabilities; Malone & Mastropieri, 1991: 2 days of training used for middle school students with learning disabilities; Rinehart et al., 1986: 45–50 minutes of instruction per day for 5 days used for sixth graders). Of course, instructors may want students to summarize material because summarization itself is a goal, not because they plan to use summarization as a study technique, and that goal may merit the efforts of training.

However, if the goal is to use summarization as a study technique, our question is whether training students would be worth the amount of time it would take, both in terms of the time required on the part of the instructor and in terms of the time taken away from students' other activities. For instance, in terms of efficacy, summarization tends to fall in the middle of the pack when compared to other techniques. In direct comparisons, it was sometimes more useful than rereading (Rewey, Dansereau, & Peel, 1991) and was as useful as note-taking (e.g., Bretzing & Kulhavy, 1979) but was less powerful than generating explanations (e.g., Bednall & Kehoe, 2011) or self-questioning (A. King, 1992).

3.5 Summarization: Overall assessment. On the basis of the available evidence, we rate summarization as low utility. It can be an effective learning strategy for learners who are already skilled at summarizing; however, many learners (including children, high school students, and even some undergraduates) will require extensive training, which makes this strategy less feasible. Our enthusiasm is further dampened by mixed findings regarding which tasks summarization actually helps. Although summarization has been examined with a wide range of text materials, many researchers have pointed to factors of these texts that seem likely to moderate the effects of summarization (e.g., length), and future research should be aimed at investigating such factors. Finally, although many studies have examined summarization training in the classroom, what are lacking are classroom studies examining the effectiveness of summarization as a technique that boosts students' learning, comprehension, and retention of course content.

4 Highlighting and underlining

Any educator who has examined students' course materials is familiar with the sight of a marked-up, multicolored textbook. More systematic evaluations of actual textbooks and other student materials have supported the claim that highlighting and underlining are common behaviors (e.g., Bell & Limber, 2010; Lonka, Lindblom-Ylänne, & Maury, 1994; Nist & Kirby, 1989). When students themselves are asked about what they do when studying, they commonly report underlining, highlighting, or otherwise marking material as they try to learn it (e.g., Cioffi, 1986; Gurung, Weidert, & Jeske, 2010). We treat these techniques as equivalent, given that, conceptually, they should work the same way (and at least one study found no differences between them; Fowler & Barker, 1974, Experiment 2). The techniques typically appeal to students because they are simple to use, do not entail training, and do not require students to invest much time beyond what is already required for reading the material. The question we ask here is, will a technique that is so easy to use actually help students learn? To understand any benefits specific to highlighting and underlining (for brevity, henceforth referred to as *highlighting*), we do not consider studies in which active marking of text was paired with other common techniques, such as note-taking (e.g., Arnold, 1942; L. B. Brown & Smiley, 1978; Mathews, 1938). Although many students report combining multiple techniques (e.g., L. Annis & Davis, 1978; Wade, Trathen, & Schraw, 1990), each technique must be evaluated independently to discover which ones are crucial for success.

4.1 General description of highlighting and underlining and why they should work. As an introduction to the relevant issues, we begin with a description of a prototypical experiment. Fowler and Barker (1974, Exp. 1) had undergraduates read articles (totaling about 8,000 words) about boredom and city life from *Scientific American* and *Science*. Students were assigned to one of three groups: a control group, in which they only read the articles; an active-highlighting group, in which they were free to highlight as much of the texts as they wanted; or a passive-highlighting group, in which they read marked texts that had been highlighted by yoked participants in the active-highlighting group. Everyone received 1 hour to study the texts (time on task was equated across groups); students in the active-highlighting condition were told to mark particularly important material. All subjects returned to the lab 1 week later and were allowed to review their original materials for 10 minutes before taking a 54-item multiple-choice test. Overall, the highlighting groups did not outperform the control group on the final test, a result that has unfortunately been echoed in much of the literature (e.g., Hoon, 1974; Idstein & Jenkins, 1972; Stordahl & Christensen, 1956).

However, results from more detailed analyses of performance in the two highlighting groups are informative about what effects highlighting might have on cognitive processing. First, within the active-highlighting group, performance was better on test items for which the relevant text had been highlighted (see Blanchard & Mikkelson, 1987; L. L. Johnson, 1988 for similar results). Second, this benefit to highlighted information was greater for the active highlighters (who selected what to highlight) than for passive highlighters (who saw the same information highlighted, but did not select it). Third, this benefit to highlighted information was accompanied by a small cost on test questions probing information that had not been highlighted.

To explain such findings, researchers often point to a basic cognitive phenomenon known as the *isolation effect*, whereby a semantically or phonologically unique item in a list is much better remembered than its less distinctive counterparts (see Hunt, 1995, for a description of this work). For instance, if students are studying a list of categorically related words (e.g., "desk," "bed," "chair," "table") and a word from a different category (e.g., "cow") is presented, the students will later be more likely to recall it than they would if it had been studied in a list of categorically related words (e.g., "goat," "pig," "horse," "chicken"). The analogy to highlighting is that a highlighted, underlined, or capitalized sentence will "pop out" of the text in the same way that the word "cow" would if it were isolated in a list of words for types of furniture. Consistent with this expectation, a number of studies have shown that reading marked text promotes later memory for the marked material: Students are more likely to remember things that the experimenter highlighted or underlined in the text (e.g., Cashen & Leicht, 1970; Crouse & Idstein, 1972; Hartley, Bartlett, & Branthwaite, 1980; Klare, Mabry, & Gustafson, 1955; see Lorch, 1989 for a review).

Actively selecting information should benefit memory more than simply reading marked text (given that the former would capitalize on the benefits of generation, Slamecka & Graf, 1978, and active processing more generally, Faw & Waller, 1976). Marked text draws the reader's attention, but additional processing should be required if the reader has to decide which material is most important. Such decisions require the reader to think about the meaning of the text and how its different pieces relate to one another (i.e., organizational processing; Hunt & Worthen, 2006). In the Fowler and Barker (1974) experiment, this benefit was reflected in the greater advantage for highlighted information among active highlighters than among passive recipients of the same highlighted text. However, active highlighting is not always better than receiving material that has already been highlighted by an experimenter (e.g., Nist & Hogrebe, 1987), probably because experimenters will usually be better than students at highlighting the most important parts of a text.

More generally, the quality of the highlighting is likely crucial to whether it helps students to learn (e.g., Wollen, Cone, Britcher, & Mindemann, 1985), but unfortunately, many studies have not contained any measure of the amount or the appropriateness of students' highlighting. Those studies that have examined the amount of marked text have found great variability in what students actually mark, with some students marking almost nothing and others marking almost everything (e.g., Idstein & Jenkins, 1972). Some intriguing data came from the active-highlighting group in Fowler and Barker (1974). Test performance was negatively correlated ($r = –.29$) with the amount of text that had been highlighted in the active-highlighting group, although this result was not significant given the small sample size ($n = 19$).

Marking too much text is likely to have multiple consequences. First, overmarking reduces the degree to which marked text is distinguished from other text, and people are less likely to remember marked text if it is not distinctive (Lorch, Lorch, & Klusewitz, 1995). Second, it likely takes less processing to mark a lot of text than to single out the most important details. Consistent with this latter idea, benefits of marking text may be more likely to be observed when experimenters impose explicit limits on the amount of text students are allowed to mark. For example, Rickards and August (1975) found that students limited to underlining a single sentence per paragraph later recalled more of a science text than did a no-underlining control group. Similarly, L. L. Johnson (1988) found that marking one sentence per paragraph helped college students in a reading class to remember the underlined information, although it did not translate into an overall benefit.

4.2 How general are the effects of highlighting and underlining? We have outlined hypothetical mechanisms by which highlighting might aid memory, and particular features of highlighting that would be necessary for these mechanisms to be effective (e.g., highlighting only important material). However, most studies have shown no benefit of highlighting (as it

is typically used) over and above the benefit of simply reading, and thus the question concerning the generality of the benefits of highlighting is largely moot. Because the research on highlighting has not been particularly encouraging, few investigations have systematically evaluated the factors that might moderate the effectiveness of the technique—for instance, we could not include a Learning Conditions (4.2a) subsection below, given the lack of relevant evidence. To the extent the literature permits, we sketch out the conditions known to moderate the effectiveness of highlighting. We also describe how our conclusion about the relative ineffectiveness of this technique holds across a wide range of situations.

4.2b Student characteristics. Highlighting has failed to help Air Force basic trainees (Stordahl & Christensen, 1956), children (e.g., Rickards & Denner, 1979), and remedial students (i.e., students who scored an average of 390 on the SAT verbal section; Nist & Hogrebe, 1987), as well as prototypical undergraduates (e.g., Todd & Kessler, 1971). It is possible that these groups struggled to highlight only relevant text, given that other studies have suggested that most undergraduates overmark text. Results from one study with airmen suggested that prior knowledge might moderate the effectiveness of highlighting. In particular, the airmen read a passage on aircraft engines that either was unmarked (control condition) or had key information underlined (Klare et al., 1955). The experimenters had access to participants' previously measured mechanical-aptitude scores and linked performance in the experiment to those scores. The marked text was more helpful to airmen who had received high scores. This study involved premarked texts and did not examine what participants would have underlined on their own, but it seems likely that students with little knowledge of a topic would struggle to identify which parts of a text were more or less important (and thus would benefit less from active highlighting than knowledgeable students would).

One other interesting possibility has come from a study in which experimenters extrinsically motivated participants by promising them that the top scorers on an exam would receive $5 (Fass & Schumacher, 1978). Participants read a text about enzymes; half the participants were told to underline key words and phrases. All participants then took a 15-item multiple-choice test. A benefit from underlining was observed among students who could earn the $5 bonus, but not among students in a control group. Thus, although results from this single study need to be replicated, it does appear that some students may have the ability to highlight effectively, but do not always do so.

4.2c Materials. Similar conclusions about marking text have come from studies using a variety of different text materials on topics as diverse as aerodynamics, ancient Greek schools, aggression, and Tanzania, ranging in length from a few hundred words to a few thousand. Todd and Kessler (1971) manipulated text length (all of the materials were relatively short, with lengths of 44, 140, or 256 words) and found that underlining was ineffective regardless of the text length. Fass and Schumacher (1978) manipulated whether a text about enzymes was easy or difficult to read; the easy version was at a seventh-grade reading level, whereas the difficult version was at high school level and contained longer sentences. A larger difference between the highlighting and control groups was found for performance on multiple-choice tests for the difficult text as opposed to the easy text.

4.2d Criterion tasks. A lack of benefit from highlighting has been observed on both immediate and delayed tests, with delays ranging from 1 week to 1 month. A variety of dependent measures have been examined, including free recall, factual multiple-choice questions, comprehension multiple-choice questions, and sentence-completion tests.

Perhaps most concerning are results from a study that suggested that underlining can be detrimental to later ability to make inferences. Peterson (1992) had education majors read a 10,000-word chapter from a history textbook; two groups underlined while studying for 90 minutes, whereas a third group was allowed only to read the chapter. One week later, all groups were permitted to review the material for 15 minutes prior to taking a test on it (the two underlining groups differed in whether they reviewed a clean copy of the original text or one containing their underlining). Everyone received the same test again 2 months later, without having another chance to review the text. The multiple-choice test consisted of 20 items that probed facts (and could be linked to specific references in the text) and 20 items that required inferences (which would have to be based on connections across the text and could not be linked to specific, underlined information). The three groups performed similarly on the factual questions, but students who had underlined (and reviewed their marked texts) were at a disadvantage on the inference questions. This pattern of results requires replication and extension, but one possible explanation for it is that standard underlining draws attention more to individual concepts (supporting memory for facts) than to connections across concepts (as required by the inference questions). Consistent with this idea, in another study, underliners who expected that a final test would be in a multiple-choice format scored higher on it than did underliners who expected it to be in a short-answer format (Kulhavy, Dyer, & Silver, 1975), regardless of the actual format of the final-test questions. Underlined information may naturally line up with the kinds of information students expect on multiple-choice tests (e.g., S. R. Schmidt, 1988), but students may be less sure about what to underline when studying for a short-answer test.

4.5 Effects in representative educational contexts. As alluded to at the beginning of this section, surveys of actual textbooks and other student materials have supported the frequency of highlighting and underlining in educational contexts (e.g., Bell & Limber, 2010; Lonka et al., 1994). Less clear are the consequences of such real-world behaviors. Classroom studies have examined whether instructor-provided markings affect examination performance. For example,

Cashen and Leicht (1970) had psychology students read *Scientific American* articles on animal learning, suicide, and group conflict, each of which contained five critical statements, which were underlined in red for half of the students. The articles were related to course content but were not covered in lectures. Exam scores on items related to the critical statements were higher when the statements had been underlined in red than when they had not. Interestingly, students in the underlining condition also scored better on exam questions about information that had been in sentences adjacent to the critical statements (as opposed to scoring worse on questions about nonunderlined information). The benefit to underlined items was replicated in another psychology class (Leicht & Cashen, 1972), although the effects were weaker. However, it is unclear whether the results from either of these studies would generalize to a situation in which students were in charge of their own highlighting, because they would likely mark many more than five statements in an article (and hence would show less discrimination between important and trivial information).

4.4 Issues for implementation. Students already are familiar with and spontaneously adopt the technique of highlighting; the problem is that the way the technique is typically implemented is not effective. Whereas the technique as it is typically used is not normally detrimental to learning (but see Peterson, 1992, for a possible exception), it may be problematic to the extent that it prevents students from engaging in other, more productive strategies.

One possibility that should be explored is whether students could be trained to highlight more effectively. We located three studies focused on training students to highlight. In two of these cases, training involved one or more sessions in which students practiced reading texts to look for main ideas before marking any text. Students received feedback about practice texts before marking (and being tested on) the target text, and training improved performance (e.g., Amer, 1994; Hayati & Shariatifar, 2009). In the third case, students received feedback on their ability to underline the most important content in a text; critically, students were instructed to underline as little as possible. In one condition, students even lost points for underlining extraneous material (Glover, Zimmer, Filbeck, & Plake, 1980). The training procedures in all three cases involved feedback, and they all had some safeguard against overuse of the technique. Given students' enthusiasm for highlighting and underlining (or perhaps overenthusiasm, given that students do not always use the technique correctly), discovering fail-proof ways to ensure that this technique is used effectively might be easier than convincing students to abandon it entirely in favor of other techniques.

4.5 Highlighting and underlining: Overall assessment. On the basis of the available evidence, we rate highlighting and underlining as having low utility. In most situations that have been examined and with most participants, highlighting does little to boost performance. It may help when students have the knowledge needed to highlight more effectively, or when texts are difficult, but it may actually hurt performance on higher-level tasks that require inference making. Future research should be aimed at teaching students how to highlight effectively, given that students are likely to continue to use this popular technique despite its relative ineffectiveness.

5 The keyword mnemonic

Develop a mental image of students hunched over textbooks, struggling with a science unit on the solar system, trying to learn the planets' names and their order in distance from the sun. Or imagine students in a class on language arts, reading a classic novel, trying to understand the motives of the main characters and how they may act later in the story. By visualizing these students in your "mind's eye," you are using one of the oldest strategies for enhancing learning—dating back to the ancient Greeks (Yates, 1966)—and arguably a powerful one: mental imagery. The earliest systematic research on imagery was begun in the late 1800s by Francis Galton (for a historical review, see Thompson, 1990); since then, many debates have arisen about its nature (e.g., Kosslyn, 1981; Pylyshyn, 1981), such as whether its power accrues from the storage of dual codes (one imaginal and one propositional) or the storage of a distinctive propositional code (e.g., Marschark & Hunt, 1989), and whether mental imagery is subserved by the same brain mechanisms as visual imagery (e.g., Goldenberg, 1998).

Few of these debates have been entirely resolved, but fortunately, their resolution is not essential for capitalizing on the power of mental imagery. In particular, it is evident that the use of imagery can enhance learning and comprehension for a wide variety of materials and for students with various abilities. A review of this entire literature would likely go beyond a single monograph or perhaps even a book, given that mental imagery is one of the most highly investigated mental activities and has inspired enough empirical research to warrant its own publication (i.e., the *Journal of Mental Imagery*). Instead of an exhaustive review, we briefly discuss two specific uses of mental imagery for improving student learning that have been empirically scrutinized: the use of the keyword mnemonic for learning foreign-language vocabulary, and the use of mental imagery for comprehending and learning text materials.

5.1 General description of the keyword mnemonic and why it works. Imagine a student struggling to learn French vocabulary, including words such as *la dent* (tooth), *la clef* (key), *revenir* (to come back), and *mourir* (to die). To facilitate learning, the student uses the keyword mnemonic, which is a technique based on interactive imagery that was developed by Atkinson and Raugh (1975). To use this mnemonic, the student would first find an English word that sounds similar to the foreign cue word, such as *dentist* for "la dent" or *cliff* for

"la clef." The student would then develop a mental image of the English keyword interacting with the English translation. So, for *la dent–tooth*, the student might imagine a dentist holding a large molar with a pair of pliers. Raugh and Atkinson (1975) had college students use the keyword mnemonic to learn Spanish-English vocabulary (e.g., *gusano–worm*): the students first learned to associate each experimenter-provided keyword with the appropriate Spanish cue (e.g., "gusano" is associated with the keyword "goose"), and then they developed interactive images to associate the keywords with their English translations. In a later test, the students were asked to generate the English translation when presented with the Spanish cue (e.g., "gusano"–?). Students who used the keyword mnemonic performed significantly better on the test than did a control group of students who studied the translation equivalents without keywords.

Beyond this first demonstration, the potential benefits of the keyword mnemonic have been extensively explored, and its power partly resides in the use of interactive images. In particular, the interactive image involves elaboration that integrates the words meaningfully, and the images themselves should help to distinguish the sought-after translation from other candidates. For instance, in the example above, the image of the "large molar" distinguishes "tooth" (the target) from other candidates relevant to dentists (e.g., gums, drills, floss). As we discuss next, the keyword mnemonic can be effectively used by students of different ages and abilities for a variety of materials. Nevertheless, our analysis of this literature also uncovered limitations of the keyword mnemonic that may constrain its utility for teachers and students. Given these limitations, we did not separate our review of the literature into separate sections that pertain to each variable category (Table 2) but instead provide a brief overview of the most relevant evidence concerning the generalizability of this technique.

5.2 a–d How general are the effects of the keyword mnemonic? The benefits of the keyword mnemonic generalize to many different kinds of material: (a) foreign-language vocabulary from a variety of languages (French, German, Italian, Latin, Russian, Spanish, and Tagalog); (b) the definitions of obscure English vocabulary words and science terms; (c) state-capital associations (e.g., Lincoln is the capital of Nebraska); (d) medical terminology; (e) people's names and accomplishments or occupations; and (f) minerals and their attributes (e.g., the mineral wolframite is soft, dark in color, and used in the home). Equally impressive, the keyword mnemonic has also been shown to benefit learners of different ages (from second graders to college students) and students with learning disabilities (for a review, see Jitendra, Edwards, Sacks, & Jacobson, 2004). Although the bulk of research on the keyword mnemonic has focused on students' retention of target materials, the technique has also been shown to improve students' performance on a variety of transfer tasks: It helps them (a) to generate appropriate sentences using newly learned English vocabulary (McDaniel & Pressley, 1984) and (b) to adapt newly acquired vocabulary to semantically novel contexts (Mastropieri, Scruggs, & Mushinski Fulk, 1990).

The overwhelming evidence that the keyword mnemonic can boost memory for many kinds of material and learners has made it a relatively popular technique. Despite the impressive outcomes, however, some aspects of these demonstrations imply limits to the utility of the keyword mnemonic. First, consider the use of this technique for its originally intended domain—the learning of foreign-language vocabulary. In the example above, *la dent* easily supports the development of a concrete keyword ("dentist") that can be easily imagined, whereas many vocabulary terms are much less amenable to the development and use of keywords. In the case of *revenir* (to come back), a student could perhaps use the keyword "revenge" (e.g., one might need "to come back" to taste its sweetness), but imaging this abstract term would be difficult and might even limit retention. Indeed, Hall (1988) found that a control group (which received task practice but no specific instructions on how to study) outperformed a keyword group in a test involving English definitions that did not easily afford keyword generation, even when the keywords were provided. Proponents of the keyword mnemonic do acknowledge that its benefits may be limited to keyword-friendly materials (e.g., concrete nouns), and in fact, the vast majority of the research on the keyword mnemonic has involved materials that afforded its use.

Second, in most studies, the keywords have been provided by the experimenters, and in some cases, the interactive images (in the form of pictures) were provided as well. Few studies have directly examined whether students can successfully generate their own keywords, and those that have have offered mixed results: Sometimes students' self-generated keywords facilitate retention as well as experimenter-provided keywords do (Shapiro & Waters, 2005), and sometimes they do not (Shriberg, Levin, McCormick, & Pressley, 1982; Thomas & Wang, 1996). For more complex materials (e.g., targets with multiple attributes, as in the wolframite example above), the experimenter-provided "keywords" were pictures, which some students may have difficulties generating even after extensive training. Finally, young students who have difficulties generating images appear to benefit from the keyword mnemonic only if keywords and an associated interactive image (in the form of a picture) are supplied during learning (Pressley & Levin, 1978). Thus, although teachers who are willing to construct appropriate keywords may find this mnemonic useful, even these teachers (and students) would be able to use the technique only for subsets of target materials that are keyword friendly.

Third, and perhaps most disconcerting, the keyword mnemonic may not produce durable retention. Some of the studies investigating the long-term benefits of the keyword mnemonic included a test soon after practice as well as one after a longer delay of several days or even weeks (e.g., Condus, Marshall, & Miller, 1986; Raugh & Atkinson, 1975). These studies

generally demonstrated a benefit of keywords at the longer delay (for a review, see Wang, Thomas, & Ouellette, 1992). Unfortunately, these promising effects were compromised by the experimental designs. In particular, all items were tested on *both* the immediate and delayed tests. Given that the keyword mnemonic yielded better performance on the immediate tests, this initial increase in successful recall could have boosted performance on the delayed tests and thus inappropriately disadvantaged the control groups. Put differently, the advantage in delayed test performance could have been largely due to the effects of retrieval practice (i.e., from the immediate test) and not to the use of keyword mnemonics per se (because retrieval can slow forgetting; see the Practice Testing section below).

This possibility was supported by data from Wang et al. (1992; see also Wang & Thomas, 1995), who administered immediate and delayed tests to *different* groups of students. As shown in Figure 4 (top panel), for participants who received the immediate test, the keyword-mnemonic group outperformed a rote-repetition control group. By contrast, this benefit vanished for participants who received only the delayed test. Even more telling, as shown in the bottom panel of Figure 4, when the researchers equated the performance of the two groups on the immediate test (by giving the rote-repetition group more practice), performance on the delayed test was significantly better for the rote-repetition group than for the keyword-mnemonic group (Wang et al., 1992).

These data suggest that the keyword mnemonic leads to accelerated forgetting. One explanation for this surprising outcome concerns decoding at retrieval: Students must decode each image to retrieve the appropriate target, and at longer delays, such decoding may be particularly difficult. For instance, when a student retrieves "a dentist holding a large molar with a pair of pliers," he or she may have difficulty deciding whether the target is "molar," "tooth," "pliers," or "enamel."

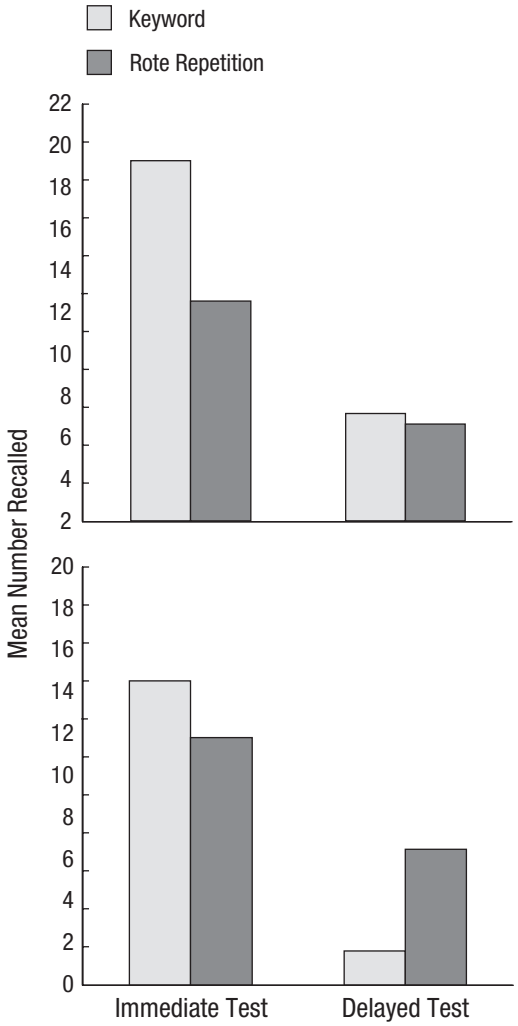


Fig. 4. Mean number of items correctly recalled on a cued-recall test occurring soon after study (immediate test) or 1 week after study (delayed test) in Wang, Thomas, and Ouellette (1992). Values in the top panel are from Experiment 1, and those in the bottom panel are from Experiment 3. Standard errors are not available.

