Journal of Experimental Psychology:
Learning, Memory, and Cognition
2010, Vol. 36, No. 5, 1118–1133

0278-7393/10/$12.00 DOI: 10.1037/a0019902

Repeated Testing Produces Superior Transfer of Learning Relative to Repeated Studying

Andrew C. Butler
Washington University in St. Louis

The present research investigated whether test-enhanced learning can be used to promote transfer. More specifically, 4 experiments examined how repeated testing and repeated studying affected retention and transfer of facts and concepts. Subjects studied prose passages and then either repeatedly restudied or took tests on the material. One week later, they took a final test that had either the same questions (Experiment 1a), new inferential questions within the same knowledge domain (Experiments 1b and 2), or new inferential questions from different knowledge domains (Experiment 3). Repeated testing produced superior retention and transfer on the final test relative to repeated studying. This finding indicates that the mnemonic benefits of test-enhanced learning are not limited to the retention of the specific response tested during initial learning but rather extend to the transfer of knowledge in a variety of contexts.

Keywords: testing, retrieval practice, transfer of learning, encoding variability, cued recall

Supplemental materials: http://dx.doi.org/10.1037/a0019902.supp

The literature on human learning and memory is rife with phenomena that have stubbornly refused to yield their secrets to psychological science. One of the oldest and greatest puzzles of all is the phenomenon of transfer of learning, or "the influence of prior learning (retained until the present) upon the learning of, or response to, new material" (McGeoch, 1942, p. 394). The theoretical and practical importance of understanding transfer of learning (also called transfer of training but hereafter referred to simply as transfer) cannot be overstated. For theories of learning and memory, explaining how and why transfer occurs represents a critical test. Transfer also has enormous practical implications for education in both schools and the workplace.

With such a clear impetus for the study of transfer, it is disappointing that the progress made toward its understanding is not commensurate with the amount of research that has been directed at the phenomenon (for an excellent recent review, see Barnett & Ceci, 2002). One factor that impedes progress is the traditional approach to studying transfer of learning. Most transfer studies focus purely on the similarities and differences between the contexts of initial learning and subsequent transfer. This approach, which has dominated the field since Thorndike and Woodworth's (1901a, 1901b, 1901c) pioneering experiments, places primary importance on the nature of the transfer context (and its similarity to the initial learning context) in determining whether or not transfer occurs. As a result of the heavy emphasis on the transfer context as a limiting factor, relatively few studies take the alternative approach of exploring how the conditions of initial learning can be arranged to better promote transfer to many different possible contexts. To be sure, the degree of similarity between learning and transfer contexts is critical. However, initial learning is equally important in that it determines the potential for transfer to occur, and this potential is then realized to varying degrees depending on the transfer context. If initial learning produces better retention of information and numerous retrieval routes to access that information, it should increase the probability of a match between the cues given in the transfer task and the stored memory trace, thereby increasing the potential for transfer to occur.

The present research investigated how the conditions of initial learning affect transfer of learning. More specifically, four experiments examined whether *test-enhanced learning,* a method that has been shown to increase long-term retention (see McDaniel, Roediger, & McDermott, 2007), can be used to promote transfer to new inferential questions about previously studied material. Test-enhanced learning is based on the finding that taking a test on previously studied material produces better retention over time relative to restudying that material for an equivalent amount of time, a result commonly called the *testing effect* (for a review, see Roediger & Karpicke, 2006a). The primary goal of the present research was to examine whether repeated testing promotes superior transfer relative to repeated studying. A secondary goal was to explore whether repeated testing using rephrased questions (i.e., a different question on each test about the same piece of information) leads to better transfer than repeated testing using the same

Andrew C. Butler, Department of Psychology, Washington University in St. Louis.

The research reported here was supported by a Collaborative Activity Award from the James S. McDonnell Foundation's 21st Century Science Initiative in Bridging Brain, Mind and Behavior (principal investigator: Henry L. Roediger III). I would like to thank the members of my dissertation committee: Roddy Roediger (chair), Dave Balota, Mark McDaniel, Jeff Zacks, Keith Sawyer, and Jim Wertsch. I would also like to thank Ileana Culcea for helping to collect and score the data and Beth Marsh for providing helpful comments on a draft of the manuscript.

Correspondence concerning this article should be addressed to Andrew C. Butler, who is now at the Department of Psychology and Neuroscience, Duke University, Box 90086, Durham, NC 27708. E-mail: andrew.butler@duke.edu

question. Repeated testing with different questions should promote encoding variability, which increases the probability of future retrieval by creating multiple retrieval routes in memory (Bower, 1972; Estes, 1955; Martin, 1968). As a result, encoding variability may also increase the probability of successful transfer. Before describing the present research, I provide the rationale for the project and then review some of the evidence that supports the efficacy of test-enhanced learning.

Rationale for the Present Research

In the literature on transfer of learning, the degree of similarity between the contexts of initial learning and transfer is an important factor in determining whether or not transfer occurs (e.g., Holyoak & Koh, 1987; see Barnett & Ceci, 2002). Research on human memory and learning provides the same conclusion: The degree of overlap between encoding and retrieval is critical to determining successful memory performance. Two different but related theories of human memory articulate this idea: the encoding specificity principle and transfer-appropriate processing. The encoding specificity principle states that a retrieval cue will be effective to the extent that it overlaps with features (or elements) in the memory trace (Tulving, 1983). Similarly, the concept of transfer-appropriate processing states that memory performance is determined by the degree of overlap between the processes engaged during encoding and those required at retrieval (Morris, Bransford, & Franks, 1977). Although the contextual nature of human memory likely precludes the formation of any general laws (Roediger, 2008), these theories are arguably the most effective at providing an explanation for the complex findings in memory research.

As these two theories state and the results of many experiments clearly show, a match between encoding and retrieval is critical to successful memory performance; however, "goodness" of encoding also matters. Some encoding tasks produce better retention of declarative knowledge than others, and the best memory performance is generally found when the processes engaged and cues given at retrieval match these encoding tasks (e.g., Fisher & Craik, 1977; Moscovitch & Craik, 1976). In the words of Moscovitch and Craik (1976), "encoding operations establish a ceiling on potential memory performance, and retrieval cues determine the extent to which that potential is utilized" (p. 455).

In the transfer literature, there is some evidence that the conditions of initial learning can influence the direction and magnitude of transfer. Numerous studies have shown that a greater degree of initial learning generally increases positive transfer (e.g., Bruce, 1933; see Ellis, 1965), as does increasing the number and variability of training problems (e.g., Bassok & Holyoak, 1989; Gick & Holyoak, 1983; see Kimball & Holyoak, 2000). These findings suggest that learning tasks that increase the retention of information and create multiple retrieval routes in memory produce better transfer. That is, when there are multiple ways to access information in memory, it increases the likelihood of a match between the memory trace and the cues presented in the transfer task.

Test-Enhanced Learning: A Potential Mechanism for Promoting Transfer

Initial learning conditions that produce long-term retention of knowledge should increase the potential for successful transfer, especially when that knowledge can be flexibly retrieved using a variety of cues. Thus, test-enhanced learning may be a highly effective method for promoting transfer. As described briefly above, test-enhanced learning is predicated on the finding that retrieving information from memory produces superior long-term retention, a robust phenomenon that has been replicated many times (for a review, see Roediger & Karpicke, 2006a). Although testing is often conceptualized as a neutral event, the act of retrieving information from memory actually changes memory (e.g., Bjork, 1975), increasing the probability of successful retrieval in the future (e.g., Karpicke & Roediger, 2008).

One idea for enhancing the mnemonic benefits of testing is introducing encoding variability during repeated testing. Encoding variability is thought to produce better retention because it increases the number of potential retrieval routes, thereby increasing the probability of a match with whatever cue is presented at retrieval (Bower, 1972; Estes, 1955; Martin, 1968). Many factors can contribute to variability in encoding of to-be-remembered material, from changes in the way in which the material is perceived (e.g., modality of presentation) or processed (e.g., experimental task) to differences in internal (e.g., neuronal activity) or external environment (e.g., location). If testing can be used to promote encoding variability, the result should be knowledge that can be accessed with a variety of retrieval cues.

Previous Research on Testing and Transfer

Within the testing-effect literature, the vast majority of studies have assessed the benefits of retrieval practice with a final test containing a verbatim re-presentation of the same questions used on the initial test. However, there are several studies that have attempted to assess whether the benefits of testing transfer to other types of questions. Some studies have found a benefit of initial testing relative to studying on a final test that consisted of rephrased versions of the questions from the initial tests (e.g., McDaniel, Anderson, Derbish, & Morrisette, 2007). Similarly, a handful of studies have shown that initial testing of paired associates in one direction (A $\rightarrow$?) leads to better performance on a final test in which the pair is tested in the opposite direction (? $\leftarrow$ B) relative to studying both members of the pair (A–B) during the initial learning phase (e.g., Carpenter, Pashler, & Vul, 2006; Kanak & Neuner, 1970). Technically speaking, the results of these studies demonstrate that testing promotes transfer; however, the new context to which knowledge is transferred is almost identical to the original context.

In a recent study, Rohrer, Taylor, and Sholar (2010) explored whether testing promoted transfer to a final test that consisted of new questions when the correct responses remained the same as those on the initial test. Rohrer et al. had elementary school students study two fictional maps that included various regions (Experiment 1) or cities (Experiment 2). After studying the maps, the students took a test on one map (matching the name of the regions or cities to the location) and restudied the other map. Later, they received both a final retention test and a final transfer test for each map. In Experiment 1, both final tests involved labeling a blank version of the map—the critical difference was that a list of the region names was presented on the retention test but not on the transfer test. In Experiment 2, the retention test involved labeling a blank version of the map from a list of the city names. In

contrast, the transfer test consisted of questions in which students were given the names of two cities and had to recall the name of the city lying along the shortest route between those two cities. Both experiments showed that taking an initial test produced better transfer than restudying the map.

