January 2009

Using repeated testing and variable encoding to promote transfer of learning

Andrew Butler
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Psychology

Dissertation Examination Committee:
Henry L. Roediger, III (Chair)
David A. Balota
Mark A. McDaniel
Jeffrey M. Zacks
R. Keith Sawyer
James V. Wertsch

USING REPEATED TESTING AND VARIABLE ENCODING TO PROMOTE TRANSFER OF LEARNING

by

Andrew Cox Butler

A dissertation presented to the
Graduate School of Arts and Sciences
of Washington University in
partial fulfillment of the
requirements for the degree
of Doctor of Philosophy

August 2009

Saint Louis, Missouri

Abstract

Within the literature on transfer of learning, relatively few studies have investigated how the conditions of initial learning can be arranged to increase the likelihood of successful transfer. The present research investigated whether test-enhanced learning can be used to promote transfer. More specifically, four experiments examined how repeated testing and repeated studying affected retention and transfer of facts and concepts. Subjects studied prose passages and then either repeatedly re-studied or took tests on the material. One week later, they took a final test that either had the same questions (Experiment 1), new inferential questions within the same knowledge domain (Experiments 2 and 3), or new inferential questions from different knowledge domains (Experiment 4). 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 a specific response, but rather extend to the retrieval of knowledge in a variety of contexts.

Acknowledgements

First and foremost, I would like to thank Roddy Roediger for all the support and guidance that he has provided me over the past six years. One could not ask for more in a doctoral advisor – he is a terrific role model, a wealth of knowledge about psychological science, and a staunch advocate for his students. I have learned many things from Roddy, but the most valuable lesson that he has taught me is about the value of relationships. He has been hugely successful in his own career, but his active interest in fostering the careers of others is what sets him apart. He makes everyone else around him better.

I would also like to thank the other members of my dissertation committee: Dave Balota, Mark McDaniel, Jeff Zacks, Keith Sawyer, and Jim Wertsch. Their feedback helped me to clarify my ideas and improve the quality of my dissertation research.

I must also thank Anne Butler, my wonderful wife, and the rest of my immediate family: my sister Louisa Butler, and my parents Pat Cox and Graham Colditz. Anne has been extremely supportive to me, and I am very grateful for the patience and understanding that she exhibited during this process.

Another big thank you goes to all the members of the extended Memory Lab group – I appreciate all the support and advice. In particular, I would like to thank Jane McConnell, who helped at so many different points. Ileana Culcea deserves much praise for collecting and scoring a sizeable portion of the data. I also thank Jake Sanches for scoring some of the data.

The research reported here was supported by a James S. McDonnell Foundation 21st Century Science Initiative in Bridging Brain, Mind and Behavior/Collaborative Award (PI: Henry L. Roediger, III).

Table of Contents

List of Tables

List of Figures

List of Appendices

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. To be successful at predicting learning, a theory must be able to account for how prior learning influences future learning. Transfer also has enormous practical implications for education in both schools and the workplace. Formal education is predicated on the assumption that what is learned in the classroom will transfer to situations outside the classroom. Similarly, vocational training is conducted to provide workers with a set of skills and knowledge that will prepare them to deal with various scenarios, both anticipated and unanticipated. The belief that transfer occurs quite frequently is critical to justifying the substantial amounts of time and money devoted to formal education and vocational training each year.

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. Although the existence of transfer is beyond question, there is little consensus among psychologists about the extent to which it occurs, let alone the underlying cognitive mechanisms that produce it. Over a century of research on the topic has generated hundreds of studies (if not more), but the literature remains a muddled mass of conflicting results. The state of the literature is so confusing

that some reviewers have concluded that transfer is ubiquitous (e.g., Halpern, 1998; Brown, 1989), while others have stated that it is extremely rare (e.g., Detterman, 1993). In some sense, the question of whether learning generalizes to new contexts is as open today as it was when the first empirical studies of transfer were conducted at the turn of the 20th century (e.g., Thorndike & Woodworth, 1901a).

Two major factors impede progress in understanding transfer. First, the field suffers from the lack of a common structure or framework within which to interpret the results of transfer studies and to identify new areas of investigation. Transfer research does not have a dominant paradigm (the one possible exception is the A-B, A-D paired associate learning task, which was once the primary method used by transfer researchers but has since fallen out of favor); rather, most transfer studies share a common definition of the phenomenon, but sharply diverge in the methods used to study it. The diversity of methods used in transfer research increases the difficulty in comparing findings across studies. However, a recent review of the transfer literature by Barnett and Ceci (2002) shows how an organizing framework can be used to resolve many of the seemingly discrepant results that exist. They proposed a taxonomy of transfer that specifies two components, content (i.e. "what is transferred") and context (i.e. "when and where it is transferred to and from"), each of which contains several dimensions along which experimental investigations can vary. As will be discussed in more depth below, their analysis indicates that significant progress has been made towards delineating the conditions under which transfer occurs, suggesting that some of the confusion in the literature may be more apparent than real.

A second major 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. To the extent that initial learning produces better retention of information and numerous retrieval routes to access that information, there should be greater potential for transfer to occur.

The present research investigated how the conditions of initial learning affect transfer of learning, using the Barnett and Ceci (2002) taxonomy to define transfer in terms of content and context. 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; e.g., Butler & Roediger, 2007; Larsen, Butler, & Roediger, in press), 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 re-

studying that material for an equivalent amount of time, a result commonly called *the testing effect* (for review see Roediger & Karpicke, 2006a). One goal of the present research was to examine whether repeated testing promotes superior transfer relative to repeated studying. Another goal was to explore whether repeated testing using re-phrased 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 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 a brief overview of the history of transfer research, describe the Barnett and Ceci (2002) taxonomy, explain the scope and rationale for the project, and review some of evidence that supports the efficacy of test-enhanced learning.

History of Transfer Research

During the 19th century, Western education was dominated by the doctrine of formal discipline, a pedagogical theory based on the age-old idea that the mind is divided into general faculties (e.g., reasoning, memory, attention, etc.), which could be strengthened through exercise like muscles. This idea, sometimes referred to as *faculty psychology*, was supported by the writings of philosophers such as Saint Thomas Aquinas (1225–1274) and John Locke (1632-1704), and later by the "scientific" system of phrenology put forth by Franz Joseph Gall (1758-1828). Influenced by contemporary views of the mind, educators believed that "faculties, like muscles, grow strong by use" (Roark, 1895; as cited in Hall, 1971) and adopted a curriculum that provided students

with mental exercise. Training that strengthened a particular faculty was thought to benefit performance on any task that relied on that faculty. For example, students memorized texts and poems because training in memorization, regardless of the content, was thought to increase general memory capacity (e.g., Winch, 1908). Thus, the doctrine of formal discipline represented an extreme view of transfer in that transfer was assumed to occur constantly and irrespective of content or context.

Empirical research on transfer of learning began as an effort to test the doctrine of formal discipline. Although faculty psychology had been discredited by scientific research, formal discipline remained popular among educators. Armed with new ideas about how the human mind works, some early psychologists challenged the current zeitgeist in education by designing experiments to examine the extent to which transfer of learning occurred. William James reported one of the first transfer experiments in his classic textbook, *Principles of Psychology* (1890), in which he investigated whether "a certain amount of daily training in learning poetry by heart will shorten the time it takes to learn an entirely different kind of poetry" (p. 666). The results obtained from a handful of subjects, including James himself, showed no evidence that training in memorization improves general "physiological retentiveness." If anything, there was some indication of people adapting certain strategies for memorization (i.e. what is now called "learning to learn"), but clearly no support for the type of transfer predicted by the doctrine of formal discipline.

Although James' results provided evidence against the doctrine of formal discipline, the most powerful refutation came ten years later in a series of articles published by two of his former students, Edward Thorndike and Robert Woodworth.

Thorndike and Woodworth (1901a; 1901b; 1901c) described experiments in which they examined transfer in subjects' ability to cross out words with certain letter combinations, memorize texts (a replication of James' earlier work), and estimate the length, area, or weight of objects. The general method was simple: they trained subjects on a task (e.g., estimating the area of paper squares), then changed some aspect of the task (e.g., from squares to triangles) and examined performance on the new task. Finding little transfer under conditions that they argued should be favorable to observing it, Thorndike and Woodworth pessimistically concluded that "improvement in any single mental function rarely brings about equal improvement in any other function, no matter how similar, for the working of every mental function-group is conditioned by the nature of the data in each particular case" (p. 250).