5.3 Effects in representative educational contexts. The keyword mnemonic has been implemented in classroom settings, and the outcomes have been mixed. On the promising side, Levin, Pressley, McCormick, Miller, and Shriberg (1979) had fifth graders use the keyword mnemonic to learn Spanish vocabulary words that were keyword friendly. Students were trained to use the mnemonic in small groups or as an entire class, and in both cases, the groups who used the keyword mnemonic performed substantially better than did control groups who were encouraged to use their own strategies while studying. Less promising are results for high school students who Levin et al. (1979) trained to use the keyword mnemonic. These students were enrolled in a 1st-year or 2nd-year language course, which is exactly the context in which one would expect the keyword mnemonic to help. However, the keyword mnemonic did not benefit recall, regardless of whether students were trained individually or in groups. Likewise, Willerman and Melvin (1979) did not find benefits of keyword-mnemonic training for college students enrolled in an elementary French course (cf. van Hell & Mahn, 1997; but see Lawson & Hogben, 1998).

5.4 Issues for implementation. The majority of research on the keyword mnemonic has involved at least some (and occasionally extensive) training, largely aimed at helping students develop interactive images and use them to subsequently retrieve targets. Beyond training, implementation also requires the development of keywords, whether by students, teachers, or textbook designers. The effort involved in generating some keywords may not be the most efficient use of time for students (or teachers), particularly given that at least one easy-to-use technique (i.e., retrieval practice, Fritz, Morris, Acton, Voelkel, & Etkind, 2007) benefits retention as much as the keyword mnemonic does.

5.5 The keyword mnemonic: Overall assessment. On the basis of the literature reviewed above, we rate the keyword mnemonic as low utility. We cannot recommend that the keyword mnemonic be widely adopted. It does show promise for keyword-friendly materials, but it is not highly efficient (in terms of time needed for training and keyword generation), and it may not produce durable learning. Moreover, it is not clear that students will consistently benefit from the keyword mnemonic when they have to generate keywords; additional research is needed to more fully explore the effectiveness of keyword generation (at all age levels) and whether doing so is an efficient use of students' time, as compared to other strategies. In one head-to-head comparison, cued recall of foreign-language vocabulary was either no different after using the keyword mnemonic (with experimenter-provided keywords) than after practice testing, or was lower on delayed criterion tests 1 week later (Fritz, Morris, Acton, et al., 2007). Given that practice testing is easier to use and more broadly applicable (as reviewed below in the Practice Testing section), it seems superior to the keyword mnemonic.

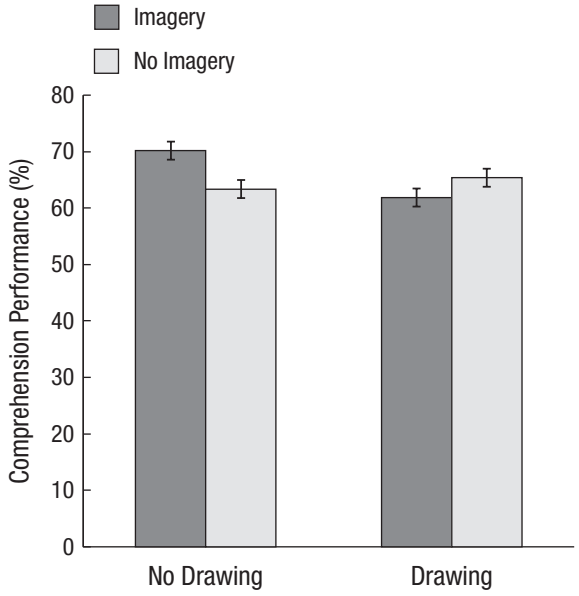


Fig. 5. Accuracy on a multiple-choice exam in which answers had to be inferred from a text in Leutner, Leopold, and Sumfleth (2009). Participants either did or did not receive instructions to use imagery while reading, and either did or did not draw pictures to illustrate the content of the text. Error bars represent standard errors.

6 Imagery use for text learning

6.1 General description of imagery use and why it should work. In one demonstration of the potential of imagery for enhancing text learning, Leutner, Leopold, and Sumfleth (2009) gave tenth graders 35 minutes to read a lengthy science text on the dipole character of water molecules. Students either were told to read the text for comprehension (control group) or were told to read the text and to mentally imagine the content of each paragraph using simple and clear mental images. Imagery instructions were also crossed with drawing: Some students were instructed to draw pictures that represented the content of each paragraph, and others did not draw. Soon after reading, the students took a multiple-choice test that included questions for which the correct answer was not directly available from the text but needed to be inferred from it. As shown in Figure 5, the instructions to mentally imagine the content of each paragraph significantly boosted the comprehension-test performance of students in the mental-imagery group, in comparison to students in the control group (Cohen's $d = 0.72$). This effect is impressive, especially given that (a) training was not required, (b) the text involved complex science content, and (c) the criterion test required learners to make inferences about the content. Finally, drawing did not improve comprehension, and it actually negated the benefits of imagery instructions. The potential for another activity to interfere with the potency of imagery is discussed further in the subsection on learning conditions (6.2a) below.

A variety of mechanisms may contribute to the benefits of imaging text material on later test performance. Developing images can enhance one's mental organization or integration of information in the text, and idiosyncratic images of particular referents in the text could enhance learning as well (cf. distinctive processing; Hunt, 2006). Moreover, using one's prior knowledge to generate a coherent representation of a narrative may enhance a student's general understanding of the text; if so, the influence of imagery use may be robust across criterion tasks that tap memory and comprehension. Despite these possibilities and the dramatic effect of imagery demonstrated by Leutner et al. (2009), our review of the literature suggests that the effects of using mental imagery to learn from text may be rather limited and not robust.

6.2 How general are the effects of imagery use for text learning? Investigations of imagery use for learning text materials have focused on single sentences and longer text materials. Evidence concerning the impact of imagery on sentence learning largely comes from investigations of other mnemonic techniques (e.g., elaborative interrogation) in which imagery instructions have been included in a comparison condition. This research has typically demonstrated that groups who receive imagery instructions have better memory for sentences than do no-instruction control groups (e.g., R. C. Anderson & Hidde, 1971; Wood, Pressley, & Winne, 1990). In the remainder of this section, we focus on the degree to which imagery instructions improve learning for longer text materials.

6.2a Learning conditions. Learning conditions play a potentially important role in moderating the benefits of imagery, so we briefly discuss two conditions here—namely, the modality of text presentation and learners' actual use of imagery after receiving imagery instructions. Modality pertains to whether students are asked to use imagery as they read a text or as they listen to a narration of a text. L. R. Brooks (1967, 1968)

reported that participants' visualization of a pathway through a matrix was disrupted when they had to read a description of it; by contrast, visualization was not disrupted when participants listened to the description. Thus, it is possible that the benefits of imagery are not fully actualized when students read text and would be most evident if they listened. Two observations are relevant to this possibility. First, the majority of imagery research has involved students *reading* texts; the fact that imagery benefits have sometimes been found indicates that reading does not entirely undermine imaginal processing. Second, in experiments in which participants either read or listened to a text, the results have been mixed. As expected, imagery has benefited performance more among students who have listened to texts than among students who have read them (De Beni & Moè, 2003; Levin & Divine-Hawkins, 1974), but in one case, imagery benefited performance similarly for both modalities in a sample of fourth graders (Maher & Sullivan, 1982).

The actual use of imagery as a learning technique should also be considered when evaluating the imagery literature. In particular, even if students are instructed to use imagery, they may not necessarily use it. For instance, R. C. Anderson and Kulhavy (1972) had high school seniors read a lengthy text passage about a fictitious primitive tribe; some students were told to generate images while reading, whereas others were told to read carefully. Imagery instructions did not influence performance, but reported use of imagery was significantly correlated with performance (see also Denis, 1982). The problem here is that some students who were instructed to use imagery did not, whereas some uninstructed students spontaneously used it. Both circumstances would reduce the observed effect of imagery instructions, and students' spontaneous use of imagery in control conditions may be partly responsible for the failure of imagery to benefit performance in some cases. Unfortunately, researchers have typically not measured imagery use, so evaluation of these possibilities must await further research.

6.2b Student characteristics. The efficacy of imagery instructions have been evaluated across a wide range of student ages and abilities. Consider data from studies involving fourth graders, given that this particular grade level has been popular in imagery research. In general, imagery instructions have tended to boost criterion performance for fourth graders, but even here the exceptions are noteworthy. For instance, imagery instructions boosted the immediate test performance of fourth graders who studied short (e.g., 12-sentence) stories that could be pictorially represented (e.g., Levin & Divine-Hawkins, 1974), but in some studies, this benefit was found only for students who were biased to use imagery or for skilled readers (Levin, Divine-Hawkins, Kerst, & Guttman, 1974). For reading longer narratives (e.g., narratives of 400 words or more), imagery instructions have significantly benefited fourth graders' free recall of text material (Gambrell & Jawitz, 1993; Rasco, Tennyson, & Boutwell, 1975; see also Lesgold, McCormick, & Golinkoff, 1975) and performance on multiple-choice questions about the text (Maher & Sullivan, 1982; this latter benefit was apparent for both high- and low-skilled readers), but even after extensive training and a reminder to use imagery, fourth graders' performance on a standardized reading-comprehension test did not improve (Lesgold et al., 1975).

Despite the promise of imagery, this patchwork of inconsistent effects for fourth graders has also been found for students of other ages. College students have been shown to reap the benefits of imagery, but these benefits depend on the nature of the criterion test (an issue we discuss below). In two studies, high school students who read a long passage did not benefit from imagery instructions (R. C. Anderson & Kulhavy, 1972; Rasco et al., 1975). Studies with fifth and sixth grade students have shown some benefits of imagery, but these trends have not all been significant (Kulhavy & Swenson, 1975) and did not arise on some criterion tests (e.g., standardized achievement tests; Miccinati, 1982). Third graders have been shown to benefit from using imagery (Oakhill & Patel, 1991; Pressley, 1976), but younger students do not appear to benefit from attempting to generate mental images when listening to a story (Guttman, Levin, & Pressley, 1977).

6.2c Materials. Similar to studies on the keyword mnemonic, investigations of imagery use for text learning have often used texts that are imagery friendly, such as narratives that can be visualized or short stories that include concrete terms. Across investigations, the specific texts have varied widely and include long passages (of 2,000 words or more; e.g., R. C. Anderson & Kulhavy, 1972; Giesen & Peeck, 1984), relatively short stories (e.g., L. K. S. Chan, Cole, & Morris, 1990; Maher & Sullivan, 1982), and brief 10-sentence passages (Levin & Divine-Hawkins, 1974; Levin et al., 1974). With regard to these variations in materials, the safest conclusion is that sometimes imagery instructions boost performance and sometimes they do not. The literature is filled with interactions whereby imagery helped for one kind of material but not for another kind of material. In these cases, failures to find an effect for any given kind of material may not be due to the material per se, but instead may reflect the effect of other, uncontrolled factors, making it is impossible to tell which (if any) characteristics of the materials predict whether imagery will be beneficial.

Fortunately, some investigators have manipulated the content of text materials when examining the benefits of imagery use. In De Beni and Moè (2003), one text included descriptions that were easy to imagine, another included a spatial description of a pathway that was easy to imagine and verbalize, and another was abstract and presumably not easy to imagine. As compared with instructions to just rehearse the texts, instructions to use imagery benefited free recall of the easy-to-imagine texts and the spatial texts but did not benefit recall of the abstract texts. Moreover, the benefits were evident only when students listened to the text, not when they read it (as discussed under "Learning Conditions," 6.2a, above). Thus, the benefits of imagery may be largely constrained to texts that directly support imaginal representations.

Although the bulk of the research on imagery has used texts that were specifically chosen to support imagery, two studies have used the Metropolitan Achievement Test, which is a standardized test that taps comprehension. Both studies used extensive training in the use of imagery while reading, and both studies failed to find an effect of imagery training on test performance (Lesgold, et al., 1975; Miccinati, 1982), even when participants were explicitly instructed to use their trained skills to complete the test (Lesgold et al., 1975).

6.2d Criterion tasks. The inconsistent benefits of imagery within groups of students can in part be explained by interactions between imagery (vs. reading) instructions and the criterion task. Consider first the results from studies involving college students. When the criterion test comprises free-recall or short-answer questions tapping information explicitly stated in the text, college students tend to benefit from instructions to image (e.g., Gyeselinck, Meneghetti, De Beni, & Pazzaglia, 2009; Hodes, 1992; Rasco et al., 1975; although, as discussed earlier, these effects may be smaller when students read the passages rather than listen to them; De Beni & Moè, 2003). By contrast, despite the fact that imagery presumably helps students develop an integrated visual model of a text, imagery instructions did not significantly help college students answer questions that required them to make inferences based on information in a text (Giesen & Peeck, 1984) or comprehension questions about a passage on the human heart (Hodes, 1992).

This pattern is also apparent from studies with sixth graders, who do show significant benefits of imagery use on measures involving the recall or summarization of text information (e.g., Kulhavy & Swenson, 1975), but show reduced or nonexistent benefits on comprehension tests and on criterion tests that require application of the knowledge (Gagne & Memory, 1978; Miccinati, 1982). In general, imagery instructions tend not to enhance students' understanding or application of the content of a text. One study demonstrated that training improved 8- and 9-year-olds' performance on inference questions, but in this case, training was extensive (three sessions), which may not be practical in some settings.

When imagery instructions do improve criterion performance, a question arises as to whether these effects are long lasting. Unfortunately, the question of whether the use of imagery protects against the forgetting of text content has not been widely investigated; in the majority of studies, criterion tests have been administered immediately or shortly after the target material was studied. In one exception, Kulhavy and Swenson (1975) found that imagery instructions benefited fifth and sixth graders' accuracy in answering questions that tapped the gist of the texts, and this effect was even apparent 1 week after the texts were initially read. The degree to which these long-term benefits are robust and generalize across a variety of criterion tasks is an open question.

6.3 Effects in representative educational contexts. Many of the studies on imagery use and text learning have involved students from real classrooms who were reading texts that were written to match the students' grade level. Most studies have used fabricated materials, and few studies have used authentic texts that students would read. Exceptions have involved the use of a science text on the dipole character of water molecules (Leutner et al., 2009) and texts on cause-effect relationships that were taken from real science and social-science textbooks (Gagne & Memory, 1978); in both cases, imagery instructions improved test performance (although the benefits were limited to a free-recall test in the latter case). Whether instructions to use imagery will help students learn materials in a manner that will translate into improved course grades is unknown, and research investigating students' performance on achievement tests has shown imagery use to be a relatively inert strategy (Lesgold et al., 1975; Miccinati, 1982; but see Rose, Parks, Androes, & McMahon, 2000, who supplemented imagery by having students act out narrative stories).

6.4 Issues for implementation. The majority of studies have examined the influence of imagery by using relatively brief instructions that encouraged students to generate images of text content while studying. Given that imagery does not appear to undermine learning (and that it does boost performance in some conditions), teachers may consider instructing students (third grade and above) to attempt to use imagery when they are reading texts that easily lend themselves to imaginal representations. How much training would be required to ensure that students consistently and effectively use imagery under the appropriate conditions is unknown.

6.5 Imagery use for learning text: Overall assessment. Imagery can improve students' learning of text materials, and the promising work by Leutner et al. (2009) speaks to the potential utility of imagery use for text learning. Imagery production is also more broadly applicable than the keyword mnemonic. Nevertheless, the benefits of imagery are largely constrained to imagery-friendly materials and to tests of memory, and further demonstrations of the effectiveness of the technique (across different criterion tests and educationally relevant retention intervals) are needed. Accordingly, we rated the use of imagery for learning text as low utility.

7 Rereading

Rereading is one of the techniques that students most frequently report using during self-regulated study (Carrier, 2003; Hartwig & Dunlosky, 2012; Karpicke, Butler, & Roediger, 2009; Kornell & Bjork, 2007; Wissman, Rawson, & Pyc, 2012). For example, Carrier (2003) surveyed college students in an upper-division psychology course, and 65% reported using rereading as a technique when preparing for course exams. More recent surveys have reported similar results. Kornell and Bjork (2007) and Hartwig and Dunlosky (2012) asked students if they typically read a textbook, article, or

other source material more than once during study. Across these two studies, 18% of students reported rereading entire chapters, and another 62% reported rereading parts or sections of the material. Even high-performing students appear to use rereading regularly. Karpicke et al. (2009) asked undergraduates at an elite university (where students' average SAT scores were above 1400) to list all of the techniques they used when studying and then to rank them in terms of frequency of use. Eighty-four percent of students included rereading textbook/notes in their list, and rereading was also the top-ranked technique (listed as the most frequently used technique by 55% of students). Students' heavy reliance on rereading during self-regulated study raises an important question: Is rereading an effective technique?

7.1 General description of rereading and why it should work. In an early study by Rothkopf (1968), undergraduates read an expository text (either a 1,500-word passage about making leather or a 750-word passage about Australian history) zero, one, two, or four times. Reading was self-paced, and rereading was *massed* (i.e., each presentation of a text occurred immediately after the previous presentation). After a 10-minute delay, a cloze test was administered in which 10% of the content words were deleted from the text and students were to fill in the missing words. As shown in Figure 6, performance improved as a function of number of readings.

Why does rereading improve learning? Mayer (1983; Bromage & Mayer, 1986) outlined two basic accounts of rereading effects. According to the *quantitative hypothesis*, rereading simply increases the total amount of information encoded, regardless of the kind or level of information within the text. In contrast, the *qualitative hypothesis* assumes that rereading differentially affects the processing of higher-level and lower-level information within a text, with particular emphasis placed on the conceptual organization and processing of main ideas during rereading. To evaluate these hypotheses, several studies have examined free recall as a function of the kind or level of text information. The results have been somewhat mixed, but the evidence appears to favor the qualitative hypothesis. Although a few studies found that rereading produced similar improvements in the recall of main ideas and of details (a finding consistent with the quantitative hypothesis), several studies have reported greater improvement in the recall of main ideas than in the recall of details (e.g., Bromage & Mayer, 1986; Kiewra, Mayer, Christensen, Kim, & Risch, 1991; Rawson & Kintsch, 2005).

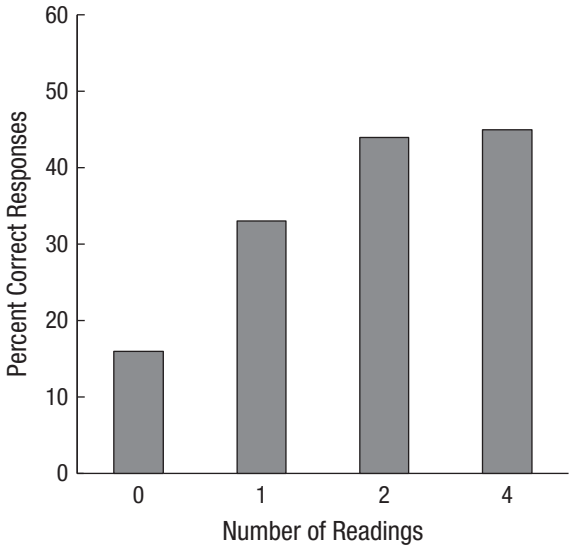


Fig. 6. Mean percentage of correct responses on a final cloze test for learners who read an expository text zero, one, two, or four times in Rothkopf (1968). Means shown are overall means for two conditions, one in which learners read a 1,500-word text and one in which learners read a 750-word text. Values are estimated from original figures in Rothkopf (1968). Standard errors are not available.

7.2 How general are the effects of rereading?

7.2a Learning conditions. Following the early work of Rothkopf (1968), subsequent research established that the effects of rereading are fairly robust across other variations in learning conditions. For example, rereading effects obtain regardless of whether learners are forewarned that they will be given the opportunity to study more than once, although Barnett and Seefeldt (1989) found a small but significant increase in the magnitude of the rereading effect among learners who were forewarned, relative to learners who were not forewarned. Furthermore, rereading effects obtain with both self-paced reading and experimenter-paced presentation. Although most studies have involved the silent reading of written material, effects of repeated presentations have also been shown when learners listen to an auditory presentation of text material (e.g., Bromage & Mayer, 1986; Mayer, 1983).[2]

One aspect of the learning conditions that does significantly moderate the effects of rereading concerns the lag between initial reading and rereading. Although advantages of rereading over reading only once have been shown with massed rereading and with *spaced* rereading (in which some amount of time passes or intervening material is presented between initial study and restudy), spaced rereading usually outperforms massed rereading. However, the relative advantage of spaced reading over massed rereading may be moderated by the length of the retention interval, an issue that we discuss further in the subsection on criterion tasks below (7.2d). The effect of spaced rereading may also depend on the length of the lag between initial study and restudy. In a recent study by Verkoeijen, Rikers, and Özsoy (2008), learners read a lengthy expository text and then reread it immediately afterward, 4 days later, or 3.5 weeks later. Two days after rereading, all participants completed a final test. Performance was greater for the group who reread after a 4-day lag than for the massed rereaders, whereas performance for the group who reread after a 3.5-week lag was intermediate and did not significantly differ from performance in either of the other two groups. With that said, spaced rereading appears to be effective at least across

moderate lags, with studies reporting significant effects after lags of several minutes, 15–30 minutes, 2 days, and 1 week.

One other learning condition that merits mention is amount of practice, or dosage. Most of the benefits of rereading over a single reading appear to accrue from the second reading: The majority of studies that have involved two levels of rereading have shown diminishing returns from additional rereading trials. However, an important caveat is that all of these studies involved massed rereading. The extent to which additional spaced rereading trials produce meaningful gains in learning remains an open question.

Finally, although learners in most experiments have studied only one text, rereading effects have also been shown when learners are asked to study several texts, providing suggestive evidence that rereading effects can withstand interference from other learning materials.

7.2b Student characteristics. The extant literature is severely limited with respect to establishing the generality of rereading effects across different groups of learners. To our knowledge, all but two studies of rereading effects have involved undergraduate students. Concerning the two exceptions, Amlund, Kardash, and Kulhavy (1986) reported rereading effects with graduate students, and O'Shea, Sindelar, and O'Shea (1985) reported effects with third graders.

The extent to which rereading effects depend on knowledge level is also woefully underexplored. In the only study to date that has provided any evidence about the extent to which knowledge may moderate rereading effects (Arnold, 1942), both high-knowledge and low-knowledge readers showed an advantage of massed rereading over outlining or summarizing a passage for the same amount of time. Additional suggestive evidence that relevant background knowledge is not requisite for rereading effects has come from three recent studies that used the same text (Rawson, 2012; Rawson & Kintsch, 2005; Verkoeijen et al., 2008) and found significant rereading effects for learners with virtually no specific prior knowledge about the main topics of the text (the charge of the Light Brigade in the Crimean War and the Hollywood film portraying the event).

Similarly, few studies have examined rereading effects as a function of ability, and the available evidence is somewhat mixed. Arnold (1942) found an advantage of massed rereading over outlining or summarizing a passage for the same amount of time among learners with both higher and lower levels of intelligence and both higher and lower levels of reading ability (but see Callender & McDaniel, 2009, who did not find an effect of massed rereading over single reading for either higher- or lower-ability readers). Raney (1993) reported a similar advantage of massed rereading over a single reading for readers with either higher or lower working-memory spans. Finally, Barnett and Seefeldt (1989) defined high- and low-ability groups by a median split of ACT scores; both groups showed an advantage of massed rereading over a single reading for short-answer factual questions, but only high-ability learners showed an effect for questions that required application of the information.

7.2c Materials. Rereading effects are robust across variations in the length and content of text material. Although most studies have used expository texts, rereading effects have also been shown for narratives. Those studies involving expository text material have used passages of considerably varying lengths, including short passages (e.g., 99–125 words), intermediate passages (e.g., 390–750 words), lengthy passages (e.g., 900–1,500 words), and textbook chapters or magazine articles with several thousand words. Additionally, a broad range of content domains and topics have been covered—an illustrative but nonexhaustive list includes physics (e.g., Ohm's law), law (e.g., legal principles of evidence), history (e.g., the construction of the Brooklyn Bridge), technology (e.g., how a camera exposure meter works), biology (e.g., insects), geography (e.g., of Africa), and psychology (e.g., the treatment of mental disorders).

7.2d Criterion tasks. Across rereading studies, the most commonly used outcome measure has been free recall, which has consistently shown effects of both massed and spaced rereading with very few exceptions. Several studies have also shown rereading effects on cue-based recall measures, such as fill-in-the-blank tests and short-answer questions tapping factual information. In contrast, the effects of rereading on recognition are less certain, with weak or nonexistent effects on sentence-verification tasks and multiple-choice questions tapping information explicitly stated in the text (Callender & McDaniel, 2009; Dunlosky & Rawson, 2005; Hinze & Wiley, 2011; Kardash & Scholes, 1995). The evidence concerning the effects of rereading on comprehension is somewhat muddy. Although some studies have shown positive effects of rereading on answering problem-solving essay questions (Mayer, 1983) and short-answer application or inference questions (Karpicke & Blunt, 2011; Rawson & Kintsch, 2005), other studies using application or inference-based questions have reported effects only for higher-ability students (Barnett & Seefeldt, 1989) or no effects at all (Callender & McDaniel, 2009; Dunlosky & Rawson, 2005; Durgunoğlu, Mir, & Ariño-Martí, 1993; Griffin, Wiley, & Thiede, 2008).