In another recent study, Johnson and Mayer (2009) investigated whether initial testing would promote superior transfer to a final test that consisted of new inferential questions that required new responses. Subjects watched a multimedia slide show about lightning formation, and then they either received a retention test, received a transfer test, or restudied the slide show. One week later, all subjects took a final test that consisted of both retention and transfer questions. All of the final test questions were the same as those that had appeared on the initial retention and transfer tests, except for two new transfer questions. The results showed that taking an initial transfer test led to superior performance on the two new transfer questions relative to restudying the slide show or taking an initial retention test (the latter two conditions produced roughly equivalent performance). However, there are two methodological issues that complicate the interpretation of these findings. First, three of the four transfer questions on the final test required slightly different applications of the same information and thus were not independent measures of transfer. Second, the retention question was always given before the transfer questions, which means that performance on the transfer questions was confounded because both types of questions tested the same material.

A few other studies have investigated whether initial testing leads to better performance on new inferential questions about previously untested material. For example, Chan, McDermott, and Roediger (2006; see too Chan, 2009, 2010) found that testing can benefit the retention of nontested but related material, a phenomenon that they termed retrieval-induced facilitation. One explanation for this finding is that people occasionally retrieve related information when answering questions on a test (but not when restudying the material) and that this covert retrieval practice enhances performance on a later test for this related information. Similarly, both Foos and Fisher (1988) and McKenzie (1972) used a final transfer test that consisted of new inferential questions about information that had not been previously tested. However, it was unclear in these two studies whether the information tested by the new questions was related to previously tested information and, if so, how the various pieces of information were related.

Only one study has examined whether encoding variability can be used to promote transfer with verbal materials. Goode, Geraci, and Roediger (2008) had subjects either repeatedly solve the same anagram (e.g., LDOOF, to which the answer is FLOOD) or repeatedly solve different variations of an anagram (e.g., DOLOF, FOLOD, and OOFLD) that was later tested. Goode et al. found that practice with different variations of an anagram led to a higher proportion of correct solutions on a final test relative to repeated practice with the same anagram, even when the anagram on the final test was one that had been repeatedly practiced. This finding suggests that encoding variability can be used to promote transfer of learning with verbal materials. Nevertheless, the evidence is still limited because relatively few studies have investigated how initial testing, either with or without variable encoding, influences performance on a subsequent transfer test.

Introduction to Experiments

The present research was designed to build upon these previous studies by investigating how the conditions of initial learning affect retention and transfer of knowledge. All four experiments used the same general procedure during the initial learning phase: Subjects studied passages about a variety of topics, and then they repeatedly restudied some passages and repeatedly took a test on other passages. The series of experiments was designed to explore progressively greater degrees of transfer. In Experiment 1a, the final test consisted of repeated questions (i.e., a verbatim representation of the questions that had been on the initial tests) to demonstrate that testing improves retention of information relative to restudying the passages. In Experiments 1b and 2, the final test consisted of new inferential questions from the same knowledge domain to assess whether testing would produce better transfer than restudying. The new inferential questions required subjects to apply the knowledge that they had learned during the initial session to answer a related question from the same domain. In Experiment 3, the final test consisted of new inferential questions from different knowledge domains to explore whether testing would promote transfer across domains.

Experiments 1a and 1b

Experiments 1a and 1b investigated whether repeated testing produces better retention and transfer, respectively, than repeated studying. The experiments also explored whether a testing procedure that promoted encoding variability by using rephrased versions of the questions would lead to better retention and transfer relative to the standard testing procedure. Both experiments consisted of two sessions, which were spaced 1 week apart. In an initial learning session, subjects studied a set of six passages about a variety of topics. Then, they repeatedly restudied two of the passages (restudy passages), repeatedly took the same test on another two passages (same test), and repeatedly took different tests on the other two passages (variable test). One week later, subjects returned to the lab for the final test that assessed retention (Experiment 1a) or transfer (Experiment 1b) of information from the passages. Table 1 contains a schematic representation of the design.

Method

Subjects and design. A total of 48 undergraduate psychology students at Washington University in St. Louis (St. Louis, MO) participated for course credit or pay (24 subjects in each experi-

Table 1
A Design Schematic of the General Procedure Used in Experiments 1a and 1b

Condition	Initial learning session				Final test
Same test	S	T_A	T_A	T_A	T
Variable test	S	T_A	T_B	T_C	T
Restudy passages	S	S	S	S	T

Note. S = study; T = test; subscript refers to the version of the test question: A, B, or C.

ment; paid subjects received $30 for participating). All subjects were treated in accordance with the "Ethical Principles of Psychologists and Code of Conduct" put forth by the American Psychological Association (2002).

Both experiments had a 3 (type of initial learning: restudy passages, same test, variable test) × 2 (type of initial test question: factual, conceptual) within-subjects design. Each variable was manipulated within subjects but between materials. The main dependent variable in Experiment 1a was performance on repeated questions (i.e., previously tested factual and conceptual questions) on the final test. The main dependent variable in Experiment 1b was performance on new inferential questions from the same knowledge domain on the final test. In addition, there was a set of control questions that was included on the final test in both experiments.

Materials and counterbalancing. The materials for both experiments consisted of six passages about a variety of topics (e.g., bats) and an associated set of questions. The passages were developed using information obtained from three online sources (www.en.wikipedia.org, www.encyclopedia.com, and www.howstuffworks.com). Each passage was approximately 1,000 words in length and arranged into eight paragraphs. Four facts and four concepts were identified in each passage. In every passage, each of the eight paragraphs contained either a single fact or a single concept. For the purposes of the present research, a fact was defined as a piece of information that was presented within a single sentence, while a concept was defined as a piece of information that was abstracted from multiple sentences. These definitions were developed in consultation with the taxonomy of educational objectives put forth by Bloom and colleagues (e.g., Bloom, 1956).

Next, a question was developed for each fact and concept. All questions were in cued-recall format, and the correct response to each question was generally between one and three sentences in length. An example of a factual question is the following: "Bats are one of the most prevalent orders of mammals. Approximately how many bat species are there in the world?" (Answer: "More than 1,000 bat species have been identified.") In contrast, an example of a conceptual question is the following: "Some bats use echolocation to navigate the environment and locate prey. How does echolocation help bats to determine the distance and size of objects?" (Answer: "Bats emit high-pitched sound waves and listen to the echoes. The distance of an object is determined by the time it takes for the echo to return. The size is calculated by the intensity of the echo: a smaller object will reflect less of the sound wave and thus produce a less intense echo.")

In addition, two rephrased versions of each question were created for use during initial testing in the variable-test condition. For each rephrased version of the question, the question stem was reworded, but the correct response remained the same. A rephrased version of the factual question given as an example above is the following: "Chiroptera is the name of the order that contains all bat species. What is the approximate number of bat species that exist?" (Answer: Same as above.) A rephrased version of the conceptual question given as an example above is the following: "Echolocation enables some bats to fly around and hunt their prey in the darkness with great precision. How can bats judge how far away an object is and how big it is through echolocation?" (Answer: Same as above.) Also, a control set of questions was created to test information contained in the passages but not tested in either the same-test or variable-test conditions. Two control questions were developed for each passage. For example, a control question was "Bats play an important role in many ecosystems by keeping insect populations in check. What other major role do they play in ecosystems?" (Answer: "Bats are also plant pollinators. Many species feed on plant nectar, gathering pollen on their bodies as they feed, which helps the plant to disperse its seed.") In terms of content, the information tested by these control questions was factual and did not overlap with the other items described above.

A set of inferential questions was developed to assess transfer on the final test in Experiment 1b. For each fact and concept, an inferential question was created that required the application of the fact or concept within the same knowledge domain (Bloom, 1956). For example, the inferential question related to the factual question given above is the following: "There are about 5,500 species of mammals in the world. Approximately what percent of all mammal species are species of bat?" (Answer: "If there are about 5,500 species of mammals and more than 1,000 species of bat, then bats account for approximately 20% of all mammal species.") The inferential question related to the conceptual question given above is the following: "An insect is moving towards a bat. Using the process of echolocation, how does the bat determine that the insect is moving towards it (i.e., rather than away from it)?" (Answer: "The bat can tell the direction that an object is moving by calculating whether the time it takes for an echo to return changes from echo to echo. If the insect is moving towards the bat, the time it takes the echo to return will get steadily shorter. Also, the intensity of the sound wave will increase because the insect will reflect more of the sound wave as it gets closer.")

The experiments were counterbalanced in two ways. First, two orders of the six passages were created to vary the position in which the passages were presented. Second, three orders of the initial learning conditions were created to ensure that each learning condition occurred equally often in each possible presentation position across subjects. These various orders were combined factorially to form six versions of each experiment. Overall, the counterbalancing ensured that, across subjects, each passage was used in each initial learning condition an equal number of times.

Procedure. Both experiments were conducted on a computer using E-Prime software (Schneider, Eschman, & Zuccolotto, 2002) and involved two sessions, spaced 1 week apart. In the first session, subjects began by studying all the passages. Each passage was presented two paragraphs at a time (approximately 250 words), with each pair of paragraphs appearing on the screen for 60 s. Thus, a total of 4 min was given to study each passage (pilot testing indicated that this amount of time was sufficient for subjects to read through the whole passage at a comfortable pace). Then, depending on the version of the experiment to which subjects were assigned, they repeatedly restudied some passages and took tests on the other passages in the same order as the passages were initially presented. Thus, each repeated study of a text and each repeated test on a text were spaced out in time. Passages that were restudied were presented in the same manner as before (i.e., 60 s per pair of paragraphs, etc.). On the tests, subjects were asked to produce a response to every question, even if they had to guess (i.e., forced report). Responses to the questions were entered into the computer using the keyboard. After each question, subjects received feedback that consisted of a re-presentation of the question and the correct response. No time limit was given to answer

each question and review the feedback, but subjects were encouraged to work quickly (and accurately).

One week after the first session, subjects returned to take a final test that was cued-recall format, self-paced, and forced report. In Experiment 1a, the final test consisted of repeated questions from the initial test conditions, questions about the passages in the restudy-passages condition, and control questions (see Materials). The version of the question that was tested on the final test was always the version that was given on the first of the three initial tests in the same-test and variable-test conditions (i.e., Version A or T_A on the schematic representation of the design; see Table 1). In Experiment 1b, the final test consisted of new inferential questions from the same domain (see Materials and Counterbalancing). After the final test, subjects were fully debriefed and dismissed.

Results

All results, unless otherwise stated, were significant at the .05 level. Pairwise comparisons were Bonferroni corrected to the .05 level. Eta squared (Pearson, 1911) and Cohen's *d* (Cohen, 1988) are the measures of effect size reported for all significant effects in the analyses of variance (ANOVAs) and *t*-test analyses, respectively. A Geisser–Greenhouse correction was used for violations of the sphericity assumption of ANOVA (Geisser & Greenhouse, 1958).

