Journal of Experimental Psychology: Applied
2017, Vol. 23, No. 4, 433–446

1076-898X/17/$12.00 http://dx.doi.org/10.1037/xap0000142

Retrieving and Applying Knowledge to Different Examples Promotes Transfer of Learning

Andrew C. Butler
University of Texas at Austin

Allison C. Black-Maier
Duke University

Nathaniel D. Raley
University of Texas at Austin

Elizabeth J. Marsh
Duke University

Introducing variability during learning often facilitates transfer to new contexts (i.e., generalization). The goal of the present study was to explore the concept of variability in an area of research where its effects have received little attention: learning through retrieval practice. In four experiments, we investigated whether retrieval practice with different examples of a concept promotes greater transfer than repeated retrieval practice with the same example. Participants watched video clips from a lecture about geological science and answered application questions about concepts: either the same question three times or three different questions. Experiments 3 and 4 also included conditions that involved repeatedly studying the information in the application questions (either the same example or three different examples). Two days later, participants took a final test with new application questions. All four experiments showed that variability during retrieval practice produced superior transfer of knowledge to new examples. Experiments 3 and 4 also showed a testing effect and a benefit from studying different examples. Overall, these findings suggest that repeatedly retrieving and applying knowledge to different examples is a powerful method for acquiring knowledge that will transfer to a variety of new contexts.

Keywords: retrieval practice, variability, transfer, learning

> ". . . The same thing recurring on different days, in different contexts, read, recited on, referred to again and again, related to other things and reviewed, gets well wrought into mental structure. This is the reason why you should enforce on your pupils habits of continuous application."
>
> —William James (1899/1962, p. 64)

Introducing variability during learning often facilitates transfer to new contexts (i.e., generalization). When learners practice with different exemplars or under varied conditions, it generally improves their subsequent performance in novel situations. Such positive effects of variability have been demonstrated in many different domains, including analogical reasoning (Bassok & Holyoak, 1989; Gick & Holyoak, 1983), speech perception (Barcroft & Sommers, 2005; Logan, Lively, & Pisoni, 1991), motor skills (Shapiro & Schmidt, 1982; Shea & Kohl, 1990), and categorization (Dukes & Bevan, 1967; Wahlheim, Finn, & Jacoby, 2012). The goal of the present study was to explore the concept of variability in an area of research where its effects have received little attention: learning through retrieval practice. Over the past decade, there has been a surge of interest in the mnemonic benefits of retrieval practice (i.e., "the testing effect"; for a review, see Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Roediger & Butler, 2011; Roediger & Karpicke, 2006; Rowland, 2014); however, among the multitude of studies that have been conducted, the effects of variability during repeated retrieval have rarely been examined.

Investigating how introducing variability during retrieval practice affects learning is important for several reasons. First, numerous studies have recently begun to explore the potential for retrieval practice to promote transfer of learning (for a review, see Carpenter, 2012), and variability is a critical factor in promoting transfer of learning in many domains. Second, knowledge about how variability affects learning through retrieval practice can inform theoretical explanations of the testing effect and, more broadly, the acquisition of declarative knowledge (e.g., facts, concepts, etc.). Third, the concept of introducing variability during retrieval practice parallels the way in which repeated practice on a particular fact or concept is often implemented in education (e.g., students are given

Andrew C. Butler, Department of Educational Psychology, University of Texas at Austin; Allison C. Black-Maier, Department of Psychology & Neuroscience, Duke University; Nathaniel D. Raley, Department of Educational Psychology, University of Texas at Austin; Elizabeth J. Marsh, Department of Psychology & Neuroscience, Duke University.

We thank Amberly Tenney, Ernst Casimir, Kaitlyn Batt, and Haley Steinman for their assistance in conducting this research and the Marsh Lab group for their comments on a draft of this article. This research was supported by a grant from the Spencer Foundation (201100093) to Andrew C. Butler and Elizabeth J. Marsh. All authors contributed to the idea for the research and designed the experiments. Andrew C. Butler, Allison C. Black-Maier, and Nathaniel D. Raley analyzed the data and drafted the manuscript. All authors edited the manuscript.

Correspondence concerning this article should be addressed to Andrew C. Butler, who is now at the Department of Education, Washington University in St. Louis, MO, 63130. E-mail: andrew.butler@wustl.edu

practice on different questions rather than the same exact question repeatedly).

To this end, we report a series of experiments that investigated how repeatedly retrieving and applying newly acquired knowledge to different examples affects learning. Of specific interest was how introducing variability during retrieval practice affects the development of deeper understanding that enables learners to transfer their knowledge to new examples. We first briefly review relevant prior research before describing the experiments that we conducted.

Background

Retrieving information just once improves subsequent memory performance relative to restudying the information (Butler & Roediger, 2007; Carpenter, 2009; Carrier & Pashler, 1992; Glover, 1989), and repeated retrieval practice is even more beneficial (Pyc & Rawson, 2009; Rawson & Dunlosky, 2011; Wheeler & Roediger, 1992). The vast majority of research on retrieval practice has focused on retention (see Dunlosky et al., 2013; Rowland, 2014); that is, the criterial test consists of retrieval cues that are identical to the cues given for practice during initial learning. However, there is growing interest in how retrieval practice affects subsequent transfer of learning to criterial tests with different retrieval cues. Recently, numerous studies have found that practicing retrieval produces superior transfer relative to various comparison activities, such as restudying information (Butler, 2010; Blunt & Karpicke, 2014; Hinze & Wiley, 2011; Johnson & Mayer, 2009; Kang, McDaniel, & Pashler, 2011; McDaniel, Howard, & Einstein, 2009; Rohrer, Taylor, & Sholar, 2010; for a review, see Carpenter, 2012). Now that the benefits of retrieval practice for subsequent transfer are well established, it is important to investigate factors that might facilitate the development of deeper understanding during retrieval practice, thus increasing transfer to new contexts. One factor that might improve subsequent transfer is variability of practice during repeated retrieval. As noted above, variability during learning yields transfer (i.e., generalization) in many other domains of learning, but this factor has received little attention with respect to retrieval practice. In most retrieval practice studies that incorporate multitrial learning, each trial almost always consists of an identical repetition of a cue–target pair (e.g., Karpicke & Roediger, 2008; Pyc & Rawson, 2009).

A better understanding of how variability during repeated retrieval practice affects learning can inform theories that attempt to explain the testing effect. Although none of these theories directly consider the effects of variability across repeated retrieval attempts, the concept of variability is a core aspect for many of them. A primary example is encoding variability theory, which is one account that has been proposed to explain the testing effect (e.g., McDaniel & Masson, 1985). Theories of encoding variability (e.g., Bower, 1972; Glenberg, 1979; Martin, 1968) posit that the information that is encoded into memory varies from trial to trial during repeated exposure to the same material because of changes in perception (e.g., presentation modality), processing (e.g., experimental task), internal context (e.g., neuronal activity), and/or external environment (e.g., location); greater variation in the information encoded increases the number of potential retrieval routes, which thereby increases the likelihood that the information will be retrieved with whatever cue is presented in the future. With respect to the testing effect, the idea is that studying and then taking a practice test introduces greater variability than studying and then restudying the same material.

Other theoretical accounts that focus specifically on explaining the testing effect also incorporate the concept of variability. The episodic context account (Karpicke, Lehman, & Aue, 2014; Lehman, Smith, & Karpicke, 2014) assumes that the temporal context in which events occur shifts over time, leading to different contextual elements being present at any given time. When people retrieve recently learned information from memory, they reinstate the prior context in which they learned that information and integrate new elements from the present context into the representation. When information is restudied, context reinstatement does not occur to the same degree and thus there is less variability in the contextual elements that are integrated into the representation. Similarly, the elaborative retrieval account (Carpenter, 2009, 2011; Carpenter & DeLosh, 2006; Glover, 1989) also suggests that retrieval produces greater variability in the elements that are integrated into the updated representation of an event relative to restudying, but the new elements are drawn from semantic knowledge rather than temporal context. This theory posits that the process of searching memory for target information produces spreading activation of related semantic knowledge, and these semantic elements are integrated into the representation of the event, thereby elaborating it. Finally, the retrieval effort account (Bjork, 1975; Bjork & Bjork, 1992) is another theoretical idea that can be interpreted in terms of the variability introduced by taking a test relative to restudying. In essence, the concept of retrieval effort captures the phenomenological experience that corresponds to the processing that occurs during retrieval and how it differs from restudying; greater effort presumably indicates a greater degree or complexity of processing, which should produce greater variability in the elements integrated into the updated representation.