The publication of Thorndike and Woodworth's (1901a; 1901b; 1901c) articles had a galvanizing effect on the field as proponents and opponents of formal discipline rushed to find evidence to support their position. Some researchers succeeded in demonstrating substantial positive transfer (e.g., Leuba & Hyde, 1905; Starch, 1911; Webb, 1917). For example, Judd (1908) had two groups of boys practice throwing darts at a target submerged in 12 inches of water, a task on which the two groups performed equally well. Boys in one of the groups knew the principle of refraction, while the boys in the other did not. On a subsequent test with a target now submerged in only 4 inches of water, the group that knew about refraction learned to hit the target quicker than the control group. Nevertheless, numerous other researchers reported studies that showed only modest positive transfer (e.g., Reed, 1917; Ruediger, 1908; Sleight, 1911) or negative transfer (e.g., Kline, 1914; Martin, 1915). Despite the claims from both sides,

this initial flurry of research activity did little to resolve the debate. The extreme version of the doctrine of formal discipline was discredited, but the argument about the extent to which transfer occurs raged on.

The disparate findings reported by researchers, not to mention the variety of tasks and materials used in their studies, made it difficult to interpret the nascent transfer literature. Thorndike, who had become a giant in the fields of psychology, education, and lexicography, proposed the first widely accepted theory of transfer. His theory of transfer by identical elements held that "one function alters any other [function] only in so far as the two functions have as factors identical elements" (Thorndike, 1913, p. 358). As for the exact meaning of the term "elements", Thorndike referred to them as "mental processes which have the same cell action in the brain as their physical correlate" (p. 359). Of course, he had no way of directly observing such processes and thus the concept of elements had to be operationalized in terms of behavior. Although Thorndike's theory can be interpreted in different ways (see McGeoch, 1942, ch. 10), he essentially argued that the degree of transfer from training to test depends upon the degree of similarity between the two contexts. This idea was instrumental in advancing subsequent research because it prompted researchers to systematically investigate how the similarity between contexts affects transfer.

The need for greater precision in manipulating the similarity between the initial learning and transfer contexts led researchers to adopt the paired-associate learning paradigm during the early 1930's (e.g., Gulliksen, 1932; McKinney, 1933; Yum, 1931). First introduced by Mary Calkins (1896), this method involved learning a list of stimulus-response pairs that generally consisted of words, nonsense syllables, or numbers (or a

mixture of these different types of materials). Learning was a multi-trial process in which the subject studied a list of stimulus-response pairs, then tried to recall the response item when presented with the stimulus, and finally received another presentation of the both stimulus and response; this process continued until a criterion was reached (e.g., one successful recall of the entire list). The paired-associate learning paradigm enabled transfer researchers to manipulate the similarity between two lists of paired-associates and then measure how learning one of the lists influenced learning of the second list.

An experiment conducted by Bruce (1933) provides a prime example of how the paired-associate learning paradigm was used in transfer research and the general pattern of results that was obtained. Using pairs of three-letter nonsense syllables, he systematically manipulated the similarity of the stimulus and response terms in both lists (i.e., S_1, R_1, S_2, R_2) as well as the number of presentations of the first list during initial learning (0, 2, 6, or 12). Table 1 illustrates the materials constructed for the nine conditions in which similarity was varied. The dependent variable was the rate of learning of the second list of paired-associates as measured by the number of trials needed to reach one perfect recall of the entire list. Three key results emerged from the experiment. First, learning to make an old response to a new stimulus (Condition V) resulted in substantial positive transfer. Second, learning to make a new response to an old stimulus (Condition I) led to marked negative transfer. Third, as the number of presentations increased (i.e. degree of initial learning), positive transfer increased and negative transfer decreased across all conditions. All comparisons were made relative to the control condition (Condition IX).

Table 1

Sample materials from Bruce (1933) in which similarity of the stimulus and response terms in both initial learning (List 1) and subsequent learning (List 2) were systematically varied in a pair-associate learning experiment.

Condition	Design	Initial Learning		Subsequent Learning		Relation of initial to subsequent material
		Stimulus (S_1)	Response (R_1)	Stimulus (S_2)	Response (R_2)	
I	A-B, A-D	req	kiv	req	zam	S_1S_2 identical, R_1R_2 unrelated
II	A-B, A-D	bij	bic	bij	tab	S_1S_2 identical, S_1R_1 similar
III	A-B, A-D	mir	ped	mir	miy	S_1S_2 identical, S_2R_2 similar
IV	A-B, A-D	tec	zox	tec	zop	S_1S_2 identical, R_1R_2 similar
V	A-B, C-B	lan	qip	fis	qip	R_1R_2 identical, S_1S_2 unrelated
VI	A-B, C-B	soj	soy	nel	soy	R_1R_2 identical, S_1R_1 similar
VII	A-B, C-B	zaf	qer	qec	qer	R_1R_2 identical, S_2R_2 similar
VIII	A-B, C-B	bes	yor	bef	yor	R_1R_2 identical, S_1S_2 similar
IX	A-B, C-D	xal	pom	cam	lup	All terms different

Reviewing the literature nine years later, McGeoch (1942) cited the results of Bruce's (1933) experiment as the typical finding in transfer research. He concluded that the "…connection of a new stimulus with a response already associated during training with some other stimulus yields positive transfer…" and that "…learning to make a new response to an old stimulus yields negative transfer" (p. 416). However, he also pointed out that manipulating similarity did not always yield straightforward effects. For example, increasing similarity between the two response terms (R_1 and R_2), while holding the stimulus terms constant, increased negative transfer. However, once the two response terms reached the point of maximal similarity or "identity" (i.e., where R_1 and R_2 are identical), negative transfer was eliminated. Despite the greater degree of control offered by the paired-associate learning paradigm, researchers still lacked a basic understanding of how similarity influenced transfer.

Osgood (1949) later proposed a resolution to the paradox posed by similarity. Through a careful analysis of the literatures on transfer and retroactive interference, he derived three empirical laws and used them to generate a three-dimensional model. Osgood's "transfer and retroaction surface" (see Figure 1) had a vertical dimension that indicated the direction (i.e. positive or negative) and degree of transfer (or retroaction). The horizontal dimensions indicated the degree of similarity between stimuli along the width of the surface and responses along the length. Similarity was conceptualized as a continuous variable ranging from "antagonistic" to "identity." Osgood's surface was a major success in that it could account for all the existing data in the literature and resolved the similarity paradox by defining "identity" as the limiting case of maximal similarity.

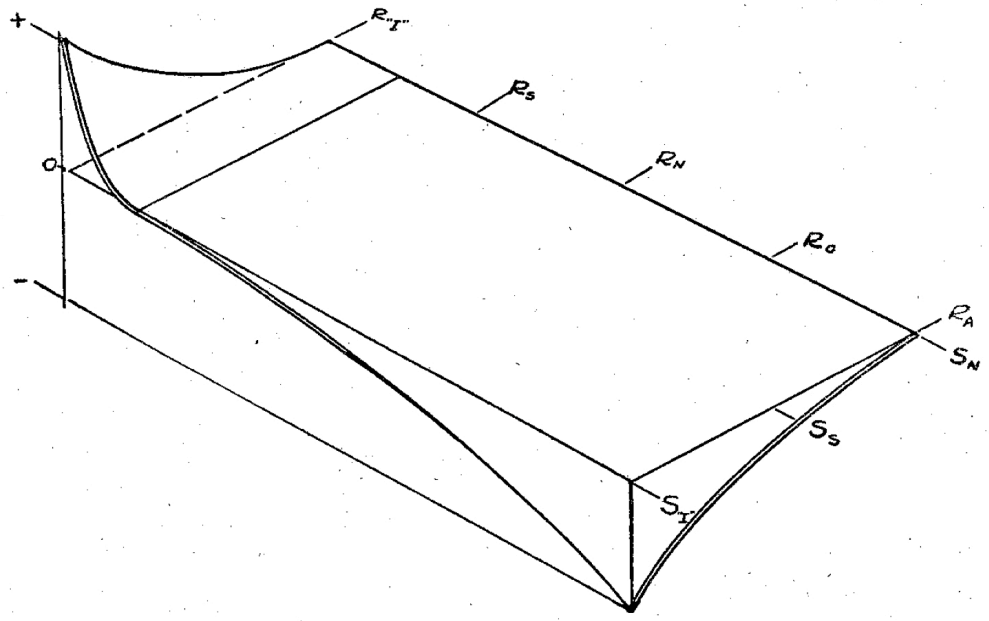

Figure 1. The Transfer and Retroaction Surface, from Osgood (1949). The vertical dimension represents the direction (i.e. positive and negative) and degree of transfer (or retroaction). The horizontal dimensions indicate the degree of similarity between stimuli along the width of the surface and responses along the length. Similarity ranges along a continuum with the following points designated: I = Identity, S = Similarity, N = Neutral, O = Opposed, and A = Antagonistic.