Scoring. A research assistant and I independently scored 20% of the cued-recall responses for each experiment. We used a coding scheme that identified the key pieces of information from the idealized correct answer that a given response must contain to be scored as correct. Each response was scored as either correct or incorrect (no partial credit was given). Both scorers were masked to condition and coded all the responses for a given question together to increase consistency in scoring. Cohen's kappa (Cohen, 1960) was calculated to assess interrater reliability. Reliability was high for both Experiment 1a ($\kappa = .88$) and Experiment 1b ($\kappa = .87$), so I resolved the few disagreements for each data set and then scored the remaining responses alone.

Initial tests. Table 2 shows the proportion of correct responses on the three initial cued-recall tests as a function of question type and initial learning condition for Experiments 1a and 1b. In both experiments, the proportion of correct responses produced by subjects increased on each successive test, presumably because they used the feedback to correct their errors. The gains in performance from Test 1 to Test 2 were larger than the gains from Test 2 to Test 3, indicating the negatively accelerated curvilinear relationship that is typically observed in multitrial learning experiments (e.g., Ebbinghaus, 1885/1964). This pattern of increasing performance held for both factual and conceptual questions in each test condition.

Performance on the factual and conceptual questions was analyzed separately via 3 (test: 1, 2, 3) $\times$ 2 (initial learning condition: same test, variable test) repeated measures ANOVAs. For the factual questions in Experiment 1a, there was a significant main effect of test, $F(2, 46) = 221.22$, $MSE = .01$, $\eta^2 = .67$, for which the quadratic trend was also significant, $F(1, 23) = 52.26$, $MSE = .01$, $\eta^2 = .24$, confirming the observation that learning increased more from Test 1 to Test 2 than from Test 2 to Test 3. Neither the main effect of initial learning condition, $F(1, 23) = 2.59$, $MSE = .06$, $p = .12$, nor the interaction was significant ($F < 1$). The same pattern of results emerged for factual questions in Experiment 1b. There was a significant main effect of test, $F(2, 46) = 110.50$, $MSE = .03$, $\eta^2 = .66$, and a significant quadratic trend confirmed the observation of a curvilinear increase in performance across the three tests, $F(1, 23) = 26.71$, $MSE = .02$, $\eta^2 = .26$. However, neither the main effect of initial learning condition, $F(1, 23) = 1.73$, $MSE = .04$, $p = .20$, nor the interaction was significant ($F < 1$).

For the conceptual questions in Experiment 1a, an ANOVA revealed a significant main effect of test, $F(2, 46) = 101.54$, $MSE = .02$, $\eta^2 = .55$, as well as a significant quadratic trend, $F(1, 23) = 12.80$, $MSE = .02$, $\eta^2 = .07$. Neither the main effect of initial learning condition nor the interaction was significant (Fs < 1). Experiment 1b yielded the same pattern of results for conceptual questions: a significant main effect of test, $F(2, 46) = 107.42$, $MSE = .02$, $\eta^2 = .48$, for which there was also a significant curvilinear trend, $F(1, 23) = 37.00$, $MSE = .01$, $\eta^2 = .09$. No other effects were significant (Fs < 1).

Final test. Figure 1 shows the proportion of correct responses on the final cued-recall test as a function of question type and initial learning condition for Experiment 1a (top panel) and Experiment 1b (bottom panel). In Experiment 1a, performance was

Table 2
Proportion of Correct Responses on the Three Initial Cued-Recall Tests as a Function of Question Type and Initial Learning Condition for Experiments 1a, 1b, 2, and 3

			Initial test		
Experiment	Question type	Learning condition	Test 1	Test 2	Test 3
1a	Factual	Same test	.40	.81	.90
		Variable test	.37	.71	.82
	Conceptual	Same test	.42	.70	.80
		Variable test	.41	.66	.79
1b	Factual	Same test	.34	.74	.88
		Variable test	.39	.78	.86
	Conceptual	Same test	.38	.66	.75
		Variable test	.39	.78	.92
2	Factual	Same test	.43	.73	.88
	Conceptual	Same test	.39	.67	.77
3	Conceptual	Same test	.38	.71	.78

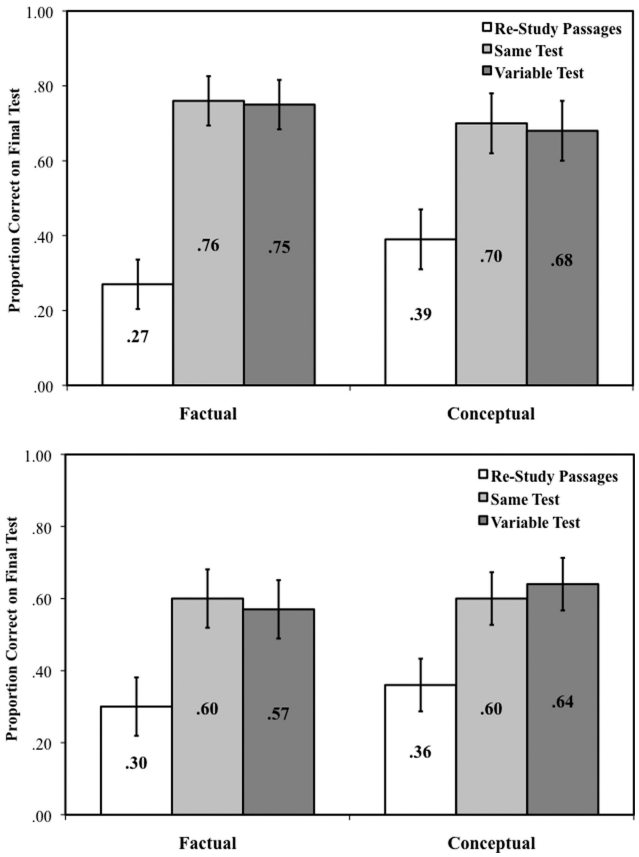

Figure 1. Proportion of correct responses on the final cued-recall test as a function of question type and initial learning condition for Experiment 1a (top panel) and Experiment 1b (bottom panel). Error bars represent 95% confidence intervals.

roughly equivalent in the same-test and variable-test conditions, but both testing conditions produced a greater proportion of correct responses than the restudy-passages condition. This pattern of performance held for both types of question. Experiment 1b yielded a similar pattern of results despite the change in the questions on the final test (i.e., new inferential questions rather than repeated questions). Performance was highest in the two initial testing conditions, both of which produced superior transfer relative to the restudy-passages condition. However, the possibility of superior transfer in the variable-test condition was not borne out: The same-test and variable-test conditions produced roughly equivalent performance. This pattern of results held for both the factual and conceptual transfer questions.

Performance on factual and conceptual questions was analyzed with separate one-way (initial learning condition: restudy passages, same test, variable test) repeated measures ANOVAs. For the factual repeated questions in Experiment 1a, there was a significant effect of initial learning condition, $F(2, 46) = 70.18$, $MSE = .03$, $\eta^2 = .75$. Planned pairwise comparisons revealed that both the same-test condition and the variable-test condition produced a significantly greater proportion of correct responses relative to the restudy-passages condition, .76 versus .27: $t(23) = 9.97$, $SEM = .05$, $d = 2.13$, and .75 versus .27: $t(23) = 10.40$, $SEM = .05$, $d = 2.09$, respectively. However, there was no significant difference between the same-test and variable-test conditions ($t < 1$). For the factual inferential questions in Experiment 1b, there was also a main effect of initial learning condition, $F(2, 46) = 16.73$, $MSE = .04$, $\eta^2 = .42$. Planned pairwise comparisons confirmed that both the same-test and variable-test conditions produced better transfer than the restudy-passages condition, .60 versus .30: $t(23) = 5.74$, $SEM = .05$, $d = 1.03$, and .57 versus .30: $t(23) = 4.38$, $SEM = .06$, $d = 0.93$, respectively. However, performance did not differ significantly between the two testing conditions ($t < 1$).

For the repeated conceptual questions in Experiment 1b, there was a significant main effect of initial learning condition, $F(2, 46) = 18.87$, $MSE = .04$, $\eta^2 = .45$. Pairwise comparisons confirmed the observation that the same-test and variable-test conditions led to significantly better performance on the final test than the restudy-passages condition, .70 versus .39: $t(23) = 5.50$, $SEM = .06$, $d = 1.29$, and .68 versus .39: $t(23) = 5.38$, $SEM = .06$, $d = 1.25$, respectively. Again, there was no significant difference between the two initial testing conditions ($t < 1$). For the conceptual inferential questions in Experiment 1b, there was also a main effect of initial learning condition, $F(2, 46) = 15.63$, $MSE = .03$, $\eta^2 = .41$. Planned pairwise comparisons revealed that the same-test and variable-test conditions led to better performance on the final transfer test relative to the restudy-passages condition, .60 versus .36: $t(23) = 4.44$, $SEM = .05$, $d = 0.74$, and .64 versus .36: $t(23) = 5.11$, $SEM = .05$, $d = 0.87$, respectively. There was no significant difference between the testing conditions ($t < 1$).

In addition to the factual and conceptual questions, the final test included a set of control questions. In Experiment 1a, performance was higher in restudy-passages condition ($M = .39$) relative to both the same-test ($M = .21$) and variable-test ($M = .24$) conditions. A one-way (initial learning condition: restudy passages, same test, variable test) repeated measures ANOVA revealed a significant main effect, $F(2, 46) = 4.10$, $MSE = .05$, $\eta^2 = .15$. Follow-up pairwise comparisons showed that the restudy-passages condition led to a significantly greater proportion of correct responses on the control questions relative to the same-test condition, .39 versus .21: $t(23) = 2.67$, $SEM = .07$, $d = 0.68$, and the variable-test condition, .39 versus .24: $t(23) = 1.98$, $SEM = .07$, $p = .06$, $d = 0.55$, although the latter difference was only marginally significant. There was no significant difference between the two testing conditions ($t < 1$). In Experiment 1b, the restudy-passages condition ($M = .36$) produced a higher proportion of correct responses than the same-test ($M = .23$) and variable-test ($M = .27$) conditions. However, a one-way (initial learning condition: restudy passages, same test, variable test) repeated measures ANOVA showed that this numerical difference was not significant, $F(2, 46) = 2.30$, $MSE = .05$, $p = .11$. The failure to detect a significant effect was likely due to insufficient power (see Results of Experiment 2).