As the description of these theories suggests, there are many ways in which variability can be introduced during repeated retrieval, and yet only a few studies have explored this topic (Butler, 2010; Glass, 2009; Goode, Geraci, & Roediger, 2008; Smith & Handy, 2014, 2016). Glass (2009) had students in a psychology course answer multiple-choice inference questions about facts (e.g., the definition of semantic priming). For each fact, they either answered the same question repeatedly or different versions of the question that had the same sentence frame but different details (e.g., "Cow will be read fastest when preceded by," "Tulip is read fastest if preceded by," "The word *cobra* will be read fastest if it is preceded by the word"). When students engaged in variable retrieval practice, they performed better on a new version of the question relative to when they practiced with the same question repeatedly. Smith and Handy (2014, 2016) had participants practice retrieving face–name pairs that were presented in front of an unrelated video background context. During repeated retrieval practice of each face–name pair, the video either stayed the same or changed. Practice with different video contexts led to better recall of the name on the final test in which the face was presented in front of novel video context. Goode et al. (2008) showed that practice with solving different arrangements of an anagram (e.g., "dolof," "folod," and "oofld" for the word "flood") led to better subsequent performance on a new arrangement of that anagram ("ldoof") relative to repeated practice with the same arrangement.

Finally, Butler (2010, Experiment 1b) manipulated whether or not a question was rephrased on successive retrieval attempts during initial learning, but the correct response was always the same. Performance on a subsequent transfer test with new inference questions showed no benefit of practice with rephrased questions during learning.

Taken as a whole, this handful of studies provides preliminary evidence to support the conclusion that variability during retrieval practice improves transfer of learning. However, it is important to note that these studies have all focused on variability in the information provided by the cue, while keeping constant both the target information that must be retrieved and the way in which the cue and target information are processed. The present set of experiments was designed to build upon these initial studies by focusing on variability in how the cue and target information are processed to answer a question. To accomplish this objective, retrieval practice consisted of application questions that presented participants with a new example and required them to make an inference based on recently learned information (see Appendix for sample materials). Thus, answering each application question involved more than just accessing information in memory in response to a new cue; rather, it required retrieving and the using that information to produce a novel inference.

Experiment 1

The goal of Experiment 1 was to investigate how variability of practice during repeated retrieval affects transfer of learning. Participants watched video clips from a lecture about geological science that covered the topics of volcanoes and earthquakes. After each video, they answered application questions about concepts covered in the clip (e.g., convection cell systems). For a given concept, they either answered the same application question three times or three different application questions. Each application question had a different correct answer, but required retrieving and using the same concept to make an inference. Two days later, participants took a final test that consisted of new application questions. Table 1 provides a schematic of this basic paradigm that was used across the four experiments.

Method

Participants. Thirty-two undergraduate students participated for payment. One additional participant who did not return for the second session was excluded. Sample size was determined before the experiment. The stopping rule was selected by heuristic with 32 participants representing exactly four times the number of counterbalance versions of the experiment (see Materials and Counterbalancing). A power analysis was not conducted because the novelty of the question being asked in the experiment meant that there was no prior research that could be used to provide an accurate estimate of the expected effect size.

Table 1
A Schematic Representation of the Paradigm Used in Experiments 1–4

Condition	Learning block				Final test
Retrieval: Same	V	R_A	R_A	R_A	R_D
Retrieval: Variable	V	R_A	R_B	R_C	R_D
Study: Same	V	S_A	S_A	S_A	R_D
Study: Variable	V	S_A	S_B	S_C	R_D

Note. The schematic illustrates the relationship between the activities in a single learning block and the final test. V = video; R = retrieval; S = study. The letters in subscript denote the specific question or example given to participants to practice retrieval or study, respectively. Experiments 1 and 2 did not include the two repeated study conditions represented in bottom half of the schematic. Participants had five learning blocks during Session 1 and the final test during Session 2.

Design. A single factor (Practice Type: same, variable) was manipulated within participants, between materials.

Materials and counterbalancing. Materials consisted of five video clips about geological science and associated questions (see Appendix for sample materials). The clips were taken from a course entitled "Nature of Earth: An Introduction to Geology" produced by The Great Courses. Each clip was approximately eight minutes long and contained two or three concepts for a total of 12 concepts. A concept was operationally defined as a piece of information that must be abstracted from multiple sentences. Four application questions in short-answer format were created for each concept. Answering these questions required participants to retrieve and apply their knowledge of a concept to a new example. The experiment was counterbalanced in two ways. Each concept was assigned equally often to the same and variable learning conditions across participants by creating two versions of the experiment. The odd numbered concepts were assigned to the same practice condition and the even numbered concepts were assigned to the variable practice condition, or vice versa. Additionally, the four questions related to each concept were counterbalanced such that they appeared equally often in initial learning tests and the final transfer test. The counterbalancing of questions was accomplished by rotating the questions through the four possible order positions (i.e., first practice, second practice, third practice, final test) to create four versions of the experiment. Crossing these two counterbalancing methods yielded eight versions of the experiment in total.

Procedure. The experiment consisted of two sessions spaced two days apart. All tasks were completed on a computer. Session 1 was separated into five blocks. In each block, participants watched one of the five video clips and then answered three questions per concept covered in that clip. The three questions consisted of either the same question three times or three different questions, depending on the type of practice to which a given concept was assigned. Due to the method of assigning concepts to condition, each block contained at least one concept from each condition. Within each block, the questions were presented one at a time in a random order and participants were instructed to generate a response to each question. If they did not know the correct answer, they were told to make a plausible guess. They were also informed that some questions would be repeated and that they should answer them again. Feedback was provided immediately after each question regardless of whether the response was correct or incorrect. Feedback consisted of a representation of the question and an idealized correct answer. Both responding and feedback viewing were self-paced. In Session 2, participants returned two days later to take a final test that consisted of 12 new application questions. The procedure and instructions for answering the final test questions was the same as for Session 1 except that participants did not receive feedback after each question. Participants were also prompted to rate their confidence in their

answer after each question on a 5-point scale (1 = *no confidence*, 5 = hi*gh confidence*); however, the analyses for these data will not be reported for any of the four experiments because none of the manipulations produced a significant difference in confidence.

Results

Eta-squared and Cohen's *d* are the measures of effect size reported in the analysis of variance (ANOVA) and *t* test analyses, respectively. A Geisser–Greenhouse correction was used for any violations of the sphericity assumption of ANOVA.

Coding. All responses were independently scored by two coders. Each response was marked as either correct or incorrect. The interrater reliability between the two coders was high ($\kappa = .85$) and the discrepancies were resolved by discussion.

Initial test performance. Table 2 contains the proportion of correct responses on the initial tests as a function of practice type and question order. As expected, performance on the first question did not differ significantly by practice type because the manipulation had not yet been implemented at this point, $t(31) = 1.11$, standard error of the mean = .04, $p = .28$, $d = .20$. Performance improved markedly from the first to second question and remained high in the same practice condition, whereas it remained relatively constant across the three questions in the variable practice condition. A 3 (Question Order) × 2 (Practice Type) repeated-measures ANOVA confirmed this observation by revealing significant main effects of practice type, $F(1, 31) = 36.55$, $MSE = .03$, $p < .0001$, $\eta^2 = .20$, and question order, $F(2, 52) = 29.45$, $MSE = .03$, $p < .0001$, $\eta^2 = .14$, which were qualified by a significant interaction, $F(2, 62) = 18.80$, $MSE = .03$, $p < .0001$, $\eta^2 = .12$.

Final test performance. Participants performed better on final test questions about concepts that they had learned through variable practice with different questions relative to same practice with a single question that was repeated three times (.64 versus .52), $t(31) = 2.77$, standard error of the mean = .04, $p = .009$, $d = .49$.

Discussion

Experiment 1 clearly showed that introducing variability during retrieval practice substantially benefited performance on the final test. Through the process of retrieving and applying knowledge of the concept to a variety of examples, participants gained deeper understanding of the concept that enabled them to transfer this knowledge to a new example when tested again two days later. Interestingly, this benefit emerged despite the fact that variability hindered gains in performance during initial learning. Whereas participants quickly reached ceiling performance by learning from feedback in the same practice condition, they struggled to apply the concepts to new examples in the variable practice condition.