Unfortunately, a number of experimenters who attempted to empirically test Osgood's (1949) surface failed to find some of the predicted results (e.g., Bugelski & Cadwallader, 1956; Wimer, 1964; but see Dallet, 1962). One problem was that Osgood (1949) treated similarity as a continuous variable, yet it is unclear whether the categories he proposed (identity, similarity, neutral, opposed, and antagonistic) truly represent a continuum. However, a bigger problem is that similarity can be varied along many different dimensions. For example, similarity can be defined in terms of the physical, semantic, or associative (i.e. pre-existing or experimentally induced) properties of the material. As Hall (1971) later pointed out, this issue "leads to the general question of whether or not consistent transfer findings can be obtained when different varieties of similarity are manipulated" (p. 378). Indeed, there is some evidence to suggest that the pattern of transfer effects observed differs as a function of the way in which similarity is operationalized in the experimental material (Ryan, 1960).

As the verbal learning tradition wound down during the 1960's, transfer research split off in different directions, forming relatively isolated sub-areas. For example, researchers have investigated the role of transfer in analogical reasoning, abstract concept learning, the teaching of intelligence and higher order skills, and motor learning (to name but a few). Despite the considerable diversity of areas in which transfer research continues, two overarching themes remain present. First, researchers have concentrated on investigating how the similarity between the contexts of initial learning and subsequent test influences the extent to which transfer occurs, largely ignoring how the conditions of initial learning might be arranged to promote better transfer. Second, transfer research has lacked an organizing framework within which to specify the

relevant dimensions along which similarity can vary, especially as materials and contexts get increasingly complex. I turn now to describing a recent attempt to provide such a framework.

A Framework for Transfer Research

The taxonomy proposed by Barnett and Ceci (2002) represents a major step towards understanding the phenomenon of transfer. Surveying the literature, they identified three critical issues in transfer research. First, the terminology used to describe the type and degree of dissimilarity between initial learning and transfer contexts varies greatly across studies. Most studies start with a basic definition of transfer similar to the example given above from McGeoch (1942), but then use modifiers to indicate the type or degree of transfer involved. For example, the term *far transfer* is used to describe a situation in which there is a large degree of dissimilarity between initial learning and transfer contexts (i.e. relative to near transfer). However, there is no agreed upon degree of dissimilarity that constitutes far transfer, and thus “far” transfer in one experiment might be “near” transfer in another. The heterogeneous use of terms is problematic because it hinders the comparison of findings across studies.

Second, the experimental tasks used in transfer studies differ in terms of the demands made on memory. For example, some tasks involve the execution of a well-learned behavior (e.g., typing a series of numbers on keyboard), whereas other tasks require more effortful retrieval of information from memory (e.g., recalling the solution to a practice problem). Third, the to-be-learned skill or knowledge can be specific (e.g., a fact) or more general (e.g., the mnemonic device "I before e, except after c" that helps

people remember how to spell certain words). This variable is often confounded with the degree of transfer involved in the experimental tasks, but it need not be.

Barnett and Ceci (2002) used these three issues to guide the development of a taxonomy that includes two factors: content (i.e. "what is transferred") and context (i.e. "when and where it is transferred to and from"). The content factor is further separated into three dimensions of learned skill, performance change, and memory demands (see upper box of Figure 2). The first dimension represents the specificity-generality of the learned skill, which is conceptualized as a continuum. At one end of the continuum are facts and routinized procedures that are relatively specific and superficial in terms of the learning required, and can be thought of as solutions to closed-problem spaces (i.e., where there is one correct solution). An example of a specific piece of knowledge is the algorithm used to solve the missionaries and cannibals problem (Reed, Ernst, & Banerji, 1974). At the other end of the continuum are general principles that require a deeper structural or causal understanding and constitute heuristics that can be applied to fuzzy-problem spaces (i.e., where the correct solution is ill-defined and a pragmatic approach is needed). For example, teaching students a study skill, such as "check your work" (Williams et al., 1996), represents a type of general heuristic that can be applied to many different situations.

A Content: What transferred			
Learned skill	Procedure	Representation	Principle or heuristic
Performance change	Speed	Accuracy	Approach
Memory demands	Execute only	Recognize and execute	Recall, recognize, and execute

B Context: When and where transferred from and to					
	Near ←				→ **Far**
Knowledge domain	Mouse vs. rat	Biology vs. botany	Biology vs. economics	Science vs. history	Science vs. art
Physical context	Same room at school	Different room at school	School vs. research lab	School vs. home	School vs. the beach
Temporal context	Same session	Next day	Weeks later	Months later	Years later
Functional context	Both clearly academic	Both academic but one nonevaluative	Academic vs. filling in tax forms	Academic vs. informal questionnaire	Academic vs. at play
Social context	Both individual	Individual vs. pair	Individual vs. small group	Individual vs. large group	Individual vs. society
Modality	Both written, same format	Both written, multiple choice vs. essay	Book learning vs. oral exam	Lecture vs. wine tasting	Lecture vs. wood carving

Figure 2. A taxonomy of transfer, from Barnett and Ceci (2002). The upper box represents the content (i.e. what is transferred) factor and the lower box represents the context (i.e. when and where it is transferred to and from) factor. The various dimensions of each factor are listed along with examples.

The second dimension of the content factor is performance change, or how transfer is measured with the experimental task. Although there are many ways to assess performance, Barnett and Ceci (2002) focus on the three most popular measures: speed, accuracy, and approach. The speed with which a skill is executed or knowledge is produced can be used to assess transfer, such as measuring typing proficiency by recording the number of words typed per minute. The accuracy and quality of the execution of a procedure or production of knowledge can also be measured, as in analogical reasoning studies when the solution to a previous problem must be used to solve a new problem (Gick & Holyoak, 1983). Finally, transfer can be assessed by looking for the use of a general approach or heuristic, such as whether or not students use the law of large numbers when trying to solve a statistical reasoning problem (Kosonen & Winne, 1995). In addition, some studies use multiple ways of assessing transfer (e.g., measuring both speed and accuracy).

The memory demands involved in the experimental task are represented in the third dimension of the content factor. Barnett and Ceci (2002) distinguish between tasks that involve 1) execution only, 2) recognition and execution, and 3) recall, recognition, and execution. In an experiment that involves acquiring a single skill or concept, such as teaching children to use the control of variables strategy in designing science experiments (Chen & Klahr, 1999), the memory demands are low and success depends entirely on execution of the skill or application of the knowledge. In an experiment that involves acquiring multiple skills or concepts, successful transfer depends on both the recognition of which skill or concept to use and the execution of the skill or application of the concept. For example, when subjects are trained on multiple practice problems and then

instructed to use the solution to a prior problem to help solve a new problem, they must recognize which problem to use and then apply that concept (Gick & Holyoak, 1983). Finally, the experimental task may require additional memory demands if subjects are required to recall a previously learned skill or concept and recognize that it is applicable to a transfer task. An example would be in analogical transfer in which no hint is given that a prior problem can be used to help solve a new problem.