Conditional analyses. Conditional analyses were conducted to explore how performance on the initial tests affected final test performance. Of interest was the extent to which successful retrieval or transfer on the final test depended upon successful retrieval on one or more of the initial tests. Table 3 shows the proportion of correct responses on the final test as a function of initial learning condition and retrieval success on the initial tests (successful on one or more tests vs. unsuccessful on all tests). In

Table 3
Proportion of Correct Responses on the Final Test as a Function of Initial Learning Condition and Retrieval Success on the Initial Tests (Successful on One or More Tests Vs. Unsuccessful on All Tests) for Experiments 1a, 1b, 2, and 3

Experiment	Learning condition	Retrieval success on initial tests	Proportion correct on final test
1a	Same test	Successful	.84
		Unsuccessful	.13
	Variable test	Successful	.79
		Unsuccessful	.15
1b	Same test	Successful	.65
		Unsuccessful	.32
	Variable test	Successful	.64
		Unsuccessful	.28
2	Same test	Successful	.57
		Unsuccessful	.31
3	Same test	Successful	.72
		Unsuccessful	.49

Experiment 1a, when subjects successfully retrieved the correct response at least once during the initial tests, the probability of producing the correct response was very high. However, when subjects did not retrieve the correct response on any of the initial tests, they generally failed to produce the correct response on the final test (even though feedback was given after each initial test). In Experiment 1b, successful retrieval on the initial tests similarly led to the production of a greater proportion of correct responses on the final transfer test. However, some transfer did occur even when subjects failed to produce the correct response on the initial tests. This result may have been due to subjects gaining some partial knowledge about the fact or concept from repeated testing (albeit not enough to constitute a correct response on the initial test) and this partial knowledge allowing them to work out the correct response to the associated transfer question on the final test.

To confirm these observations, a 2 (retrieval success: successful, unsuccessful) $\times$ 2 (initial learning condition: same test, variable test) repeated measures ANOVA was conducted for each experiment. For Experiment 1a, the ANOVA revealed a significant main effect of retrieval success, $F(1, 15) = 166.31$, $MSE = .04$, $\eta^2 = .82$, but neither the main effect of initial learning condition nor the interaction was significant (Fs < 1).[1] For Experiment 1b, the ANOVA showed a significant main effect of retrieval success, $F(1, 16) = 31.04$, $MSE = .07$, $\eta^2 = .51$, which confirmed the observation that retrieval success during initial learning led to superior transfer on the final test. Neither the main effect of initial learning condition nor the interaction was significant (Fs < 1).[2]

Discussion

Experiments 1a and 1b produced several important results. In both experiments, performance increased across the three initial tests in a curvilinear manner, and this pattern held for both testing conditions and for both types of questions. In Experiment 1a, repeated testing led to better performance than repeated studying on the main set of factual and conceptual questions that were repeated verbatim on the final test; however, there was no difference in performance between the two testing conditions. In Experiment 1b, the final test performance on the new inferential questions showed that repeated testing produced superior transfer of both factual and conceptual information relative to repeated studying of the passages. However, the variable-test condition did not produce better transfer than the same-test condition. The conditional analyses conducted on both experiments revealed that subjects retained and transferred a high proportion of the information that they successfully retrieved at least once on the initial tests but otherwise generally failed to produce the correct response on the final test. Finally, repeated studying of the passages led to better performance than repeated testing for the control questions on the final test, but this effect was only significant in Experiment 1a.

The most important result that emerged from these experiments is the finding that repeated testing produced better transfer than repeated studying in Experiment 1b. The vast majority of previous studies on the testing effect used a final test with questions repeated verbatim from the initial tests (e.g., Butler & Roediger, 2008; Carrier & Pashler, 1992; Karpicke & Roediger, 2008). Since these prior studies focused on the retention of a specific response, the question of whether retrieval practice promotes the acquisition of knowledge that can be transferred to new contexts was left open. Thus, the results of Experiment 1b are exciting because they indicate that the mnemonic benefits of retrieval practice extend beyond the retention of a specific response. Relative to repeated studying of passages, repeated testing led to better performance on new inferential questions that required the application of previously learned information to produce a new response. In addition, repeated testing led to better transfer of both factual and conceptual information.

Why did repeated testing produce better retention and transfer than repeated studying of the passages? The results of the conditional analyses suggest that the successful retrieval of information from memory during the initial learning session may be the critical mechanism. When a fact or concept was retrieved at least once on the initial tests, there was a high probability that it would be successfully retrieved again (Experiment 1a) or transferred (Experiment 1b) on the final test. As indicated by the curvilinear increase in the proportion of correct responses across the initial tests in both experiments, the feedback provided after each test was also important because it enabled subjects to correct their errors and successfully retrieve the correct response on a subsequent test. By the third test, subjects were able to retrieve about 80% of the facts and concepts at least once, and they retained or transferred much of that information on the final test 1 week later.

A final result of note from Experiments 1a and 1b is that the variable-test condition did not lead to superior final test performance relative to the same-test condition in either experiment. Encoding variability should increase the probability of future

[1] Eight subjects were excluded from this analysis because they correctly answered every question on the initial tests at least once in one or more of the two testing conditions and thus did not produce a mean for unsuccessful retrieval on the initial tests.

[2] Seven subjects were excluded from this analysis because they correctly answered every question on the initial tests at least once in one or more of the two testing conditions and thus did not produce a mean for unsuccessful retrieval on the initial tests.

retrieval because it creates multiple retrieval routes to a particular memory. One potential reason that the variable-test condition did not produce superior retention in Experiment 1a is the nature of the final test; encoding variability might not confer any mnemonic benefit on a final test with questions that were repeated verbatim from the initial tests because the additional features that are encoded in the variable-test condition are superfluous (but see Goode et al., 2008). However, the final test in Experiment 1b consisted of new inferential questions and thus presented a situation in which encoding variability could be expected to help. One possibility is that the greater variety of features encoded in the variable-test condition did not match the features in the retrieval cues any better than the features encoded in the same-test condition.

Alternatively, the way in which the questions were rephrased in the variable-test condition may not have made them different enough to induce encoding variability. Still another possibility is that there was a substantial amount of encoding variability in both the same-test and variable-test conditions due to other factors (e.g., the spaced presentation of the questions, the random ordering of questions within a test), and thus, the rephrasing manipulation only added a small degree of variability. The failure to find support for the encoding variability hypothesis is briefly discussed in the General Discussion, but as a result of the null effects found in first two experiments, the variable-test condition was dropped for Experiments 2 and 3.

Experiment 2

There were two main goals in conducting Experiment 2. The first goal was to replicate the finding from Experiment 1b that repeated testing led to better transfer relative to repeated studying of the passages. The second goal was to compare repeated testing with a more stringent control condition: repeated studying of the isolated facts and concepts. In this new restudy-isolated-sentences control condition, subjects were presented with the individual facts and concepts and told to study them in anticipation of a test (for a similar procedure, see Butler & Roediger, 2008). Thus, the information processed in the restudy-isolated-sentences condition was essentially the same as that processed in the repeated-testing condition, except that there was no attempt to retrieve the information in the former condition. Critically, this new control condition also allowed for the evaluation of an alternative explanation for the results of Experiments 1a and 1b: that the differences in final test performance were due to differences in total time on task during initial learning. The standard repeated restudy-passages condition of Experiments 1a and 1b was also included in Experiment 2. The design, materials, and procedure were the same as in Experiment 1b, except that the variable-test condition was dropped to include the restudy-isolated-sentences condition.

Method

Subjects and design. Twenty-four undergraduate psychology students at Washington University in St. Louis participated for course credit or pay (subjects were paid $30 to participate). The design was a 3 (type of initial learning: restudy passages, restudy isolated sentences, same test) × 2 (type of initial test question: factual, conceptual) within-subjects design. Both variables were manipulated within subjects but between materials. As in Experiment 1b, the main dependent variable was new inferential questions on the final transfer test.

Materials and counterbalancing. The materials from Experiment 1b were used.

Procedure. The procedure was the same as in Experiment 1b with the exception that the variable-test condition was replaced by the restudy-isolated-sentences condition. In the restudy-isolated-sentences condition, subjects studied each fact and concept for 30 s. There was a total of four facts and four concepts per passage, so the restudy-isolated-sentences condition and restudy-passages condition were equated in terms of total time on task (4 min).

Results

Scoring. A research assistant and I each scored 20% of the cued-recall responses independently in the same manner as in the previous experiments. Interrater reliability was high ($\kappa = .90$), so I resolved the few disagreements and then scored the remaining responses alone.

Initial tests. Table 2 shows the proportion of correct responses on the three initial cued-recall tests as a function of question type for the same-test condition. As expected, the overall pattern of results mirrored those observed in Experiments 1a and 1b. The proportion of correct responses increased across successive tests in a curvilinear manner. Separate one-way (test: 1, 2, 3) repeated measures ANOVAs were used to analyze performance on the factual and conceptual questions. For factual questions, there was a significant main effect of test, $F(2, 46) = 81.81$, $MSE = .02$, $\eta^2 = .78$, for which the quadratic trend was also significant, $F(1, 23) = 11.09$, $MSE = .02$, $\eta^2 = .33$. Likewise, there was a main effect of test for conceptual questions, $F(2, 46) = 71.26$, $MSE = .01$, $\eta^2 = .76$, and a significant quadratic trend, $F(1, 23) = 18.27$, $MSE = .02$, $\eta^2 = .44$.

Response times. The mean number of seconds that subjects spent on each question (i.e., both responding and reviewing feedback) across the three initial tests was computed for the same-test condition. Overall, subjects spent more time on conceptual questions than on factual questions, 42.1 versus. 26.2: $t(23) = 7.84$, $SEM = 2.03$, $d = 1.12$. The average time spent on each question was also compared to the time spent restudying the isolated facts and concepts. Subjects spent significantly more time answering the conceptual questions than restudying the concepts, 42.1 versus 30.0: $t(23) = 4.02$, $SEM = 5.88$, $d = 1.01$, but they spent significantly more time restudying the facts than answering the fact questions, 30.0 versus 26.2: $t(23) = 2.36$, $SEM = 7.11$, $d = 0.65$.