Why was there relatively little improvement in performance across questions during variable retrieval practice? This finding is somewhat puzzling given that other studies investigating variability during retrieval practice have found substantial improvement in performance across different questions, even though the degree of improvement was less than repeated practice with the same question (Butler, 2010; Glass, 2009; Goode et al., 2008; Smith & Handy, 2014). One potential explanation for the lack of improvement during variable practice in Experiment 1 is that, as intended, the present study is asking a fundamentally different question about the effects of variability during learning. That is, whereas prior studies have focused on the retrieval of the same target information in response to different cues, participants in Experiment 1 had to process the cue and target information differently in order answer each application question. The process of transferring knowledge of a concept to a new example is difficult, and people often fail if given just a single opportunity (i.e., compared to three opportunities in the same practice condition).

Another possible explanation for the lack of improvement is that participants did not recognize that the questions in the variable practice condition were related, and thus they did not draw upon what they had learned from prior questions. This possibility seems unlikely given that every question related to a particular concept included a few key terms to help participants recognize the concept that needed to be applied (see Appendix). Nevertheless, participants in Experiment 1 were never explicitly instructed to pay attention to the relationship among different questions about the same concept. Research in many different literatures (e.g., concept learning, analogical reasoning, etc.) has demonstrated the importance of connecting subsequent events to prior events in learning from repeated presentations of material (e.g., Madigan, 1969; for a review, see Benjamin & Ross, 2010).

Table 2
Proportion Correct on the Initial Test in Experiments 1, 2, and 3 as a Function of Practice Type, Question Order, and Concept Labeling (Only Manipulated in Experiment 2)

Experiment	Concept labeling	Practice type	Question order		
			First	Second	Third
Experiment 1	Unlabeled	Same	.58	.92	.95
		Variable	.62	.68	.66
Experiment 2	Unlabeled	Same	.46	.81	.85
		Variable	.51	.64	.66
	Labeled	Same	.57	.86	.92
		Variable	.57	.69	.65
Experiment 3	Unlabeled	Same	.58	.88	.93
		Variable	.57	.68	.65
Experiment 4	Unlabeled	Same	.51	.86	.92
		Variable	.53	.57	.60

Experiment 2

The main goal of Experiment 2 was to examine the hypothesis that participants often failed to recognize the relationship among the three questions during variable practice. We manipulated the presence of a concept label above each question to draw participants' attention to the relationship among questions about the same concept. The concept label did not present any information that was not already given in the question, but it clearly signaled the relationship of the question to other questions and the accompanying instructions directed participants to attend to these relationships. A secondary goal was to replicate the novel finding in Experiment 1 that variable retrieval practice produced superior transfer on a final test with new application questions.

Method

Participants. Forty-eight undergraduate students participated for payment. Eight additional participants were excluded because they did not return for the second session. As in Experiment 1, sample size of was determined by heuristic before the experiment with 48 participants representing exactly six times the number of counterbalance versions of the experiment. The sample size was increased relative to Experiment 1 because of the addition of a second independent variable.

Design. The experiment had a 2 (Practice Type: same, variable) × 2 (Concept Labeling: labeled, unlabeled) mixed-factorial design. Concept labeling was manipulated between participants and practice type was manipulated within participants, between materials.

Materials and counterbalancing. The materials from Experiment 1 were used again. In addition, a short label was created for each concept. The label did not contain any new information and it did not provide any hints about the answer to the question. The experiment was counterbalanced in the same manner as Experiment 1 except for the question order. Instead of randomizing the order of presentation of the questions in the initial learning session, the four questions related to each concept were rotated through each position (first, second, third, and final test) such that they appeared equally often in each position across participants.

Procedure. The procedure was the same as Experiment 1 except for the following changes. In Session 1, participants were randomly assigned to one of the two concept labeling conditions. In the labeled condition, the concept label always appeared above each question and its corresponding feedback message. Participants were instructed to pay attention to the concept label above each question because there would be multiple questions related to the same concept.

Results

Coding. Responses were coded in the same manner as Experiment 1. The interrater reliability was high ($\kappa = .84$) and the discrepancies were resolved by discussion.

Initial test performance. Table 2 shows the proportion of correct responses on the initial tests as a function of practice type, question order, and concept labeling condition. Performance on the first question was approximately equal across the four conditions with a small numerical advantage for the labeled relative to the unlabeled condition. A 2 (Practice Type) × 2 (Concept Labeling) repeated-measures ANOVA showed no main effect of practice type ($F < 1$) or concept labeling, $F(1, 46) = 2.89$, $MSE = .06$, $p = .096$, $\eta^2 = .06$, and no interaction ($F < 1$).

Overall, performance on the initial test questions replicated the findings of Experiment 1. The proportion of correct responses increased substantially from the first question to the second and third questions in the same practice condition, while it increased only slightly across the three questions in the variable practice condition. Concept labeling did not appear to affect the pattern of performance. A 3 (Question Order) × 2 (Practice Type) × 2 (Concept Labeling) repeated-measures ANOVA revealed main effects of question order, $F(2, 92) = 66.99$, $MSE = .03$, $p < .0001$, $\eta^2 = .29$, and practice type, $F(1, 46) = 27.45$, $MSE = .04$, $p < .0001$, $\eta^2 = .09$, but not concept labeling, $F(1, 46) = 2.09$, $MSE = .11$, $p = .155$, $\eta^2 = .04$. The interaction between practice type and question order was also significant, $F(2, 92) = 22.19$, $MSE = .02$, $p < .0001$, $\eta^2 = .07$, but none of the other interactions were significant (all Fs < 1).

Final test performance. Figure 1 shows the proportion of correct responses on the final test as a function of practice type and concept labeling. Replicating the findings of Experiment 1, variable practice produced superior performance relative to same practice. Labeling the questions produced slightly better final test performance in the variable practice conditions, but concept labeling did not affect performance in the same practice conditions. Confirming these observations, a 2 (Practice Type) × 2 (Concept Labeling) repeated-measures ANOVA showed a significant main effect of practice type, $F(1, 46) = 9.96$, $MSE = .03$, $p = .003$, $\eta^2 = .18$. Neither the main effect of concept labeling nor the interaction was significant (Fs < 1).

Discussion

Replicating the results of Experiment 1, Experiment 2 showed that variable practice produced superior transfer to new application questions relative to same practice. However, concept labeling did not have a significant effect on initial or final test performance, which suggests that participants often recognized the relationship among the three different questions in the variable practice condition even in the absence of a concept label and instructions to attend to the relationship among questions. Nevertheless, it is important to consider that the related questions were presented in close succession with only one or two unrelated questions in between. A clear signal of the relationship among the questions may become increasingly important with greater spacing among questions.

Experiment 3

In Experiment 3, we sought to compare the relative effects of introducing variability during repeated retrieval practice and repeated study of information, a common control activity in testing

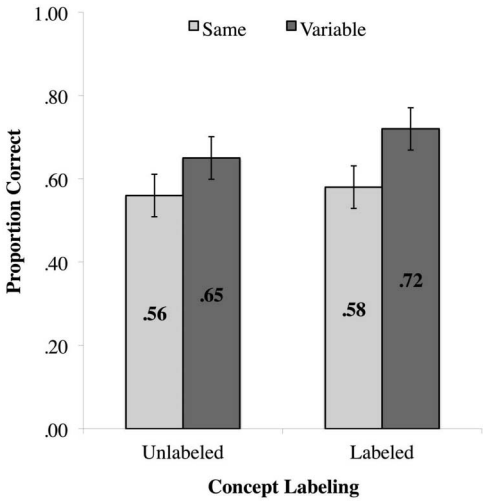

Figure 1. Proportion of correct responses to new application questions on the final test as a function of practice type and concept labeling in Experiment 2. Error bars represent 95% confidence intervals.

effect research (e.g., Glover, 1989; see Roediger & Butler, 2011; Roediger & Karpicke, 2006). The effects of introducing variability during repeated study have been extensively investigated within the literature on encoding variability (Bower, 1972; Estes, 1955; Martin, 1968; Glenberg, 1979). As described above, theories of encoding variability posit that variability across exposures to the same material results in the integration of new elements into the representation in memory, which increases the probability of a match with a subsequent retrieval cue. Although some studies have produced results that support encoding variability ideas (e.g., McDaniel & Masson, 1985; McFarland, Rhodes, & Frey, 1979), other studies have found that variability during study has no effect or a negative effect on retention (e.g., Maskarinec & Thompson, 1976; Postman & Knecht, 1983). Although the findings have been mixed, it is important to note that the vast majority of the research on encoding variability during repeated study has used basic materials (e.g., lists of words) and assessed retention by using cues that were highly similar or identical to the material present during initial learning. That is, no research has looked at how variability during repeated study affects understanding of the type of complex concepts used in the present experiments.