The second factor of the Barnett and Ceci (2002) taxonomy is context, which includes six dimensions: knowledge domain, physical context, temporal context, functional context, social context, and modality. Each dimension represents a continuum ranging from similar to dissimilar in terms of context or near to far in terms of transfer (see lower box in Figure 2 for examples from each dimension). Knowledge domain refers to the general area to which the skill or knowledge is to be applied. Physical context involves both the general location (e.g., school, home, outdoors, etc.) as well as the specific features present in the location (e.g., lighting, size of room, decorations, etc.). Temporal context primarily concerns the elapsed time between training and transfer tasks, but also might include the amount of time given to perform the task. Functional context pertains to the way in which the subject perceives the task (e.g., an academic assessment, a household chore, a leisure activity, etc.). Social context refers to whether the task is learned and performed alone or with other people. Finally, modality involves both macro-aspects (e.g., visual vs. auditory, written vs. spoken, etc.) and micro-aspects (e.g., multiple-choice test vs. essay, typing vs. writing with pen, etc.) of the task.

After describing their taxonomy, Barnett and Ceci (2002) proceeded to categorize existing transfer studies along these dimensions. Their analysis showed that the

framework was a useful tool for resolving some of the disparate findings in the literature and in identifying areas in which little or no research had been conducted. Importantly, Barnett and Ceci described their taxonomy as an initial attempt to organize the literature with the understanding that additional dimensions may be needed to better specify differences among transfer studies. In addition, they briefly discussed the possibility of interactions among the various dimensions.

Based on this framework, we may ask: what constitutes far transfer? Must a study investigate far transfer along all six dimensions or is one dimension sufficient? Barnett and Ceci (2002) left this question open, but they suggested that it is more fruitful to precisely define each experiment within the framework than settle for a broad label. With this thought in mind, I will now attempt to frame the present research within the context of their taxonomy.

Scope of the Present Research

In light of the many ways in which transfer can be investigated and what is feasible in a dissertation project, it was necessary to constrain the scope of the present research. I focused on the acquisition, retention, and transfer of declarative knowledge (e.g., facts, concepts, etc.). In terms of transfer, I was interested in how this knowledge is used to make inferences in response to new questions within the same knowledge domain as well as different knowledge domains. Within the context of Barnett and Ceci's (2002) taxonomy, the learned skill in the present experiments was declarative knowledge, which was relatively specific in nature and constituted a closed-problem space in which there was one correct answer. Performance was measured in terms of the accuracy of responses to cued recall questions, rather than speed of execution or overall approach. The memory

demands consisted of retrieving previously learned knowledge (i.e., recall) and applying it to answer new inferential questions (i.e., execution).

Overall, the progression of experiments was designed to gradually extend the distance between initial learning and subsequent transfer along the knowledge domain dimension (i.e. from near to far). Experiment 1 examined the retention of previously learned information on a final test with repeated questions (i.e. the same questions that were given on the initial tests). Experiments 2 and 3 investigated the transfer of previously learned information to new inferential questions within the same knowledge domain. Experiment 4 explored transfer to new inferential questions in a different knowledge domain. A variety of knowledge domains were used in the present research – subjects learned about a different topic from each passage that they studied (see Method below for more details). The other dimensions of the context factor were held constant within and across experiments: physical context (psychology laboratory, same room), temporal context (one week between learning and transfer), functional context (psychology experiment), social context (individual), and modality (a verbal task that is presented visually using a computer).

Rationale for the Present Research

With the scope of the present research constrained to the acquisition, retention, and transfer of declarative knowledge, I will now describe the rationale for the project. Much as similarity between the contexts of initial learning and transfer is important in the broader transfer literature (e.g., Holyoak & Koh, 1987), 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. 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.

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; Tulving & Thomson, 1973). Thus, depending on which features of the perceived event are encoded, a given retrieval cue will be more or less effective. For example, Barclay et al. (1974) had subjects study sentences (e.g., "*the man lifted the piano*" or "*the man tuned the piano*") that emphasized a particular attribute of to-be-remembered items (e.g., the weight of the piano or the sound of the piano, respectively). After a short break, subjects were asked to recall items from the sentences using a list of cues that were either congruent or incongruent with the attributes of the items that were emphasized during encoding. If subjects had encoded *the man lifted the piano*, then the cue "*something heavy*" would be congruent and the cue "*something with a nice sound*" would be incongruent (and vice versa for the other orienting sentence). The results showed higher levels of recall when the retrieval cue matched the way in which the item was initially encoded, demonstrating that the similarity of encoding and retrieval operations is an important determinant of memory performance.

Whereas the encoding specificity principle focuses on the contents of memory (i.e. the static trace), the related concept of transfer-appropriate processing takes a more dynamic approach by stating 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; Jacoby, 1975; McDaniel, Friedman, & Bourne, 1978; Roediger, Weldon, & Challis, 1989). For example, Jacoby (1983) had subjects encode a list of words in which each word was presented with no context (XXX – NIGHT), context (DAY – NIGHT), or required the subject to generate the word (DAY – ???). On a subsequent recognition test, subjects correctly remembered more words that were generated than words that were read with context, which in turn were better remembered than words that were read without context. However, a perceptual identification test in which individual words were flashed for a very brief period of time (25 msec) showed the opposite effect: words read without context were most likely to be identified relative to words read with context, which were better identified than generated words. One explanation for this pattern of results is that generating a word from an antonym requires conceptual or "top-down" processing, which matches the type of conceptual processing partly required by a recognition test. Likewise, reading a word without context taps perceptual or "bottom-up" processing, which matches with the type of processing required by a perceptual identification test.

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. 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)."

For example, Fisher and Craik (1977) presented subjects with a series of 64 to-be-remembered words, each of which was paired with a context word that either rhymed with or was semantically related to the to-be-remembered word. On an immediate recall test, subjects received either the rhyme word or the semantically associated word as a cue. The results of Fisher and Craik's (1977) experiment indicated that compatibility between encoding and retrieval led to better memory performance. When the to-be-remembered words were encoded in the context of rhyme words, subjects recalled a greater proportion of words when cued with a rhyme word (M = .26) than with a semantically associated word (M = .17). Likewise, words encoded in the context of a semantic associate were better recalled with a semantically associated word (M = .44) than with a rhyme word (M = .17). However, the greatest proportion of words was recalled in the condition in which semantically associated words served as encoding context and retrieval cues. Fisher and Craik concluded that "both the qualitative nature of the encoding and the degree of compatibility between encoding and cue are apparently necessary to give an adequate account of memory processes" (p. 710).

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 by promoting encoding variability may produce better transfer. Indeed, to the extent that initial learning increases both the

availability and *accessibility* of knowledge (Tulving & Pearlstone, 1966), there should be greater potential for successful retrieval of that knowledge when needed in a future transfer context.

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 (see Roediger & Karpicke, 2006a). When retrieval practice is given repeatedly over time and is coupled with feedback, the mnemonic benefits of testing increase substantially (e.g., Karpicke & Roediger, 2008). One additional, but previously untested, idea is that introducing encoding variability during repeated testing will increase retention and transfer by producing knowledge that can be accessed through several different retrieval routes. In this section, I describe evidence from previous studies that supports the efficacy of testing, repeated testing, feedback, and encoding variability in promoting retention and transfer.

Testing

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 (Karpicke & Roediger, 2008). The testing effect is a robust phenomenon, which has been replicated many times (see Roediger & Karpicke, 2006a). The mnemonic benefits of testing were first demonstrated by a handful

of early researchers working at the intersection of psychology and education (Gates, 1917; Jones, 1923-1924; Spitzer, 1939). However, over the remainder of the 20th century, interest in the testing effect was sporadic, leading one author to title his article "The 'testing' phenomenon: Not gone but nearly forgotten" (Glover, 1989). More recently, there has been a resurgence of interest in the phenomenon that has generated a substantial amount of research (see Marsh et al., 2007; McDaniel, Roediger, et al., 2007; Pashler et al., 2007) and calls for the use of testing as a learning tool in the classroom (e.g., Glover, 1989; Leeming, 2002; Roediger & Karpicke, 2006a)

The generalizability of the testing effect is well established. Traditionally, the phenomenon was investigated in laboratory settings using discrete verbal materials, such as lists of individual words or word pairs (e.g., Allen, Mahler, & Estes, 1969; Hogan & Kintsch, 1971; Izawa, 1970; Thompson, Wenger, & Bartling, 1978; Tulving, 1967). However, recent research has demonstrated testing effects with educationally relevant materials, such as prose passages and textbooks (e.g., Butler & Roediger, 2008; Foos & Fisher, 1988; Kang, in press; Kang, McDermott, & Roediger, 2007; Roediger & Karpicke, 2006b). In addition, other studies have shown strong, positive effects of testing in real-world educational contexts using retention intervals of up to six months (Gates, 1917; Jones, 1923-1924; Larsen et al., in press; McDaniel, Anderson, et al., 2007; Spitzer, 1939; see too Bangert-Drowns, Kulik, & Kulik, 1991). Thus, a large body of research supports the conclusion that testing promotes long-term retention with a variety of materials across many different contexts.