The average total time spent on each test during the initial learning phase was calculated for the same-test condition by multiplying the average time spent on factual and conceptual questions by the total number of each type of question per passage (four factual and four conceptual). On average, subjects spent 273 s (4.6 min) to complete a test on each passage, which was slightly more time than the 240 s (4.0 min) that they spent in the two repeated study conditions (i.e., either restudying a passage or restudying all the isolated facts and concepts from a passage). A one-way (initial learning condition: restudy passages, restudy isolated sentences, same test) repeated measures ANOVA showed a significant difference among the conditions, $F(2, 46) = 3.62$, $MSE = 2{,}450.76$, $\eta^2 = .14$. However, follow-up pairwise comparisons only yielded

marginally significant differences between the same-test condition and the restudy-passages and restudy-isolated-sentences conditions, 273 versus 240: $t(23) = 1.90$, $SEM = 17.50$, $p = .07$; the results were the same for both comparisons.

Final test. Figure 2 shows the proportion of correct responses on the final cued-recall test as a function of question type and initial learning condition. The same-test condition produced higher performance than both the restudy conditions, and this pattern held for both factual and conceptual questions. Interestingly, restudying the isolated facts and concepts did not lead to better transfer relative to restudying the entire passage. Subjects presumably spent more time processing each fact and concept in the restudy-isolated-sentences condition than in the restudy-passages condition, yet this additional study time did not lead to greater transfer.

Performance on the factual and conceptual inferential questions was analyzed separately by one-way (initial learning condition: restudy passages, restudy isolated sentences, same test) repeated measures ANOVAs. There was a significant main effect of initial learning condition for factual inferential questions, $F(2, 46) = 10.21$, $MSE = .03$, $\eta^2 = .31$. Pairwise comparisons showed that same-test condition led to significantly higher final test performance than the restudy-passages condition, .53 versus .31: $t(23) = 5.74$, $SEM = .05$, $d = 1.03$, and the restudy-isolated-sentences condition, .53 versus .33: $t(23) = 3.22$, $SEM = .06$, $d = 0.85$. There was no significant difference between the two restudy conditions ($t < 1$). For the conceptual inferential questions, there was also a significant main effect of initial learning condition, $F(2, 46) = 4.13$, $MSE = .05$, $\eta^2 = .15$. Pairwise comparisons confirmed that the same-test condition produced significantly better transfer than the restudy-passages condition, .58 versus .41: $t(23) = 2.27$, $SEM = .06$, $d = 0.63$, and the restudy-isolated-sentences condition, .58 versus .44: $t(23) = 2.61$, $SEM = .07$, $d = 0.54$. Again, the two restudy conditions did not differ significantly ($t < 1$).

The control questions on the final test were also analyzed to determine whether testing might benefit other (untested) material from the same passage. The restudy-passages condition ($M = .30$) produced a higher proportion of correct responses on the control questions relative to the same-test ($M = .24$) condition, replicating the results of Experiments 1a and 1b. The restudy-passages condition also led to better performance than the restudy-isolated-sentences condition ($M = .20$); this result makes sense because in the latter condition, subjects were only reexposed to the facts and concepts that were tested in the same-test condition rather than the whole passages that contained the information needed to answer the control questions. Despite the numerical superiority of the restudy-passages condition, a one-way (initial learning condition: restudy passages, restudy isolated sentences, same test) repeated measures ANOVA did not show a significant difference among the conditions, $F(2, 46) = 1.76$, $MSE = .04$, $p = .19$. Much like Experiment 1b, this null result is likely due to insufficient power. To address this issue, the data from Experiments 1a, 1b, and 2 were collapsed across experiment to compare performance on the control questions in the restudy-passages and same-test conditions. This analysis yielded a significant result: restudying the passages produced better performance than repeated testing, .35 versus .23: $t(71) = 3.69$, $SEM = .03$, $d = 0.50$.

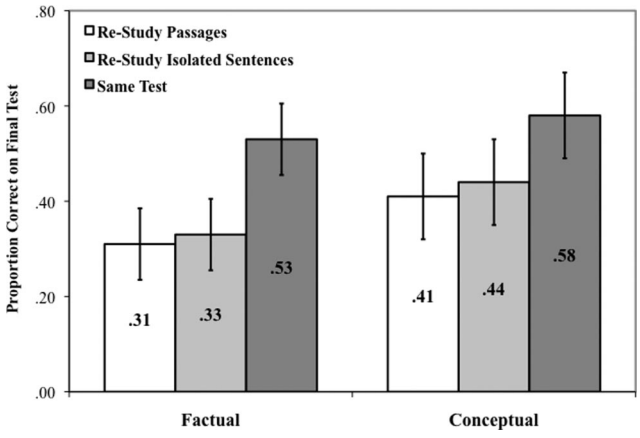

Figure 2. Proportion of correct responses on the final cued-recall test as a function of question type and initial learning condition for Experiment 2. Error bars represent 95% confidence intervals.

Conditional analyses. The relationship between performance on the initial test and performance on the final test was examined through conditional analyses. The proportion of correct responses on the final test was calculated as a function of retrieval success on the initial tests (successful on one or more tests vs. unsuccessful on all tests; see Table 3). Successful retrieval on the initial tests led to a significantly greater proportion of correct responses on the final test relative to when subjects were unsuccessful on the initial tests, .57 versus .31: $t(16) = 3.26$, $SEM = .08$, $d = 0.97$.[3]

Discussion

Experiment 2 replicated and extended the key findings of Experiment 1b by incorporating a more stringent control condition. As in both previous experiments, performance increased on each successive test in a curvilinear fashion for both factual and conceptual questions. On the final test, repeated testing led to better performance than repeated studying of the passages (replicating Experiment 1b) and repeated studying of the isolated facts and concepts. The latter two conditions did not differ. Conditional analyses again indicated that a greater proportion of correct responses were produced on the final transfer test when the related fact or concept had been successfully retrieved at least once on the initial tests. Finally, repeated studying of the passages produced a higher proportion of correct responses on the control questions than repeated testing or repeated studying of the isolated facts and concepts, but the effect was not reliable.

The major finding that emerged from Experiment 2 is that repeated testing produced better transfer than both repeated studying of the passages and repeated studying of the isolated facts and concepts. Any finding must be viewed with some degree of skepticism until it is replicated, and thus, it was important to demonstrate that the principal result from Experiment 1b could be obtained again. In addition, the comparison of the same-test and restudy-isolated-sentences conditions provided a more stringent

[3] Seven subjects were excluded from this analysis because they correctly answered every question on the initial tests at least once in one or more of the two testing conditions and thus did not produce a mean for unsuccessful retrieval on the initial tests.

assessment of whether retrieval might be the critical mechanism that produced the superior transfer in Experiment 1b. The restudy-isolated-sentences condition arguably represents a better control condition because subjects repeatedly studied the same facts and concepts that were repeatedly tested in the same-test condition without being reexposed to the additional information that was contained in each passage. In other words, these two conditions were well matched except for one major difference: The same-test condition provided the opportunity for retrieval, whereas the restudy-isolated-sentences condition did not. Thus, the finding that repeated testing produced superior final test performance relative to repeated study of isolated facts and concepts provides strong support for the idea that retrieval of information from memory promotes transfer of learning.

Alternatively, it is possible that the differences in final test performance resulted from differences in the total time spent on task during the initial learning session. Subjects spent more time taking the tests in the same-test condition than they did restudying the passages or restudying the facts and concepts. However, this difference was mainly due to the large amount of time required to type in the responses to the conceptual questions on each test. Subjects spent significantly less time completing the factual questions in the same-test condition (26.2 s) than they did studying the facts in the restudy-isolated-sentences condition (30.0 s). If the total time explanation is correct, then the restudy-isolated-sentences condition should have produced better final test performance on the factual items relative to the same-test condition (or at least equivalent performance). However, as the results clearly show, repeated testing produced substantially more transfer on the inferential questions related to the facts than repeatedly studying the isolated facts. In addition, the restudy-isolated-sentences condition did not lead to better transfer than the restudy-passages condition for either type of question, even though subjects presumably spent more time processing each fact and concept in the former condition. Indeed, there are many studies that show that increasing the amount of time spent processing material does not always improve retention (e.g., Amlund, Kardash, & Kulhavy, 1986; Callender & McDaniel, 2009). Overall, this evidence indicates that the total time hypothesis is not a viable explanation for the present results.

Experiment 3

Far transfer is difficult to obtain in both laboratory and applied studies, but it is very important to understand (see Barnett & Ceci, 2002). Indeed, Detterman (1993) argued that experimental investigations of transfer should be considered trivial unless they demonstrate far transfer, and his criterion essentially requires far transfer along multiple dimensions in Barnett and Ceci's (2002) framework (e.g., knowledge domain, physical context, temporal context, modality, etc.). With such a stringent criterion, only a small number of studies would qualify as having demonstrated far transfer (e.g., Adey & Shayer, 1993; Chen & Klahr, 1999; Fong, Krantz, & Nisbett, 1986; Herrnstein, Nickerson, de Sanchez, & Swets, 1986; Kosonen & Winne, 1995). In contrast with Detterman's criterion, the main goal of Experiment 3 was relatively modest: to explore whether retrieval practice could be used to promote far transfer along a single dimension in Barnett and Ceci's framework. To this end, the experiment included a final test that assessed transfer of learning to new inferential questions in different knowledge domains, which constitutes far transfer along the knowledge domain dimension.

The design, materials, and procedure were similar to those used in the previous experiments, except for a few critical changes. The primary change was that new final test questions were developed, each of which required subjects to use a concept that they had acquired in the initial learning session to make inferences about a related concept in a completely different domain. Second, the factual items were dropped because they were so specific that it was impossible to find a related fact in a different domain for many items. In the first three experiments, each passage consisted of eight paragraphs: four paragraphs that each contained one of the four critical facts and another four paragraphs that each contained one of the four critical concepts (see Method for Experiments 1a and 1b). The paragraphs that contained the critical facts were dropped for Experiment 3, making the passages shorter. Third, only two initial learning conditions were used: Subjects were repeatedly tested on some passages and repeatedly studied other passages.

Method

Subjects and design. Twenty undergraduate psychology students at Washington University in St. Louis participated for course credit or pay (subjects were paid $25 to participate). The sole independent variable was type of initial learning (restudy passages, same test), which was manipulated within subjects but between materials. The main dependent variable was new inferential questions within different knowledge domains.