Concerning the durability of learning, most of the studies that have shown significant rereading effects have administered criterion tests within a few minutes after the final study trial, and most of these studies reported an advantage of massed rereading over a single reading. The effects of massed rereading after longer delays are somewhat mixed. Agarwal, Karpicke, Kang, Roediger, and McDermott (2008; see also Karpicke & Blunt, 2011) reported massed rereading effects after 1 week, but other studies have failed to find significant effects after 1–2 days (Callender & McDaniel, 2009; Cranney, Ahn, McKinnon, Morris, & Watts, 2009; Hinze & Wiley, 2011; Rawson & Kintsch, 2005).

Fewer studies have involved spaced rereading, although a relatively consistent advantage for spaced rereading over a single reading has been shown both on immediate tests and on tests administered after a 2-day delay. Regarding the comparison of massed rereading with spaced rereading, neither

schedule shows a consistent advantage on immediate tests. A similar number of studies have shown an advantage of spacing over massing, an advantage of massing over spacing, and no differences in performance. In contrast, spaced rereading consistently outperforms massed rereading on delayed tests. We explore the benefits of spacing more generally in the Distributed Practice section below.

7.3 Effects in representative educational contexts. Given that rereading is the study technique that students most commonly report using, it is perhaps ironic that no experimental research has assessed its impact on learning in educational contexts. Although many of the topics of the expository texts used in rereading research are arguably similar to those that students might encounter in a course, none of the aforementioned studies have involved materials taken from actual course content. Furthermore, none of the studies were administered in the context of a course, nor have any of the outcome measures involved course-related tests. The only available evidence involves correlational findings reported in survey studies, and it is mixed. Carrier (2003) found a nonsignificant negative association between self-reported rereading of textbook chapters and exam performance but a significantly positive association between self-reported review of lecture notes and exam performance. Hartwig and Dunlosky (2012) found a small but significant positive association between self-reported rereading of textbook chapters or notes and self-reported grade point average, even after controlling for self-reported use of other techniques.

7.4 Issues for implementation. One advantage of rereading is that students require no training to use it, other than perhaps being instructed that rereading is generally most effective when completed after a moderate delay rather than immediately after an initial reading. Additionally, relative to some other learning techniques, rereading is relatively economical with respect to time demands (e.g., in those studies permitting self-paced study, the amount of time spent rereading has typically been less than the amount of time spent during initial reading). However, in head-to-head comparisons of learning techniques, rereading has not fared well against some of the more effective techniques discussed here. For example, direct comparisons of rereading to elaborative interrogation, self-explanation, and practice testing (described in the Practice Testing section below) have consistently shown rereading to be an inferior technique for promoting learning.

7.5 Rereading: Overall assessment. Based on the available evidence, we rate rereading as having low utility. Although benefits from rereading have been shown across a relatively wide range of text materials, the generality of rereading effects across the other categories of variables in Table 2 has not been well established. Almost no research on rereading has involved learners younger than college-age students, and an insufficient amount of research has systematically examined the extent to which rereading effects depend on other student characteristics, such as knowledge or ability. Concerning criterion tasks, the effects of rereading do appear to be durable across at least modest delays when rereading is spaced. However, most effects have been shown with recall-based memory measures, whereas the benefit for comprehension is less clear. Finally, although rereading is relatively economical with respect to time demands and training requirements when compared with some other learning techniques, rereading is also typically much less effective. The relative disadvantage of rereading to other techniques is the largest strike against rereading and is the factor that weighed most heavily in our decision to assign it a rating of low utility.

8 Practice testing

Testing is likely viewed by many students as an undesirable necessity of education, and we suspect that most students would prefer to take as few tests as possible. This view of testing is understandable, given that most students' experience with testing involves high-stakes summative assessments that are administered to evaluate learning. This view of testing is also unfortunate, because it overshadows the fact that testing also *improves* learning. Since the seminal study by Abbott (1909), more than 100 years of research has yielded several hundred experiments showing that practice testing enhances learning and retention (for recent reviews, see Rawson & Dunlosky, 2011; Roediger & Butler, 2011; Roediger, Putnam, & Smith, 2011). Even in 1906, Edward Thorndike recommended that "the active recall of a fact from within is, as a rule, better than its impression from without" (p. 123, Thorndike, 1906). The century of research on practice testing since then has supported Thorndike's recommendation by demonstrating the broad generalizability of the benefits of practice testing.

Note that we use the term *practice testing* here (a) to distinguish testing that is completed as a low-stakes or no-stakes practice or learning activity outside of class from summative assessments that are administered by an instructor in class, and (b) to encompass any form of practice testing that students would be able to engage in on their own. For example, practice testing could involve practicing recall of target information via the use of actual or virtual flashcards, completing practice problems or questions included at the end of textbook chapters, or completing practice tests included in the electronic supplemental materials that increasingly accompany textbooks.

8.1 General description of practice testing and why it should work. As an illustrative example of the power of testing, Runquist (1983) presented undergraduates with a list of word pairs for initial study. After a brief interval during which participants completed filler tasks, half of the pairs were tested via cued recall and half were not. Participants completed a final cued-recall test for all pairs either 10 minutes or 1 week later. Final-test performance was better for pairs that were practice tested than pairs that were not (53% versus 36% after

10 minutes, 35% versus 4% after 1 week). Whereas this study illustrates the method of comparing performance between conditions that do and do not involve a practice test, many other studies have compared a practice-testing condition with more stringent conditions involving additional presentations of the to-be-learned information. For example, Roediger and Karpicke (2006b) presented undergraduates with a short expository text for initial study followed either by a second study trial or by a practice free-recall test. One week later, free recall was considerably better among the group that had taken the practice test than among the group that had restudied (56% versus 42%). As another particularly compelling demonstration of the potency of testing as compared with restudy, Karpicke and Roediger (2008) presented undergraduates with Swahili-English translations for cycles of study and practice cued recall until items were correctly recalled once. After the first correct recall, items were presented only in subsequent study cycles with no further testing, or only in subsequent test cycles with no further study. Performance on a final test 1 week later was substantially greater after continued testing (80%) than after continued study (36%).

Why does practice testing improve learning? Whereas a wealth of studies have established the generality of testing effects, theories about why it improves learning have lagged behind. Nonetheless, theoretical accounts are increasingly emerging to explain two different kinds of testing effects, which are referred to as *direct effects* and *mediated effects* of testing (Roediger & Karpicke, 2006a). Direct effects refer to changes in learning that arise from the act of taking a test itself, whereas mediated effects refer to changes in learning that arise from an influence of testing on the amount or kind of encoding that takes place after the test (e.g., during a subsequent restudy opportunity).

Concerning direct effects of practice testing, Carpenter (2009) recently proposed that testing can enhance retention by triggering elaborative retrieval processes. Attempting to retrieve target information involves a search of long-term memory that activates related information, and this activated information may then be encoded along with the retrieved target, forming an elaborated trace that affords multiple pathways to facilitate later access to that information. In support of this account, Carpenter (2011) had learners study weakly related word pairs (e.g., "mother"–"child") followed either by additional study or a practice cued-recall test. On a later final test, recall of the target word was prompted via a previously unpresented but strongly related word (e.g., "father"). Performance was greater following a practice test than following restudy, presumably because the practice test increased the likelihood that the related information was activated and encoded along with the target during learning.

Concerning mediated effects of practice testing, Pyc and Rawson (2010, 2012b) proposed a similar account, according to which practice testing facilitates the encoding of more effective mediators (i.e., elaborative information connecting cues and targets) during subsequent restudy opportunities. Pyc and Rawson (2010) presented learners with Swahili-English translations in an initial study block, which was followed by three blocks of restudy trials; for half of the participants, each restudy trial was preceded by practice cued recall. All learners were prompted to generate and report a keyword mediator during each restudy trial. When tested 1 week later, compared with students who had only restudied, students who had engaged in practice cued recall were more likely to recall their mediators when prompted with the cue word and were more likely to recall the target when prompted with their mediator.

Recent evidence also suggests that practice testing may enhance how well students mentally organize information and how well they process idiosyncratic aspects of individual items, which together can support better retention and test performance (Hunt, 1995, 2006). Zaromb and Roediger (2010) presented learners with lists consisting of words from different taxonomic categories (e.g., vegetables, clothing) either for eight blocks of study trials or for four blocks of study trials with each trial followed by a practice free-recall test. Replicating basic testing effects, final free recall 2 days later was greater when items had received practice tests (39%) than when they had only been studied (17%). Importantly, the practice test condition also outperformed the study condition on secondary measures primarily tapping organizational processing and idiosyncratic processing.

8.2 How general are the effects of practice testing? Given the volume of research on testing effects, an exhaustive review of the literature is beyond the scope of this article. Accordingly, our synthesis below is primarily based on studies from the past 10 years (which include more than 120 articles), which we believe represent the current state of the field. Most of these studies compared conditions involving practice tests with conditions not involving practice tests or involving only restudy; however, we also considered more recent work pitting different practice-testing conditions against one another to explore when practice testing works best.

8.2a Learning conditions. The majority of research on practice testing has used test formats that involve cued recall of target information from memory, but some studies have also shown testing effects with other recall-based practice-test formats, including free recall, short-answer questions, and fill-in-the-blank questions. A growing number of studies using multiple-choice practice tests have also reported testing effects. Across these formats, most prior research has involved practice tests that tap memory for explicitly presented information. However, several studies have also shown testing effects for practice tests that tap comprehension, including short-answer application and multiple-choice inference-based questions (e.g., Agarwal & Roediger, 2011; Butler, 2010; C. I. Johnson & Mayer, 2009). Testing effects have also been shown in a study in which practice involved predicting (vs. studying) input-output values in an inductive function learning task (Kang, McDaniel, & Pashler, 2011) and a study in which participants practiced (vs. restudied) resuscitation procedures (Kromann,

Jensen, & Ringsted, 2009). Some research has demonstrated testing effects even when practice tests are open book (Agarwal et al., 2008; Weinstein, McDermott, & Roediger, 2010).

It is important to note that practice tests can benefit learning even when the format of the practice test does not match the format of the criterion test. For example, research has shown cross-format effects of multiple-choice practice tests on subsequent cued recall (Fazio, Agarwal, Marsh, & Roediger, 2010; Marsh, Agarwal, & Roediger, 2009; Roediger & Marsh, 2005), practice free recall on subsequent multiple-choice and short-answer inference tests (McDaniel, Howard, & Einstein, 2009), and practice cued recall on subsequent free recall and recognition (Carpenter, Pashler, & Vul, 2006; Vaughn & Rawson, 2011).

Although various practice-test formats work, some work better than others. Glover (1989) presented students with a short expository text for initial study and then manipulated the format of the practice test (free recall, fill in the blank, or recognition) and the format of the final test (free recall, fill in the blank, or recognition). On all three final-test formats, performance was greater following free-recall practice than following fill-in-the-blank practice, which in turn was greater than performance following recognition practice. Similarly, Carpenter and DeLosh (2006) found that free-recall practice outperformed cued-recall and recognition practice regardless of whether the final test was in a free-recall, cued-recall, or recognition format, and Hinze and Wiley (2011) found that performance on a multiple-choice final test was better following cued recall of paragraphs than following fill-in-the-blank practice. Further work is needed to support strong prescriptive conclusions, but the available evidence suggests that practice tests that require more generative responses (e.g., recall or short answer) are more effective than practice tests that require less generative responses (e.g., fill in the blank or recognition).

In addition to practice-test format, two other conditions of learning that strongly influence the benefits of practice testing are dosage and timing. Concerning dosage, the simplest conclusion is that more is better. Some studies supporting this conclusion have manipulated the number of practice tests, and final-test performance has consistently been better following multiple practice tests than following a single practice test (e.g., Karpicke & Roediger, 2007a, 2010; Logan & Balota, 2008; Pavlik & Anderson, 2005). In other studies, experimenters have varied the number of practice tests to manipulate the level of success achieved during practice. For example, Vaughn and Rawson (2011) observed significantly greater final-test performance when students engaged in cued-recall practice until target items were recalled four to five times versus only once. Several other studies have shown that final-test performance improves as the number of correct responses during practice increases (e.g., Karpicke & Roediger, 2007b, 2008; Pyc & Rawson, 2009, 2012a; Rawson & Dunlosky, 2011), albeit with diminishing returns as higher criterion levels are achieved. Whereas these studies have involved manipulations of dosage within a practice session, other studies that have manipulated the number of practice sessions have also found that more is better (Bahrick, 1979; Bahrick, Bahrick, Bahrick, & Bahrick, 1993; Morris & Fritz, 2002; Rawson & Dunlosky, 2011).

However, the benefit of repeated practice testing in turn depends on the timing of the practice tests. Several studies have increased the number of tests presented in immediate succession within a session and have found minimal or nonexistent effects, in contrast to the sizable benefits observed when repeated tests are spaced (e.g., Carpenter & DeLosh, 2005; Cull, 2000; Glover, 1989; Karpicke & Bauernschmidt, 2011). Concerning the time intervals involved with spacing, longer is better. Repeated practice testing produces greater benefits when lags between trials within a session are longer rather than shorter (e.g., Pashler, Zarow, & Triplett, 2003; Pavlik & Anderson, 2005; Pyc & Rawson, 2009, 2012b), when trials are completed in different practice sessions rather than all in the same session (e.g., Bahrick, 1979; Bahrick & Hall, 2005; Kornell, 2009; Rohrer, 2009; Rohrer & Taylor, 2006), and when intervals between practice sessions are longer rather than shorter (Bahrick et al., 1993; Carpenter, Pashler, & Cepeda, 2009, although the optimal lag between sessions may depend on retention interval—see Cepeda et al., 2009; Cepeda, Vul, Rohrer, Wixted, & Pashler, 2008). We discuss lag effects further in the Distributed Practice section below.

8.2b Student characteristics. A large majority of studies have involved college students as participants, but testing effects have also been demonstrated across participants of widely varying ages. Studies involving nonundergraduate samples have differed somewhat in the kind, dosage, or timing of practice testing involved, but some form of testing effect has been demonstrated with preschoolers and kindergartners (Fritz, Morris, Nolan, & Singleton, 2007; Kratochwill, Demuth, & Conzemius, 1977), elementary school students (Atkinson & Paulson, 1972; Bouwmeester & Verkoeijen, 2011; Fishman, Keller, & Atkinson, 1968; Gates, 1917; Metcalfe & Kornell, 2007; Metcalfe, Kornell, & Finn, 2009; Myers, 1914; Rea & Modigliani, 1985; Rohrer, Taylor, & Sholar, 2010; Spitzer, 1939), middle school students (Carpenter et al., 2009; Fritz, Morris, Nolan, et al., 2007; Glover, 1989; McDaniel, Agarwal, Huelser, McDermott, & Roediger, 2011; Metcalfe, Kornell, & Son, 2007; Sones & Stroud, 1940), high school students (Duchastel, 1981; Duchastel & Nungester, 1982; Marsh et al., 2009; Nungester & Duchastel, 1982), and more advanced students, such as 3rd- and 4th-year medical-school students (Kromann et al., 2009; Rees, 1986; Schmidmaier et al., 2011). On the other end of the continuum, testing effects have also been shown with middle-aged learners and with older adults (Balota, Duchek, Sergent-Marshall, & Roediger, 2006; Bishara & Jacoby, 2008; Logan & Balota, 2008; Maddox, Balota, Coane, & Duchek, 2011; Sumowski, Chiaravalloti, & DeLuca, 2010; Tse, Balota, & Roediger, 2010).

In contrast to the relatively broad range of ages covered in the testing-effect literature, surprisingly minimal research has examined testing effects as a function of individual differences

in knowledge or ability. In the only study including groups of learners with different knowledge levels, Carroll, Campbell-Ratcliffe, Murnane, and Perfect (2007) presented first-year undergraduates and advanced psychology majors with two passages from an abnormal-psychology textbook. Students completed a short-answer practice test on one of the passages and then took a final test over both passages either 15 minutes or 1 day later. Both groups showed similar testing effects at both time points (with 33% and 38% better accuracy, respectively, on the material that had been practice tested relative to the material that had not). Although these initial results provide encouraging evidence that testing effects may be robust across knowledge levels, further work is needed before strong conclusions can be drawn about the extent to which knowledge level moderates testing effects.

Likewise, minimal research has examined testing effects as a function of academically relevant ability levels. In a study by Spitzer (1939), 3,605 sixth graders from 91 different elementary schools read a short text and took an immediate test, to provide a baseline measure of reading comprehension ability. In the groups of interest here, all students read an experimental text, half completed a practice multiple-choice test, and then all completed a multiple-choice test either 1 or 7 days later. Spitzer reported final-test performance for the experimental text separately for the top and bottom thirds of performers on the baseline measure. As shown in Figure 7, taking the practice test benefited both groups of students. With that said, the testing effect appeared to be somewhat larger for higher-ability readers than for lower-ability readers (with approximately 20%, vs. 12%, improvements in accuracy), although Spitzer did not report the relevant inferential statistics.

Finally, evidence from studies involving patient populations is at least suggestive with respect to the generality of testing effects across different levels of learning capacity. For example, Balota et al. (2006) found that spaced practice tests improved retention over short time intervals not only for younger adults and healthy older adults but also for older adults with Alzheimer's disease. Similarly, Sumowski et al. (2010) found that a practice test produced larger testing effects for memory-impaired, versus memory-intact, subsets of middle-aged individuals with multiple sclerosis (d = 0.95 vs. d = 0.54, respectively, with grouping based on performance on a baseline measure of memory). In sum, several studies have suggested that practice testing may benefit individuals with varying levels of knowledge or ability, but the extent to which the magnitude of the benefit depends on these factors remains an open question.

8.2c Materials. Many of the studies that have demonstrated testing effects have involved relatively simple verbal materials, including word lists and paired associates. However, most of the sets of materials used have had some educational relevance. A sizable majority of studies using paired-associate materials have included foreign-language translations (including Chinese, Iñupiaq, Japanese, Lithuanian, Spanish, and Swahili) or vocabulary words paired with synonyms. Other studies have extended effects to paired book titles and author names, names and faces, objects and names, and pictures and foreign-language translations (e.g., Barcroft, 2007; Carpenter & Vul, 2011; Morris & Fritz, 2002; Rohrer, 2009).

A considerable number of studies have also shown testing effects for factual information, including trivia facts and general knowledge questions (e.g., Butler, Karpicke, & Roediger,

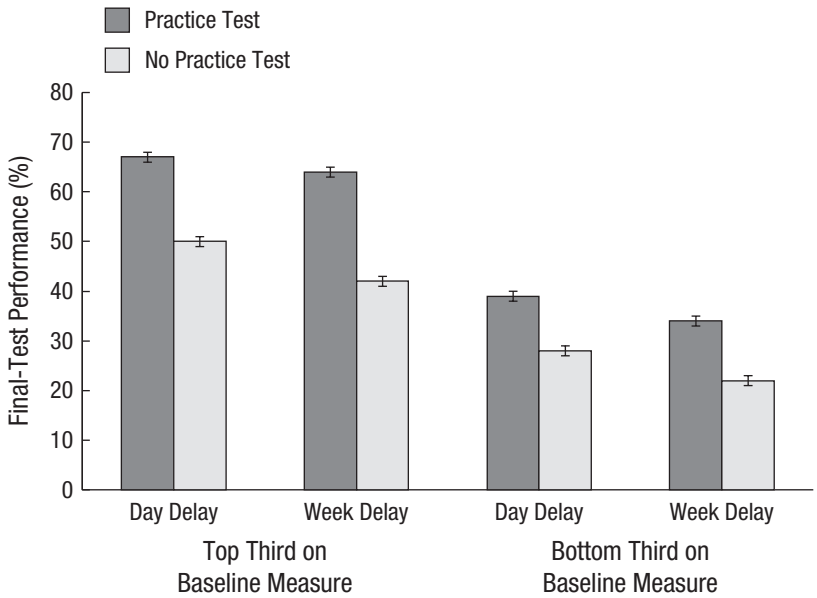


Fig. 7. Mean accuracy on a final test administered 1 day or 1 week after a learning session that either did or did not include a practice test, for the top and bottom thirds of scorers on a baseline measure of ability, in Spitzer (1939). Error bars represent standard errors.

2008; T. A. Smith & Kimball, 2010) and facts drawn from classroom units in science, history, and psychology (e.g., Carpenter et al., 2009; McDaniel et al., 2011; McDaniel, Wildman, & Anderson, 2012). Earlier research showed that practice tests helped children learn multiplication facts and spelling lists (Atkinson & Paulson, 1972; Fishman et al., 1968; Rea & Modigliani, 1985), and recent studies have reported enhanced learning of definitions of vocabulary words (Metcalfe et al., 2007) and definitions of key term concepts from classroom material (Rawson & Dunlosky, 2011).

An increasing number of studies have shown benefits for learning from text materials of various lengths (from 160 words to 2,000 words or more), of various text genres (e.g., encyclopedia entries, scientific journal articles, textbook passages), and on a wide range of topics (e.g., Civil War economics, bat echolocation, sea otters, the big bang theory, fossils, Arctic exploration, toucans). Practice tests have improved learning from video lectures and from narrated animations on topics such as adult development, lightning, neuroanatomy, and art history (Butler & Roediger, 2007; Cranney et al., 2009; Vojdanoska, Cranney, & Newell, 2010).

Although much of the work on testing effects has used verbal materials, practice testing has also been shown to support learning of materials that include visual or spatial information, including learning of features and locations on maps (Carpenter & Pashler, 2007; Rohrer et al., 2010), identifying birds (Jacoby, Wahlheim, & Coane, 2010), naming objects (Cepeda et al., 2009; Fritz et al., 2007), associating names with faces (Helder & Shaughnessy, 2008; Morris & Fritz, 2002), learning spatial locations of objects (Sommer, Schoell, & Büchel, 2008), learning symbols (Coppens, Verkoeijen, & Rikers, 2011), and identifying depicted parts of a flower (Glover, 1989). Finally, recent work has extended testing effects to nondeclarative learning, including the learning of resuscitation skills (Kromann et al., 2009) and inductive learning of input-output functions (Kang, McDaniel, et al., 2011).

8.2d Criterion tasks. Although cued recall is the most commonly used criterion measure, testing effects have also been shown with other forms of memory tests, including free-recall, recognition, and fill-in-the-blank tests, as well as short-answer and multiple-choice questions that tap memory for information explicitly stated in text material.

Regarding transfer, the modal method in testing-effect research has involved using the same questions tapping the same target information (e.g., the same cued-recall prompts or multiple-choice questions) on practice tests and criterion tests. However, as described in the subsection on learning conditions (8.2a) above, many studies have also shown testing effects when learning of the same target information is evaluated using different test formats for practice and criterion tests. Furthermore, an increasing number of studies have shown that practice testing a subset of information influences memory for related but untested information (J. C. K. Chan, 2009, 2010; J. C. K. Chan, McDermott, & Roediger, 2006; Cranney et al., 2009), although benefits have not always accrued to related information (see Carroll et al., 2007; Duchastel, 1981).

Although most research has involved memory-based practice tests and criterion measures, several recent studies have also reported encouraging results concerning the extent to which practice testing can benefit comprehension. Positive effects have been shown on criterion tests that require inferences or the application of previously learned information (Agarwal & Roediger, 2011; Butler, 2010; Foos & Fisher, 1988; C. I. Johnson & Mayer, 2009; Karpicke & Blunt, 2011; McDaniel et al., 2009), including criterion tests that used different questions or different test formats than those used during practice. For example, Karpicke and Blunt (2011) found that practicing free recall of text material facilitated performance on a subsequent criterion test involving inference-based short-answer questions, as well as on a concept-mapping test. In fact, concept-mapping performance was better following free-recall practice during study than following concept mapping during study. Similarly, Butler (2010) presented students with expository texts for initial study, which was followed either by repeated restudy or by repeated practice short-answer tests (with feedback) tapping key facts and concepts from the texts. One week later, performance on new inference-based short-answer questions tapping the key facts and concepts was better following practice testing than following restudy (see Fig. 8). The outcomes of a follow-up experiment are particularly striking, given that the criterion test involved far transfer, in that questions required the concepts from one domain to be applied in a novel domain (e.g., students had to apply information learned about bat wings to make inferences about the development of new kinds of aircraft).

Finally, recent studies have also shown testing effects involving other forms of transfer. Jacoby et al. (2010)

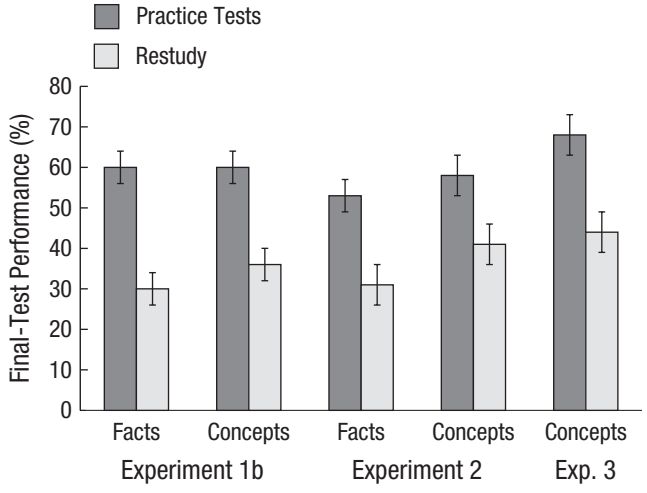


Fig. 8. Accuracy on final tests that consisted of inference-based transfer questions tapping key facts or concepts, administered 1 week after a learning session that involved either practice tests or restudy, in Butler (2010). Error bars represent standard errors.

presented learners with pictures of birds and their family names for initial study, which was followed either by additional study of the picture-name pairs or by practice tests in which learners were shown each picture and attempted to retrieve the appropriate family name prior to being shown the correct answer. The subsequent criterion test involved the same families of birds but included new pictures of birds from those families. Learners were more accurate in classifying new birds following practice testing than following restudy only. Similarly, Kang, McDaniel, & Pashler (2011) examined inductive function learning under conditions in which learners either studied pairs of input-output values or predicted output for a given input value prior to being shown the correct output. The prediction group outperformed the study-only group on a criterion test for both trained pairs and untrained extrapolation pairs.