Several theoretical explanations have been proposed for the mnemonic benefits of testing. One early hypothesis held that taking a test after studying constituted an

additional exposure to the material and therefore the superior retention was the result of an increase in total study time (e.g., Thompson et al., 1978). However, this so called total time hypothesis (also referred to as the amount-of-processing hypothesis) was refuted by subsequent studies that showed that taking a test led to better retention relative to re-studying the material for an equivalent amount of time (e.g., Butler & Roediger, 2007; Carrier & Pashler, 1992; Glover, 1989; Roediger & Karpicke, 2006b). As a result, some researchers have argued that the cognitive effort required by retrieval enhances retention (e.g., Gardiner, Craik, & Bleasdale, 1973), while others have hypothesized that the act of retrieval elaborates the existing memory trace and / or creates additional retrieval routes to that trace (Bjork, 1975; McDaniel & Masson, 1985). Another possibility is that taking an initial test leads to better performance on a subsequent test relative to re-studying because the processes engaged on an initial test better match the processes required by the final test (i.e. transfer-appropriate processing; Morris et al., 1977). Finally, studying and taking a test represent distinct encoding events, therefore testing after studying may increase encoding variability, which should lead to better retention by creating multiple retrieval routes to that memory (Bower, 1972; Estes, 1955; Martin, 1968). No one theory has emerged as the dominant explanation of the testing effect and it may be best to consider these last four theories as complementary.

Repeated Testing

Although a single test confers a substantial mnemonic benefit, repeated testing leads to even better retention. For example, Wheeler and Roediger (1992) had subjects listen to a story while viewing a series of 60 related pictures. After hearing the story, subjects received one, three, or no tests on the names of the pictures. When they returned

to the lab one week later for a final test, the group that took three initial tests recalled a greater proportion of the names (M = .42) than the group that took only one test (M = .34), which recalled a greater proportion of names than the no test group (M = .28). In Wheeler and Roediger's procedure, the repeated tests were given consecutively and there were no breaks between tests. However, repeated testing is even more effective if it is spaced out over time rather than massed together (e.g., Bahrick & Hall, 2005). Spaced practice generally produces better long-term retention than massed practice, a finding called the spacing effect, which dates to the first empirical investigations of human memory (Ebbinghaus 1885/1964; see Roediger, 1985) and has been replicated many times (e.g., Glenberg, 1976; Melton, 1970; for a review, see Cepeda et al., 2006; Dempster, 1989).

Feedback

The critical mechanism in learning from tests is successful retrieval; however, if test-takers do not retrieve the correct response, then the benefits of testing are limited (e.g., Kang et al., 2007). Thus, providing feedback is one way to enhance learning from tests because it enables test-takers to correct errors (Bangert-Drowns, Kulik, Kulik, & Morgan, 2001; Kulhavy, 1977; Kulhavy & Stock, 1989) and maintain correct responses (Butler, Karpicke, & Roediger, 2008). For example, Butler et al. (2008) gave subjects an initial multiple-choice test on general knowledge information and provided feedback on half of the questions. On a final test one week later, subjects correctly answered a greater proportion of the questions for which they had received feedback (M = .83) than questions for which no feedback was given (M = .47). This result illustrates the powerful

effect that feedback can have on learning from tests, especially when initial test performance is low (see too Butler & Roediger, 2008).

Many studies have investigated the various factors that influence the effectiveness of feedback after a test, such as the content of the feedback message and the timing of feedback (for reviews see Butler & Winne, 1995; Hattie & Timperley, 2007; Kulhavy & Stock, 1989). In terms of the feedback message, the critical piece of information seems to be the correct response (for a meta-analysis see Bangert-Drowns, Kulik, Kulik, & Morgan, 2001). Many studies have shown that presenting the correct response is more effective than simply indicating whether the response is correct or incorrect (e.g., Gilman, 1969; Pashler et al., 2005; Roper, 1977), presumably because test-takers have no way to correct their errors without knowing the correct response. Interestingly, elaborations of the feedback message (e.g., providing an explanation of why the answer is correct) have not proven to be more effective than simply providing the correct response (see Kulhavy & Stock, 1989). However, this finding may be an artifact of how the effect of feedback is assessed on the final test, which generally consists of a verbatim re-presentation of the original questions and therefore only requires retention of the correct response.

There is substantial disagreement among researchers about whether it is better to give feedback immediately or after a delay. Most reviewers of the literature have noted the variety of findings that exist to support both sides of the debate, but still draw the conclusion that feedback should be given as soon as possible (e.g., Ammons, 1956; Azevedo & Bernard, 1995; Kulik & Kulik, 1988; Mory, 2004). Contrary to the perceived superiority of immediate feedback, which is a legacy of the behaviorist approach to learning (e.g., Skinner, 1954), there is much evidence to support the notion that delayed

feedback leads to superior retention (e.g., Butler, Karpicke, & Roediger, 2007; Kulhavy & Anderson, 1972). Even delaying feedback until after the end of the test provides a significant boost to retention relative to immediate feedback (e.g., Butler & Roediger, 2008), possibly due to its spaced presentation in the former condition. However, one critical assumption is that the feedback is fully processed after the delay. If full processing cannot be guaranteed, then giving feedback immediately may be better because people will be more motivated to engage the feedback just after taking the test.

Encoding Variability

As described above, one potential explanation for the testing effect is that taking a test after studying material creates variability in the encoding of that material. 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.

Test-Enhanced Learning and Transfer

Within the testing effect literature, the vast majority of studies have assessed the benefits of retrieval practice with a final test that contains a verbatim representation of the same questions used on the initial test. However, there are a handful of studies that have attempted to assess whether the benefits of testing transfer to other types of questions.

Several studies have investigated whether testing produces better retention than studying on a final test that consists of re-phrased versions of the questions from the initial tests. For example, McDaniel, Anderson et al. (2007) conducted an experiment in which students read a textbook chapter and then were re-exposed to facts from the chapter by either re-reading them or taking a fill-in-the-blank quiz on them. Each quiz item consisted of the same factual statement that was presented in the re-reading condition except that a critical word had been deleted. McDaniel and colleagues found a testing effect on a final test consisting of new fill-in-the-blank questions that had a different critical word deleted (see too McDaniel & Sun, submitted).

Similar to the McDaniel, Anderson, et al. (2007), a handful of paired-associate learning studies have shown that retrieval practice increases retention of more than just the specific response to the question or cue given during initial testing because associated information also seems to be better retained. For example, initial testing of paired associates in one direction (A → ?) leads to better performance on a final test in which the pair is tested in the opposite direction (? ← 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). Many of these paired associate learning experiments were conducted to test ideas that stem from the principle of associative symmetry put forth by Asch and Ebenholtz (1962; see Kahana, 2002 for a recent discussion of these ideas). 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.

Other studies have shown that initial testing with inferential questions leads to better performance on new inferential questions about the same material on a final test (Foos & Fisher, 1988; McKenzie, 1972); however, it is unclear how (or if) the inferential questions on the initial test and the transfer test were related. In addition, Chan, McDermott, and Roediger (2006) found that testing can benefit the retention of non-tested, but related material, a phenomenon they termed retrieval-induced facilitation. Finally, there is also some evidence from the analogical reasoning literature that attempting to generate solutions to training problems leads to better performance on subsequent transfer problems relative to studying the training problems (Needham & Begg, 1991).