Materials and counterbalancing. The materials from the first three experiments were used with some modifications. Only the material related to the concepts was used because the facts were too specific to allow the creation of related inferential questions from different knowledge domains. The six passages were reduced in length from 1,000 to 500 words each by cutting out the paragraphs associated with the facts, and the questions about the facts were dropped from the tests. For each concept, a new inferential question was created to assess transfer to a different knowledge domain. For example, the following concept was tested on the initial test (or restudied in the passage): "A bat has a very different wing structure from a bird. What is the wing structure of a bat like relative to that of a bird?" (Answer: "A bird's wing has fairly rigid bone structure that is efficient at providing lift, whereas a bat has a much more flexible wing structure that allows for greater maneuverability.") The related inferential question about a different domain was the following: "The U.S. Military is looking at bat wings for inspiration in developing a new type of aircraft. How would this new type of aircraft differ from traditional aircrafts like fighter jets?" (Answer: "Traditional aircrafts are modeled after bird wings, which are rigid and good for providing lift. Bat wings are more flexible, and thus an aircraft modeled on bat wings would have greater maneuverability.")

Each inferential question included some mention of the relevant concept from the initial learning session. Whether or not subjects spontaneously recognize that prior learning is relevant to a new situation is an important determinant of transfer (see Gick & Holyoak, 1987). Obviously, if subjects do not spontaneously recognize that prior learning is relevant, it would be impossible for

transfer to occur. Thus, the purpose of giving subjects a hint was to negate the need for them to recognize that a previously learned concept was relevant (for a similar procedure, see Gick & Holyoak, 1980; Reed, Ernst, & Banerji, 1974), focusing instead on their ability to recall and apply that concept to answer the inferential question. For counterbalancing purposes, two orders of initial learning condition were crossed factorially with two orders of the passages to create four versions of the experiment.

Procedure. The procedure was same as that used in the first three experiments with a few exceptions. During the initial learning session, subjects studied all six of the passages and then either repeatedly took a test on the passages or repeatedly restudied them. The final test consisted of new inferential questions about different domains. Subjects were explicitly instructed that the test would require them to think about the information that they learned in the previous session and use that information to infer the answers to the final test questions.

Results

Scoring. The cued-recall responses were scored in the same manner as in the previous experiments, and interrater reliability was high ($\kappa = .91$).

Initial tests. As in the first three experiments, the proportion of correct responses on the initial cued-recall tests increased in a curvilinear fashion from Test 1 ($M = .38$) to Test 2 ($M = .71$) to Test 3 ($M = .78$) in the same-test condition (see Table 2). A one-way (test: 1, 2, 3) repeated measures ANOVA revealed a significant main effect of test, $F(2, 38) = 99.89$, $MSE = .01$, $\eta^2 = .84$, for which there was also a significant quadratic trend, $F(1, 19) = 58.41$, $MSE = .02$, $\eta^2 = .76$.

Final test. The same-test condition produced better transfer relative to the restudy-passages condition, and this observation was confirmed by a paired-samples t test, .68 versus .44: $t(19) = 5.23$, $SEM = .05$, $d = 0.99$.

Conditional analyses. Conditional analyses were conducted to examine whether final test performance was correlated with initial test performance (see Table 3). Subjects produced a significantly greater proportion of correct responses on the final test when they had successfully retrieved the concept at least once on the initial tests relative to when they had not retrieved the concept, .72 versus .49: $t(16) = 2.61$, $SEM = .09$, $d = 0.73$.[4]

Discussion

The results of Experiment 3 replicated many of the findings of the first three experiments but also produced an important new finding: Repeated testing produced better transfer to new inferential questions from different domains relative to repeated studying of the passages. The results of the conditional analyses indicated that the retrieval of information from memory may be the critical mechanism that produced the difference in final test performance. When subjects successfully retrieved a concept on at least one of the initial tests, they were more likely to correctly answer the related transfer question on the final test than if they failed to retrieve it on all three tests. This new finding is important because it extends the mnemonic benefits of retrieval practice to situations in which knowledge must be transferred to a different context. The results of Experiment 3 are discussed further in the General Discussion.

General Discussion

In a series of four experiments, I investigated how repeated testing and repeated studying affect the retention and transfer of facts and concepts contained in prose passages. Experiment 1a showed that repeated testing led to better retention of facts and concepts than repeated studying of passages. However, repeated testing with different versions of a question did not lead to better final test performance than repeated testing with the same version of the question. Experiment 1b built upon Experiment 1a by demonstrating that repeated testing also led to better transfer to new questions within the same knowledge domain relative to repeated studying of passages. Again, repeated testing with different versions of a question did not lead to better transfer than repeated testing with the same version of the question. Experiment 2 replicated Experiment 1b by showing that repeated testing led to better transfer than both repeated studying of passages and repeated studying of the isolated facts and concepts relevant to the questions. Experiment 3 extended the findings of Experiments 1b and 2 by showing that repeated testing produced better transfer even to new questions in different knowledge domains relative to repeated studying of passages.

Overall, the findings of the present study clearly demonstrate the effectiveness of retrieval practice in promoting both retention and transfer of knowledge. I now turn to discussing these findings in more depth. First, I consider the significance of the findings within the broader memory literature. Second, I examine why the retrieval of information from memory produced superior transfer by discussing some possible theoretical explanations for this novel finding. Third, I briefly reassess the encoding variability hypothesis in light of the results of Experiments 1a and 1b. Finally, I close with some remarks about the implications of the present findings for educational practice and a few ideas for future research.

Retrieval Practice Produces Superior Retention and Transfer

The most important finding that emerged from the present research is that repeated practice at retrieving information from memory produced better transfer to several different types of questions than repeatedly studying the same information. Relatively few studies have investigated whether the benefits of testing extend beyond the retention of a specific response. For the most part, researchers have focused on evaluating various theoretical explanations of the testing effect (e.g., Glover, 1989; Pyc & Rawson, 2009) and establishing its generalizability to various materials (e.g., Carpenter & Pashler, 2007) and applied contexts (e.g., Larsen, Butler, & Roediger, 2009; McDaniel, Anderson, et al., 2007). The possibility that retrieval practice could promote superior transfer has been largely ignored in the testing-effect literature despite the importance of demonstrating transfer to theories of memory and learning as well as for educational practice.

Experiments 1b and 2 showed that repeated testing produced better transfer to new inferential questions within the same knowl-

[4] Five subjects were excluded from this analysis because they correctly answered every question on the initial tests at least once in one or more of the two testing conditions and thus did not produce a mean for unsuccessful retrieval on the initial tests.

edge domain than either repeatedly studying passages or repeatedly studying isolated facts and concepts. A few recent studies have investigated whether retrieving information from memory promotes transfer to new inferential questions within the same knowledge domain. As discussed in the introduction, Johnson and Mayer (2009) found that taking an initial transfer test led to superior transfer relative to restudying a multimedia slide show. However, there are some methodological issues that preclude drawing strong conclusions from their findings. Rohrer et al. (2010) found that taking an initial test on a map produced better transfer than restudying the map. Yet their results differ from the present findings in that the same set of correct responses (the region or city names) was used for initial test, the retention test, and the transfer test.

In addition, McDaniel, Howard, and Einstein (2009, Experiment 2) had subjects use one of three study strategies while reading complex passages that described mechanical devices: (a) read the passage, attempt to recall it from memory, and then reread the passage; (b) read the passage twice; or (c) read the passage twice and take notes while reading. On a final test 1 week later, subjects who had attempted to recall the passages between readings were significantly better at answering inferential questions than subjects who had repeatedly read the passages only (however, the reading-with-note-taking condition produced equivalent performance to the testing condition).

Experiment 3 of the present research showed that repeated testing produced better transfer to new inferential questions in different knowledge domains relative to repeatedly studying passages, thus extending the difference between the initial learning and subsequent transfer contexts farther than in any previous testing-effect study. This result is impressive because transfer to a different knowledge domain constitutes far transfer along a single dimension in Barnett and Ceci's (2002) taxonomy, and far transfer has been notoriously difficult to obtain in many laboratory experiments (see Barnett & Ceci, 2002). Only one other study has reported a similar result, albeit within a very different paradigm. In a series of five experiments on analogical reasoning, Needham and Begg (1991) presented subjects with training problems and then had them either attempt to generate a solution (before hearing the correct solution) or study the correct solution. The authors labeled the generate condition as *problem-oriented training* and the study condition as *memory-oriented training* (which is perhaps somewhat ironic in hindsight). Attempting to generate solutions to the training problems led to significantly better performance on the subsequent transfer problems relative to studying the solutions. Interestingly, this result was obtained even though subjects rarely succeeded in generating the correct solution to the initial training problems. Thus, the findings of Needham and Begg differ in an important way from the findings of the present research in which retrieval of the correct response occurred frequently during the initial learning phase.

Performance on the control questions that were included on the final test also provided an interesting set of results. The purpose of the control questions was to explore whether the benefits of repeated testing extended to other (untested) information contained in the same passages and to examine any potential differences in retention that result from studying a text four times versus just one time. In Experiments 1a, 1b, and 2, repeated studying of the passages led to better performance on the control questions relative to repeated testing. Due to the small number of control items, there was insufficient power to detect a significant difference in Experiments 1b and 2. However, when the control question data were collapsed across the three experiments, performance was significantly higher in the restudy-passages condition than in the same-test condition. Nevertheless, studying a passage four times only improved performance by 12% relative to studying a passage once, a relatively small gain given the large amount of additional time spent studying.

One potential explanation for the small magnitude of this effect is that subjects may not have been making the effort to restudy the passages. The potential for lack of effort during restudy tasks is always a possibility in this type of experiment. Aside from monitoring subjects to make sure that they are attending to the passages (which was done in the present set of experiments), there is no way to guarantee that they are carefully restudying the passages without changing the nature of the task. Still, the spaced presentation of the passages and the experimenter control of study time in the present research should have made it more likely that subjects would expend the effort to restudy the passages (i.e., relative to massed presentation and self-paced study).