To compare the effects of variability on repeated retrieval practice and repeated study, we manipulated both learning activity and practice type within the same experiment (i.e., the two factors were crossed; see Table 1). In the two repeated study conditions (i.e., same study and variable study), participants studied the exact same material that appeared in the application questions given in the two retrieval practice conditions (i.e., same retrieval practice or variable retrieval practice). The material from each application question and corresponding feedback was edited into paragraph form so that it could be studied as an example. Participants in the repeated study condition either studied the same example three times or three different examples about the same concept. We also manipulated practice type between participants to see whether the findings from Experiments 1 and 2 would generalize from a within-participants design.

Method

Participants. Forty-eight undergraduate students participated for payment. One additional participant was excluded for failing to follow instructions. The sample size of 48 participants was determined before the experiment based on the sample size of Experiment 2.

Design. The experiment had a 2 (Practice Type: same, variable) × 2 (Learning Activity: study, retrieval practice) mixed-factorial design. Practice type was manipulated between participants and learning activity was manipulated within participants, between materials.

Materials and counterbalancing. The materials from Experiment 1 were used again. In addition, each question and correct answer was edited into a paragraph to create materials for the study condition (see Appendix for sample materials). Thus, participants in the retrieval practice and study conditions were exposed to the exact same information. The experiment was counterbalanced in the same general manner as Experiment 2 with each concept also assigned equally often to the study and retrieval practice conditions across participants.

Procedure. The procedure was the same as Experiment 1 except for the following changes. In Session 1, participants were randomly assigned to either the same or variable practice condition. Regardless of practice type, six of the concepts were repeatedly studied and the other six concepts were assigned to repeated retrieval practice. In the study condition, participants were required to study each example for a minimum of 30 s, but could continue to study for as long as they wanted.

Results

Coding. Responses were coded in the same manner as Experiments 1 and 2. The interrater reliability was high ($\kappa = .82$) and the discrepancies were resolved through discussion.

Initial test performance. Table 2 shows the proportion of correct responses on the initial tests as a function of question order and practice type. Performance on the first question was nearly equivalent in the same and variable retrieval practice conditions ($t < 1$). The overall pattern of performance replicated Experiments 1 and 2. The proportion of correct responses increased greatly from the first question to the second and third questions in the same retrieval practice condition, but only increased slightly in the variable retrieval practice condition. A 3 (Question Order) × 2 (Practice Type) repeated-measures ANOVA revealed significant main effects of question order, $F(2, 78) = 26.64$, $MSE = .03$, $p < .0001$, $\eta^2 = .33$, and practice type, $F(1, 46) = 10.32$, $MSE = .10$, $p = .002$, $\eta^2 = .18$, as well as a significant interaction, $F(2, 92) = 8.94$, $MSE = .03$, $p = .0002$, $\eta^2 = .11$.

Final test performance. Figure 2 depicts the proportion of correct responses on the final test as a function of practice type and learning activity. Clearly, both factors affected participants' ability to transfer their learning on the final test. Repeated retrieval during initial learning led to superior performance relative to repeated studying, and variable practice produced better performance than same practice. A 2 (Practice Type) × 2 (Learning Activity) repeated-measures ANOVA showed main effects of practice type,

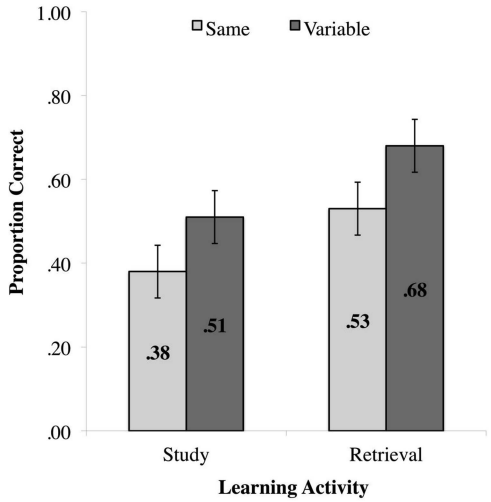

Figure 2. Proportion of correct responses to new application questions on the final test as a function of practice type and learning activity in Experiment 3. Error bars represent 95% confidence intervals.

$F(1, 46) = 6.14$, $MSE = .10$, $p = .017$, $\eta^2 = .12$, and learning activity $F(1, 46) = 20.10$, $MSE = .03$, $p < .0001$, $\eta^2 = .30$. The interaction was not significant ($F < 1$).

Discussion

Experiment 3 showed that retrieval practice with different questions produced superior transfer to new application questions on the final test relative to repeated retrieval practice with the same question. This result replicates the findings from Experiments 1 and 2 and demonstrates that it generalizes to a between-subjects design. Collapsing across practice type, repeated retrieval practice also produced superior transfer relative to repeated study, a testing effect that is consistent with the findings of most other studies on retrieval practice and transfer (e.g., Butler, 2010; see Carpenter, 2012). Interestingly, the effects of learning activity and practice type were additive but not superadditive, which suggests that there is no special synergy that results from combining retrieval practice with variable practice.

Another novel result from Experiment 3 was that introducing variability during repeated study produced superior transfer to new application questions on the final test. Studying three different examples allowed learners to develop a deeper understanding of the complex concepts presented in the videos relative to studying the same example three times. This finding is important given that most studies on encoding variability use basic materials and focus on retention as the primary outcome measure. Given the mixed findings in the encoding variability literature, the results of Experiment 3 suggest that it may be more fruitful to explore the effects of encoding variability with more complex materials and transfer as the outcome measure.

Experiment 4

The purpose of Experiment 4 was to replicate the findings from Experiment 3 and further investigate the mechanism(s) by which variable practice and retrieval practice promote superior transfer of knowledge. Transfer can be conceptualized as a three-step process (see Barnett & Ceci, 2002) in which the learner needs to (a) recognize that knowledge acquired in prior context is relevant to a new context, (b) recall that knowledge, and (c) apply that knowledge to the new context. Retention of knowledge drives the first two steps in this process (i.e., recognizing and recalling), whereas understanding enables the third step. Thus, in this conceptualization, retention is a necessary but not sufficient condition for transfer to occur. Retention is obviously critical, especially when the ability to transfer knowledge is assessed after a delay, such as in the present set of experiments; however, transfer will not occur without sufficient understanding to enable the application of knowledge to the new context.

In Experiment 4, we wanted to investigate whether variable practice and retrieval practice, respectively, produce better retention, deeper understanding, or both. Recognition was unlikely to have been a problem for participants in any of the experiments because they were instructed that the questions on the final test were about material from the videos, and thus they were aware that they needed to recall and apply the knowledge that they had acquired during the initial learning session. However, the differences that emerged on the final test used in Experiments 1–3 could be due to improved ability to recall knowledge (i.e., retention), to apply knowledge (i.e., understanding), or both. Thus, we adopted a two-phase final test procedure that has been used in prior research to isolate the unique contributions of retention and understanding to successful transfer (Butler, Godbole, & Marsh, 2013).

Experiment 4 used the same design and procedure as Experiment 3 except for the addition of a second phase to the final test in which participants reanswered the new application questions with a description of the relevant concept present (see Appendix for sample materials). The purpose of this second phase was to separate the recall and application steps in the transfer process. On the first phase, which was the same as the final test in Experiments 1–3, differences in performance on the new application questions could be due to participants' ability to recall their knowledge of the concepts, apply this knowledge, or both. However, presenting a description of the concept during the second reanswer phase eliminates the need to recall this knowledge; as a result, any differences in performance in the second phase should reflect participants' ability to apply their knowledge. Table 3 explains the logic of this approach for examining how variable retrieval practice and same retrieval practice affect the three steps of the transfer process (recognition, recall, and application); however, the same logic applies to any comparisons that can be made among the four conditions.