In addition to establishing testing effects across an array of outcome measures, studies have also demonstrated testing effects across many retention intervals. Indeed, in contrast to literatures on other learning techniques, contemporary research on testing effects has actually used short retention intervals *less* often than longer retention intervals. Although a fair number of studies have shown testing effects after short delays (0–20 minutes), the sizable majority of recent research has involved delays of at least 1 day, and the modal retention interval used is 1 week. The preference for using longer retention intervals may be due in part to outcomes from several studies reporting that testing effects are larger when final tests are administered after longer delays (J. C. K. Chan, 2009; Coppens et al., 2011; C. I. Johnson & Mayer, 2009; Kornell, Bjork, & Garcia, 2011; Roediger & Karpicke, 2006b; Runquist, 1983; Schmidmaier et al., 2011; Toppino & Cohen, 2009; Wenger, Thompson, & Bartling, 1980; Wheeler, Ewers, & Buonanno, 2003). It is impressive that testing effects have been observed after even longer intervals, including intervals of 2 to 4 weeks (e.g., Bahrick & Hall, 2005; Butler & Roediger, 2007; Carpenter, Pashler, Wixted, & Vul, 2008; Kromann et al., 2009; Rohrer, 2009), 2 to 4 months (e.g., McDaniel, Anderson, Derbish, & Morrisette, 2007; Morris & Fritz, 2002; Rawson & Dunlosky, 2011), 5 to 8 months (McDaniel et al., 2011; Rees, 1986), 9-11 months (Carpenter et al., 2009), and even 1 to 5 years (Bahrick et al., 1993). These findings are great news for students and educators, given that a key educational goal is durable knowledge and not just temporary improvements in learning.

8.3 Effects in representative educational contexts. As described above, much of the research on testing effects has involved educationally relevant materials, tasks, and retention intervals. Additionally, several studies have reported testing effects using authentic classroom materials (i.e., material taken from classes in which student participants were enrolled; Carpenter et al., 2009; Cranney et al., 2009; Kromann et al., 2009; McDaniel et al., 2007; Rawson & Dunlosky, 2011; Rees, 1986; Vojdanoska et al., 2010). Whereas the criterion measures in these studies involved experimenter-devised tests or no-stakes pop quizzes, research has also shown effects of practice testing on actual summative course assessments (Balch, 1998; Daniel & Broida, 2004; Lyle & Crawford, 2011; McDaniel et al., 2011; McDaniel et al., 2012).

For example, a study by McDaniel et al. (2012) involved undergraduates enrolled in an online psychology course on the brain and behavior. Each week, students could earn course points by completing an online practice activity up to four times. In the online activity, some information was presented for practice testing with feedback, some information was presented for restudy, and some information was not presented. Subsequent unit exams included questions that had been presented during the practice tests and also new, related questions focusing on different aspects of the practiced concepts. As shown in Figure 9, grades on unit exams were higher for information that had been practice tested than for restudied information or unpracticed information, for both repeated questions and for new related questions.

8.4 Issues for implementation. Practice testing appears to be relatively reasonable with respect to time demands. Most research has shown effects of practice testing when the amount of time allotted for practice testing is modest and is equated with the time allotted for restudying. Another merit of practice testing is that it can be implemented with minimal training. Students can engage in recall-based self-testing in a relatively straightforward fashion. For example, students can self-test via cued recall by creating flashcards (free and low-cost flashcard software is also readily available) or by using the Cornell

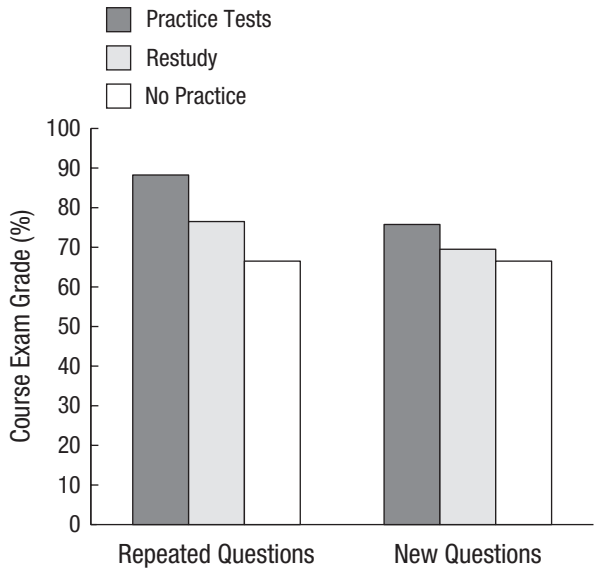


Fig. 9. Grades on course exams covering items that were presented for practice testing, presented for restudy, or not presented during online learning activities that students completed for course points. The course exam included some questions that had been presented during practice tests as well as new questions tapping the same information. For simplicity, outcomes reported here are collapsed across two experiments reported by McDaniel, Wildman, and Anderson (2012).

note-taking system (which involves leaving a blank column when taking notes in class and entering key terms or questions in it shortly after taking notes to use for self-testing when reviewing notes at a later time; for more details, see Pauk & Ross, 2010). More structured forms of practice testing (e.g., multiple-choice, short-answer, and fill-in-the-blank tests) are often readily available to students via practice problems or questions included at the end of textbook chapters or in the electronic supplemental materials that accompany many textbooks. With that said, students would likely benefit from some basic instruction on how to most effectively use practice tests, given that the benefits of testing depend on the kind of test, dosage, and timing. As described above, practice testing is particularly advantageous when it involves retrieval and is continued until items are answered correctly more than once within and across practice sessions, and with longer as opposed to shorter intervals between trials or sessions.

Concerning the effectiveness of practice testing relative to other learning techniques, a few studies have shown benefits of practice testing over concept mapping, note-taking, and imagery use (Fritz et al., 2007; Karpicke & Blunt, 2011; McDaniel et al., 2009; Neuschatz, Preston, Toglia, & Neuschatz, 2005), but the most frequent comparisons have involved pitting practice testing against unguided restudy. The modal outcome is that practice testing outperforms restudying, although this effect depends somewhat on the extent to which practice tests are accompanied by feedback involving presentation of the correct answer. Although many studies have shown that testing alone outperforms restudy, some studies have failed to find this advantage (in most of these cases, accuracy on the practice test has been relatively low). In contrast, the advantage of practice testing with feedback over restudy is extremely robust. Practice testing with feedback also consistently outperforms practice testing alone.

Another reason to recommend the implementation of feedback with practice testing is that it protects against perseveration errors when students respond incorrectly on a practice test. For example, Butler and Roediger (2008) found that a multiple-choice practice test increased intrusions of false alternatives on a final cued-recall test when no feedback was provided, whereas no such increase was observed when feedback was given. Fortunately, the corrective effect of feedback does not require that it be presented immediately after the practice test. Metcalfe et al. (2009) found that final-test performance for initially incorrect responses was actually better when feedback had been delayed than when it had been immediate. Also encouraging is evidence suggesting that feedback is particularly effective for correcting high-confidence errors (e.g., Butterfield & Metcalfe, 2001). Finally, we note that the effects of practice-test errors on subsequent performance tend to be relatively small, often do not obtain, and are heavily outweighed by the positive benefits of testing (e.g., Fazio et al., 2010; Kang, Pashler, et al., 2011; Roediger & Marsh, 2005). Thus, potential concerns about errors do not constitute a serious issue for implementation, particularly when feedback is provided.

Finally, although we have focused on students' use of practice testing, in keeping with the purpose of this monograph, we briefly note that instructors can also support student learning by increasing the use of low-stakes or no-stakes practice testing in the classroom. Several studies have also reported positive outcomes from administering summative assessments that are shorter and more frequent rather than longer and less frequent (e.g., one exam per week rather than only two or three exams per semester), not only for learning outcomes but also on students' ratings of factors such as course satisfaction and preference for more frequent testing (e.g., Keys, 1934; Kika, McLaughlin, & Dixon, 1992; Leeming, 2002; for a review, see Bangert-Drowns, Kulik, & Kulik, 1991).

8.5 Practice testing: Overall assessment. On the basis of the evidence described above, we rate practice testing as having high utility. Testing effects have been demonstrated across an impressive range of practice-test formats, kinds of material, learner ages, outcome measures, and retention intervals. Thus, practice testing has broad applicability. Practice testing is not particularly time intensive relative to other techniques, and it can be implemented with minimal training. Finally, several studies have provided evidence for the efficacy of practice testing in representative educational contexts. Regarding recommendations for future research, one gap identified in the literature concerns the extent to which the benefits of practice testing depend on learners' characteristics, such as prior knowledge or ability. Exploring individual differences in testing effects would align well with the aim to identify the broader generalizability of the benefits of practice testing. Moreover, research aimed at more thoroughly identify the causes of practice-test effects may provide further insights into maximizing these effects.

9 Distributed practice

To-be-learned material is often encountered on more than one occasion, such as when students review their notes and then later use flashcards to restudy the materials, or when a topic is covered in class and then later studied in a textbook. Even so, students mass much of their study prior to tests and believe that this popular cramming strategy is effective. Although cramming is better than not studying at all in the short term, given the same amount of time for study, would students be better off spreading out their study of content? The answer to this question is a resounding "yes." The term *distributed-practice effect* refers to the finding that distributing learning over time (either within a single study session or across sessions) typically benefits long-term retention more than does massing learning opportunities back-to-back or in relatively close succession.

Given the volume of research on distributed practice, an exhaustive review of the literature is beyond the scope of this article. Fortunately, this area of research benefits from extensive review articles (e.g., Benjamin & Tullis, 2010; Cepeda, Pashler, Vul, Wixted, & Rohrer, 2006; Delaney, Verkoeijen, & Spirgel, 2010; Dempster & Farris, 1990; Donovan & Radosevich, 1999; Janiszewski, Noel, & Sawyer, 2003), which provided foundations for the current review. In keeping with recent reviews (Cepeda et al., 2006; Delaney et al., 2010), we use the term *distributed practice* to encompass both *spacing effects* (i.e., the advantage of spaced over massed practice) and *lag effects* (i.e., the advantage of spacing with longer lags over spacing with shorter lags), and we draw on both literatures for our summary.

9.1 General description of distributed practice and why it should work. To illustrate the issues involved, we begin with a description of a classic experiment on distributed practice, in which students learned translations of Spanish words to criterion in an original session (Bahrick, 1979). Students then participated in six additional sessions in which they had the chance to retrieve and relearn the translations (feedback was provided). Figure 10 presents results from this study. In the zero-spacing condition (represented by the circles in Fig. 10), the learning sessions were back-to-back, and learning was rapid across the six massed sessions. In the 1-day condition (represented by the squares in Fig. 10), learning sessions were spaced 1 day apart, resulting in slightly more forgetting across sessions (i.e., lower performance on the initial test in each session) than in the zero-spacing condition, but students in the 1-day condition still obtained almost perfect accuracy by the sixth session. In contrast, when learning sessions were separated by 30 days, forgetting was much greater across sessions, and initial test performance did not reach the level observed in the other two conditions, even after six sessions (see triangles in Fig. 10). The key point for our present purposes is that the pattern reversed on the final test 30 days later, such that the best retention of the translations was observed in the condition in which relearning sessions had been separated by 30 days. That is, the condition with the most intersession forgetting yielded the greatest long-term retention. Spaced practice (1 day or 30 days) was superior to massed practice (0 days), and the benefit was greater following a longer lag (30 days) than a shorter lag (1 day).

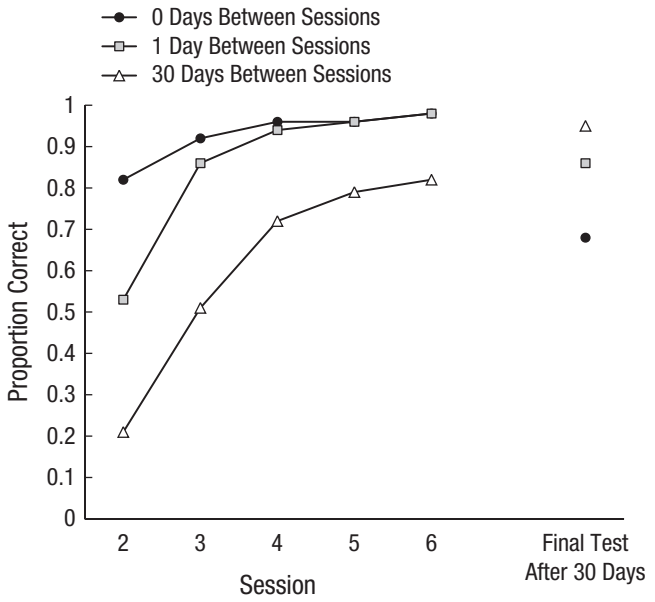


Fig. 10. Proportion of items answered correctly on an initial test administered in each of six practice sessions (prior to actual practice) and on the final test 30 days after the final practice session as a function of lag between sessions (0 days, 1 day, or 30 days) in Bahrick (1979).

Many theories of distributed-practice effects have been proposed and tested. Consider some of the accounts currently under debate (for in-depth reviews, see Benjamin & Tullis, 2010; Cepeda et al., 2006). One theory invokes the idea of deficient processing, arguing that the processing of material during a second learning opportunity suffers when it is close in time to the original learning episode. Basically, students do not have to work very hard to reread notes or retrieve something from memory when they have just completed this same activity, and furthermore, they may be misled by the ease of this second task and think they know the material better than they really do (e.g., Bahrick & Hall, 2005). Another theory involves reminding; namely, the second presentation of to-be-learned material serves to remind the learner of the first learning opportunity, leading it to be retrieved, a process well known to enhance memory (see the Practice Testing section above). Some researchers also draw on consolidation in their explanations, positing that the second learning episode benefits from any consolidation of the first trace that has already happened. Given the relatively large magnitude of distributed-practice effects, it is plausible that multiple mechanisms may contribute to them; hence, particular theories often invoke different combinations of mechanisms to explain the effects.

9.2 How general are the effects of distributed practice? The distributed-practice effect is robust. Cepeda et al. (2006) reviewed 254 studies involving more than 14,000 participants altogether; overall, students recalled more after spaced study (47%) than after massed study (37%). In both Donovan and Radosevich's (1999) and Janiszewski et al.'s (2003) meta-analyses, distributed practice was associated with moderate effect sizes for recall of verbal stimuli. As we describe below, the distributed-practice effect generalizes across many of the categories of variables listed in Table 2.

9.2a Learning conditions. Distributed practice refers to a particular *schedule* of learning episodes, as opposed to a particular *kind* of learning episode. That is, the distributed-practice effect refers to better learning when learning episodes are spread out in time than when they occur in close succession, but those learning episodes could involve restudying material, retrieving information from memory, or practicing skills. Because our emphasis is on educational applications, we will not

draw heavily on the skill literature, given that tasks such as ball tossing, gymnastics, and music memorization are less relevant to our purposes. Because much theory on the distributed-practice effect is derived from research on the spacing of study episodes, we focus on that research, but we also discuss relevant studies on distributed retrieval practice. In general, distributed practice testing is better than distributed study (e.g., Carpenter et al., 2009), as would be expected from the large literature on the benefits of practice testing.

One of the most important questions about distributed practice involves how to space the learning episodes—that is, how should the multiple encoding opportunities be arranged? Cepeda et al. (2006) noted that most studies have used relatively short intervals (less than 1 day), whereas we would expect the typical interval between educational learning opportunities (e.g., lecture and studying) to be longer. Recall that the classic investigation by Bahrick (1979) showed a larger distributed-practice effect with 30-day lags between sessions than with 1-day lags (Fig. 10); Cepeda et al. (2006) noted that "every study examined here with a retention interval longer than 1 month demonstrated a benefit from distribution of learning across weeks or months" (p. 370; "retention interval" here refers to the time between the last study opportunity and the final test).

However, the answer is not as simple as "longer lags are better"—the answer depends on how long the learner wants to retain information. Impressive data come from Cepeda, Vul, Rohrer, Wixted, and Pashler (2008), who examined people's learning of trivia facts in an internet study that had 26 different conditions, which combined different between-session intervals (from no lag to a lag of 105 days) with different retention intervals (up to 350 days). In brief, criterion performance was best when the lag between sessions was approximately 10–20% of the desired retention interval. For example, to remember something for 1 week, learning episodes should be spaced 12 to 24 hours apart; to remember something for 5 years, the learning episodes should be spaced 6 to 12 months apart. Of course, when students are preparing for examinations, the degree to which they can space their study sessions may be limited, but the longest intervals (e.g., intervals of 1 month or more) may be ideal for studying core content that needs to be retained for cumulative examinations or achievement tests that assess the knowledge students have gained across several years of education.

Finally, the distributed-practice effect may depend on the type of processing evoked across learning episodes. In the meta-analysis by Janiszewski et al. (2003), intentional processing was associated with a larger effect size ($M = .35$) than was incidental processing ($M = .24$). Several things should be noted. First, the distributed-practice effect is sometimes observed with incidental processing (e.g., R. L. Greene, 1989; Toppino, Fearnow-Kenney, Kiepert, & Teremula, 2009); it is not eliminated across the board, but the average effect size is slightly (albeit significantly) smaller. Second, the type of processing learners engage in may covary with the intentionality of their learning, with students being more likely to extract meaning from materials when they are deliberately trying to learn them. In at least two studies, deeper processing yielded a distributed-practice effect whereas more shallow processing did not (e.g., Challis, 1993; Delaney & Knowles, 2005). Whereas understanding how distributed-practice effects change with strategy has important theoretical implications, this issue is less important when considering applications to education, because when students are studying, they presumably are intentionally trying to learn.

9.2b Student characteristics. The majority of distributed-practice experiments have tested undergraduates, but effects have also been demonstrated in other populations. In at least some situations, even clinical populations can benefit from distributed practice, including individuals with multiple sclerosis (Goverover, Hillary, Chiaravalloti, Arango-Lasprilla, & DeLuca, 2009), traumatic brain injuries (Goverover, Arango-Lasprilla, Hillary, Chiaravalloti, & DeLuca, 2009), and amnesia (Cermak, Verfaellie, Lanzoni, Mather, & Chase, 1996). In general, children of all ages benefit from distributed study. For example, when learning pictures, children as young as preschoolers recognize and recall more items studied after longer lags than after shorter lags (Toppino, 1991; Toppino, Kasserman, & Mracek, 1991). Similarly, 3-year-olds are better able to classify new exemplars of a category if the category was originally learned through spaced rather than massed study (Vlach, Sandhofer, & Kornell, 2008). Even 2-year-olds show benefits of distributed practice, such that it increases their later ability to produce studied words (Childers & Tomasello, 2002). These benefits of spacing for language learning also occur for children with specific language impairment (Riches, Tomasello, & Conti-Ramsden, 2005).

At the other end of the life span, older adults learning paired associates benefit from distributed practice as much as young adults do (e.g., Balota, Duchek, & Paullin, 1989). Similar conclusions are reached when spacing involves practice tests rather than study opportunities (e.g., Balota et al., 2006; Logan & Balota, 2008) and when older adults are learning to classify exemplars of a category (as opposed to paired associates; Kornell, Castel, Eich, & Bjork, 2010). In summary, learners of different ages benefit from distributed practice, but an open issue is the degree to which the distributed-practice effect may be moderated by other individual characteristics, such as prior knowledge and motivation.

9.2c Materials. Distributed-practice effects have been observed with many types of to-be-learned materials, including definitions (e.g., Dempster, 1987), face-name pairs (e.g., Carpenter & DeLosh, 2005), translations of foreign vocabulary words (e.g., Bahrick & Hall, 2005), trivia facts (e.g., Cepeda et al., 2008), texts (e.g., Rawson & Kintsch, 2005), lectures (e.g., Glover & Corkill, 1987), and pictures (e.g., Hintzman & Rogers, 1973). Distributed study has also yielded improved performance in a range of domains, including biology (Reynolds & Glaser, 1964) and advertising (e.g., Appleton-Knapp, Bjork, & Wickens, 2005). If we include practice testing and practice of

skills, then the list of domains in which benefits of distributed practice have been successfully demonstrated can be expanded to include mathematics (e.g., Rickard, Lau, & Pashler, 2008; Rohrer, 2009), history (Carpenter et al., 2009), music (e.g., Simmons, 2011), and surgery (e.g., Moulton et al., 2006), among others.

Not all tasks yield comparably large distributed-practice effects. For instance, distributed-practice effects are large for free recall but are smaller (or even nonexistent) for tasks that are very complex, such as airplane control (Donovan & Radosevich, 1999). It is not clear how to map these kinds of complex tasks, which tend to have a large motor component, onto the types of complex tasks seen in education. The U.S. Institute of Education Sciences guide on organizing study to improve learning explicitly notes that "one limitation of the literature is that few studies have examined acquisition of complex bodies of structured information" (Pashler et al., 2007, p. 6). The data that exist (which are reviewed below) have come from classroom studies and are promising.

9.2d Criterion tasks. We alluded earlier to the fact that distributed-practice effects are robust over long retention intervals, with Cepeda and colleagues (2008) arguing that the ideal lag between practice sessions would be approximately 10–20% of the desired retention interval. They examined learning up to 350 days after study; other studies have shown benefits of distributed testing after intervals lasting for months (e.g., Cepeda et al., 2009) and even years (e.g., Bahrick et al., 1993; Bahrick & Phelps, 1987). In fact, the distributed-practice effect is often stronger on delayed tests than immediate ones, with massed practice (cramming) actually benefitting performance on immediate tests (e.g., Rawson & Kintsch, 2005).

Much research has established the durability of distributed-practice effects over time, but much less attention has been devoted to other kinds of criterion tasks used in educational contexts. The Cepeda et al. (2009) meta-analysis, for example, focused on studies in which the dependent measure was verbal free recall. The distributed-practice effect has been generalized to dependent measures beyond free recall, including multiple-choice questions, cued-recall and short-answer questions (e.g., Reynolds & Glaser, 1964), frequency judgments (e.g., Hintzman & Rogers, 1973), and, sometimes, implicit memory (e.g., R. L. Greene, 1990; Jacoby & Dallas, 1981). More generally, although studies using these basic measures of memory can inform the field by advancing theory, the effects of distributed practice on these measures will not necessarily generalize to all other educationally relevant measures. Given that students are often expected to go beyond the basic retention of materials, this gap is perhaps the largest and most important to fill for the literature on distributed practice. With that said, some relevant data from classroom studies are available; we turn to these in the next section.

9.3 Effects in representative educational contexts. Most of the classroom studies that have demonstrated distributed-practice effects have involved spacing of more than just study opportunities. It is not surprising that real classroom exercises would use a variety of techniques, given that the goal of educators is to maximize learning rather than to isolate the contributions of individual techniques. Consider a study by Sobel, Cepeda, and Kapler (2011) in which fifth graders learned vocabulary words. Each learning session had multiple steps: A teacher read and defined words; the students wrote down the definitions; the teacher repeated the definitions and used them in sentences, and students reread the definitions; finally, the students wrote down the definitions again and created sentences using the words. Several different kinds of study (including reading from booklets and overheads, as well as teacher instruction) and practice tests (e.g., generating definitions and sentences) were spaced in this research. The criterion test was administered 5 weeks after the second learning session, and students successfully defined a greater proportion of GRE vocabulary words (e.g., *accolade*) learned in sessions spaced a week apart than vocabulary words learned in sessions spaced a minute apart (Sobel et al., 2011). A mix of teacher instruction and student practice was also involved in a demonstration of the benefits of distributed practice for learning phonics in first graders (Seabrook, Brown, & Solity, 2005).

Another study examined learning of statistics across two sections of the same course, one of which was taught over a 6-month period and the other of which covered the same material in an 8-week period (Budé, Imbos, van de Wiel, & Berger, 2011). The authors took advantage of a curriculum change at their university that allowed them to compare learning in a class taught before the university reduced the length of the course with learning in a class taught after the change. The curriculum change meant that lectures, problem-based group meetings, and lab sessions (as well as student-driven study, assignments, etc.) were implemented within a much shorter time period; in other words, a variety of study and retrieval activities were more spaced out in time in one class than in the other. Students whose course lasted 6 months outperformed students in the 8-week course both on an open-ended test tapping conceptual understanding (see Fig. 11) and on the final exam (Fig. 12). Critically, the two groups performed similarly on a control exam from another course (Fig. 12), suggesting that the effects of distributed practice were not due to ability differences across classes.

Finally, a number of classroom studies have examined the benefits of distributed practice tests. Distributed practice testing helps students in actual classrooms learn history facts (Carpenter et al., 2009), foreign language vocabulary (K. C. Bloom & Shuell, 1981), and spelling (Fishman et al., 1968).