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. They found that practice with different variations of an anagram led to 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 consists of four experiments that investigated 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 re-studied some passages and repeatedly took a test on other passages. As explained above, the series of experiments was designed to explore progressively greater degrees of transfer. In Experiment 1, the final test consisted of repeated questions (i.e. a verbatim re-presentation of the questions that were on the initial tests) in order to demonstrate that testing improves retention of information relative to re-studying the passages. In Experiments 2 and 3, the final test consisted of new inferential questions from the same knowledge domain in order to assess whether testing would produce better transfer than re-studying. The new inferential questions required subjects to apply the knowledge that they learned during the initial session to answer a related question from the same domain. In Experiment 4, the final test consisted of new inferential questions from different knowledge domains to explore whether testing would promote transfer of a concept across domains. As in Experiments 2 and 3, the final test questions in Experiment 4 required subjects to apply knowledge from the initial learning session; however, the questions were about phenomena in new domains that operated on the same underlying concept, making the final test similar to an analogical reasoning task.

Experiment 1

Experiment 1 investigated whether repeated testing leads to better retention on a final test with repeated questions relative to re-studying. The experiment was designed to accomplish three goals. First, it was important to demonstrate the basic testing effect with the new set of materials developed for this set of experiments. If testing leads to better retention of the materials relative to re-studying, then any failure to find differences on the transfer tests in the subsequent experiments could not be attributed to the materials. Second, the experiment examined whether the benefits of retrieval practice held for both facts and concepts (see Method below for definitions and examples). Most testing effect studies have either used facts as the to-be-remembered materials or used both types of items without distinguishing between them. Third, the experiment explored whether a testing procedure that promoted encoding variability would lead to better retention relative to the standard testing procedure. To test this hypothesis, a repeated testing condition in which the questions were re-phrased on each initial test was compared to a repeated testing condition in which the same question was repeated verbatim on each initial test.

Experiment 1 consisted of two sessions, which were spaced one week apart. In an initial learning session, subjects studied a set of six passages about a variety of topics. Then, they repeatedly re-studied two of the passages (re-study passages), repeatedly took the same test on another two passages (same test), and repeatedly took different tests on the other two passages (variable test). To maximize the potential for learning from the tests, feedback was given after each test question. One week later, subjects returned to the lab for the final test in which retention of the material was assessed. This final test

included questions about all six passages, some of which had been previously tested during the initial learning session. Of the questions that had not been previously tested, some of these questions were about the passages that had been repeatedly re-studied. In addition, there was also a set of control questions that were never tested during the initial learning session. The control questions were included to examine whether the benefits of testing were limited to the specific items that were initially tested or whether these benefits extended to other (untested) information contained in the same passages. Table 2 contains a schematic representation of the design.

Based on previous testing effect studies, several predictions were made about the results of Experiment 1. On the initial tests, performance was expected to increase across successive tests because subjects would use the feedback to correct their errors. On the final test, repeated testing was predicted to produce better retention than re-studying the passages. However, it was unclear whether final test performance would differ between the two repeated testing conditions. One possibility was that the encoding variability promoted by the re-phrasing of questions might lead to better retention relative to repeated testing with the same question. However, it was also possible that encoding variability might not produce any benefit in retention because the final test consisted of repeated items. That is, being tested three times on the same item might lead to better or equivalent final test performance on that item relative to being tested once on that item and then twice more with re-phrased versions of the item. Finally, it was predicted that the re-study passages condition would lead to better performance on the control questions relative to the two testing conditions because subjects would have four opportunities to study the information, whereas they would only study it once in the testing conditions.

Table 2

A design schematic of the general procedure used in Experiment 1.

Condition	Initial Learning Session				Final Test
Same Test	S	T_A	T_A	T_A	T_A
Variable Test	S	T_A	T_B	T_C	T_A
Re-Study Passages	S	S	S	S	T_A

Note. S = Study, T = Test. Subscript refers to the version of the test question: A, B, or C.

Method

Subjects & Design. Twenty-four undergraduate psychology students at Washington University in St. Louis participated for course credit or pay. All subjects were treated in accordance with the "Ethical Principles of Psychologists and Code of Conduct" put forth by the APA (2002).

The design was a 3 (Type of Initial Learning: Re-Study Passages, Same Test, Variable Test) x 2 (Type of Initial Test Question: Factual, Conceptual) within-subjects design. Both variables were manipulated within-subjects, but between-materials. The main dependent variable was performance on repeated questions (i.e. previously tested factual and conceptual questions) on the final test. In addition, there was a set of control questions that were included on the final test to explore whether the benefits of repeated testing would spread to information from the same passage that was not initially tested.

Materials & Counterbalancing. The materials 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 1000 words in length and arranged into eight paragraphs (see Appendix A for prose passages). 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 is presented within a single sentence, while a concept was defined as a piece of information that must be abstracted from multiple sentences. These definitions

were developed in consultation with the taxonomy of educational objectives put forth by Bloom and colleagues (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 of the object 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*). Appendix B contains all the questions used.

In addition, two re-phrased versions of each question were created for use during initial testing in the variable test condition. For each re-phrased version of the question, the question stem was re-worded, but the correct response remained the same. A re-phrased version of the factual question given as an example above was 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 re-phrased version of the conceptual question given as an example above was 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) (see Appendix B for additional questions). Also, a control set of questions was created to test information contained in the passages, but not tested in either the same 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*) (see Appendix B for additional questions). In terms of content, the information tested by these control questions was factual and had minimal overlap with the other items described above.

The experiment was 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 the experiment. Overall, the counterbalancing ensured that, across subjects, each passage was used in each initial learning condition an equal number of times.

Procedure. The entire experiment was conducted on a computer using E-Prime software (Schneider, Eschman, & Zuccolotto, 2002) and it involved two sessions, spaced one week apart. In session one, subjects began by studying all the passages. Each passage was presented two paragraphs at a time (approx. 250 words) with each pair of paragraphs appearing on the screen for 60 seconds. Thus, a total of four minutes was given to study each passage. Then, depending on the version of the experiment to which subjects were

assigned, they repeatedly re-studied some passages and took tests on the other passages in the same order as the passages were initially presented. Passages that were re-studied were presented in the same manner as before (i.e. 60 seconds 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 consisted of repeated questions from the initial test conditions, questions about the passages in the re-study 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 2). Again, the test was cued recall format, self-paced, and forced report. After entering each response, subjects were asked to rate their confidence on a scale of 0 to 100. At the end of the test, they were fully debriefed and dismissed.

Results

All results, unless otherwise stated, were significant at the .05 level. Pair-wise 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 ANOVA and *t*-test analyses, respectively. A Geisser-Greenhouse correction was used for violations of the sphericity assumption of ANOVA (Geisser & Greenhouse, 1958).

Scoring. The author and a research assistant each scored 20% of the cued recall responses independently. Both scorers were blind to condition and coded all the responses for a given question together in order to increase consistency in scoring. Cohen's kappa (Cohen, 1960) was calculated to assess inter-rater reliability. Reliability was high ($\kappa = .88$), so the author resolved the few disagreements and then scored the remaining responses alone.

Initial Tests. Table 3 shows the proportion of correct responses on the three initial cued recall tests as a function of question type and initial learning condition. Overall, 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 multi-trial learning experiments (e.g., Ebbinghaus 1885/1964). This pattern of increasing performance held for both factual and conceptual questions, as well as for the same test and variable test conditions.

Table 3

Mean proportion of correct responses on the three initial cued recall tests as a function of question type and initial learning condition for Experiment 1.

Question Type	Learning Condition	Test 1	Test 2	Test 3
Factual	Same Test	.40	.81	.90
	Variable Test	.37	.71	.82
Conceptual	Same Test	.42	.70	.80
	Variable Test	.41	.66	.79

Performance on the factual and conceptual questions were analyzed separately via 3 (Test: 1, 2, 3) x 2 (Initial Learning Condition: Same Test, Variable Test) repeated measures ANOVAs. For the factual questions, 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 from the ANOVA for the conceptual questions. There was a significant main effect of test [$F(2,46) = 101.54$, $MSE = .02$, $\eta^2 = .55$] as well as a significant quadratic trend [$F(1,23) = 12.80$, $MSE = .02$, $\eta^2 = .07$]. Again, neither the main effect of initial learning condition nor the interaction was significant (Fs < 1).