Theoretical Explanations for the Mnemonic Benefits of Retrieval Practice

Why did repeated testing produce better transfer than repeated studying? As Barnett and Ceci (2002) argued, the memory demands involved in the process of transfer can be broken down into three components: recognition, recall, and execution. First, a person must recognize that prior learning is relevant to a new context. Second, the person must successfully recall the knowledge that was learned earlier. Third, the person must use or apply that knowledge to successfully execute the transfer task. In the present study, there were no memory demands with respect to the recognition component because subjects were explicitly told that the questions on the final test were related to the information they had learned in the previous session. However, there were significant memory demands with respect to the recall component. Given the fact that retrieval practice produces better retention than restudying, the recall component is probably one locus of the superior transfer produced by repeated testing relative to repeated studying. Of course, retrieval practice may have also affected the execution component by enhancing subjects' ability to apply the knowledge they had learned earlier to answer the inference questions. Attempting to produce a response from memory to answer a question may foster better understanding of the information relative to restudying it. For example, McDaniel et al. (2009) argued that retrieval practice promotes deep learning of the material more than restudying the material does. Unfortunately, the recall and execution components cannot be separated in the present study, and thus, additional research is required to determine whether retrieval practice influences both components of the transfer process.

Given that repeated testing likely improved retention and thereby affected the recall component of transfer, how can the mnemonic benefits of retrieval practice be explained? A number of different explanations have been put forth to account for the testing effect, most of which focus on retrieval as the critical mechanism. One idea is that the act of retrieving information from memory leads to the elaboration of existing retrieval routes and/or the

creation of additional retrieval routes (e.g., Bjork, 1975; McDaniel & Masson, 1985). Taking a test after studying may result in the encoding of additional features or the formation of alternative routes to access the memory trace, whereas restudying the material does not. Thus, this explanation for the testing effect incorporates the concept of encoding variability (Bower, 1972; Estes, 1955; Martin, 1968), which is discussed further in the next section.

A related idea is that the effort involved in retrieval is responsible for the testing effect (e.g., Gardiner, Craik, & Bleasdale, 1973). Retrieval that requires greater effort is assumed to produce better retention than less effortful retrieval, similar to the idea of depth of processing at encoding (e.g., Craik & Tulving, 1975). One piece of evidence that supports this hypothesis is the finding that production tests generally produce superior retention relative to recognition tests on a final test given later (e.g., Butler & Roediger, 2007). Additional support comes from the finding that increasing the spacing of initial tests leads to better retention (e.g., Jacoby, 1978; Modigliani, 1976). Several recent studies that directly tested the retrieval effort hypothesis also support this explanation (e.g., Carpenter & DeLosh, 2006; Pyc & Rawson, 2009).

Another idea that may help to explain the benefits of testing is the concept of transfer-appropriate processing (Morris et al., 1977). According to this hypothesis, memory performance is enhanced to the extent that the processes during encoding match those required during retrieval. In most testing-effect studies, retention is generally assessed with a final test, and thus, an encoding condition in which memory is tested may provide a better match. That is, the processes engaged during an initial test are highly similar to the processes required on the final test, whereas the processes engaged while restudying the material are different. Indeed, some researchers have argued that retrieving information from memory strengthens the process of retrieval itself, rather than the specific representation or trace in memory (Runquist, 1983; Wheeler, Ewers, & Buonanno, 2003).

The "new theory of disuse" proposed by Bjork and Bjork (1992) incorporates many of the ideas into a more formal theoretical explanation for the testing effect (as well as other memory phenomena). According to their theory, each item or representation in memory has two strengths: (a) *storage strength*, which reflects how well the item is learned, and (b) *retrieval strength,* which reflects how easy it is to retrieve the item at any given point in time given the cues provided. Storage strength is assumed to grow with each study or retrieval opportunity, and the accumulated strength is never lost. Retrieval strength also grows with each study or retrieval opportunity, but the accumulated strength is gradually lost as a function of subsequent study and retrieval of other items. Thus, storage capacity is assumed to be unlimited, whereas retrieval capacity is limited. That is, an infinite number of events can be stored, but only a finite number will be retrievable at any given point. The distinction between storage strength and retrieval strength is a similar to the distinction between *habit strength* and *response strength* in Estes's (1955) stimulus sampling theory. The new theory of disuse also incorporates his idea of stimulus fluctuation that motivated later encoding variability theories (e.g., Bower, 1972).

Bjork and Bjork's (1992) theory provides an explanation for the testing effect by assuming that retrieving information from memory produces greater increases in storage and retrieval strength than does studying the information again. An item's retrieval strength and storage strength increase whenever that item is either studied or retrieved from memory. However, the magnitude of the increases in retrieval strength and storage strength depend upon the current retrieval strength; the higher the current retrieval strength, the smaller the increases will be in magnitude. Thus, successful retrieval of an item with low retrieval strength produces greater increments in retrieval strength and storage strength than successful retrieval of an item with high retrieval strength. This assumption incorporates the retrieval effort hypothesis discussed above (e.g., Gardiner et al., 1973) and explains both the finding that production tests produce better retention on a later test than do recognition tests (e.g., Butler & Roediger, 2007) and the finding that increasing the spacing of tests increases retention (e.g., Jacoby, 1978; Modigliani, 1976).

The new theory of disuse (Bjork & Bjork, 1992) can also account for the common finding that restudying often produces equivalent or better performance than taking a test when retention is assessed with an immediate final test, whereas testing produces better retention on delayed tests (e.g., Roediger & Karpicke, 2006b; Runquist, 1983; Toppino & Cohen, 2009; Wheeler et al., 2003). This retention interval interaction is explained by reasoning that taking an initial test produces greater increases in storage and retrieval strength than restudying the items, but only for the items that are successfully retrieved; restudying produces smaller increases in storage and retrieval strength for all the items. If retention is assessed immediately, restudying will result in a greater or equivalent number of items being accessible relative to prior testing. However, if retention is assessed after a delay, the retrieval strength of the restudied items will have decreased faster than the retrieval strength of the tested items, resulting in testing producing superior performance relative to restudying on the final test.

Despite the emphasis on successful retrieval as the critical mechanism, it is clear that the feedback provided after each question also played an important role in producing superior retention and transfer in the present research. First and foremost, feedback enabled test takers to correct errors (Bangert-Drowns, Kulik, Kulik, & Morgan, 1991) and maintain correct responses (Butler, Karpicke, & Roediger, 2008) during initial learning, increasing the probability that successful retrieval would occur on the next test. In addition, there is some evidence that unsuccessful retrieval attempts can enhance future learning (e.g., Kornell, Hays, & Bjork, 2009; Slamecka & Fevreiski, 1983), a finding that is sometimes referred to as *test potentiation* (Izawa, 1970). Unsuccessful retrieval attempts may increase deep processing of the question and subsequent feedback or activate related knowledge that enhances processing of the feedback. Although the relative contributions of testing and feedback in producing the superior retention and transfer cannot be determined in the present research because of the procedure used, it is an important question for future research.

As a final note, it is important to stress that the total time hypothesis does not provide a valid explanation for the superior retention and transfer produced by repeated testing in the present research. Thompson, Wenger, and Barling (1978; see too Kolers, 1973) were the first to suggest that simply the additional exposure to material provided by taking a test is responsible for producing the testing effect. However, several subsequent studies have directly tested the total time hypothesis and found no support for it (e.g., Carrier & Pashler, 1992; Glover, 1989; Roediger &

Karpicke, 2006b; Toppino & Cohen, 2009). Numerous reviewers of the testing-effect literature have also evaluated the total time hypothesis in light of existing evidence and determined that it is not satisfactory (e.g., Dempster, 1996; Roediger & Karpicke, 2006a).

Within the present research, three findings argue against total time on task as an explanation. First and most important, subjects spent significantly less time completing the factual questions in the same-test condition in Experiment 3 than they did studying the facts in the restudy-isolated-sentences condition, yet repeated testing produced substantially more transfer on the inferential questions related to the facts than repeated studying of the facts. Second, subjects spent more time processing the critical facts and concepts in the restudy-isolated-sentences condition of Experiment 3 than they did in the restudy-passages condition, and yet these two restudy conditions yielded equivalent performance on the final transfer test. Third, performance on the control questions in Experiments 1–3 showed that studying a passage four times produced only modest gains in retention relative to studying it once. Clearly, the total time hypothesis can be eliminated as a potential explanation for the findings of the present research.

Encoding Variability Failed to Produce Superior Retention and Transfer

A secondary goal of the present research was to explore whether repeated testing using rephrased questions would lead to better retention and transfer than repeated testing using the same question. The hypothesis was that repeated testing with different questions should induce encoding variability, which would create multiple retrieval routes in memory. As the number of retrieval routes increased, the probability of successful retrieval in the future should have also increased, resulting in superior retention and transfer. However, the results of Experiments 1a and 1b do not support this hypothesis. Although the results of Experiments 1a and 1b do not support the encoding variability hypothesis, they do not invalidate the hypothesis either. The broader literature contains mixed results: Some studies have found evidence to support the notion of encoding variability (e.g., McDaniel & Masson, 1985; McFarland, Rhodes, & Frey, 1979), whereas others have failed (e.g., Maskarinec & Thompson, 1976; Postman & Knecht, 1983).

One improvement that could be made in future experiments would be to provide greater specification of the features that would be encoded as a result of answering the different versions of the initial test questions, the features that would comprise the retrieval cues given on the final transfer test, and the relationship between these sets of features. Encoding variability theories have been criticized for being vague with respect to the features that are being varied from trial to trial (e.g., Hintzman, 1974, 1976). However, it is possible to specify in greater detail the features involved in encoding variability (e.g., Glenberg, 1979), and thus, researchers should try to include such specification in future studies. On the whole, encoding variability theories retain great explanatory power, so further research that tests the predictions of these theories is certainly warranted.

Practical Application to Education

The findings of the present research also have implications for educational practice and vocational training, as well as any other situation in which transfer is desirable. The substantial literature on the testing effect has already led many researchers to advocate for the use of testing as a learning tool (e.g., Glover, 1989; Leeming, 2002; Roediger & Karpicke, 2006a). However, one major criticism that has been leveled at testing-effect research is that testing only promotes the learning of a specific response, which is not the primary goal of education or vocational training. The results obtained in this study and other recent investigations (e.g., Johnson & Mayer, 2009; McDaniel et al., 2009) suggest that the mnemonic benefits of retrieving information from memory extend well beyond the retention of a specific response. At the very least, testing produces superior retention of information, which represents an important component of the transfer process.