9.4 Issues for implementation. Several obstacles may arise when implementing distributed practice in the classroom. Dempster and Farris (1990) made the interesting point that many textbooks do not encourage distributed learning, in that they lump related material together and do not review previously covered material in subsequent units. At least one formal content analysis of actual textbooks (specifically, elementary-school mathematics textbooks; Stigler, Fuson, Ham, & Kim, 1986) supported this claim, showing that American textbooks

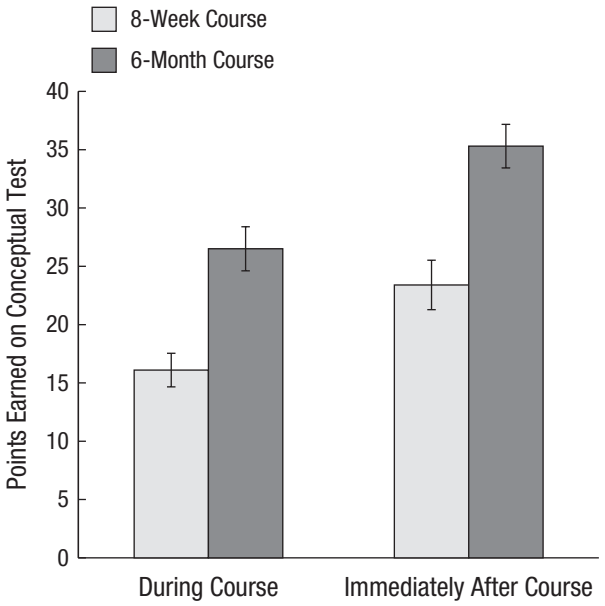


Fig. 11. Points earned on an open-ended test tapping conceptual understanding of content from two sections of a course, one taught over an 8-week period and the other taught over a 6-month period, in Budé, Imbos, van de Wiel, and Berger (2011). Error bars represent standard errors.

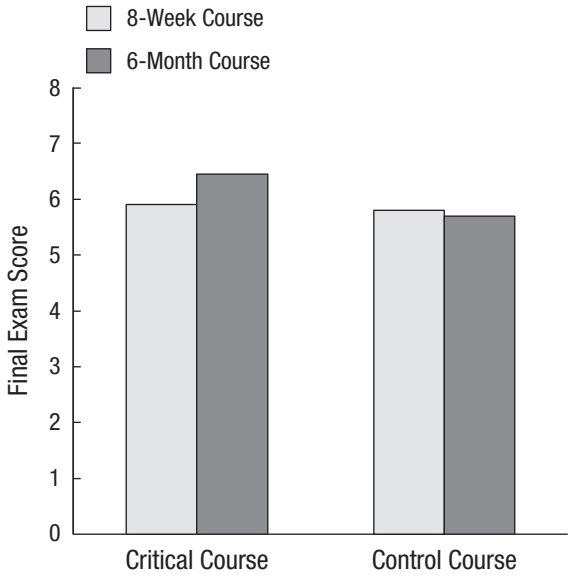


Fig. 12. Final-exam scores in a critical course and a control course as a function of the length of the course (8 weeks or 6 months); data drawn from Budé, Imbos, van de Wiel, and Berger (2011). Standard errors are not available.

grouped to-be-worked problems together (presumably at the end of chapters) as opposed to distributing them throughout the pages. These textbooks also contained less variability in sets of problems than did comparable textbooks from the former Soviet Union. Thus, one issue students face is that their study materials may not be set up in a way that encourages distributed practice.

A second issue involves how students naturally study. Michael (1991) used the term *procrastination scallop* to describe the typical study pattern—namely, that time spent studying increases as an exam approaches. Mawhinney, Bostow, Laws, Blumenfield, and Hopkins (1971) documented this pattern using volunteers who agreed to study in an observation room that allowed their time spent studying to be recorded. With daily testing, students studied for a consistent amount of time across sessions. But when testing occurred only once every 3 weeks, time spent studying increased across the interval, peaking right before the exam (Mawhinney et al., 1971). In other words, less frequent testing led to massed study immediately before the test, whereas daily testing effectively led to study that was distributed over time. The implication is that students will not necessarily engage in distributed study unless the situation forces them to do so; it is unclear whether this is because of practical constraints or because students do not understand the memorial benefits of distributed practice.

With regard to the issue of whether students understand the benefits of distributed practice, the data are not entirely definitive. Several laboratory studies have investigated students' choices about whether to mass or space repeated studying of paired associates (e.g., GRE vocabulary words paired with their definitions). In such studies, students typically choose between restudying an item almost immediately after learning (massing) or restudying the item later in the same session (spacing). Although students do choose to mass their study under some conditions (e.g., Benjamin & Bird, 2006; Son, 2004), they typically choose to space their study of items (Pyc & Dunlosky, 2010; Toppino, Cohen, Davis, & Moors, 2009). This bias toward spacing does not necessarily mean that students understand the benefits of distributed practice per se (e.g., they may put off restudying a pair because they do not want to see it again immediately), and one study has shown that students rate their overall level of learning as higher after massed study than after spaced study, even when the students had experienced the benefits of spacing (e.g., Kornell & Bjork, 2008). Other recent studies have provided evidence that students are unaware of the benefits of practicing with longer, as opposed to shorter, lags (Pyc & Rawson, 2012b; Wissman et al., 2012).

In sum, because of practical constraints and students' potential lack of awareness of the benefits of this technique, students may need some training and some convincing that distributed practice is a good way to learn and retain information. Simply experiencing the distributed-practice effect may not always be sufficient, but a demonstration paired with instruction about the effect may be more convincing to students (e.g., Balch, 2006).

9.5 Distributed practice: Overall assessment. On the basis of the available evidence, we rate distributed practice as having high utility: It works across students of different ages, with a wide variety of materials, on the majority of standard laboratory measures, and over long delays. It is easy to implement

(although it may require some training) and has been used successfully in a number of classroom studies. Although less research has examined distributed-practice effects using complex materials, the existing classroom studies have suggested that distributed practice should work for complex materials as well. Future research should examine this issue, as well as possible individual differences beyond age and criterion tasks that require higher-level cognition. Finally, future work should isolate the contributions of distributed study from those of distributed retrieval in educational contexts.

10 Interleaved practice

In virtually every kind of class at every grade level, students are expected to learn content from many different subtopics or problems of many different kinds. For example, students in a neuroanatomy course would learn about several different divisions of the nervous system, and students in a geometry course would learn various formulas for computing properties of objects such as surface area and volume. Given that the goal is to learn all of the material, how should a student schedule his or her studying of the different materials? An intuitive approach, and one we suspect is adopted by most students, involves *blocking* study or practice, such that all content from one subtopic is studied or all problems of one type are practiced before the student moves on to the next set of material. In contrast, recent research has begun to explore *interleaved practice*, in which students alternate their practice of different kinds of items or problems. Our focus here is on whether interleaved practice benefits students' learning of educationally relevant material.

Before we present evidence of the efficacy of this technique, we should point out that, in contrast to the other techniques we have reviewed in this monograph, many fewer studies have investigated the benefits of interleaved practice on measures relevant to student achievement. Nonetheless, we elected to include this technique in our review because (a) plenty of evidence indicates that interleaving can improve motor learning under some conditions (for reviews, see Brady, 1998; R. A. Schmidt & Bjork, 1992; Wulf & Shea, 2002) and (b) the growing literature on interleaving and performance on cognitive tasks is demonstrating the same kind of promise.

10.1 General description of interleaved practice and why it should work. Interleaved practice, as opposed to blocked practice, is easily understood by considering a method used by Rohrer and Taylor (2007), which involved teaching college students to compute the volumes of different geometric solids. Students had two practice sessions, which were separated by 1 week. During each practice session, students were given tutorials on how to find the volume for four different kinds of geometric solids and completed 16 practice problems (4 for each solid). After the completion of each practice problem, the correct solution was shown for 10 seconds. Students in a blocked-practice condition first read a tutorial on finding the volume of a given solid, which was immediately followed by the four practice problems for that kind of solid. Practice solving volumes for a given solid was then followed by the tutorial and practice problems for the next kind of solid, and so on. Students in an interleaved-practice group first read all four tutorials and then completed all the practice problems, with the constraint that every set of four consecutive problems included one problem for each of the four kinds of solids. One week after the second practice session, all students took a criterion test in which they solved two novel problems for each of the four kinds of solids. Students' percentages of correct responses during the practice sessions and during the criterion test are presented in Figure 13, which illustrates a typical interleaving effect: During practice, performance was better with blocked practice than interleaved practice, but this advantage dramatically reversed on the criterion test, such that interleaved practice boosted accuracy by 43%.

One explanation for this impressive effect is that interleaving gave students practice at identifying which solution method (i.e., which of several different formulas) should be used for a given solid (see also, Mayfield & Chase, 2002). Put differently, interleaved practice helps students to discriminate between the different kinds of problems so that they will be more likely to use the correct solution method for each one. Compelling evidence for this possibility was provided by Taylor and Rohrer (2010). Fourth graders learned to solve mathematical problems involving prisms. For a prism with a given number of base sides (b), students learned to solve for the number of faces ($b + 2$), edges ($b \times 3$), corners ($b \times 2$), or angles ($b \times 6$). Students first practiced *partial* problems: A term for a single component of a prism was presented (e.g., corners), the student had to produce the correct formula (i.e., for corners, the correct response would be "$b \times 2$"), and then

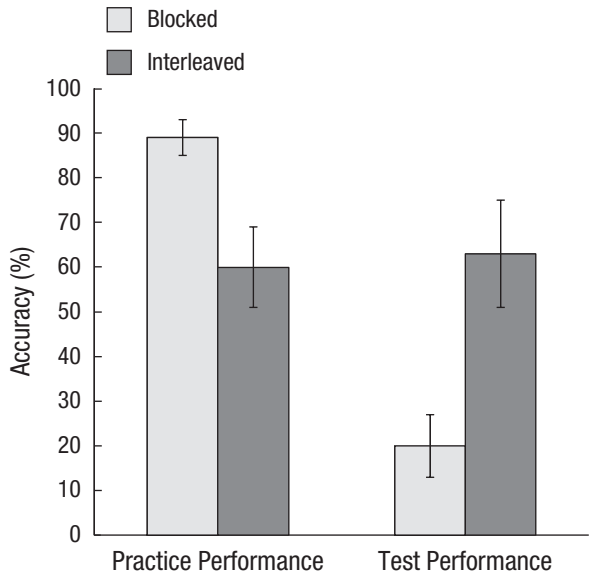


Fig. 13. Percentage of correct responses on sets of problems completed in practice sessions and on a delayed criterion test in Rohrer and Taylor (2007). Error bars represent standard errors.

feedback (the correct answer) was provided. After practicing partial problems, students practiced *full* problems, in which they were shown a prism with a number of base sides (e.g., 14 sides) and a term for a single component (e.g., edges). Students had to produce the correct formula ($b \times 3$) and solve the problem by substituting the appropriate value of b (14×3). Most important, students in a blocked-practice group completed all partial- and full-practice problems for one prism feature (e.g., angles) before moving onto the next. For students in an interleaved-practice group, each block of four practice problems included one problem for each of the four prism features. One day after practice, a criterion test was administered in which students were asked to solve full problems that had not appeared during practice.

Accuracy during practice was greater for students who had received blocked practice than for students who had received interleaved practice, both for partial problems (99% vs. 68%, respectively) and for full problems (98% vs. 79%). By contrast, accuracy 1 day later was substantially higher for students who had received interleaved practice (77%) than for students who had received blocked practice (38%). As with Rohrer and Taylor (2006), a plausible explanation for this pattern is that interleaved practice helped students to discriminate between various kinds of problems and to learn the appropriate formula to apply for each one. This explanation was supported by a detailed analysis of errors the fourth graders made when solving the full problems during the criterion task. *Fabrication errors* involved cases in which students used a formula that was not originally trained (e.g., $b \times 8$), whereas *discrimination errors* involved cases in which students used one of the four formulas that had been practiced but was not appropriate for a given problem. As shown in Figure 14, the two groups did not differ in fabrication errors, but discrimination errors were more common after blocked practice than after interleaved practiced. Students who received interleaved practice apparently were better at discriminating among the kinds of problems and consistently applied the correct formula to each one.

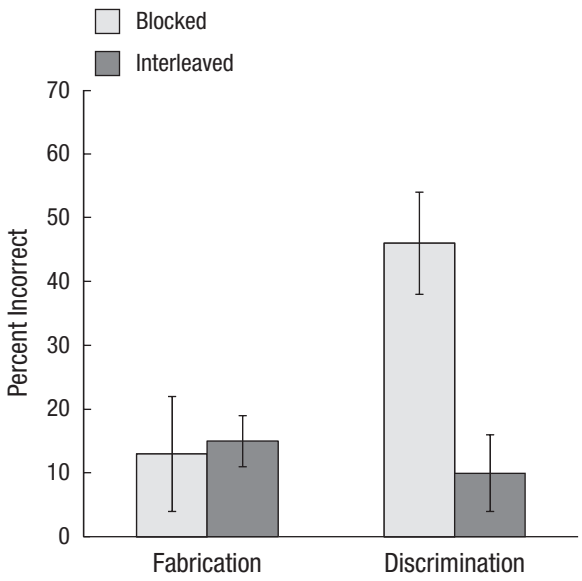


Fig. 14. Types of errors made by fourth graders while solving mathematical problems on a delayed criterion test in Taylor and Rohrer (2010). Error bars represent standard errors.

How does interleaving produce these benefits? One explanation is that interleaved practice promotes organizational processing and item-specific processing because it allows students to more readily compare different kinds of problems. For instance, in Rohrer and Taylor (2007), it is possible that when students were solving for the volume of one kind of solid (e.g., a wedge) during interleaved practice, the solution method used for the immediately prior problem involving a different kind of solid (e.g., a spheroid) was still in working memory and hence encouraged a comparison of the two problems and their different formulas. Another possible explanation is based on the distributed retrieval from long-term memory that is afforded by interleaved practice. In particular, for blocked practice, the information relevant to completing a task (whether it be a solution to a problem or memory for a set of related items) should reside in working memory; hence, participants should not have to retrieve the solution. So, if a student completes a block of problems solving for volumes of wedges, the solution to each new problem will be readily available from working memory. By contrast, for interleaved practice, when the next type of problem is presented, the solution method for it must be retrieved from long-term memory. So, if a student has just solved for the volume of a wedge and then must solve for the volume of a spheroid, he or she must retrieve the formula for spheroids from memory. Such delayed practice testing would boost memory for the retrieved information (for details, see the Practice Testing section above). This retrieval-practice hypothesis and the discriminative-contrast hypothesis are not mutually exclusive, and other mechanisms may also contribute to the benefits of interleaved practice.

10.2 How general are the effects of interleaved practice?

10.2a Learning conditions. Interleaved practice itself represents a learning condition, and it naturally covaries with distributed practice. For instance, if the practice trials for tasks of a given kind are blocked, the practice for the task is massed. By contrast, by interleaving practice across tasks of different kinds, any two instances of a task from a given set (e.g., solving for the volume of a given type of geometrical solid) would be separated by practice of instances from other tasks. Thus, at least some of the benefits of interleaved practice may reflect the benefits of distributed practice. However, some researchers have investigated the benefits of interleaved practice with spacing held constant (e.g., Kang & Pashler, 2012; Mitchell, Nash, & Hall, 2008), and the results suggested that spacing is not responsible for interleaving effects. For instance, Kang and Pashler (2012) had college students study paintings by various artists with the goal of developing a concept of each artists' style, so that the students could later correctly identify the artists who had produced paintings that had not been

presented during practice. During practice, the presentation of paintings was either blocked by artist (e.g., all paintings by Jan Blencowe were presented first, followed by all paintings by Richard Lindenberg, and so on) or interleaved. Most important, a third group received blocked practice, but instead of viewing the paintings one right after another in a massed fashion, a cartoon drawing was presented in between the presentation of each painting (the cartoons were presented so that the temporal spacing in this spaced-block-practice group was the same as that for the interleaved group). Criterion performance was best after interleaved practice and was significantly better than after either standard or temporally spaced blocked practice. No differences occurred in performance between the two blocked-practice groups, which indicates that spacing alone will not consistently benefit concept formation.

This outcome is more consistent with the discriminative-contrast hypothesis than the retrieval-practice hypothesis. In particular, on each trial, the group receiving temporally spaced blocked practice presumably needed to retrieve (from long-term memory) what they had already learned about a painters' style, yet doing so did not boost their performance. That is, interleaved practice encouraged students to identify the critical differences among the various artists' styles, which in turn helped students discriminate among the artists' paintings on the criterion test. According to this hypothesis, interleaved practice may further enhance students' ability to develop accurate concepts (e.g., a concept of an artist's style) when exemplars of different concepts are presented simultaneously. For instance, instead of paintings being presented separately but in an interleaved fashion, a set of paintings could be presented at the same time. In this case, a student could more readily scan the paintings of the various artists to identify differences among them. Kang and Pashler (2012) found that simultaneous presentation of paintings from different artists yielded about the same level of criterion performance (68%) as standard interleaving did (65%), and that both types of interleaved practice were superior to blocked practice (58%; for a similar finding involving students learning to classify birds, see Wahlheim, Dunlosky, & Jacoby, 2011).

Finally, the amount of instruction and practice that students initially receive with each task may influence the degree to which interleaving all tasks enhances performance. In fact, in educational contexts, introducing a new concept or problem type (e.g., how to find the volume of a spheroid) would naturally begin with initial instruction and blocked practice with that concept or problem type, and most of the studies reported in this section involved an introduction to all tasks before interleaving began. The question is how much initial practice is enough, and whether students with low skill levels (or students learning to solve more difficult tasks) will require more practice before interleaving begins. Given that skill level and task difficulty have been shown to moderate the benefits of interleaving in the literature on motor learning (e.g., Brady, 1998; Wulf & Shea, 2002), it seems likely that they do the same for cognitive tasks. If so, the dosage of initial instruction and blocked practice should interact with the benefits of interleaving, such that more pretraining should be required for younger and less skilled students, as well as for more complex tasks.

Consistent with this possibility are findings from Rau, Aleven, and Rummel (2010), who used various practice schedules to help teach fifth and sixth graders about fractions. During practice, students were presented with different ways to represent fractions, such as with pie charts, line segments, and set representations. Practice was either blocked (e.g., students worked with pie charts first, then line segments, and so on), interleaved, or first blocked and then interleaved. The prepractice and postpractice criterion tests involved fractions. Increases in accuracy from the prepractice test to the postpractice test occurred only after blocked and blocked-plus-interleaved practice (students in these two groups tended to perform similarly), and then, these benefits were largely shown only for students with low prior knowledge. This outcome provides partial support for the hypothesis that interleaved practice may be most beneficial only after a certain level of competency has been achieved using blocked practice with an individual concept or problem type.

10.2b Student characteristics. The majority of studies on interleaved practice have included college-aged students, and across these studies, sometimes interleaved practice has boosted performance, and sometimes it has not. Even so, differences in the effectiveness of interleaved practice for this age group are likely more relevant to the kind of task employed or, perhaps, to the dosage of practice, factors that we discuss in other sections. Some studies have included college students who were learning tasks relevant to their career goals—for instance, engineering students who were learning to diagnose system failures (e.g., de Croock, van Merriënboer, & Paas, 1998) and medical students who were learning to interpret electrocardiograms (Hatala, Brooks, & Norman, 2003). We highlight outcomes from these studies in the Materials subsection (10.2c) below. Finally, Mayfield and Chase (2002) conducted an extensive intervention to train algebra to college students with poor math skills; interleaving was largely successful, and we describe this experiment in detail in the Effects in Representative Educational Contexts subsection (10.3) below.

Concerning younger students, as reported above, Taylor and Rohrer (2010) reported that fourth graders benefited from interleaved practice when they were learning how to solve mathematical problems. In contrast, Rau et al. (2010) used various practice schedules to help teach fifth and sixth graders about fractions and found that interleaved practice did not boost performance. Finally, Olina, Reiser, Huang, Lim, and Park (2006) had high school students learn various rules for comma usage with interleaved or blocked practice; higher-skill students appeared to be hurt by interleaving (although pretests scores favored those in the blocked group, and that advantage may have carried through to the criterion test), and interleaving did not help lower-skill students.

10.2c Materials. The benefits of interleaved practice have been explored using a variety of cognitive tasks and materials, from the simple (e.g., paired associate learning) to the relatively complex (e.g., diagnosing failures of a complicated piece of machinery). Outcomes have been mixed. Schneider, Healy, and Bourne (1998, 2002) had college students learn French vocabulary words from different categories, such as body parts, dinnerware, and foods. Across multiple studies, translation equivalents from the same category were blocked during practice or were interleaved. Immediately after practice, students who had received blocked practice recalled more translations than did students who had received interleaved practice (Schneider et al., 2002). One week after practice, correct recall was essentially the same in the blocked-practice group as in the interleaved-practice group. In another study (Schneider et al., 1998, Experiment 2), interleaved practice led to somewhat better performance than blocked practice on a delayed test, but this benefit was largely due to a slightly lower error rate. Based on these two studies, it does not appear that interleaved practice of vocabulary boosts retention.

More promising are results from studies that have investigated students' learning of mathematics. We have already described some of these studies above (Rohrer & Taylor, 2007; Taylor & Rohrer, 2010; but see Rau et al., 2010). Other math skills that have been trained include the use of Boolean functions (Carlson & Shin, 1996; Carlson & Yaure, 1990) and algebraic skills (Mayfield & Chase, 2002). For the former, interleaved practice improved students' speed in solving multistep Boolean problems, especially when students could preview the entire multistep problem during solution (Carlson & Shin, 1996). For the latter, interleaving substantially boosted students' ability to solve novel algebra problems (as we discuss in detail below).

Van Merriënboer and colleagues (de Croock & van Merriënboer, 2007; de Croock et al., 1998; van Merriënboer, de Croock, & Jelsma, 1997; van Merriënboer, Schuurman, de Croock, & Paas, 2002) trained students to diagnose problems that occurred in a distiller system in which different components could fail; practice at diagnosing failures involving each component was either blocked or interleaved during practice. Across their studies, interleaved practice sometimes led to better performance on transfer tasks (which involved new combinations of system failures), but it did not always boost performance, leading the authors to suggest that perhaps more practice was needed to demonstrate the superiority of interleaved practice (de Croock & van Merriënboer, 2007). Blocked and interleaved practice have also been used to train students to make complex multidimensional judgments (Helsdingen, van Gog, & van Merriënboer, 2011a, 2011b), with results showing that decision making on criterion tests was better after interleaved than blocked practice. One impressive outcome was reported by Hatala et al. (2003), who trained medical students to make electrocardiogram diagnoses for myocardial infarction, ventricular hypertrophy, bundle branch blocks, and ischemia. The criterion test was on novel diagnoses, and accuracy was substantially greater after interleaved practice (47%) than after blocked practice (30%).

Finally, interleaved practice has been shown to improve the formation of concepts about artists' painting styles (Kang & Pashler, 2012; Kornell & Bjork, 2008) and about bird classifications (Wahlheim et al., 2011). The degree to which the benefits of interleaving improve concept formation across different kinds of concepts (and for students of different abilities) is currently unknown, but research and theory by Goldstone (1996) suggest that interleaving will not always be better. In particular, when exemplars within a category are dissimilar, blocking may be superior, because it will help learners identify what the members of a category have in common. By contrast, when exemplars from different categories are similar (as with the styles of artists and the classifications of birds used in the prior interleaving studies on concept formation), interleaving may work best because of discriminative contrast (e.g., Carvalho & Goldstone, 2011). These possibilities should be thoroughly explored with naturalistic materials before any general recommendations can be offered concerning the use of interleaved practice for concept formation.

10.2d Criterion tasks. In the literature on interleaving, the materials that are the focus of instruction and practice are used as the criterion task. Thus, if students practice solving problems of a certain kind, the criterion task will involve solving different versions of that kind of problem. For this reason, the current section largely reflects the analysis of the preceding section on materials (10.2c). One remaining issue, however, concerns the degree to which the benefits of interleaved practice are maintained across time. Although the delay between practice and criterion tests for many of the studies described above was minimal, several studies have used retention intervals as long as 1 to 2 weeks. In some of these cases, interleaved practice benefited performance (e.g., Mayfield & Chase, 2002; Rohrer & Taylor, 2007), but in others, the potential benefits of interleaving did not manifest after the longer retention interval (e.g., de Croock & van Merriënboer, 2007; Rau et al., 2010). In the latter cases, interleaved practice may not have been potent at any retention interval. For instance, interleaved practice may not be potent for learning foreign-language vocabulary (Schneider et al., 1998) or for students who have not received enough practice with a complex task (de Croock & van Merriënboer, 2007).

10.3 Effects in representative educational contexts. It seems plausible that motivated students could easily use interleaving without help. Moreover, several studies have used procedures for instruction that could be used in the classroom (e.g., Hatala et al., 2003; Mayfield & Chase, 2002; Olina et al., 2006; Rau et al, 2010). We highlight one exemplary study here. Mayfield and Chase (2002) taught algebra rules to college students with poor math skills across 25 sessions. In different sessions, either a single algebra rule was introduced or previously introduced rules were reviewed. For review

sessions, either the rule learned in the immediately previous session was reviewed (which was analogous to blocking) or the rule learned in the previous session was reviewed along with the rules from earlier sessions (which was analogous to interleaved practice). Tests were administered prior to training, during the session after each review, and then 4 to 9 weeks after practice ended. On the tests, students had to apply the rules they had learned as well as solve problems by using novel combinations of the trained rules. The groups performed similarly at the beginning of training, but by the final tests, performance on both application and problem-solving items was substantially better for the interleaved group, and these benefits were still evident (albeit no longer statistically significant) on the delayed retention test.