Response Times on Initial Tests. Table 4 shows the mean number of seconds that subjects spent on each item (i.e. both responding and reviewing feedback) as a function of question type and initial test condition (see Appendix C for full response time data). In both the same test and variable test conditions, subjects spent less time on factual questions than on conceptual questions. This observation was confirmed by a 2 (Question Type: Factual, Conceptual) x 2 (Initial Learning Condition: Same Test, Variable Test) repeated measures ANOVA, which yielded a significant main effect of question type [$F(1,23) = 122.76$, $MSE = 56.89$, $\eta^2 = .61$]. However, neither the main effect of initial learning condition [$F(1,23) = 1.22$, $MSE = 42.29$, $p = .28$] nor the interaction was significant [$F(1,23) = 1.89$, $MSE = 83.34$, $p = .18$].

Table 4

Mean number of seconds that subjects spent on each item (i.e. both responding and reviewing feedback) as a function of question type and initial learning condition for Experiment 1.

Question Type	Learning Condition	Mean Time Per Item
Factual	Same Test	27.1
	Variable Test	28.2
Conceptual	Same Test	46.7
	Variable Test	43.7

Note. Means represent the average of all three initial tests. See Appendix C for full results.

Of additional interest was how much time was spent on task in each the three initial learning conditions. Total time spent on each test was computed by multiplying the average time spent on factual and conceptual questions by the total number of each type of question per passage (4 factual and 4 conceptual). Subjects spent an average of 295 seconds (4.9 minutes) and 284 seconds (4.7 minutes) completing the test for each passage in the same test and variable test conditions, respectively. In contrast, each passage was studied for a total of 240 seconds (4.0 minutes) in the re-study passages condition. Thus, subjects spent somewhat more time on task in the testing conditions than the re-study passages condition.

A one-way (Initial Learning Condition: Re-Study Passages, Same Test, Variable Test) repeated measures ANOVA revealed a significant main effect [$F(2,31) = 7.13$, $MSE = 4246.71$, $\eta^2 = .24$]. Follow up pair-wise comparisons confirmed that subjects spent more time on task in the same test and variable test conditions than in the re-study passages condition [295 sec vs. 240 sec: $t(23) = 2.78$, $SEM = 19.88$, $d = .75$; 284 sec vs. 240 sec: $t(23) = 3.04$, $SEM = 14.34$, $d = .81$, respectively]. However, there was no significant difference in total time spent per passage between the same test and variable test conditions ($t < 1$).

Final Test. Figure 3 shows the proportion of correct responses on the final cued recall test as a function of question type and initial learning condition. For both factual and conceptual questions, performance was roughly equivalent in the same test and variable test conditions, but both testing conditions produced a greater proportion of correct responses than the re-study passages condition.

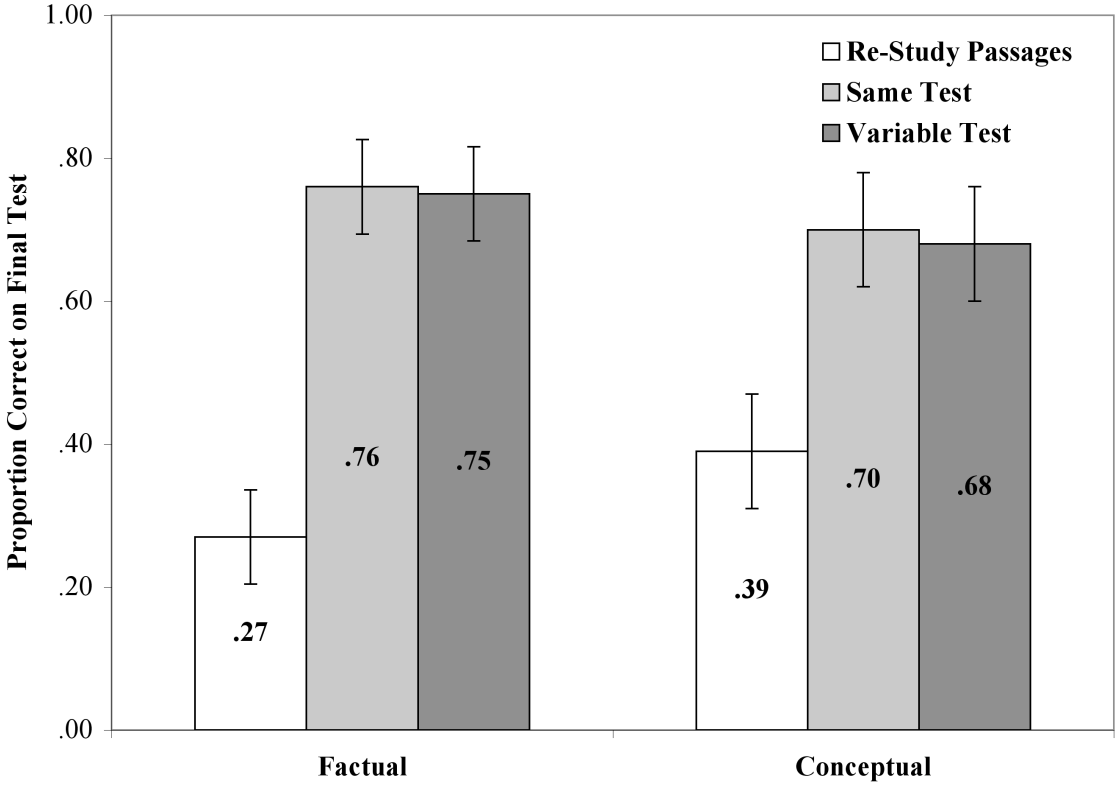

Figure 3. Mean proportion of correct responses on the final cued recall test as a function of question type and initial learning condition for Experiment 1. Error bars represent 95% confidence intervals.

Performance on factual and conceptual questions was analyzed with separate one-way (Initial Learning Condition: Re-Study Passages, Same Test, Variable Test) repeated measures ANOVAs. For factual questions, there was a significant effect of initial learning condition [$F(2,46) = 70.18$, $MSE = .03$, $\eta^2 = .75$]. Planned pair-wise comparisons revealed that both the same test condition and the variable test condition produced a significantly greater proportion of correct responses relative to the re-study passages condition [.76 vs. .27: $t(23) = 9.97$, $SEM = .05$, $d = 2.13$ and .75 vs. .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 conceptual questions, there was also a significant main effect of initial learning condition [$F(2,46) = 18.87$, $MSE = .04$, $\eta^2 = .45$]. Pair-wise comparisons confirmed the observation that the same test and variable test conditions led to significantly better performance on the final test than the re-study passages condition [.70 vs. .39: $t(23) = 5.50$ $SEM = .06$, $d = 1.29$ and .68 vs. .39: $t(23) = 5.38$, $SEM = .06$, $d = 1.25$, respectively]. Again, there was no significant difference between the two initial testing conditions ($t < 1$). Final test performance in the three initial learning conditions (collapsed across question type) was also analyzed with an ANCOVA that included total time on task as a covariate. The same test and variable test conditions led to significantly better performance on the final test than the re-study passages condition even after controlling for differences in total time on task.

In addition to the factual and conceptual questions, the final test included a set of control questions. Performance was higher in re-study passages condition ($M = .39$) relative to both the same test ($M = .21$) and variable test ($M = .24$) conditions. A one-way (Initial Learning Condition: Re-Study Passages, Same Test, Variable Test) repeated

measure ANOVA revealed a significant main effect [$F(2,46) = 4.10$, $MSE = .05$, $\eta^2 = .15$]. Follow up pair-wise comparisons showed that the re-study passages condition led to a significantly greater proportion of correct responses on the control questions relative to the same test condition [.39 vs. .21: $t(23) = 2.67$, $SEM = .07$, $d = .68$] and the variable test condition [.39 vs. .24: $t(23) = 1.98$, $SEM = .07$, $p = .06$, $d = .55$], although the latter difference was only marginally significant. There was no significant difference between the two testing conditions ($t < 1$). One possible explanation for differential performance on the control questions is that it was due to retrieval-induced forgetting: retrieval practice on a subset of studied material can inhibit later retrieval of the remaining material (M. C. Anderson, 2003; M. C. Anderson, Bjork, & Bjork, 1994). However, this explanation is unlikely because retrieval-induced forgetting is a short-lived phenomenon that is eliminated after a 24-hour delay (MacLeod & Macrae, 2001; Saunders & MacLeod, 2002) let alone the week delay used in the present experiment.