Concluding Remarks

The findings reported in this article have important implications for future research on both transfer of learning and the testing effect. The traditional approach to studying transfer of learning has been to focus purely on the similarities and differences between the contexts of initial learning and subsequent transfer. Although the match between contexts is important in determining whether transfer occurs, the present research shows that it is also important to consider how the conditions of initial learning can be arranged to better promote transfer. More specifically, the finding that retrieval practice was highly effective in promoting transfer in the present study suggests that it may enhance transfer in other paradigms too (e.g., Needham & Begg, 1991). Future research on transfer of learning should investigate how testing can be used to optimize subsequent performance in a range of transfer contexts.

Concomitantly, future research on the testing effect needs to continue to explore whether the mnemonic benefits of retrieval practice extend beyond the retention of a specific response. Although the further development of theory is also clearly a priority, exploring how testing can be used to promote transfer should be a primary area of investigation in testing-effect research. In addition, it will be important to determine why retrieval practice promotes superior transfer. The findings of the present research suggest that testing may promote transfer because it increases the retention of information, which makes the recall component of transfer possible (within the framework proposed by Barnett & Ceci, 2002). However, repeated testing may also improve people's understanding of the material, enabling them to better perform the execution component of the transfer process (i.e., the ability to apply the knowledge to a new situation).

References

Adey, P., & Shayer, M. (1993). An exploration of long-term far-transfer effects following an extended intervention program in the high school science curriculum. *Cognition and Instruction, 11,* 1–29.

American Psychological Association. (2002). Ethical principles of psychologists and code of conduct. *American Psychologist, 57,* 1060–1074.

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

Bangert-Drowns, R. L., Kulik, C. C., Kulik, J. A., & Morgan, M. (1991). The instructional effect of feedback in test-like events. *Review of Educational Research, 61,* 213–238.

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we

learn? A taxonomy for far transfer. *Psychological Bulletin, 128,* 612–637.

Bassok, M., & Holyoak, K. J. (1989). Interdomain transfer between isomorphic topics in algebra and physics. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 153–166.

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In R. L. Solso (Ed.), *Information processing and cognition* (pp. 123–144). New York, NY: Wiley.

Bjork, R. A., & Bjork, E. L. (1992). A new theory of disuse and an old theory of stimulus fluctuation. In A. Healy, S. Kosslyn, & R. Shiffrin (Eds.), *From learning processes to cognitive processes: Essays in honor of William K. Estes* (Vol. 2, pp. 35–67). Hillsdale, NJ: Erlbaum.

Bloom, B. S. (1956). *Taxonomy of educational objectives: The classification of educational goals.* Essex, England: Harlow.

Bower, G. H. (1972). Stimulus sampling theory of encoding variability. In A. W. Melton & E. Martin (Eds.), *Coding processes in human memory* (pp. 85–123). New York, NY: Wiley.

Bruce, R. W. (1933). Conditions of transfer of training. *Journal of Experimental Psychology, 16,* 343–361.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III (2008). Correcting a metacognitive error: Feedback enhances 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.

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

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

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

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,* 533–571.

Chen, Z., & Klahr, D. (1999). All other things being equal: Acquisition and transfer of the control of variables strategy. *Child Development, 70,* 1098–1120.

Cohen, J. (1960). A coefficient of agreement for nominal scales. *Educational and Psychological Measurement, 20,* 37–46.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum.

Craik, F. I. M., & Tulving, E. (1975). Depth of processing and the retention of words in episodic memory. *Journal of Experimental Psychology: General, 104,* 268–294.

Dempster, F. N. (1996). Distributing and managing the conditions of encoding and practice. In E. L. Bjork & R. A. Bjork (Eds.), *Human memory* (pp. 197–236). San Diego, CA: Academic Press.

Detterman, D. K. (1993). The case for the prosecution: Transfer as an epiphenomenon. In D. K. Detterman & R. J. Sternberg (Eds.), *Transfer on trial: Intelligence, cognition, and instruction* (pp. 1–24). Westport, CT: Ablex Publishing.

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). New York, NY: Dover. (Original work published 1885)

Ellis, H. (1965). *The transfer of learning.* Oxford, England: Macmillan.

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review, 62,* 369–377.

Fisher, R. P., & Craik, F. I. M. (1977). Interaction between encoding and retrieval operations in cued recall. *Journal of Experimental Psychology: Human Learning and Memory, 3,* 701–711.

Fong, G. T., Krantz, D. H., & Nisbett, R. E. (1986). The effects of statistical training on thinking about everyday problems. *Cognitive Psychology, 18,* 253–292.

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

Gardiner, J. M., Craik, F. I. M., & Bleasdale, F. A. (1973). Retrieval difficulty and subsequent recall. *Memory & Cognition, 1,* 213–216.

Geisser, S., & Greenhouse, S. W. (1958). An extension of Box's results on the use of *F* distribution in multivariate analysis. *Annals of Mathematical Statistics, 29,* 885–891.

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12,* 306–355.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1–38.

Gick, M. L., & Holyoak, K. J. (1987). The cognitive basis of knowledge transfer. In S. M. Cormier & J. D. Hagman (Eds.), *Transfer of training: Contemporary research and applications* (pp. 9–46). New York, NY: Academic Press.

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition, 7,* 95–112.

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

Goode, M. K., Geraci, L., & Roediger, H. L., III (2008). Superiority of variable to repeated practice in transfer on anagram solution. *Psychonomic Bulletin & Review, 15,* 662–666.

Herrnstein, R. J., Nickerson, R. S., de Sanchez, M., & Swets, J. A. (1986). Teaching thinking skills. *American Psychologist, 41,* 1279–1289.

Hintzman, D. L. (1974). Theoretical implications of the spacing effect. In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola Symposium* (pp. 77–99). Oxford, England: Erlbaum.

Hintzman, D. L. (1976). Repetition and memory. In G. H. Bower (Ed.), *The psychology of learning and motivation* (Vol. 10, pp. 47–91). New York, NY: Academic Press.

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition, 15,* 332–340.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83,* 340–344.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17,* 649–667.

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

Kanak, N. J., & Neuner, S. D. (1970). Associative symmetry and item availability as a function of five methods of paired-associate acquisition. *Journal of Experimental Psychology, 86,* 288–295.

Karpicke, J. D., & Roediger, H. L., III (2008, February 15). The critical importance of retrieval for learning. *Science, 15,* 966–968.

Kimball, D., & Holyoak, K. J. (2000). Transfer and expertise. In E. Tulving & F. I. M. Craik (Eds.), *The Oxford handbook of memory* (pp. 109–122). New York, NY: Oxford University Press.

Kolers, P. A. (1973). Remembering operations. *Memory & Cognition, 1,* 347–355.

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

Kosonen, P., & Winne, P. H. (1995). Effects of teaching statistical laws on reasoning about everyday problems. *Journal of Educational Psychology, 87,* 33–46.

Larsen, D. P., Butler, A. C., & Roediger, H. L., III (2009). Repeated testing improves long-term retention relative to repeated study: A randomized, controlled trial. *Medical Education, 43,* 1174–1181.

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

Martin, E. (1968). Stimulus meaningfulness and paired-associate transfer: An encoding variability hypothesis. *Psychological Review, 75,* 421–441.

Maskarinec, A. S., & Thompson, C. P. (1976). The within-list distributed practice effect: Tests of the varied context and varied encoding hypotheses. *Memory & Cognition, 41,* 741–746.

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

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 11,* 371–385.

McDaniel, M. A., Roediger, H. L., III, & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic Bulletin & Review, 14,* 200–206.

McFarland, C. E., Rhodes, D. D., & Frey, T. J. (1979). Semantic-feature variability and the spacing effect. *Journal of Verbal Learning and Verbal Behavior, 18,* 163–172.

McGeoch, J. A. (1942). *The psychology of human learning: An introduction.* New York, NY: Longmans, Green.

McKenzie, G. R. (1972). Some effects of frequent quizzes on inferential thinking. *American Educational Research Journal, 9,* 231–240.

Modigliani, V. (1976). Effects on a later recall by delaying initial recall. *Journal of Experimental Psychology: Human Learning and Memory, 2,* 609–622.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer-appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16,* 519–533.

Moscovitch, M., & Craik, F. I. M. (1976). Depth of processing, retrieval cues, and uniqueness of encoding as factors in recall. *Journal of Verbal Learning and Verbal Behavior, 15,* 447–458.

Needham, D. R., & Begg, I. M. (1991). Problem-oriented training promotes spontaneous analogical transfer: Memory-oriented training promotes memory for training. *Memory & Cognition, 19,* 543–557.

Pearson, K. (1911). On a correction needful in the case of the correlation ratio. *Biometrika, 8,* 254–256.

Postman, L., & Knecht, K. (1983). Encoding variability and retention. *Journal of Verbal Learning and Verbal Behavior, 22,* 133–152.

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.

Reed, S. K., Ernst, G. W., & Banerji, R. (1974). The role of analogy in transfer between similar problem states. *Cognitive Psychology, 6,* 436–450.

Roediger, H. L., III (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology, 59,* 225–254.

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.

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

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

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime reference guide.* Pittsburgh, PA: Psychology Software Tools.

Slamecka, N. J., & Fevreiski, J. (1983). The generation effect when generation fails. *Journal of Verbal Learning and Verbal Behavior, 22,* 153–163.

Thompson, C. P., Wenger, S. K., & Bartling, C. A. (1978). How recall facilitates subsequent recall: A reappraisal. *Journal of Experimental Psychology: Human Learning and Memory, 4,* 210–221.

Thorndike, E. L., & Woodworth, R. S. (1901a). The influence of improvement in one mental function upon the efficiency of other functions. (I). *Psychological Review, 8,* 247–261.

Thorndike, E. L., & Woodworth, R. S. (1901b). The influence of improvement in one mental function upon the efficiency of other functions: II. The estimation of magnitudes. *Psychological Review, 8,* 384–395.

Thorndike, E. L., & Woodworth, R. S. (1901c). The influence of improvement in one mental function upon the efficiency of other functions: III. Functions involving attention, observation, and discrimination. *Psychological Review, 8,* 553–564.

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

Tulving, E. (1983). *Elements of episodic memory.* New York, NY: Oxford University Press.

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

Received August 30, 2009
Revision received April 2, 2010
Accepted April 7, 2010 ■