10.4 Issues for implementation. Not only is the result from Mayfield and Chase (2002) promising, their procedure offers a tactic for the implementation of interleaved practice, both by teachers in the classroom and by students regulating their study (for a detailed discussion of implementation, see Rohrer, 2009). In particular, after a given kind of problem (or topic) has been introduced, practice should first focus on that particular problem. After the next kind of problem is introduced (e.g., during another lecture or study session), that problem should first be practiced, but it should be followed by extra practice that involves interleaving the current type of problem with others introduced during previous sessions. As each new type of problem is introduced, practice should be interleaved with practice for problems from other sessions that students will be expected to discriminate between (e.g., if the criterion test will involve a mixture of several types of problems, then these should be practiced in an interleaved manner during class or study sessions). Interleaved practice may take a bit more time to use than blocked practice, because solution times often slow during interleaved practice; even so, such slowing likely indicates the recruitment of other processes—such as discriminative contrast—that boost performance. Thus, teachers and students could integrate interleaved practice into their schedules without too much modification.

10.5 Interleaved practice: Overall recommendations. On the basis of the available evidence, we rate interleaved practice as having moderate utility. On the positive side, interleaved practice has been shown to have relatively dramatic effects on students' learning and retention of mathematical skills, and teachers and students should consider adopting it in the appropriate contexts. Also, interleaving does help (and rarely hinders) other kinds of cognitive skills. On the negative side, the literature on interleaved practice is currently small, but it contains enough null effects to raise concern. Although the null effects may indicate that the technique does not consistently work well, they may instead reflect that we do not fully understand the mechanisms underlying the effects of interleaving and therefore do not always use it appropriately. For instance, in some cases, students may not have had enough instruction or practice with individual tasks to reap the benefits of interleaved practice. Given the promise of interleaved practice for improving student achievement, there is a great need for research that systematically evaluates how its benefits are moderated by dosage during training, student abilities, and the difficulty of materials.

Closing Remarks

Relative utility of the learning techniques

Our goal was to provide reviews that were extensive enough to allow anyone interested in using a particular technique to judge its utility for his or her own instructional or learning goals. We also realized that offering some general ratings (and the reasons behind them) might be useful to readers interested in quickly obtaining an overview on what technique may work best. To do so, we have provided an assessment of how each technique fared with respect to the generalizability of its benefits across the four categories of variables listed in Table 2, issues for implementation, and evidence for its effectiveness from work in representative educational contexts (see Table 4). Our goal for these assessments was to indicate both (a) whether sufficient evidence is available to support conclusions about the generalizability of a technique, issues for its implementation, or its efficacy in educational contexts, and, if sufficient evidence does exist, (b) whether it indicates that the technique works.[3] For instance, practice testing received an assessment of *Positive* (P) for criterion tasks; this rating indicates that we found enough evidence to conclude that practice testing benefits student performance across a wide range of criterion tasks and retention intervals. Of course, it does not mean that further work in this area (i.e., testing with different criterion tasks) would not be valuable, but the extent of the evidence is promising enough to recommend it to teachers and students.

A *Negative* (N) rating indicates that the available evidence shows that the learning technique does not benefit performance for the particular category or issue. For instance, despite its popularity, highlighting did not boost performance across a variety of criterion tasks, so it received a rating of N for this variable.

A *Qualified* (Q) rating indicates that both positive and negative evidence has been reported with respect to a particular category or issue. For instance, the keyword mnemonic received a Q rating for materials, because evidence indicates that this technique does work for learning materials that are imagery friendly but does not work well for materials that cannot be easily imagined.

A rating of *Insufficient* (I) indicates that insufficient evidence is available to draw conclusions about the effects of a given technique for a particular category or issue. For instance, elaborative interrogation received an I rating for criterion tasks because we currently do not know whether its effects are durable across educationally relevant retention intervals. Any cell

Table 4. Utility Assessment and Ratings of Generalizability for Each of the Learning Techniques

Technique	Utility	Learners	Materials	Criterion tasks	Issues for implementation	Educational contexts
Elaborative interrogation	Moderate	P-I	P	I	P	I
Self-explanation	Moderate	P-I	P	P-I	Q	I
Summarization	Low	Q	P-I	Q	Q	I
Highlighting	Low	Q	Q	N	P	N
The keyword mnemonic	Low	Q	Q	Q-I	Q	Q-I
Imagery use for text learning	Low	Q	Q	Q-I	P	I
Rereading	Low	I	P	Q-I	P	I
Practice testing	High	P-I	P	P	P	P
Distributed practice	High	P-I	P	P-I	P	P-I
Interleaved practice	Moderate	I	Q	P-I	P	P-I

Note: A positive (P) rating indicates that available evidence demonstrates efficacy of a learning technique with respect to a given variable or issue. A negative (N) rating indicates that a technique is largely ineffective for a given variable. A qualified (Q) rating indicates that the technique yielded positive effects under some conditions (or in some groups) but not others. An insufficient (I) rating indicates that there is insufficient evidence to support a definitive assessment for one or more factors for a given variable or issue.

in Table 4 with an I rating highlights the need for further systematic research.

Finally, some cells include more than one rating. In these cases, enough evidence exists to evaluate a technique on one dimension of a category or issue, yet insufficient evidence is available for some other dimension. For instance, self-explanation received a P-I rating for criterion tasks because the available evidence is positive on one dimension (generalizability across a range of criterion tasks) but is insufficient on another key dimension (whether the benefit of self-explanation generalizes across longer retention intervals). As another example, rereading received a Q-I rating for criterion tasks because evidence for the effectiveness of this technique over long retention intervals is qualified (i.e., under some learning conditions, it does not produce an effect for longer retention intervals), and insufficient evidence is available that is relevant to its effectiveness across different kinds of criterion tasks (e.g., rereading does boost performance on recall tasks, but little is known as to its benefits for comprehension). When techniques have multiple ratings for one or more variables, readers will need to consult the reviews for details.

Finally, we used these ratings to develop an overall utility assessment for each of the learning techniques. The utility assessments largely reflect how well the benefits of each learning technique generalize across the different categories of variables (e.g., for how many variables the technique received a P rating). For example, the keyword mnemonic and imagery use for text learning were rated low in utility in part because their effects are limited to materials that are amenable to imagery and because they may not work well for students of all ages. Even so, some teachers may decide that the benefits of techniques with low-utility ratings match their instructional goals for their students. Thus, although we do offer these easy-to-use assessments of each learning technique, we also encourage interested teachers and students to carefully read each review to make informed decisions about which techniques will best meet their instructional and learning goals.

Implications for research on learning techniques

A main goal of this monograph was to develop evidence-based recommendations for teachers and students about the relative utility of various learning techniques. A related goal was to identify areas that have been underinvestigated and that will require further research before evidence-based recommendations for their use in education can be made. A number of these gaps are immediately apparent upon inspection of Table 4. To highlight a few, we do not yet know the extent to which many of the learning techniques will benefit students of various ages, abilities, and levels of prior knowledge. Likewise, with a few exceptions (e.g., practice testing and distributed practice), the degree to which many of the techniques support durable learning (e.g., over a number of weeks) is largely unknown, partly because investigations of these techniques have typically involved a single session that included both practice and criterion tests (for a discussion of the limitations of such single-session research, see Rawson & Dunlosky, 2011). Finally, few techniques have been evaluated in representative educational contexts.

This appraisal (along with Table 4) suggests two directions for future research that could have immediate implications for education. First, more research is needed to fully explore the degree to which the benefits of some techniques generalize to the variables listed in Table 2. Particularly important will be investigations that evaluate the degree to which interactions among the variables limit or magnify the benefits of a given technique. Second, the benefit of most of the techniques in representative educational settings needs to be more fully explored. Easy-to-use versions of the most promising tech-

niques should be developed and evaluated in controlled investigations conducted in educationally representative contexts. Ideally, the criterion measures would include high-stakes tests, such as performance on in-class exams and on achievement tests. We realize that such research efforts can be time-consuming and costly, but conducting them will be crucial for recommending educational changes that will have a reasonable likelihood of improving student learning and achievement.

Implications for students, teachers, and student achievement

Pressley and colleagues (Pressley, 1986; Pressley, Goodchild, et al., 1989) developed a good-strategy-user model, according to which being a sophisticated strategy user involves "knowing the techniques that accomplish important life goals (i.e., strategies), knowing when and how to use those methods . . . and using those methods in combination with a rich network of nonstrategic knowledge that one possesses about the world" (p. 302). However, Pressley, Goodchild, et al. (1989) also noted that "many students are committed to ineffective strategies . . . moreover, there is not enough professional evaluation of techniques that are recommended in the literature, with many strategies oversold by proponents" (p. 301). We agree and hope that the current reviews will have a positive impact with respect to fostering further scientific evaluation of the techniques.

Concerning students' commitment to ineffective strategies, recent surveys have indicated that students most often endorse the use of rereading and highlighting, two strategies that we found to have relatively low utility. Nevertheless, some students do report using practice testing, and these students appear to benefit from its use. For instance, Gurung (2005) had college students describe the strategies they used in preparing for classroom examinations in an introductory psychology course. The frequency of students' reported use of practice testing was significantly correlated with their performance on a final exam (see also Hartwig & Dunlosky, 2012). Given that practice testing is relatively easy to use, students who do not currently use this technique should be able to incorporate it into their study routine.

Why don't many students consistently use effective techniques? One possibility is that students are not instructed about which techniques are effective or how to use them effectively during formal schooling. Part of the problem may be that teachers themselves are not told about the efficacy of various learning techniques. Given that teachers would most likely learn about these techniques in classes on educational psychology, it is revealing that most of the techniques do not receive sufficient coverage in educational-psychology textbooks. We surveyed six textbooks (cited in the Introduction), and, except for mnemonics based on imagery (e.g., the keyword mnemonic), none of the techniques was covered by all of the books. Moreover, in the subset of textbooks that did describe one or more of these techniques, the coverage in most cases was relatively minimal, with a brief description of a given technique and relatively little guidance on its use, effectiveness, and limitations. Thus, many teachers are unlikely getting a sufficient introduction to which techniques work best and how to train students to use them.

A second problem may be that a premium is placed on teaching students content and critical-thinking skills, whereas less time is spent teaching students to develop effective techniques and strategies to guide learning. As noted by McNamara (2010), "there is an overwhelming assumption in our educational system that the most important thing to *deliver* to students is content" (p. 341, italics in original). One concern here is that students who do well in earlier grades, in which learning is largely supervised, may struggle later, when they are expected to regulate much of their own learning, such as in high school or college. Teaching students to use these techniques would not take much time away from teaching content and would likely be most beneficial if the use of the techniques was consistently taught across multiple content areas, so that students could broadly experience their effects on learning and class grades. Even here, however, recommendations on how to train students to use the most effective techniques would benefit from further research. One key issue concerns the earliest age at which a given technique could (or should) be taught. Teachers can expect that upper elementary students should be capable of using many of the techniques, yet even these students may need some guidance on how to most effectively implement them. Certainly, identifying the age at which students have the self-regulatory capabilities to effectively use a technique (and how much training they would need to do so) is an important objective for future research. Another issue is how often students will need to be retrained or reminded to use the techniques to ensure that students will continue to use them when they are not instructed to do so. Given the promise of some of the learning techniques, research on professional development that involves training teachers to help students use the techniques would be valuable.

Beyond training students to use these techniques, teachers could also incorporate some of them into their lesson plans. For instance, when beginning a new section of a unit, a teacher could begin with a practice test (with feedback) on the most important ideas from the previous section. When students are practicing problems from a unit on mathematics, recently studied problems could be interleaved with related problems from previous units. Teachers could also harness distributed practice by re-presenting the most important concepts and activities over the course of several classes. When introducing key concepts or facts in class, teachers could engage students in explanatory questioning by prompting them to consider how the information is new to them, how it relates to what they already know, or why it might be true. Even homework assignments could be designed to take advantage of many of these techniques. In these examples (and in others provided in the Issues for Implementation subsections), teachers could

implement a technique to help students learn, regardless of whether students are themselves aware that a particular technique is being used.

We realize that many factors are responsible whenever any one student fails to achieve in school (Hattie, 2009) and hence that a change to any single factor may have a relatively limited effect on student learning and achievement. The learning techniques described in this monograph will not be a panacea for improving achievement for all students, and perhaps obviously, they will benefit only students who are motivated and capable of using them. Nevertheless, when used properly, we suspect that they will produce meaningful gains in performance in the classroom, on achievement tests, and on many tasks encountered across the life span. It is obvious that many students are not using effective learning techniques but could use the more effective techniques without much effort, so teachers should be encouraged to more consistently (and explicitly) train students to use learning techniques as they are engaged in pursuing various instructional and learning goals.

Acknowledgments

We thank Reed Hunt, Mark McDaniel, Roddy Roediger, and Keith Thiede for input about various aspects of this monograph. We appreciate constructive comments about a draft from Sean Kang, Alexsa MacKendrick, Richard Mayer, Hal Pashler, Dan Robinson, and Doug Rohrer. Thanks also to Robert Goldstone, Detlev Leutner, and Doug Rohrer for providing details of their research, and to Melissa Bishop and Cindy Widuck for technical support.

Declaration of Conflicting Interests

The authors declared that they had no conflicts of interest with respect to their authorship or the publication of this article.

Funding

This research was supported by a Bridging Brain, Mind and Behavior Collaborative Award through the James S. McDonnell Foundation's 21st Century Science Initiative.

Notes

1. We also recommend a recent practice guide from the U.S. Institute of Education Sciences (Pashler et al., 2007), which discusses some of the techniques described here. The current monograph, however, provides more in-depth and up-to-date reviews of the techniques and also reviews some techniques not included in the practice guide.
2. Although this presentation mode does not involve reading per se, reading comprehension and listening comprehension processes are highly similar aside from differences at the level of decoding the perceptual input (Gernsbacher, Varner, & Faust, 1990).
3. We did not include learning conditions as a category of variable in this table because the techniques vary greatly with respect to relevant learning conditions. Please see the reviews for assessments of how well the techniques generalized across relevant learning conditions.

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.

Agarwal, P. K., & Roediger, H. L., III. (2011). Expectancy of an open-book test decreases performance on a delayed closed-book test. *Memory*, *19*, 836–852.

Ainsworth, S., & Burcham, S. (2007). The impact of text coherence on learning by self-explanation. *Learning and Instruction*, *17*, 286–303.

Aleven, V., & Koedinger, K. R. (2002). An effective metacognitive strategy: Learning by doing and explaining with a computer-based cognitive tutor. *Cognitive Science*, *26*, 147–179.

Amer, A. A. (1994). The effect of knowledge-map and underlining training on the reading comprehension of scientific texts. *English for Specific Purposes*, *13*, 35–45.

Amlund, J. T., Kardash, C. A. M., & Kulhavy, R. W. (1986). Repetitive reading and recall of expository text. *Reading Research Quarterly*, *21*, 49–58.

Anderson, L. W., & Krathwohl, D. R. (Eds.). (2001). *A taxonomy for learning, teaching and assessing: A revision of Bloom's taxonomy of educational objectives: Complete edition*. New York, NY: Longman.

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

Anderson, R. C., & Hidde, J. L. (1971). Imagery and sentence learning. *Journal of Educational Psychology*, *62*, 526–530.

Anderson, R. C., & Kulhavy, R. W. (1972). Imagery and prose learning. *Journal of Educational Psychology*, *63*, 242–243.

Anderson, T. H., & Armbruster, B. B. (1984). Studying. In R. Barr (Ed.), *Handbook of reading research* (pp. 657–679). White Plains, NY: Longman.

Annis, L., & Davis, J. K. (1978). Study techniques: Comparing their effectiveness. *American Biology Teacher*, *40*, 108–110.

Annis, L. F. (1985). Student-generated paragraph summaries and the information-processing theory of prose learning. *Journal of Experimental Education*, *51*, 4–10.

Appleton-Knapp, S. L., Bjork, R. A., & Wickens, T. D. (2005). Examining the spacing effect in advertising: Encoding variability, retrieval processes, and their interaction. *Journal of Consumer Research*, *32*, 266–276.

Armbruster, B. B., Anderson, T. H., & Ostertag, J. (1987). Does text structure/summarization instruction facilitate learning from expository text? *Reading Research Quarterly*, *22*, 331–346.

Arnold, H. F. (1942). The comparative effectiveness of certain study techniques in the field of history. *Journal of Educational Psychology*, *32*, 449–457.

Atkinson, R. C., & Paulson, J. A. (1972). An approach to the psychology of instruction. *Psychological Bulletin*, *78*, 49–61.

Atkinson, R. C., & Raugh, M. R. (1975). An application of the mnemonic keyword method to the acquisition of a Russian vocabulary. *Journal of Experimental Psychology: Human Learning and Memory*, *104*, 126–133.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General*, *108*, 296–308.

Bahrick, H. P., Bahrick, L. E., Bahrick, A. S., & Bahrick, P. E. (1993). Maintenance of foreign language vocabulary and the spacing effect. *Psychological Science*, *4*, 316–321.

Bahrick, H. P., & Hall, L. K. (2005). The importance of retrieval failures to long-term retention: A metacognitive explanation of the spacing effect. *Journal of Memory and Language*, *52*, 566–577.

Bahrick, H. P., & Phelps, E. (1987). Retention of Spanish vocabulary over 8 years. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 344–349.

Balch, W. R. (1998). Practice versus review exams and final exam performance. *Teaching of Psychology*, *25*, 181–185.

Balch, W. R. (2006). Encouraging distributed study: A classroom experiment on the spacing effect. *Teaching of Psychology*, *33*, 249–252.

Balota, D. A., Duchek, J. M., & Paullin, R. (1989). Age-related differences in the impact of spacing, lag, and retention interval. *Psychology and Aging*, *4*, 3–9.

Balota, D. A., Duchek, J. M., Sergent-Marshall, S. D., & Roediger, H. L., III. (2006). Does expanded retrieval produce benefits over equal-interval spacing? Explorations of spacing effects in healthy aging and early stage Alzheimer's disease. *Psychology and Aging*, *21*, 19–31.

Bangert-Drowns, R. L., Kulik, J. A., & Kulik, C.-L. C. (1991). Effects of frequent classroom testing. *Journal of Educational Research*, *85*, 89–99.

Barcroft, J. (2007). Effect of opportunities for word retrieval during second language vocabulary learning. *Language Learning*, *57*, 35–56.

Barnett, J. E., & Seefeldt, R. W. (1989). Read something once, why read it again? Repetitive reading and recall. *Journal of Reading Behavior*, *21*, 351–360.

Bean, T. W., & Steenwyk, F. L. (1984). The effect of three forms of summarization instruction on sixth graders' summary writing and comprehension. *Journal of Reading Behavior*, *16*, 297–306.

Bednall, T. C., & Kehoe, E. J. (2011). Effects of self-regulatory instructional aids on self-directed study. *Instructional Science*, *39*, 205–226.

Bell, K. E., & Limber, J. E. (2010). Reading skill, textbook marking, and course performance. *Literacy Research and Instruction*, *49*, 56–67.

Benjamin, A. S., & Bird, R. D. (2006). Metacognitive control of the spacing of study repetitions. *Journal of Memory and Language*, *55*, 126–137.

Benjamin, A. S., & Tullis, J. (2010). What makes distributed practice effective? *Cognitive Psychology*, *61*, 228–247.

Berry, D. C. (1983). Metacognitive experience and transfer of logical reasoning. *Quarterly Journal of Experimental Psychology, 35A*, 39–49.

Bishara, A. J., & Jacoby, L. L. (2008). Aging, spaced retrieval, and inflexible memory performance. *Psychonomic Bulletin & Review*, *15*, 52–57.

Blanchard, J., & Mikkelson, V. (1987). Underlining performance outcomes in expository text. *Journal of Educational Research*, *80*, 197–201.

Bloom, B. S., Engelhart, M., Furst, E. J., Hill, W., & Krathwohl, D. R. (1956). *Taxonomy of educational objectives, Handbook I: Cognitive domain*. New York, NY: Longman.

Bloom, K. C., & Shuell, T. J. (1981). Effects of massed and distributed practice on the learning and retention of second-language vocabulary. *Journal of Educational Research*, *74*, 245–248.

Bouwmeester, S., & Verkoeijen, P. P. J. L. (2011). Why do some children benefit more from testing than others? Gist trace processing to explain the testing effect. *Journal of Memory and Language*, *65*, 32–41.

Brady, F. (1998). A theoretical and empirical review of the contextual interference effect and the learning of motor skills. *QUEST*, *50*, 266–293.

Bransford, J. D., & Franks, J. J. (1971). The abstraction of linguistic ideas. *Cognitive Psychology*, *2*, 331–350.

Bretzing, B. H., & Kulhavy, R. W. (1979). Notetaking and depth of processing. *Contemporary Educational Psychology*, *4*, 145–153.

Bretzing, B. H., & Kulhavy, R. W. (1981). Note-taking and passage style. *Journal of Educational Psychology*, *73*, 242–250.

Bromage, B. K., & Mayer, R. E. (1986). Quantitative and qualitative effects of repetition on learning from technical text. *Journal of Educational Psychology*, *78*, 271–278.

Brooks, L. R. (1967). The suppression of visualization by reading. *The Quarterly Journal of Experimental Psychology*, *19*, 289–299.

Brooks, L. R. (1968). Spatial and verbal components of the act of recall. *Canadian Journal of Psychology*, *22*, 349–368.

Brooks, L. W., Dansereau, D. F., Holley, C. D., & Spurlin, J. E. (1983). Generation of descriptive text headings. *Contemporary Educational Psychology*, *8*, 103–108.

Brown, A. L., Campione, J. C., & Day, J. D. (1981). Learning to learn: On training students to learn from texts. *Educational Researcher*, *10*, 14–21.

Brown, A. L., & Day, J. D. (1983). Macrorules for summarizing texts: The development of expertise. *Journal of Verbal Learning and Verbal Behavior*, *22*, 1–14.

Brown, A. L., Day, J. D., & Jones, R. S. (1983). The development of plans for summarizing texts. *Child Development*, *54*, 968–979.

Brown, L. B., & Smiley, S. S. (1978). The development of strategies for studying texts. *Child Development*, *49*, 1076–1088.

Brozo, W. G., Stahl, N. A., & Gordon, B. (1985). Training effects of summarizing, item writing, and knowledge of information sources on reading test performance. *Issues in Literacy: A Research Perspective—34th Yearbook of the National Reading Conference* (pp. 48–54). Rochester, NY: National Reading Conference.

Budé, L., Imbos, T., van de Wiel, M. W., & Berger, M. P. (2011). The effect of distributed practice on students' conceptual understanding of statistics. *Higher Education*, *62*, 69–79.

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

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III. (2008). Correcting a metacognitive error: Feedback increases retention of low-confidence correct responses. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 918–928.

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.

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

Butterfield, B., & Metcalfe, J. (2001). Errors committed with high confidence are hypercorrected. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *27*, 1491–1494.

Callender, A. A., & McDaniel, M. A. (2007). The benefits of embedded question adjuncts for low and high structure builders. *Journal of Educational Psychology*, *99*, 339–348.

Callender, A. A., & McDaniel, M. A. (2009). The limited benefits of rereading educational texts. *Contemporary Educational Psychology*, *34*, 30–41.

Carlson, R. A., & Shin, J. C. (1996). Practice schedules and subgoal instantiation in cascaded problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *22*, 157–168.

Carlson, R. A., & Yaure, R. G. (1990). Practice schedules and the use of component skills in problem solving. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 484–496.

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and 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, and Cognition*, *37*, 1547–1552.

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology*, *19*, 619–636.

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*, 268–276.

Carpenter, S. K., & Pashler, H. (2007). Testing beyond words: Using tests to enhance visuospatial map learning. *Psychonomic Bulletin & Review*, *14*, 474–478.

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., Pashler, H., & Vul, E. (2006). What types of learning are enhanced by a cued recall test? *Psychonomic Bulletin & Review*, *13*, 826–830.

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., & Vul, E. (2011). Delaying feedback by three seconds benefits retention of face-name pairs: The role of active anticipatory processing. *Memory & Cognition*, *39*, 1211–1221.

Carr, E., Bigler, M., & Morningstar, C. (1991). The effects of the CVS strategy on children's learning. *Learner Factors/Teacher Factors: Issues in Literacy Research and Instruction—40th Yearbook of the National Reading Conference* (pp. 193–200). Rochester, NY: National Reading Conference.

Carrier, L. M. (2003). College students' choices of study strategies. *Perceptual & Motor Skills*, *96*, 54–56.

Carroll, M., Campbell-Ratcliffe, J., Murnane, H., & Perfect, T. (2007). Retrieval-induced forgetting in educational contexts: Monitoring, expertise, text integration, and test format. *European Journal of Cognitive Psychology*, *19* , 580–606.

Carvalho, P. F., & Goldstone, R. L. (2011, November). *Comparison between successively presented stimuli during blocked and interleaved presentations in category learning*. Paper presented at the 52nd Annual Meeting of the Psychonomic Society, Seattle, WA.

Cashen, M. C., & Leicht, K. L. (1970). Role of the isolation effect in a formal educational setting. *Journal of Educational Psychology*, *61*, 484–486.

Cepeda, N. J., Coburn, N., Rohrer, D., Wixted, J. T., Mozer, M. C., & Pashler, H. (2009). Optimizing distributed practice: Theoretical analysis and practical implications. *Experimental Psychology*, *56*, 236–246.