Method

Participants. Sixty undergraduate students participated for course credit. Four additional participants were excluded: one who did not return for the second session and three who did not follow instructions. The sample size was determined by heuristic before the experiment with 60 participants representing exactly 10 times the six counterbalance versions of the experiment. The sample size was increased relative to Experiments 3 and 4 because of the addition of the second phase of the final test.

Design. The design was the same as Experiment 3.

Materials and counterbalancing. The materials from Experiment 3 were used again. The counterbalancing was the same as

Table 3
The Logic Behind the Two-Phase Final Test Used in Experiment 4

Step in the transfer process	Final test phase	
	Initial answer to new application question	Reanswer after concept represented
Recognition	Variable RP = Same RP	Variable RP = Same RP
Recall	Variable RP > Same RP	Variable RP = Same RP
Application	Variable RP > Same RP	Variable RP > Same RP

Note. RP = retrieval practice. When initially answering the new application questions, there should be no difference between the same and variable retrieval practice conditions with respect to the recognition component of the feedback process; however, the variable retrieval practice condition could lead to better recall and/or application. In the reanswer phase, both recognition and recall are equated between the same and variable retrieval practice conditions by presenting the concept description; thus, a superiority of variable retrieval practice over same retrieval practice must be due to the application component. The same logic can be applied to compare the two practice types within the repeated study condition, and also to compare repeated retrieval practice and repeated study.

Experiment 3 with one exception. To avoid the possibility of ceiling effects on the second phase of the final test, the most difficult question for each concept was identified by looking at the data from Experiments 1–3 and then assigned to be presented on the final test. That is, out of the four questions related to each concept, the question that yielded the lowest proportion correct was reserved for the final test, while the other three questions were rotated through each position during initial learning (first, second, third).

Procedure. The procedure was the same as Experiment 3 except for the addition of a second phase to the final test. After participants completed the final test by answering all 12 of the new application questions, they were unexpectedly asked to reanswer each question in the presence of a description of the relevant concept (see Appendix for example). The description of the concept was presented at the top of the screen with the question directly below. Participants were instructed that they could provide the same answer as they did in the first phase or change their answer based on the information in description.

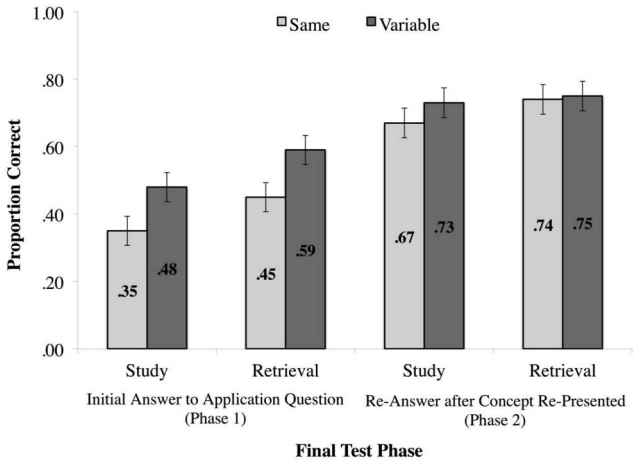

Figure 3. Proportion of correct responses to new application questions on the initial answer (left side) and reanswer (right side) phases of the final test as a function of practice type and learning activity in Experiment 4. Error bars represent 95% confidence intervals.

Results

Coding. Responses were coded in the same manner as the other experiments. The interrater reliability was high ($\kappa = .90$) and the discrepancies were resolved through discussion.

Initial test performance. Table 2 shows the proportion of correct responses on the initial tests as a function of question order and practice type. As expected, performance on the first question was approximately equal in the same and variable test conditions ($t < 1$). Just as in Experiments 1–3, performance increased substantially across the three questions in the same retrieval practice condition, but only increased slightly in the variable retrieval practice condition. A 3 (Question Order) × 2 (Practice Type) repeated-measures ANOVA revealed significant main effects of question order, $F(2, 116) = 42.13$, $MSE = .02$, $p < .00001$, $\eta^2 = .42$, and practice type, $F(1, 58) = 14.29$, $MSE = .12$, $p = .0003$, $\eta^2 = .20$, as well as a significant interaction, $F(2, 116) = 23.95$, $MSE = .02$, $p < .00001$, $\eta^2 = .29$.

Final test performance: Phase 1. The left side of Figure 3 depicts the proportion of correct responses on the initial answer phase of the final test as a function of practice type and learning activity. The pattern of performance replicated the results of Experiment 3. Variable produced superior transfer relative to same practice for both repeated retrieval practice and repeated study. In addition, repeated retrieval practice led to better transfer than repeated study. A 2 (Practice Type) × 2 (Learning Activity) repeated-measures ANOVA showed significant main effects of practice type, $F(1, 58) = 4.61$, $MSE = .10$, $p = .036$, $\eta^2 = .07$, and learning activity, $F(1, 58) = 12.39$, $MSE = .03$, $p = .001$, $\eta^2 = .18$. The interaction was not significant ($F < 1$).

Final test performance: Phase 2. The right side of Figure 3 depicts the proportion of correct responses on the reanswer phase of the final test as a function of practice type and learning activity. As expected, overall performance increased substantially from Phase 1 to Phase 2 because participants could use the concept descriptions when reanswering the application questions. In addition, the pattern of performance across the four conditions changed substantially. Performance was approximately equivalent in the variable retrieval practice, same retrieval practice, and variable study conditions, whereas performance in the same study condition was slightly lower. Despite this numerical difference, a 2 (Practice Type) × 2 (Learning Activity) repeated-measures ANOVA showed no main effects of practice type ($F < 1$) and learning activity, $F(1, 58) = 2.21$, $MSE = .03$, $p = .14$, $\eta^2 = .04$, and the interaction was not significant ($F < 1$). A follow-up paired samples t test revealed no significant difference between performance in the variable study and same study conditions (.73 vs. .67), $t(58) = .935$, standard error of the difference (SED) = .07, $p = .35$, $d = .24$.

To compare Phase 1 to Phase 2, a 2 (Practice Type) × 2 (Learning Activity) × 2 (Phase) repeated-measures ANOVA was conducted. This analysis revealed main effects of learning activity, $F(1, 58) = 8.03$, $MSE = .05$, $p = .006$, $\eta^2 = .12$, and phase, $F(1, 58) = 116.84$, $MSE = .03$, $p = .0001$, $\eta^2 = .67$, but no main effect of practice type, $F(1, 58) = 2.37$, $MSE = .19$, $p = .13$, $\eta^2 = .04$. However, these main effects were qualified by significant interactions between phase and practice type, $F(1, 58) = 4.58$, $MSE = .03$, $p = .037$, $\eta^2 = .07$, as well as a phase and learning activity, $F(1, 58) = 4.38$, $MSE = .04$, $p = .013$, $\eta^2 = .07$. The three-way interaction was not significant, $F(1, 58) = 1.28$, $MSE = .01$, $p = .262$, $\eta^2 = .02$.

Discussion

Experiment 4 replicated all of the main findings from Experiment 3. The novel question being asked was why do variable practice and retrieval practice produce superior transfer of learning—is it the result of better retention, deeper understanding, or both? The key to answering this question was to observe any potential changes in the pattern of performance across the four conditions from the initial answer phase to the reanswer phase of the final test. As expected, the initial answer phase yielded the same pattern of performance as in Experiment 3 with both variable practice and retrieval practice producing better transfer to the new application questions. However, this pattern largely disappeared on the second reanswer phase, as evidenced by the significant

interactions between phase and both independent variables. Numerically, the same study condition produced the lower performance than three of the other conditions, which were all approximately equal. However, none of the effects were significant when performance in Phase 2 was analyzed in isolation.

The change in the pattern of performance across the two phases of the final test suggests that variable practice and retrieval practice produce superior transfer mainly because they improve the retention of knowledge. This interpretation corresponds well to the literatures on both of these factors. As described in the introduction, numerous studies have found that variability increases retention when introduced during repeated retrieval or repeated study (e.g., McDaniel & Masson, 1985; McFarland et al., 1979). Likewise, there is a wealth of research that demonstrates the power of repeated retrieval practice to promote long-term retention relative to repeated study (e.g., Karpicke & Roediger, 2008; see Roediger & Butler, 2011). Nevertheless, the fact that the same study condition produced the lowest performance on the second phase of the final test leaves open the possibility that both variable practice and retrieval practice promote deeper understanding as well. Any contribution by these two factors to the development of deeper understanding would seem to be redundant (i.e., not additive), so a fruitful next step might be to further investigate this question with each factor independently. In addition, it is possible that the pattern of performance in the second phase of the final test was the result of performance reaching functional ceiling. If so, it is possible that a slight change in the methodology, such as a between-subjects manipulation of the two phases or introducing a delay between the phases, would reduce overall performance across the various conditions and thereby produce a clearer answer to this question of interest.

General Discussion

Taken as a whole, the findings of these four experiments provide solid evidence that both variable practice and retrieval practice yield superior transfer of knowledge to new examples. Each experiment showed that repeated retrieval practice with different application questions produced superior performance on new applications questions relative to repeated retrieval practice with the same application question. Indeed, this finding seems to be quite robust—the average effect size for the advantage of variable over same retrieval practice was .51, Experiment 1 = .49; Experiment 2 (label) = .61; Experiment 2 (no label) = .32; Experiment 3 = .61; Experiment 4 = .53. Interestingly, this benefit was obtained despite the fact that variability of practice hindered performance during initial learning. Experiments 3 and 4 also demonstrated that the beneficial effects of variable practice on subsequent transfer extend to the repeated study of complex materials (i.e., an encoding variability effect). The final two experiments also showed that repeated retrieval practice produced superior transfer relative to repeated study.

Why did introducing variability during repeated retrieval practice produce superior transfer of knowledge to new examples? One possibility is that variability of practice may enable learners to differentiate the core elements of a complex concept from the superficial elements that are unique to each instantiation of that concept. Although memory for superficial elements can play an important role in the transfer process (e.g., recognizing that prior learning is relevant; Holyoak & Koh, 1987), it is the core elements that are critical to the successful application of knowledge. This distinction between core and superficial elements is common to many other learning literatures, including analogical reasoning (Kimball & Holyoak, 2000), speech perception (Miller & Eimas, 1995), motor skills (Schmidt, 1975), and categorization (Nosofsky, Clark, & Shin, 1989). Indeed, much research in these literatures shows that variability during learning promotes transfer to novel situations (i.e., generalization; e.g., Barcroft & Sommers, 2005; Bassok & Holyoak, 1989; Dukes & Bevan, 1967; Gick & Holyoak, 1983; Logan et al., 1991; Shapiro & Schmidt, 1982; Shea & Kohl, 1990; Wahlheim et al., 2012).

Of course, the differentiation of core elements from superficial elements is not necessarily unique to retrieval practice. Indeed, we found benefits of variability for repeated study as well, suggesting that studying different examples facilitates this process as well. This finding is important given that previous research on encoding variability has yielded mixed results. Of course, almost all of the previous studies on encoding variability have focused on assessing retention of simple materials (e.g., homographs, word pairs) with standard memory tests, such as recognition, cued recall, and free recall (e.g., Dellarosa & Bourne, 1985; Postman & Knecht, 1983; Slamecka & Barlow, 1979). Based on our results and the evidence that variability during learning promotes generalization in other domains, future research on encoding variability should utilize materials of greater complexity and criterial tests that require transfer of knowledge to new contexts.

Our findings also have direct implications for theories that attempt to explain the mnemonic benefits of retrieval practice. The present research was not designed to test any particular theory because none of the existing theories specifically address the concepts of variability during repeated retrieval practice or transfer of knowledge. Nevertheless, as discussed in the introduction, several theoretical explanations for the testing effect include variability as a mechanism. Encoding variability (e.g., Bower, 1972; Glenberg, 1979; Martin, 1968) and retrieval effort (Bjork, 1975; Bjork & Bjork, 1992) accounts treat variability broadly, whereas the episodic context (Karpicke et al., 2014; Lehman, Smith, & Karpicke, 2014) and elaborative retrieval (Carpenter, 2009, 2011; Carpenter & DeLosh, 2006; Glover, 1989) accounts posit specific types of variability (see Introduction). Our findings suggest the need for a broad conceptualization of variability that incorporate many types of variability (e.g., perception, processing, internal context, external environment, etc.). An important goal within such a framework would be to specify the nature of the new elements that are integrated into representations of an event as a result of reprocessing during retrieval. Depending on the nature of retrieval task, variability could produce a change in what is activated in general knowledge, integration of new temporal elements from the present context, or a host of other possibilities.

Indeed, the idea that retrieval practice tasks can differ in terms of the processing that they induce (and hence the new elements that are integrated into representations in memory) is implicit in some testing effect theories. For example, the retrieval effort account assumes that retrieval effort varies across retrieval tasks (Bjork, 1975; Bjork & Bjork, 1992), which suggests that different tasks can promote qualitatively different types of processing (e.g., Whitten, 1978). The present research provides further support for this view of retrieval practice. If retrieval practice always induced

a single way of processing information, then a greater number of successful retrievals should always produce better subsequent performance. Using the data from all four experiments, we performed a conditional analysis to assess how performance during the initial learning affected transfer to the new application questions on the final test. Table 4 shows performance on the final test as a function of the number of correct responses during initial learning and practice type. When participants answered the same application question repeatedly during initial learning, final test performance improved as the number of correct responses during initial learning increased. A similar pattern of improvement was observed when participants answered three different application questions, but the relative improvement was substantially greater. Importantly, answering two different application questions correctly (while getting the third incorrect) yielded a similar level of transfer to answering the same application question correctly three times (.63 vs. .62).

Given this view about the multitude of different ways in which retrieval practice can induce people to process information, one broader conclusion is that the effects of variability are contextual, like many other phenomena in human memory and learning (Roediger, 2008). Introducing variability during repeated practice represents a "desirable difficulty" in learning in that it slows the acquisition of knowledge, but enhances subsequent use of that knowledge (Bjork, 1994; Schmidt & Bjork, 1992). However, the desirability of any difficulty depends on the context (see McDaniel & Butler, 2010), and accordingly variability during retrieval practice may not always be desirable if the new elements integrated into the representation of the event do not confer an advantage for the subsequent retrieval and use of that knowledge.

Conclusion

In closing, we would be remiss not to discuss the implications of our findings for educational practice. Experiments 3 and 4 demonstrate the potential of retrieval practice and variability as mechanisms for fostering learning that promotes subsequent transfer to new examples. Repeatedly retrieving and applying knowledge to different questions during initial learning improved subsequent performance on new application questions on the final test by the equivalent of about one letter grade relative to repeatedly studying the same example (Experiment 3 = .68 vs. .38, d = 1.34; Experiment 4 = .59 vs. 35, d = .91). We think that this comparison highlights the difference between what commonly occurs and what is possible in education. Students often repeatedly study the same examples, such as when they use the common learning strategy of reading a textbook, highlighting it, and the rereading the text that they highlighted (e.g., Karpicke, Butler, & Roediger, 2009). Our findings suggest that engaging this type of learning activity has limited benefits—repeatedly retrieving and applying knowledge to different examples represents a much better way to acquire knowledge that will transfer to a variety of new contexts. Of course, this advice depends upon the educator having access to additional questions and examples; simply rewording existing materials may not introduce enough variability (see Butler, 2010).

That said, when considering the generalizability of these findings to authentic educational contexts, it is important to consider some the potential constraints on these effects and the follow-up questions that need to be investigated. For example, one potential factor that bears further scrutiny is the provision of feedback after each retrieval attempt. Although feedback is routinely provided in the classroom and other educational contexts, there may be circumstances under which giving learners feedback is undesirable or impossible. Prior research suggests that the mnemonic benefits of retrieval practice are robust even in the absence of feedback (Roediger & Butler, 2011), but only when learners succeed in retrieving the information from memory (e.g., Kang, McDermott, & Roediger, 2007). Thus, a more important constraint is likely to be the level of performance during retrieval practice, especially when variability is introduced. If learners fail to successfully retrieve and apply their knowledge, then they will not benefit from such practice without feedback. Relatedly, another factor that requires further research is the lag between practice opportunities. Greater lags may reduce the magnitude of the benefit of variable practice relative to same practice because of the increased variability in other contextual elements when same practice is spaced over time and/or the failure to connect variable practice opportunities when longer lags intervene (see discussion of Experiment 2). Finally, yet another factor that needs to be considered is the time interval over which the effects observed in the present research can be expected to persist. The benefits of retrieval practice often increase over time (Roediger & Butler, 2011; Rowland, 2014), but there is presumably a point where forgetting will eliminate the effects produced in the present research. The solution to improving the durability of these effects may be found in modifying the way in which retrieval practice is structured—that is, increasing the number and spacing of the opportunities to practice retrieval (Rawson & Dunlosky, 2011). Despite the fact that many questions remain to be answered, the findings produced by the present research suggest that repeatedly retrieving and applying knowledge to different examples is a highly beneficial learning activity.