Cepeda, N. J., Pashler, H., Vul, E., Wixted, J. T., & Rohrer, D. (2006). Distributed practice in verbal recall tasks: A review and quantitative synthesis. *Psychological Bulletin*, *132*, 354–380.

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological Science*, *19*, 1095–1102.

Cermak, L. S., Verfaellie, M., Lanzoni, S., Mather, M., & Chase, K. A. (1996). Effect of spaced repetitions on amnesia patients' recall and recognition performance. *Neuropsychology*, *10*, 219–227.

Challis, B. H. (1993). Spacing effects on cued-memory tests depend on level of processing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *19*, 389–396.

Chan, J. C. K. (2009). When does retrieval induce forgetting and when does it induce facilitation? Implications for retrieval inhibition, testing effect, and text processing. *Journal of Memory and Language*, *61*, 153–170.

Chan, J. C. K. (2010). Long-term effects of testing on the recall of nontested materials. *Memory*, *18*, 49–57.

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

Chan, L. K. S., Cole, P. G., & Morris, J. N. (1990). Effects of instruction in the use of a visual-imagery strategy on the reading-comprehension competence of disabled and average readers. *Learning Disability Quarterly*, *13*, 2–11.

Chi, M. T. H. (2009). Active-constructive-interactive: A conceptual framework for differentiating learning activities. *Topics in Cognitive Science*, *1*, 73–105.

Chi, M. T. H., de Leeuw, N., Chiu, M.-H., & LaVancher, C. (1994). Eliciting self-explanations improves understanding. *Cognitive Science*, *18*, 439–477.

Childers, J. B., & Tomasello, M. (2002). Two-year-olds learn novel nouns, verbs, and conventional actions from massed or distributed exposures. *Developmental Psychology*, *38*, 967–978.

Cioffi, G. (1986). Relationships among comprehension strategies reported by college students. *Reading Research and Instruction*, *25*, 220–231.

Condus, M. M., Marshall, K. J., & Miller, S. R. (1986). Effects of the keyword mnemonic strategy on vocabulary acquisition and maintenance by learning disabled children. *Journal of Learning Disabilities*, *19*, 609–613.

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.

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

Cranney, J., Ahn, M., McKinnon, R., Morris, S., & Watts, K. (2009). The testing effect, collaborative learning, and retrieval-induced facilitation in a classroom setting. *European Journal of Cognitive Psychology*, *21*, 919–940.

Crawford, C. C. (1925a). The correlation between college lecture notes and quiz papers. *Journal of Educational Research*, *12*, 282–291.

Crawford, C. C. (1925b). Some experimental studies of the results of college note-taking. *Journal of Educational Research*, *12*, 379–386.

Crouse, J. H., & Idstein, P. (1972). Effects of encoding cues on prose learning. *Journal of Educational Psychology*, *68*, 309–313.

Cull, W. L. (2000). Untangling the benefits of multiple study opportunities and repeated testing for cued recall. *Applied Cognitive Psychology*, *14*, 215–235.

Daniel, D. B., & Broida, J. (2004). Using web-based quizzing to improve exam performance: Lessons learned. *Teaching of Psychology*, *31*, 207–208.

De Beni, R., & Moè, A. (2003). Presentation modality effects in studying passages. Are mental images always effective? *Applied Cognitive Psychology*, *17*, 309–324.

de Bruin, A. B. H., Rikers, R. M. J. P., & Schmidt, H. G. (2007). The effect of self-explanation and prediction on the development of principled understanding of chess in novices. *Contemporary Educational Psychology*, *32*, 188–205.

de Croock, M. B. M., & van Merriënboer, J. J. G. (2007). Paradoxical effect of information presentation formats and contextual interference on transfer of a complex cognitive skill. *Computers in Human Behavior*, *23*, 1740–1761.

de Croock, M. B. M., van Merriënboer, J. J. G., & Pass, F. (1998). High versus low contextual interference in simulation-based training of troubleshooting skills: Effects on transfer performance and invested mental effort. *Computers in Human Behavior*, *14*, 249–267.

de Koning, B. B., Tabbers, H. K., Rikers, R. M. J. P., & Paas, F. (2011). Improved effectiveness of cueing by self-explanations when learning from a complex animation. *Applied Cognitive Psychology*, *25*, 183–194.

Delaney, P. F., & Knowles, M. E. (2005). Encoding strategy changes and spacing effects in free recall of unmixed lists. *Journal of Memory and Language*, *52*, 120–130.

Delaney, P. F., Verkoeijen, P. P. J. L., & Spirgel, A. (2010). Spacing and the testing effects: A deeply critical, lengthy, and at times discursive review of the literature. *Psychology of Learning and Motivation*, *53*, 63–147.

Dempster, F. N. (1987). Effects of variable encoding and spaced presentations on vocabulary learning. *Journal of Educational Psychology*, *79*, 162–170.

Dempster, F. N., & Farris, R. (1990). The spacing effect: Research and practice. *Journal of Research and Development in Education*, *23*, 97–101.

Denis, M. (1982). Imaging while reading text: A study of individual differences. *Memory & Cognition*, *10*, 540–545.

Didierjean, A., & Cauzinille-Marmèche, E. (1997). Eliciting self-explanations improves problem solving: What processes are involved? *Current Psychology of Cognition*, *16*, 325–351.

Di Vesta, F. J., & Gray, G. S. (1972). Listening and note taking. *Journal of Educational Psychology*, *63*, 8–14.

Doctorow, M., Wittrock, M. C., & Marks, C. (1978). Generative processes in reading comprehension. *Journal of Educational Psychology*, *70*, 109–118.

Donovan, J. J., & Radosevich, D. J. (1999). A meta-analytic review of the distribution of practice effect: Now you see it, now you don't. *Journal of Applied Psychology*, *84*, 795–805.

Dornisch, M. M., & Sperling, R. A. (2006). Facilitating learning from technology-enhanced text: Effects of prompted elaborative interrogation. *Journal of Educational Research*, *99*, 156–165.

Duchastel, P. C. (1981). Retention of prose following testing with different types of test. *Contemporary Educational Psychology*, *6*, 217–226.

Duchastel, P. C., & Nungester, R. J. (1982). Testing effects measured with alternate test forms. *Journal of Educational Research*, *75*, 309–313.

Dunlosky, J., & Rawson, K. A. (2005). Why does rereading improve metacomprehension accuracy? Evaluating the levels-of-disruption hypothesis for the rereading effect. *Discourse Processes*, *40*, 37–56.

Durgunoğlu, A. Y., Mir, M., & Ariño-Martí, S. (1993). Effects of repeated readings on bilingual and monolingual memory for text. *Contemporary Educational Psychology*, *18*, 294–317.

Dyer, J. W., Riley, J., & Yekovich, F. R. (1979). An analysis of three study skills: Notetaking, summarizing, and rereading. *Journal of Educational Research*, *73*, 3–7.

Einstein, G. O., Morris, J., & Smith, S. (1985). Note-taking, individual differences, and memory for lecture information. *Journal of Educational Psychology*, *77*, 522–532.

Fass, W., & Schumacher, G. M. (1978). Effects of motivation, subject activity, and readability on the retention of prose materials. *Journal of Educational Psychology*, *70*, 803–807.

Faw, H. W., & Waller, T. G. (1976). Mathemagenic behaviors and efficiency in learning from prose materials: Review, critique and recommendations. *Review of Educational Research*, *46*, 691–720.

Fazio, L. K., Agarwal, P. K., Marsh, E. J., & Roediger, H. L., III. (2010). Memorial consequences of multiple-choice testing on immediate and delayed tests. *Memory & Cognition*, *38*, 407–418.

Fishman, E. J., Keller, L., & Atkinson, R. C. (1968). Massed versus distributed practice in computerized spelling drills. *Journal of Educational Psychology*, *59*, 290–296.

Foos, P. W. (1995). The effect of variations in text summarization opportunities on test performance. *Journal of Experimental Education*, *63*, 89–95.

Foos, P. W., & Fisher, R. P. (1988). Using tests as learning opportunities. *Journal of Educational Psychology*, *80*, 179–183.

Fowler, R. L., & Barker, A. S. (1974). Effectiveness of highlighting for retention of text material. *Journal of Applied Psychology*, *59*, 358–364.

Friend, R. (2001). Effects of strategy instruction on summary writing of college students. *Contemporary Educational Psychology*, *26*, 3–24.

Fritz, C. O., Morris, P. E., Acton, M., Voelkel, A. R., & Etkind, R. (2007). Comparing and combining retrieval practice and the keyword mnemonic for foreign vocabulary learning. *Applied Cognitive Psychology*, *21*, 499–526.

Fritz, C. O., Morris, P. E., Nolan, D., & Singleton, J. (2007). Expanding retrieval practice: An effective aid to preschool children's learning. *Quarterly Journal of Experimental Psychology*, *60*, 991–1004.

Gagne, E. D., & Memory, D. (1978). Instructional events and comprehension: Generalization across passages. *Journal of Reading Behavior*, *10*(4), 321–335.

Gajria, M., & Salvia, J. (1992). The effects of summarization instruction on text comprehension of students with learning disabilities. *Exceptional Children*, *58*, 508–516.

Gambrell, L. B., & Jawitz, P. B. (1993). Mental imagery, text illustrations, and children's story comprehension and recall. *Reading Research Quarterly*, *28*(3), 265–276.

Garner, R. (1982). Efficient text summarization: Costs and benefits. *Journal of Educational Research*, *75*, 275–279.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology*, *6*, 1–104.

Gernsbacher, M. A., Varner, K. R., & Faust, M. E. (1990). Investigating differences in general comprehension skill. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 430–445.

Giesen, C., & Peeck, J. (1984). Effects of imagery instruction on reading and retaining a literary text. *Journal of Mental Imagery*, *8*, 79–90.

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology*, *81*, 392–399.

Glover, J. A., & Corkill, A. J. (1987). Influence of paraphrased repetitions on the spacing effect. *Journal of Educational Psychology*, *79*, 198–199.

Glover, J. A., Zimmer, J. W., Filbeck, R. W., & Plake, B. S. (1980). Effects of training students to identify the semantic base of prose materials. *Journal of Applied Behavior Analysis*, *13*, 655–667.

Goldenberg, G. (1998). Is there a common substrate for visual recognition and visual imagery? *Neurocase*, 141–147.

Goldstone, R. L. (1996). Isolated and interrelated concepts. *Memory & Cognition*, *24*, 608–628.

Goverover, Y., Arango-Lasprilla, J. C., Hillary, F. G., Chiaravalloti, N., & DeLuca, J. (2009). Application of the spacing effect to improve learning and memory for functional tasks in traumatic brain injury: A pilot study. *The American Journal of Occupational Therapy*, *63*, 543–548.

Goverover, Y., Hillary, F. G., Chiaravalloti, N., Arango-Lasprilla, J. C., & DeLuca, J. (2009). A functional application of the spacing effect to improve learning and memory in persons with multiple sclerosis. *Journal of Clinical and Experimental Neuropsychology*, *31*, 513–522.

Greene, C., Symons, S., & Richards, C. (1996). Elaborative interrogation effects for children with learning disabilities: Isolated facts versus connected prose. *Contemporary Educational Psychology*, *21*, 19–42.

Greene, R. L. (1989). Spacing effects in memory: Evidence for a two-process account. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 371–377.

Greene, R. L. (1990). Spacing effects on implicit memory tests. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *16*, 1004–1011.

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

Gurung, R. A. R. (2005). How do students really study (and does it matter)? *Teaching of Psychology*, *32*, 239–241.

Gurung, R. A. R., Weidert, J., & Jeske, A. (2010). Focusing on how students study. *Journal of the Scholarship of Teaching and Learning*, *10*, 28–35.

Guttman, J., Levin, J. R., & Pressley, M. (1977). Pictures, partial pictures, and young children's oral prose learning. *Journal of Educational Psychology*, *69*(5), 473–480.

Gyeselinck, V., Meneghetti, C., De Beni, R., & Pazzaglia, F. (2009). The role of working memory in spatial text processing: What benefit of imagery strategy and visuospatial abilities? *Learning and Individual Differences*, *19*, 12–20.

Hall, J. W. (1988). On the utility of the keyword mnemonic for vocabulary learning. *Journal of Educational Psychology*, *80*, 554–562.

Hamilton, R. J. (1997). Effects of three types of elaboration on learning concepts from text. *Contemporary Educational Psychology*, *22*, 299–318.

Hare, V. C., & Borchardt, K. M. (1984). Direct instruction of summarization skills. *Reading Research Quarterly*, *20*, 62–78.

Hartley, J., Bartlett, S., & Branthwaite, A. (1980). Underlining can make a difference: Sometimes. *Journal of Educational Research*, *73*, 218–224.

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.

Hatala, R. M., Brooks, L. R., & Norman, G. R. (2003). Practice makes perfect: The critical role of mixed practice in the acquisition of ECG interpretation skills. *Advanced in Health Sciences Education*, *8*, 17–26.

Hattie, J. A. C. (2009). *Visible learning: A synthesis of over 800 meta-analyses relating to achievement*. London, England: Routledge.

Hayati, A. M., & Shariatifar, S. (2009). Mapping Strategies. *Journal of College Reading and Learning*, *39*, 53–67.

Head, M. H., Readence, J. E., & Buss, R. R. (1989). An examination of summary writing as a measure of reading comprehension. *Reading Research and Instruction*, *28*, 1–11.

Helder, E., & Shaughnessy, J. J. (2008). Retrieval opportunities while multitasking improve name recall. *Memory*, *16*, 896–909.

Helsdingen, A., van Gog, T., & van Merriënboer, J. J. G. (2011a). The effects of practice schedule and critical thinking prompts on learning and transfer of a complex judgment task. *Journal of Educational Psychology*, *103*, 383–398.

Helsdingen, A., van Gog, T., & van Merriënboer, J. J. G. (2011b). The effects of practice schedule on learning a complex judgment task. *Learning and Instruction*, *21*, 126–136.

Hidi, S., & Anderson, V. (1986). Producing written summaries: Task demands, cognitive operations, and implications for instruction. *Review of Educational Research*, *56*, 473–493.

Hintzman, D. L., & Rogers, M. K. (1973). Spacing effects in picture memory. *Memory & Cognition*, *1*, 430–434.

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

Hodes, C. L. (1992). The effectiveness of mental imagery and visual illustrations: A comparison of two instructional variables. *Journal of Research and Development in Education*, *26*, 46–58.

Hoon, P. W. (1974). Efficacy of three common study methods. *Psychological Reports*, *35*, 1057–1058.

Hunt, R. R. (1995). The subtlety of distinctiveness: What von Restorff really did. *Psychonomic Bulletin & Review*, *2*, 105–112.

Hunt, R. R. (2006). The concept of distinctiveness in memory research. In R. R. Hunt & J. B. Worthen (Eds.), *Distinctiveness and memory* (pp. 3–25). New York, NY: Oxford University Press.

Hunt, R. R., & Smith, R. E. (1996). Accessing the particular from the general: The power of distinctiveness in the context of organization. *Memory & Cognition*, *24*, 217–225.

Hunt, R. R., & Worthen, J. B. (Eds.). (2006). *Distinctiveness and memory*. New York, NY: Oxford University Press.

Idstein, P., & Jenkins, J. R. (1972). Underlining versus repetitive reading. *The Journal of Educational Research*, *65*, 321–323.

Jacoby, L. L., & Dallas, M. (1981). On the relationship between autobiographical memory and perceptual learning. *Journal of Experimental Psychology: General*, *110*, 306–340.

Jacoby, L. L., Wahlheim, C. N., & Coane, J. H. (2010). Test-enhanced learning of natural concepts: Effects on recognition memory, classification, and metacognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 1441–1451.

Janiszewski, C., Noel, H., & Sawyer, A. G. (2003). A meta-analysis of the spacing effect in verbal learning: Implications for research on advertising repetition and consumer memory. *Journal of Consumer Research*, *30*, 138–149.

Jenkins, J. J. (1979). Four points to remember: A tetrahedral model and memory experiments. In L. S. Cermak & I. M. Craik (Eds.), *Levels of processing in human memory* (pp. 429–446). Hillsdale, NJ: Lawrence Erlbaum.

Jitendra, A. K., Edwards, L. L., Sacks, G., & Jacobson, L. A. (2004). What research says about vocabulary instruction for students with learning disabilities. *Exceptional Children*, 70, 299–322.

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*, 621–629.

Johnson, L. L. (1988). Effects of underlining textbook sentences on passage and sentence retention. *Reading Research and Instruction*, *28*, 18–32.

Kahl, B., & Woloshyn, V. E. (1994). Using elaborative interrogation to facilitate acquisition of factual information in cooperative learning settings: One good strategy deserves another. *Applied Cognitive Psychology*, *8*, 465–478.

Kang, S. H. K., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review*, *18*, 998–1005.

Kang, S. H. K., & Pashler, H. (2012). Learning painting styles: Spacing is advantageous when it promotes discriminative contrast. *Applied Cognitive Psychology*, *26*, 97–103.

Kang, S. H. K., Pashler, H., Cepeda, N. J., Rohrer, D., Carpenter, S. K., & Mozer, M. C. (2011). Does incorrect guessing impair fact learning? *Journal of Educational Psychology*, *103*, 48–59.

Kardash, C. M., & Scholes, R. J. (1995). Effects of preexisting beliefs and repeated readings on belief change, comprehension, and recall of persuasive text. *Contemporary Educational Psychology*, *20*, 201–221.

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: Absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *37*, 1250–1257.

Karpicke, J. D., & Blunt, J. R. (2011). Retrieval practice produces more learning than elaborative studying with concept mapping. *Science*, *331*, 772–775.

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practice retrieval when they study on their own? *Memory*, *17*, 471–479.

Karpicke, J. D., & Roediger, H. L., III. (2007a). Expanding retrieval practice promotes short-term retention, but equally spaced retrieval enhances long-term retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *33*, 704–719.

Karpicke, J. D., & Roediger, H. L., III. (2007b). Repeated retrieval during learning is the key to long-term retention. *Journal of Memory and Language*, *57*, 151–162.

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

Karpicke, J. D., & Roediger, H. L., III. (2010). Is expanding retrieval a superior method for learning text materials? *Memory & Cognition*, *38*, 116–124.

Keys, N. (1934). The influence on learning and retention of weekly as opposed to monthly tests. *Journal of Educational Psychology*, 427–436.

Kiewra, K. A., Mayer, R. E., Christensen, M., Kim, S.-I., & Risch, N. (1991). Effects of repetition on recall and note-taking: Strategies for learning from lectures. *Journal of Educational Psychology*, *83*, 120–123.

Kika, F. M., McLaughlin, T. F., & Dixon, J. (1992). Effects of frequent testing of secondary algebra students. *Journal of Educational Research*, *85*, 159–162.

King, A. (1992). Comparison of self-questioning, summarizing, and notetaking-review as strategies for learning from lectures. *American Education Research Journal*, *29*, 303–323.

King, J. R., Biggs, S., & Lipsky, S. (1984). Students' self-questioning and summarizing as reading study strategies. *Journal of Reading Behavior*, *16*, 205–218.

Klare, G. R., Mabry, J. E., & Gustafson, L. M. (1955). The relationship of patterning (underlining) to immediate retention and to acceptability of technical material. *Journal of Applied Psychology*, *39*, 40–42.

Kornell, N. (2009). Optimising learning using flashcards: Spacing is more effective than cramming. *Applied Cognitive Psychology*, *23*, 1297–1317.

Kornell, N., & Bjork, R. A. (2007). The promise and perils of self-regulated study. *Psychonomic Bulletin & Review*, *14*, 219–224.

Kornell, N., & Bjork, R. A. (2008). Learning, concepts, and categories: Is spacing the "enemy of induction?" *Psychological Science*, *19*, 585–592.

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

Kornell, N., Castel, A. D., Eich, T. S., & Bjork, R. A. (2010). Spacing as the friend of both memory and induction in young and older adults. *Psychology and Aging*, *25*, 498–503.

Kosslyn, S. M. (1981). The medium and the message in mental imagery: A theory. *Psychological Review*, *88*, 46–66.

Kratochwill, T. R., Demuth, D. M., & Conzemius, W. C. (1977). The effects of overlearning on preschool children's retention of sight vocabulary words. *Reading Improvement*, *14*, 223–228.

Kromann, C. B., Jensen, M. L., & Ringsted, C. (2009). The effects of testing on skills learning. *Medical Education*, *43*, 21–27.

Kulhavy, R. W., Dyer, J. W., & Silver, L. (1975). The effects of notetaking and test expectancy on the learning of text material. *Journal of Educational Research*, *68*, 363–365.

Kulhavy, R. W., & Swenson, I. (1975). Imagery instructions and the comprehension of text. *British Journal of Educational Psychology*, *45*, 47–51.

Lawson, M. J., & Hogben, D. (1998). Learning and recall of foreign-language vocabulary: Effects of a keyword strategy for immediate and delayed recall. *Learning and Instruction*, *8*(2), 179–194.

Lee, H. W., Lim, K. Y., & Grabowski, B. L. (2010). Improving self-regulation, learning strategy use, and achievement with metacognitive feedback. *Educational Technology Research Development*, *58*, 629–648.

Leeming, F. C. (2002). The exam-a-day procedure improves performance in psychology classes. *Teaching of Psychology*, *29*, 210–212.

Leicht, K. L., & Cashen, V. M. (1972). Type of highlighted material and examination performance. *Journal of Educational Research*, *65*, 315–316.

Lesgold, A. M., McCormick, C., & Golinkoff, R. M. (1975). Imagery Training and children's prose learning. *Journal of Educational Psychology*, *67*(5), 663–667.

Leutner, D., Leopold, C., & Sumfleth, E. (2009). Cognitive load and science text comprehension: Effects of drawing and mentally imagining text content. *Computers in Human Behavior*, *25*, 284–289.

Levin, J. R., & Divine-Hawkins, P. (1974). Visual imagery as a prose-learning process. *Journal of Reading Behavior*, *6*, 23–30.

Levin, J. R., Divine-Hawkins, P., Kerst, S. M., & Guttman, J. (1974). Individual differences in learning from pictures and words: The development and application of an instrument. *Journal of Educational Psychology*, *66*(3), 296–303.

Levin, J. R., Pressley, M., McCormick, C. B., Miller, G. E., & Shriberg, L. K. (1979). Assessing the classroom potential of the keyword method. *Journal of Educational Psychology*, *71*(5), 583–594.

Logan, J. M., & Balota, D. A. (2008). Expanded vs. equal interval spaced retrieval practice: Exploring different schedules of spacing and retention interval in younger and older adults. *Aging, Neuropsychology, and Cognition*, *15*, 257–280.

Lonka, K., Lindblom-Ylänne, S., & Maury, S. (1994). The effect of study strategies on learning from text. *Learning and Instruction*, *4*, 253–271.

Lorch, R. F. (1989). Text-signaling devices and their effects on reading and memory processes. *Educational Psychology Review*, *1*, 209–234.

Lorch, R. F., Jr., Lorch, E. P., & Klusewitz, M. A. (1995). Effects of typographical cues on reading and recall of text. *Contemporary Educational Psychology*, *20*, 51–64.

Lyle, K. B., & Crawford, N. A. (2011). Retrieving essential material at the end of lectures improves performance on statistics exams. *Teaching of Psychology*, *38*, 94–97.

Maddox, G. B., Balota, D. A., Coane, J. H., & Duchek, J. M. (2011). The role of forgetting rate in producing a benefit of expanded over equal spaced retrieval in young and older adults. *Psychology of Aging*, *26*, 661–670.

Magliano, J. P., Trabasso, T., & Graesser, A. C. (1999). Strategic processing during comprehension. *Journal of Educational Psychology*, *91*, 615–629.

Maher, J. H., & Sullivan, H. (1982). Effects of mental imagery and oral and print stimuli on prose learning of intermediate grade children. *Educational Technology Research & Development*, *30*, 175–183.

Malone, L. D., & Mastropieri, M. A. (1991). Reading comprehension instruction: Summarization and self-monitoring training for students with learning disabilities. *Exceptional Children*, *58*, 270–283.

Marschark, M., & Hunt, R. R. (1989). A reexamination of the role of imagery in learning and memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *15*, 710–720.

Marsh, E. J., Agarwal, P. K., & Roediger, H. L., III. (2009). Memorial consequences of answering SAT II questions. *Journal of Experimental Psychology: Applied*, *15*, 1–11.

Marsh, E. J., & Butler, A. C. (in press). Memory in educational settings. In D. Reisberg (Ed.), *Oxford handbook of cognitive psychology*.

Mastropieri, M. A., Scruggs, T. E., & Mushinski Fulk, B. J. (1990). Teaching abstract vocabulary with the keyword method: Effects on recall and comprehension. *Journal of Learning Disabilities*, *23*, 92–107.

Mathews, C. O. (1938). Comparison of methods of study for immediate and delayed recall. *Journal of Educational Psychology*, *29*, 101–106.

Matthews, P., & Rittle-Johnson, B. (2009). In pursuit of knowledge: Comparing self-explanations, concepts, and procedures as pedagogical tools. *Journal of Experimental Child Psychology*, *104*, 1–21.

Mawhinney, V. T., Bostow, D. E., Laws, D. R., Blumenfield, G. J., & Hopkins, B. L. (1971). A comparison of students studying behavior produced by daily, weekly, and three-week testing schedules. *Journal of Applied Behavior Analysis*, *4*, 257–264.

Mayer, R. E. (1983). Can you repeat that? Qualitative effects of repetition and advance organizers on learning from science prose. *Journal of Educational Psychology*, *75*, 40–49.