Table 4
Proportion of Correct Responses to New Application Questions on the Final Test in Experiments 1–4 as a Function of the Number of Correct Responses During Initial Learning and Practice Type

Practice type	Number of correct responses during initial learning				Grand mean (total)
	0	1	2	3	
Same	.28 (43)	.54 (52)	.44 (271)	.62 (414)	.53 (780)
Variable	.50 (88)	.52 (184)	.63 (275)	.85 (246)	.66 (793)
Grand mean (total)	.43 (131)	.52 (236)	.54 (546)	.70 (660)	.60 (1,573)

Note. Data is from the retrieval practice conditions only; data from the study conditions in Experiments 3 and 4 were excluded. In parentheses is the number of items (i.e. concepts nested within participants) that contributed to the mean. Across the four experiments, data related to 25 items were excluded from the analysis because a participant failed to enter a response for one or more questions during initial learning or on the final test.

References

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition, 27,* 387–414. http://dx.doi.org/10.1017/S0272263105050175

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. http://dx.doi.org/10.1037/0033-2909.128.4.612

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. http://dx.doi.org/10.1037/0278-7393.15.1.153

Benjamin, A. S., & Ross, B. H. (2010). The causes and consequences of reminding. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: A Festschrift in honor of Robert A. Bjork* (pp. 71–87). New York, NY: Psychology Press.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1994). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

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.

Blunt, J. R., & Karpicke, J. D. (2014). Learning with retrieval-based concept mapping. *Journal of Educational Psychology, 106,* 849–858. http://dx.doi.org/10.1037/a0035934

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.

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 1118–1133. http://dx.doi.org/10.1037/a0019902

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology, 105,* 290–298. http://dx.doi.org/10.1037/a0031026

Butler, A. C., & Roediger, H. L., III. (2007). Testing improves long-term retention in a simulated classroom setting. *The European Journal of Cognitive Psychology, 19,* 514–527. http://dx.doi.org/10.1080/09541440701326097

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35,* 1563–1569. http://dx.doi.org/10.1037/a0017021

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37,* 1547–1552. http://dx.doi.org/10.1037/a0024140

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current Directions in Psychological Science, 21,* 279–283. http://dx.doi.org/10.1177/0963721412452728

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. http://dx.doi.org/10.3758/BF03193405

Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory & Cognition, 20,* 633–642. http://dx.doi.org/10.3758/BF03202713

Dellarosa, D., & Bourne, L. E., Jr. (1985). Surface form and the spacing effect. *Memory & Cognition, 13,* 529–537. http://dx.doi.org/10.3758/BF03198324

Dukes, W. F., & Bevan, W. (1967). Stimulus variation and repetition in the acquisition of naming responses. *Journal of Experimental Psychology, 74,* 178–181. http://dx.doi.org/10.1037/h0024575

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques promising directions from cognitive and educational psychology. *Psychological Science in the Public Interest, 14,* 4–58. http://dx.doi.org/10.1177/1529100612453266

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review, 62,* 369–377. http://dx.doi.org/10.1037/h0046888

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1–38. http://dx.doi.org/10.1016/0010-0285(83)90002-6

Glass, A. L. (2009). The effect of distributed questioning with varied examples on exam performance on inference questions. *Educational Psychology, 29,* 831–848. http://dx.doi.org/10.1080/01443410903310674

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition, 7,* 95–112. http://dx.doi.org/10.3758/BF03197590

Glover, J. A. (1989). The "testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81,* 392–399. http://dx.doi.org/10.1037/0022-0663.81.3.392

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. http://dx.doi.org/10.3758/PBR.15.3.662

Hinze, S. R., & Wiley, J. (2011). Testing the limits of testing effects using completion tests. *Memory, 19,* 290–304. http://dx.doi.org/10.1080/09658211.2011.560121

Holyoak, K. J., & Koh, K. (1987). Surface and structural similarity in analogical transfer. *Memory & Cognition, 15,* 332–340. http://dx.doi.org/10.3758/BF03197035

James, W. (1962). *Talks to teachers on psychology: And to students on some of life's ideals*. New York, NY: Dover Publications, Inc. (Original work published 1899)

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology, 101,* 621–629. http://dx.doi.org/10.1037/a0015183

Kang, S. H. K., McDaniel, M. A., & Pashler, H. (2011). Effects of testing on learning of functions. *Psychonomic Bulletin & Review, 18,* 998–1005. http://dx.doi.org/10.3758/s13423-011-0113-x

Kang, S. H., McDermott, K. B., & Roediger, H. L., III. (2007). Test format and corrective feedback modify the effect of testing on long-term retention. *The European Journal of Cognitive Psychology, 19,* 528–558. http://dx.doi.org/10.1080/09541440601056620

Karpicke, J. D., Butler, A. C., & Roediger, H. L., III. (2009). Metacognitive strategies in student learning: Do students practise retrieval when they study on their own? *Memory, 17,* 471–479. http://dx.doi.org/10.1080/09658210802647009

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. *Psychology of Learning and Motivation, 61,* 237–284. http://dx.doi.org/10.1016/B978-0-12-800283-4.00007-1

Karpicke, J. D., & Roediger, H. L., III. (2008). The critical importance of retrieval for learning. *Science, 319,* 966–968. http://dx.doi.org/10.1126/science.1152408

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.

Lehman, M., Smith, M. A., & Karpicke, J. D. (2014). Toward an episodic context account of retrieval-based learning: Dissociating retrieval practice and elaboration. *Journal of Experimental Psychology: Learning,*

Memory, and Cognition, 40, 1787–1794. http://dx.doi.org/10.1037/xlm0000012

Logan, J. S., Lively, S. E., & Pisoni, D. B. (1991). Training Japanese listeners to identify English /r/ and /l/: A first report. *The Journal of the Acoustical Society of America, 89,* 874–886. http://dx.doi.org/10.1121/1.1894649

Madigan, S. A. (1969). Intraserial repetition and coding processes in free recall. *Journal of Verbal Learning and Verbal Behavior, 8,* 828–835. http://dx.doi.org/10.1016/S0022-5371(69)80050-2

Martin, E. (1968). Stimulus meaningfulness and paired-associate transfer: An encoding variability hypothesis. *Psychological Review, 75,* 421–441. http://dx.doi.org/10.1037/h0026301

Maskarinec, A. S., & Thompson, C. P. (1976). The within-list distributed practice effect: Tests of the varied context and varied encoding hypotheses. *Memory & Cognition, 4,* 741–746. http://dx.doi.org/10.3758/BF03213242

McDaniel, M. A., & Butler, A. C. (2010). A contextual framework for understanding when difficulties are desirable. In A. S. Benjamin (Ed.), *Successful remembering and successful forgetting: Essays in honor of Robert A. Bjork* (pp. 175–199). New York, NY: Psychology Press.

McDaniel, M. A., Howard, D. C., & Einstein, G. O. (2009). The read-recite-review study strategy: Effective and portable. *Psychological Science, 20,* 516–522. http://dx.doi.org/10.1111/j.1467-9280.2009.02325.x

McDaniel, M. A., & Masson, M. E. J. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 11,* 371–385. http://dx.doi.org/10.1037/0278-7393.11.2.371

McFarland, C. E., Jr., Rhodes, D. D., & Frey, T. J. (1979). Semantic-feature variability and the spacing effect. *Journal of Verbal Learning and Verbal Behavior, 18,* 163–172. http://dx.doi.org/10.1016/S0022-5371(79)90100-2

Miller, J. L., & Eimas, P. D. (1995). Speech perception: From signal to word. *Annual Review of Psychology, 46,* 467–492. http://dx.doi.org/10.1146/annurev.ps.46.020195.002343

Nosofsky, R. M., Clark, S. E., & Shin, H. J. (1989). Rules and exemplars in categorization, identification, and recognition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 15,* 282–304. http://dx.doi.org/10.1037/0278-7393.15.2.282

Postman, L., & Knecht, K. (1983). Encoding variability and retention. *Journal of Verbal Learning and Verbal Behavior, 22,* 133–152. http://dx.doi.org/10.1016/S0022-5371(83)90101-9

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. http://dx.doi.org/10.1016/j.jml.2009.01.004

Rawson, K. A., & Dunlosky, J. (2011). Optimizing schedules of retrieval practice for durable and efficient learning: How much is enough? *Journal of Experimental Psychology: General, 140,* 283–302. http://dx.doi.org/10.1037/a0023956

Roediger, H. L., III. (2008). Relativity of remembering: Why the laws of memory vanished. *Annual Review of Psychology, 59,* 225–254. http://dx.doi.org/10.1146/annurev.psych.57.102904.190139

Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences, 15,* 20–27. http://dx.doi.org/10.1016/j.tics.2010.09.003

Roediger, H. L., III, & Karpicke, J. D. (2006). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science, 1,* 181–210. http://dx.doi.org/10.1111/j.1745-6916.2006.00012.x

Rohrer, D., Taylor, K., & Sholar, B. (2010). Tests enhance the transfer of learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36,* 233–239. http://dx.doi.org/10.1037/a0017678

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140,* 1432–1463. http://dx.doi.org/10.1037/a0037559

Schmidt, R. A. (1975). A schema theory of discrete motor skill learning. *Psychological Review, 82,* 225–260. http://dx.doi.org/10.1037/h0076770

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3,* 207–217. http://dx.doi.org/10.1111/j.1467-9280.1992.tb00029.x

Shapiro, D. C., & Schmidt, R. A. (1982). The schema theory: Recent evidence and developmental implications. In J. A. S. Keiso & J. E. Clark (Eds.), *The development of movement control and co-ordination* (pp. 113–150). New York, NY: Wiley.

Shea, C. H., & Kohl, R. M. (1990). Specificity and variability of practice. *Research Quarterly for Exercise and Sport, 61,* 169–177. http://dx.doi.org/10.1080/02701367.1990.10608671

Slamecka, N. J., & Barlow, W. (1979). The role of semantic and surface features in word repetition effects. *Journal of Verbal Learning and Verbal Behavior, 18,* 617–627. http://dx.doi.org/10.1016/S0022-5371(79)90344-X

Smith, S. M., & Handy, J. D. (2014). Effects of varied and constant environmental contexts on acquisition and retention. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 40,* 1582–1593. http://dx.doi.org/10.1037/xlm0000019

Smith, S. M., & Handy, J. D. (2016). The crutch of context-dependency: Effects of contextual support and constancy on acquisition and retention. *Memory, 24,* 1134–1141. http://dx.doi.org/10.1080/09658211.2015.1071852

Wahlheim, C. N., Finn, B., & Jacoby, L. L. (2012). Metacognitive judgments of repetition and variability effects in natural concept learning: Evidence for variability neglect. *Memory & Cognition, 40,* 703–716. http://dx.doi.org/10.3758/s13421-011-0180-2

Wheeler, M. A., & Roediger, H. L., III. (1992). Disparate effects of repeated testing: Reconciling Ballard's (1913) and Bartlett's (1932) results. *Psychological Science, 3,* 240–245. http://dx.doi.org/10.1111/j.1467-9280.1992.tb00036.x

Whitten, W. B. (1978). Initial-retrieval "depth" and the negative recency effect. *Memory & Cognition, 6,* 590–598. http://dx.doi.org/10.3758/BF03198248

Appendix

Sample Materials Used in Experiments 1-4

Concept Label

"Earth's structure"

Concept Description

Traditionally, the Earth is divided into three major layers: the core, the mantle, and the crust. These layers vary in density—the core is the densest, the mantle is slightly less dense, and the crust is the least dense. As a result of the varying densities, the upper layers "float" on the lower layers because they are less dense.

Application Questions Used for Retrieval Practice

Question 1: In an article about the Earth's structure, a major U.S. newspaper reports the following densities for its three major layers: the core (2.2 g/cm^3), the mantle (4.4 g/cm^3), and the crust (11.5 g/cm^3). What is wrong with this characterization of the Earth's structure?

Answer 1: The article reports that the core is the least dense and the crust is the densest. However, the core should be the densest, the mantle should be slightly less dense, and the crust should be the least dense. Earth's structure results from the upper layers "floating" on the lower layers because they are less dense.

Question 2: In 1692, Edmund Halley theorized that the Earth had a hollow center. However, his theory was considered logically impossible once the average density of the Earth (5.5 g/cm^3) was calculated and the density of rock near the surface (2.2 g/cm^3) was measured. Why did these findings make the "hollow earth" theory logically impossible?

Answer 2: If the rock near the surface (2.2 g/cm^3) is less dense than the average density of the Earth (5.5 g/cm^3), then the center must be denser than the average density of the Earth and thus cannot be hollow. Indeed, the Earth's core is the densest layer, and the upper layers "float" on the lower layers because of a difference in density.

Question 3: One theory suggests that planets form after the collapse of a nebula (an interstellar cloud of dust). The dust particles accumulate mass through gravitational attraction to form ever-larger bodies, and these concentrations differentiate by density to form the interior of a planet. How does the Earth's structure provide support for this theory?

Answer 3: The Earth's structure provides support for this theory because it is consistent with the planetary structure that would be produced through such a process of formation. Earth is composed of three major layers that vary in density with less dense upper layers "floating" on more dense lower layers.

Question 4: A new planet is discovered that is composed of following elements: silica dust (2.2 g/cm^3), carbon dioxide (.0018 g/cm^3), hermatite (4.5 g/cm^3), iron and nickel mixture (7.6 g/cm^3), water (.98 g/cm^3), and amphibolite (2.9 g/cm^3). What does your knowledge of the Earth's structure tell us about the structure of this new planet?

Answer 4: Based on Earth's structure, the densest materials (iron and nickel) are probably at the core and then the other elements are layered on top in decreasing densities: hematite, amphibolite, silica dust, water, and carbon dioxide.

Examples Used for Study

Study 1: In an article about the Earth's structure, a major U.S. newspaper reports the following densities for its three major layers: the core (2.2 g/cm^3), the mantle (4.4 g/cm^3), and the crust (11.5 g/cm^3). This characterization of the Earth's structure is wrong because it states that the core is the least dense and the crust is the densest. However, the core should be the densest, the mantle should be slightly less dense, and the crust should be the least dense. Earth's structure results from the upper layers "floating" on the lower layers because they are less dense.

Study 2: In 1692, Edmund Halley theorized that the Earth had a hollow center. However, his theory was considered logically impossible once the average density of the Earth (5.5 g/cm^3) was calculated and the density of rock near the surface (2.2 g/cm^3) was measured. These findings made the "hollow earth" theory logically impossible because if the rock near the surface (2.2 g/cm^3) is less dense than the average density of the Earth (5.5 g/cm^3), then the center must be denser than the average density of the Earth and thus cannot be hollow. Indeed, the Earth's core is the densest layer, and the upper layers "float" on the lower layers because of a difference in density.

(*Appendix continues*)

Study 3: One theory suggests that planets form after the collapse of a nebula (an interstellar cloud of dust). The dust particles accumulate mass through gravitational attraction to form ever-larger bodies, and these concentrations differentiate by density to form the interior of a planet. The Earth's structure provides support for this theory because it is consistent with the planetary structure that would be produced through such a process of formation. Earth is composed of three major layers that vary in density with less dense upper layers "floating" on more dense lower layers.

Study 4: A new planet is discovered that is composed of following elements: silica dust (2.2 g/cm^3), carbon dioxide (.0018 g/cm^3), hematite (4.5 g/cm^3), iron and nickel mixture (7.6 g/cm^3), water (.98 g/cm^3), and amphibolite (2.9 g/cm^3). Our knowledge of the Earth's structure tells us about the structure of this new planet. Based on Earth's structure, the densest materials (iron and nickel) are probably at the core and then the other elements are layered on top in decreasing densities: hematite, amphibolite, silica dust, water, and carbon dioxide.

Note. Each question and answer pair (e.g., Question 1 and Answer 1) corresponds to the study example with the same number (e.g., Study 1).

Received August 31, 2016
Revision received June 29, 2017
Accepted July 13, 2017 ■