Mayfield, K. H., & Chase, P. N. (2002). The effects of cumulative practice on mathematics problem solving. *Journal of Applied Behavior Analysis*, *35*, 105–123.

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, H. L., III. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology*, *103*, 399–414.

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., & Donnelly, C. M. (1996). Learning with analogy and elaborative interrogation. *Journal of Educational Psychology*, *88*, 508–519.

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science*, *20*, 516–522.

McDaniel, M. A., & Pressley, M. (1984). Putting the keyword method in context. *Journal of Educational Psychology*, *76*, 598–609.

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 and Cognition*, *1*, 18–26.

McNamara, D. S. (2010). Strategies to read and learn: Overcoming learning by consumption. *Medical Education*, *44*, 240–346.

Metcalfe, J., & Kornell, N. (2007). Principles of cognitive science in education: The effects of generation, errors, and feedback. *Psychonomic Bulletin & Review*, *14*, 225–229.

Metcalfe, J., Kornell, N., & Finn, B. (2009). Delayed versus immediate feedback in children's and adults' vocabulary learning. *Memory & Cognition*, *37*, 1077–1087.

Metcalfe, J., Kornell, N., & Son, L. K. (2007). A cognitive-science based programme to enhance study efficacy in a high and low risk setting. *European Journal of Cognitive Psychology*, *19*, 743–768.

Miccinati, J. L. (1982). The influence of a six-week imagery training program on children's reading comprehension. *Journal of Reading Behavior*, *14*(2), 197–203.

Michael, J. (1991). A behavioral perspective on college teaching. *The Behavior Analyst*, *14*, 229–239.

Miller, G. E., & Pressley, M. (1989). Picture versus question elaboration on young children's learning of sentences containing high- and low-probability content. *Journal of Experimental Child Psychology*, *48*, 431–450.

Mitchell, C., Nash, S., & Hall, G. (2008). The intermixed-blocked effect in human perceptual learning is not the consequence of trial spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *34*, 237–242.

Morris, P. E., & Fritz, C. O. (2002). The improved name game: Better use of expanding retrieval practice. *Memory*, *10*, 259–266.

Moulton, C.-A., Dubrowski, A. E., MacRae, H., Graham, B., Grober, E., & Reznick, R. (2006). Teaching surgical skills: What kind of practice makes perfect? *Annals of Surgery*, *244*, 400–409.

Myers, G. C. (1914). Recall in relation to retention. *Journal of Educational Psychology*, *5*, 119–130.

Neuschatz, J. S., Preston, E. L., Toglia, M. P., & Neuschatz, J. S. (2005). Comparison of the efficacy of two name-learning techniques: Expanding rehearsal and name-face imagery. *American Journal of Psychology*, *118*, 79–102.

Nist, S. L., & Hogrebe, M. C. (1987). The role of underlining and annotating in remembering textual information. *Reading Research and Instruction*, *27*, 12–25.

Nist, S. L., & Kirby, K. (1989). The text marking patterns of college students. *Reading Psychology: An International Quarterly*, *10*, 321–338.

Nungester, R. J., & Duchastel, P. C. (1982). Testing versus review: Effects on retention. *Journal of Educational Psychology*, *74*, 18–22.

Oakhill, J., & Patel, S. (1991). Can imagery training help children who have comprehension problems? *Journal of Research in Reading*, *14*(2), 106–115.

Olina, Z., Resier, R., Huang, X., Lim, J., & Park, S. (2006). Problem format and presentation sequence: Effects on learning and mental effort among US high school students. *Applied Cognitive Psychology*, *20*, 299–309.

O'Reilly, T., Symons, S., & MacLatchy-Gaudet, H. (1998). A comparison of self-explanation and elaborative interrogation. *Contemporary Educational Psychology*, *23*, 434–445.

Ormrod, J. E. (2008). *Educational psychology: Developing learners* (6th ed.). Upper Saddle River, NJ: Pearson Education.

O'Shea, L. J., Sindelar, P. T., & O'Shea, D. J. (1985). The effects of repeated readings and attentional cues on reading fluency and comprehension. *Journal of Reading Behavior*, *17*, 129–142.

Ozgungor, S., & Guthrie, J. T. (2004). Interactions among elaborative interrogation, knowledge, and interest in the process of constructing knowledge from text. *Journal of Educational Psychology*, *96*, 437–443.

Palincsar, A. S., & Brown, A. L. (1984). Reciprocal teaching of comprehension-fostering and comprehension-monitoring activities. *Cognition and Instruction*, *1*, 117–175.

Pashler, H., Bain, P., Bottge, B., Graesser, A., Koedinger, K., McDaniel, M., & Metcalfe, J. (2007). *Organizing instruction and study to improve student learning* (NCER 2007–2004). Washington, DC: National Center for Education Research, Institute of Education Sciences, U.S. Department of Education.

Pashler, H., Zarow, G., & Triplett, B. (2003). Is temporal spacing of tests helpful even when it inflates error rates? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *29*, 1051–1057.

Pauk, W., & Ross, J. Q. (2010). *How to study in college* (10th ed.). Boston, MA: Wadsworth.

Pavlik, P. I., Jr., & Anderson, J. R. (2005). Practice and forgetting effects on vocabulary memory: An activation-based model of the spacing effect. *Cognitive Science*, *29*, 559–586.

Peterson, S. E. (1992). The cognitive functions of underlining as a study technique. *Reading Research and Instruction*, *31*, 49–56.

Pressley, M. (1976). Mental imagery helps eight-year-olds remember what they read. *Journal of Educational Psychology*, *68*, 355–359.

Pressley, M. (1986). The relevance of the good strategy user model to the teaching of mathematics. *Educational Psychologists*, *21*, 139–161.

Pressley, M., Goodchild, F., Fleet, F., Zajchowski, R., & Evans, E. D. (1989). The challenges of classroom strategy instruction. *The Elementary School Journal*, *89*, 301–342.

Pressley, M., Johnson, C. J., Symons, S., McGoldrick, J. A., & Kurita, J. A. (1989). Strategies that improve children's memory and comprehension of text. *The Elementary School Journal*, *90*, 3–32.

Pressley, M., & Levin, J. R. (1978). Developmental constraints associated with children's use of the keyword method of foreign language vocabulary learning. *Journal of Experimental Child Psychology*, *26*, 359–372.

Pressley, M., McDaniel, M. A., Turnure, J. E., Wood, E., & Ahmad, M. (1987). Generation and precision of elaboration: Effects on intentional and incidental learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *13*, 291–300.

Pyc, M. A., & Dunlosky, J. (2010). Toward an understanding of students' allocation of study time: Why do they decide to mass or space their practice? *Memory & Cognition*, *38*, 431–440.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language*, *60*, 437–447.

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

Pyc, M. A., & Rawson, K. A. (2012a). Are judgments of learning made after correct responses during retrieval practice sensitive to lag and criterion level effects? *Memory & Cognition*, *40*, 976–988.

Pyc, M. A., & Rawson, K. A. (2012b). Why is test-restudy practice beneficial for memory? An evaluation of the mediator shift hypothesis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *38*, 737–746.

Pylyshyn, Z. W. (1981). The imagery debate: Analogue media versus tacit knowledge. *Psychological Review*, *88*, 16–45.

Ramsay, C. M., Sperling, R. A., & Dornisch, M. M. (2010). A comparison of the effects of students' expository text comprehension strategies. *Instructional Science*, *38*, 551–570.

Raney, G. E. (1993). Monitoring changes in cognitive load during reading: An event-related brain potential and reaction time analysis. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *1*, 51–69.

Rasco, R. W., Tennyson, R. D., & Boutwell, R. C. (1975). Imagery instructions and drawings in learning prose. *Journal of Educational Psychology*, *67*, 188–192.

Rau, M. A., Aleven, V., & Rummel, N. (2010). Blocked versus interleaved practice with multiple representations in an intelligent tutoring system for fractions. In V. Aleven, J. Kay, & J. Mostow (Eds.), *Intelligent tutoring systems* (pp. 413–422). Berlin/Heidelberg, Germany: Springer-Verlag.

Raugh, M. R., & Atkinson, R. C. (1975). A mnemonic method for learning a second-language vocabulary. *Journal of Educational Psychology*, *67*, 1–16.

Rawson, K. A. (2012). Why do rereading lag effects depend on test delay? *Journal of Memory and Language*, *66*, 870–884.

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.

Rawson, K. A., & Kintsch, W. (2005). Rereading effects depend upon the time of test. *Journal of Educational Psychology*, *97*, 70–80.

Rawson, K. A., & Van Overschelde, J. P. (2008). How does knowledge promote memory? The distinctiveness theory of skilled memory. *Journal of Memory and Language*, *58*, 646–668.

Rea, C. P., & Modigliani, V. (1985). The effect of expanded versus massed practice on the retention of multiplication facts and spelling lists. *Human Learning*, *4*, 11–18.

Recht, D. R., & Leslie, L. (1988). Effect of prior knowledge on good and poor readers' memory of text. *Journal of Educational Psychology*, *80*, 16–20.

Rees, P. J. (1986). Do medical students learn from multiple choice examinations? *Medical Education*, *20*, 123–125.

Rewey, K. L., Dansereau, D. F., & Peel, J. L. (1991). Knowledge maps and information processing strategies. *Contemporary Educational Psychology*, *16*, 203–214.

Reynolds, J. H., & Glaser, R. (1964). Effects of repetition and spaced review upon retention of a complex learning task. *Journal of Educational Psychology*, *55*, 297–308.

Riches, N. G., Tomasello, M., & Conti-Ramsden, G. (2005). Verb learning in children with SLI: Frequency and spacing effects. *Journal of Speech, Language, and Hearing Research*, *48*, 1397–1411.

Rickard, T. C., Lau, J. S.-H., & Pashler, H. (2008). Spacing and the transition from calculation to retrieval. *Psychonomic Bulletin & Review*, *15*, 656–661.

Rickards, J. P., & August G. J. (1975). Generative underlining strategies in prose recall. *Journal of Educational Psychology*, *67*, 860–865.

Rickards, J. P., & Denner, P. R. (1979). Depressive effects of underlining and adjunct questions on children's recall of text. *Instructional Science*, *8*, 81–90.

Rinehart, S. D., Stahl, S. A., & Erickson, L. G. (1986). Some effects of summarization training on reading and studying. *Reading Research Quarterly*, *21*, 422–438.

Rittle-Johnson, B. (2006). Promoting transfer: Effects of self-explanation and direct instruction. *Child Development*, *77*, 1–15.

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. (2006a). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, *1*, 181–210.

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

Roediger, H. L., III, & Marsh, E. J. (2005). The positive and negative consequences of multiple-choice testing. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *31*, 1155–1159.

Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their applications to educational practice. *Psychology of Learning and Motivation*, *44*, 1–36.

Rohrer, D. (2009). The effects of spacing and mixing practice problems. *Journal for Research in Mathematics Education*, *40*, 4–17.

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practice on the retention of mathematics knowledge. *Applied Cognitive Psychology*, *20*, 1209–1224.

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*, 481–498.

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 233–239.

Rose, D. S., Parks, M., Androes, K., & McMahon, S. D. (2000). Imagery-based learning: Improving elementary students' reading comprehension with drama techniques. *The Journal of Educational Research*, *94*(1), 55–63.

Ross, S. M., & Di Vesta, F. J. (1976). Oral summary as a review strategy for enhancing recall of textual material. *Journal of Educational Psychology*, *68*, 689–695.

Rothkopf, E. Z. (1968). Textual constraint as a function of repeated inspection. *Journal of Educational Psychology*, *59*, 20–25.

Runquist, W. N. (1983). Some effects of remembering on forgetting. *Memory & Cognition*, *11*, 641–650.

Santrock, J. (2008). *Educational psychology*. New York, NY: McGraw-Hill.

Schmidmaier, R., Ebersbach, R., Schiller, M., Hege, I., Hlozer, M., & Fischer, M. R. (2011). Using electronic flashcards to promote learning in medical students: Retesting versus restudying. *Medical Education*, *45*, 1101–1110.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science*, *3*, 207–217.

Schmidt, S. R. (1988). Test expectancy and individual-item versus relational processing. *The American Journal of Psychology*, *101*, 59–71.

Schneider, V. I., Healy, A. F., & Bourne, L. E. (1998). Contextual interference effects in foreign language vocabulary acquisition and retention. In A. F. Healy & L. E Bourne (Eds.), *Foreign language learning* (pp. 77–90). Mahwah, NJ: Lawrence Erlbaum.

Schneider, V. I., Healy, A. F., & Bourne, L. E. (2002). What is learned under difficult conditions is hard to forget: Contextual interference effects in foreign vocabulary acquisition, retention, and transfer. *Journal of Memory and Language*, *46*, 419–440.

Schworm, S., & Renkl, A. (2006). Computer-supported example-based learning: When instructional explanations reduce self-explanations. *Computers & Education*, *46*, 426–445.

Scruggs, T. E., Mastropieri, M. A., & Sullivan, G. S. (1994). Promoting relational thinking: Elaborative interrogation for students with mild disabilities. *Exceptional Children*, *60*, 450–457.

Scruggs, T. E., Mastropieri, M. A., Sullivan, G. S., & Hesser, L. S. (1993). Improving reasoning and recall: The differential effects of elaborative interrogation and mnemonic elaboration. *Learning Disability Quarterly*, *16*, 233–240.

Seabrook, R., Brown, G. D. A., & Solity, J. E. (2005). Distributed and massed practice: From laboratory to classroom. *Applied Cognitive Psychology*, *19*, 107–122.

Seifert, T. L. (1993). Effects of elaborative interrogation with prose passages. *Journal of Educational Psychology*, *85*, 642–651.

Seifert, T. L. (1994). Enhancing memory for main ideas using elaborative interrogation. *Contemporary Educational Psychology*, *19*, 360–366.

Shapiro, A. M., & Waters, D. L. (2005). An investigation of the cognitive processes underlying the keyword method of foreign vocabulary learning. *Language Teaching Research*, *9*(2), 129–146.

Shriberg, L. K., Levin, J. R., McCormick, C. B., & Pressley, M. (1982). Learning about "famous" people via the keyword method. *Journal of Educational Psychology*, *74*(2), 238–247.

Simmons, A. L. (2011). Distributed practice and procedural memory consolidation in musicians' skill learning. *Journal of Research in Music Education*, *59*, 357–368.

Slamecka, N. J., & Graf, P. (1978). The generation effect: Delineation of a phenomenon. *Journal of Experimental Psychology: Human Learning and Memory*, *4*, 592–604.

Slavin, R. E. (2009). *Educational psychology: Theory and practice* (9th ed.). Upper Saddle River, NJ: Pearson Education.

Smith, B. L., Holliday, W. G., & Austin, H. W. (2010). Students' comprehension of science textbooks using a question-based reading strategy. *Journal of Research in Science Teaching*, *47*, 363–379.

Smith, T. A., & Kimball, D. R. (2010). Learning from feedback: Spacing and the delay-retention effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *36*, 80–95.

Snowman, J., McCown, R., & Biehler, R. (2009). *Psychology applied to teaching* (12th ed.). Boston, MA: Houghton Mifflin.

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*, 763–767.

Sommer, T., Schoell, E., & Büchel, C. (2008). Associative symmetry of the memory for object-location associations as revealed by the testing effect. *Acta Psychologica*, *128*, 238–248.

Son, L. K. (2004). Spacing one's study: Evidence for a metacognitive control strategy. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *30*, 601–604.

Sones, A. M., & Stroud, J. B. (1940). Review, with special reference to temporal position. *Journal of Educational Psychology*, 665–676.

Spitzer, H. F. (1939). Studies in retention. *Journal of Educational Psychology*, *30*, 641–656.

Spörer, N., Brunstein, J. C., & Kieschke, U. (2009). Improving students' reading and comprehension skills: Effects of strategy instruction and reciprocal teaching. *Learning and Instruction*, *19*, 272–286.

Spurlin, J. E., Dansereau, D. F., O'Donnell, A., & Brooks, L. W. (1988). Text processing: Effects of summarization frequency on text recall. *Journal of Experimental Education*, *56*, 199–202.

Stein, B. L., & Kirby, J. R. (1992). The effects of text absent and text present conditions on summarization and recall of text. *Journal of Reading Behavior*, *24*, 217–231.

Stein, B. S., & Bransford, J. D. (1979). Constraints on effective elaboration: Effects of precision and subject generation. *Journal of Verbal Learning and Verbal Behavior*, *18*, 769–777.

Sternberg, R. J., & Williams, W. M. (2010). *Educational psychology* (2nd ed.). Upper Saddle River, NJ: Pearson.

Stigler, J. W., Fuson, K. C., Ham, M., & Kim, M. S. (1986). An analysis of addition and subtraction word problems in American and Soviet elementary mathematics textbooks. *Cognition and Instruction*, *3*, 153–171.

Stordahl, K. E., & Christensen, C. M. (1956). The effect of study techniques on comprehension and retention. *Journal of Educational Research*, *49*, 561–570.

Sumowski, J. F., Chiaravalloti, N., & DeLuca, J. (2010). Retrieval practice improves memory in multiple sclerosis: Clinical application of the testing effect. *Neuropsychology*, *24*, 267–272.

Taylor, K., & Rohrer, D. (2010). The effects of interleaved practice. *Applied Cognitive Psychology*, *24*, 837–848.

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

Thomas, M. H., & Wang, A. Y. (1996). Learning by the keyword mnemonic: Looking for long-term benefits. *Journal of Experimental Psychology: Applied*, *2*, 330–342.

Thompson, S. V. (1990). Visual imagery: A discussion. *Educational Psychology*, *10*(2), 141–182.

Thorndike, E. L. (1906). *The principles of teaching based on psychology*. New York, NY: A.G. Seiler.

Todd, W. B., & Kessler, C. C., III. (1971). Influence of response mode, sex, reading ability, and level of difficulty on four measures of recall of meaningful written material. *Journal of Educational Psychology*, *62*, 229–234.

Toppino, T. C. (1991). The spacing effect in young children's free recall: Support for automatic-process explanations. *Memory & Cognition*, *19*, 159–167.

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

Toppino, T. C., Cohen, M. S., Davis, M. L., & Moors, A. C. (2009). Metacognitive control over the distribution of practice: When is spacing preferred? *Journal of Experimental Psychology*, *35*, 1352–1358.

Toppino, T. C., Fearnow-Kenney, M. D., Kiepert, M. H., & Teremula, A. C. (2009). The spacing effect in intentional and incidental free recall by children and adults: Limits on the automaticity hypothesis. *Memory & Cognition*, *37*, 316–325.

Toppino, T. C., Kasserman, J. E., & Mracek, W. A. (1991). The effect of spacing repetitions on the recognition memory of young children and adults. *Journal of Experimental Child Psychology*, *51*, 123–138.

Tse, C.-S., Balota, D. A., & Roediger, H. L., III. (2010). The benefits and costs of repeated testing on the learning of face-name pairs in healthy older adults. *Psychology and Aging*, *25*, 833–845.

van Hell, J. G., & Mahn, A. C. (1997). Keyword mnemonics versus rote rehearsal: Learning concrete and abstract foreign words by experienced and inexperienced learners. *Language Learning*, *47*(3), 507–546.

van Merriënboer, J. J. G., de Croock, M. B. M., & Jelsma, O. (1997). The transfer paradox: Effects of contextual interference on retention and transfer performance of a complex cognitive skill. *Perceptual & Motor Skills*, *84*, 784–786.

van Merriënboer, J. J. G., Schuurman, J. G., de Croock, M. B. M., & Paas, F. G. W. C. (2002). Redirecting learners' attention during training: Effects on cognitive load, transfer test performance and training efficiency. *Learning and Instruction*, *12*, 11–37.

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.

Verkoeijen, P. P. J. L., Rikers, R. M. J. P., & Özsoy, B. (2008). Distributed rereading can hurt the spacing effect in text memory. *Applied Cognitive Psychology*, *22*, 685–695.

Vlach, H. A., Sandhofer, C. M., & Kornell, N. (2008). The spacing effect in children's memory and category induction. *Cognition*, *109*, 163–167.

Vojdanoska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology*, *24*, 1183–1195.

Wade, S. E., Trathen, W., & Schraw, G. (1990). An analysis of spontaneous study strategies. *Reading Research Quarterly*, *25*, 147–166.

Wade-Stein, D., & Kintsch, W. (2004). Summary street: Interactive computer support for writing. *Cognition and Instruction*, *22*, 333–362.

Wahlheim, C. N., Dunlosky, J., & Jacoby, L. L. (2011). Spacing enhances the learning of natural concepts: An investigation of mechanisms, metacognition, and aging. *Memory & Cognition*, *39*, 750–763.

Wang, A. Y., & Thomas, M. H. (1995). Effects of keywords on long-term retention: Help or hindrance? *Journal of Educational Psychology*, *87*, 468–475.

Wang, A. Y., Thomas, M. H., & Ouellette, J. A. (1992). Keyword mnemonic and retention of second-language vocabulary words. *Journal of Educational Psychology*, *84*, 520–528.

Weinstein, Y., McDermott, K. B., & Roediger, H. L., III. (2010). A comparison of study strategies for passages: Rereading, answering questions, and generating questions. *Journal of Experimental Psychology: Applied*, *16*, 308–316.

Wenger, S. K., Thompson, C. P., & Bartling, C. A. (1980). Recall facilitates subsequent recognition. *Journal of Experimental Psychology: Human Learning and Memory*, *6*, 135–144.

Wheeler, M. A., Ewers, M., & Buonanno, J. F. (2003). Different rates of forgetting following study versus test trials. *Memory*, *11*, 571–580.

Willerman, B., & Melvin, B. (1979). Reservations about the keyword mnemonic. *The Canadian Modern Language Review*, *35*(3), 443–453.

Willoughby, T., Waller, T. G., Wood, E., & MacKinnon, G. E. (1993). The effect of prior knowledge on an immediate and delayed associative learning task following elaborative interrogation. *Contemporary Educational Psychology*, *18*, 36–46.

Willoughby, T., & Wood, E. (1994). Elaborative interrogation examined at encoding and retrieval. *Learning and Instruction*, *4*, 139–149.

Wissman, K. T., Rawson, K. A., & Pyc, M. A. (2012). How and when do students use flashcards? *Memory*, *20*, 568–579.

Wittrock, M. C. (1990). Generative processes of comprehension. *Educational Psychologist*, *24*, 345–376.

Wollen, K. A., Cone, R. S., Britcher, J. C., & Mindemann, K. M. (1985). The effect of instructional sets upon the apportionment of study time to individual lines of text. *Human Learning*, *4*, 89–103.

Woloshyn, V. E., Paivio, A., & Pressley, M. (1994). Use of elaborative interrogation to help students acquire information consistent with prior knowledge and information inconsistent with prior knowledge. *Journal of Educational Psychology*, *86*, 79–89.

Woloshyn, V. E., Pressley, M., & Schneider, W. (1992). Elaborative-interrogation and prior-knowledge effects on learning of facts. *Journal of Educational Psychology*, *84* , 115–124.

Woloshyn, V. E., & Stockley, D. B. (1995). Helping students acquire belief-inconsistent and belief-consistent science facts: Comparisons between individual and dyad study using elaborative interrogation, self-selected study and repetitious-reading. *Applied Cognitive Psychology*, *9*, 75–89.

Woloshyn, V. E., Willoughby, T., Wood, E., & Pressley, M. (1990). Elaborative interrogation facilitates adult learning of factual paragraphs. *Journal of Educational Psychology*, *82*, 513–524.

Wong, B. Y. L., Wong, R., Perry, N., & Sawatsky, D. (1986). The efficacy of a self-questioning summarization strategy for use by underachievers and learning disabled adolescents in social studies. *Learning Disabilities Focus*, *2*, 20–35.

Wong, R. M. F., Lawson, M. J., & Keeves, J. (2002). The effects of self-explanation training on students' problem solving in high-school mathematics. *Learning and Instruction*, *12*, 233–262.

Wood, E., & Hewitt, K. L. (1993). Assessing the impact of elaborative strategy instruction relative to spontaneous strategy use in high achievers. *Exceptionality*, *4*, 65–79.

Wood, E., Miller, G., Symons, S., Canough, T., & Yedlicka, J. (1993). Effects of elaborative interrogation on young learners' recall of facts. *Elementary School Journal*, *94*, 245–254.

Wood, E., Pressley, M., & Winne, P. H. (1990). Elaborative interrogation effects on children's learning of factual content. *Journal of Educational Psychology*, *82*(4), 741–748.

Wood, E., Willoughby, T., Bolger, A., Younger, J., & Kaspar, V. (1993). Effectiveness of elaboration strategies for grade school children as a function of academic achievement. *Journal of Experimental Child Psychology*, *56*, 240–253.

Woolfolk, A. (2007). *Educational psychology* (10th ed.). Boston, MA: Allyn & Bacon.

Wulf, G., & Shea, C. H. (2002). Principles derived from the study of simple skills do not generalize to complex skill learning. *Psychonomic Bulletin & Review*, *9*, 185–211.

Yates, F. A. (1966). *The art of memory*. London, England: Pimlico.

Yu, G. (2009). The shifting sands in the effects of source text summarizability on summary writing. *Assessing Writing*, *14*, 116–137.

Zaromb, F. M., & Roediger, H. L., III. (2010). The testing effect in free recall is associated with enhanced organizational processes. *Memory & Cognition*, *3*(8), 995–1008.