Confidence. The purpose of including confidence judgments on the final test was to explore whether the initial learning conditions had a differential effect on subjects' metacognitive monitoring. The confidence judgments were used to assess calibration, the absolute correspondence between test performance and confidence, and resolution, the relative correspondence between performance and confidence (for elaboration, see Koriat & Goldsmith, 1996; Nelson, 1984; Nelson & Dunlosky, 1991). Table 5 shows the mean confidence judgment on the final cued recall tests as a function of question type and initial learning condition (as well as the mean proportion of correct responses and the difference between confidence and proportion correct). Mean confidence judgments were converted to proportions for comparison with the mean proportions of correct responses.

Table 5

Mean confidence judgment, mean proportion of correct responses, and mean difference between confidence and proportion correct on the final cued recall tests as a function of question type and initial learning condition for Experiment 1.

Question Type	Learning Condition	Confidence Judgment	Proportion Correct	Difference (CJ – PC)
Factual	Same Test	.79	.76	+ .03
	Variable Test	.79	.75	+ .04
	Re-Study Passages	.42	.27	+ .15
Conceptual	Same Test	.72	.70	+ .02
	Variable Test	.74	.68	+ .06
	Re-Study Passages	.52	.39	+ .13

Note. Mean confidence judgments were converted to proportions for ease of comparison with the mean proportions of correct responses and statistical analyses.

On both types of question, subjects were relatively well calibrated on items in the two testing conditions, but were overconfident on items in the re-study passages condition. Statistical analyses were performed on the difference scores (confidence minus proportion correct). Since the pattern of results was the same for both types of question, the data were collapsed across question type for analysis. A one-way (Initial Learning Condition: Re-Study Passages, Same Test, Variable Test) repeated measures ANOVA revealed a significant main effect [$F(2,46) = 10.22$, $MSE = .01$, $\eta^2 = .31$]. Pair-wise comparisons confirmed that subjects were significantly more overconfident on items in the re-study passages condition relative to the same test condition [.14 vs. .03: $t(23) = 3.82$, $SEM = .03$, $d = .73$] and the variable test condition [.14 vs. .05: $t(23) = 3.38$, $SEM = .03$, $d = .64$]. The two testing conditions did not differ ($t < 1$).

Next, within-subject Goodman-Kruskal gamma correlations between accuracy (i.e. correct vs. incorrect) and confidence judgments on the final cued recall test were computed to assess resolution. Whereas calibration is a global measure of metacognitive monitoring, resolution indicates how well subjects' are able to differentiate between correct and incorrect responses on an item-by-item basis. Across all three conditions, the mean gamma correlations were positive and relatively large, indicating that subjects were good at monitoring the accuracy of their responses. The re-study passages condition produced the highest mean gamma correlation ($M = .57$), followed by the variable test condition ($M = .52$) and the same test condition ($M = .51$). However, a one-way (Initial Learning Condition: Re-Study Passages, Same Test, Variable Test) repeated measures ANOVA did not show any significant differences among the means ($F < 1$). Three

subjects were excluded from this analysis because a gamma correlation could not be calculated for one or more of the initial learning conditions.

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 on the final test depended upon successful retrieval on one or more of the initial tests. Table 6 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). 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).

To confirm these observations, a 2 (Retrieval Success: Successful, Unsuccessful) x 2 (Initial Learning Condition: Same Test, Variable Test) repeated measures ANOVA was conducted. 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 therefore did not produce a mean for unsuccessful retrieval on the initial tests. 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).

Table 6

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 Experiment 1.

Learning Condition	Retrieval Success on Initial Tests	Proportion Correct on Final Test
Same Test	Successful	.84
	Unsuccessful	.13
Variable Test	Successful	.79
	Unsuccessful	.15

Discussion

Experiment 1 produced several important results. During the initial learning session, performance increased across the three tests in a curvilinear manner, and this pattern held for both testing conditions and for both types of questions. Subjects spent more time answering questions and studying the feedback for the conceptual questions than the factual questions, and more total time on task in the two testing conditions relative to the re-study passages condition. For the main set of factual and conceptual questions that were repeated verbatim on the final test, repeated testing led to better performance than repeated studying; however, there was no difference in performance between the two testing conditions. Conditional analyses revealed that subjects retained 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. In terms of metacognitive monitoring, subjects were well-calibrated in the testing conditions, but overconfident in the re-study passages condition. Finally, repeated studying of the passages led to better performance than repeated testing for the control questions on the final test, which were about information contained in the passages that was not initial tested.

The most important result that emerged from Experiment 1 was the robust testing effect for both factual and conceptual information. Repeated testing during the initial learning session produced a higher proportion of correct responses on the final cued recall test relative to repeatedly studying the passage. Although this finding was expected based on prior studies (e.g., Karpicke & Roediger, 2008; Roediger & Karpicke, 2006b), it is also novel in that it shows that the mnemonic benefits of retrieval practice extend to

more complex conceptual information. Most testing effect studies have used factual information as the to-be-learned materials. In addition, this finding is important because it demonstrates that testing effects can be obtained with this new set of materials.

Why did repeated testing produce better retention 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 on the final test. In contrast, when subjects failed to retrieve a fact or concept on the initial tests, it was relatively rare that they would produce the correct response on the final test. As indicated by the curvilinear increase in the proportion of correct responses across the initial tests, the feedback provided after each test was 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 almost all of that information until the final test one week later.

Alternatively, it is possible that repeated testing produced superior retention because subjects spent a greater amount of time processing the material in this condition. On average, subjects took about 4.8 minutes to complete each test on a passage in the repeated testing conditions, which was more than the 4.0 minutes that they spent studying each passage in the re-study passages condition. Still, there are at least two reasons why it is unlikely that these differences in total time on task contribute to the final test performance. First, subjects spent much more time on the conceptual questions than on the factual questions, and yet retained both the facts and concepts equally well. Second,

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). Performance on the control questions in Experiment 1 supports the idea that the way in which the materials are processed is more important than the amount of time spent processing them. Studying a passage four times led to a greater proportion of correct responses on the control questions relative to studying the passage once. However, the benefit of repeated studying over studying once (effect sizes of .68 and .55 for factual and conceptual items, respectively) was small compared to the benefit of repeated testing relative to repeated studying (effect sizes of 1.29 and 1.25). Nevertheless, total time on task cannot be completely ruled out as an explanation for the final test results, and this issue is revisited in Experiment 3.

A final result of note from Experiment 1 was that the variable test condition did not lead to superior final test performance relative to the same test condition. One possible outcome discussed during the Introduction to Experiment 1 was that the encoding variability presumably induced through repeated testing with re-phrased questions might produce better retention than repeated testing with the same question. However, given that retention was assessed on the final test with questions that were repeated verbatim from the initial tests, it was also possible that encoding variability might not confer any mnemonic benefit. Encoding variability increases the probability of future retrieval because it creates multiple retrieval routes to a particular memory. In other words, when a greater variety of features are encoded, it increases the potential for a match with the features in the retrieval cue. In the same test condition, subjects received three retrieval opportunities with the version of the question that appeared on the final

test, whereas they received only one retrieval opportunity with that version in the variable test condition (as well as two more attempts with re-phrased versions). Perhaps a single retrieval opportunity with the version of the question that later appeared on the final test was sufficient to encode the features of that question, thus rendering any additional features that were encoded in the variable test condition superfluous. Several alternative explanations are possible as well and they will be discussed after Experiment 2.

Experiment 2

Experiment 1 showed that repeated testing produced better retention than repeated studying of the passages as measured by a final test on which the initial test questions were repeated verbatim. The goal of Experiment 2 was to build upon this finding by exploring whether retrieval practice would produce better transfer as well. The same overall design, materials, and procedure were used, except for the questions on the final test which consisted of new inferential questions that required the application of previously learned knowledge within the same knowledge domain.

Method

Subjects & Design. Twenty-four undergraduate psychology students at Washington University in St. Louis participated for course credit or pay.

The design was the same as in Experiment 1, but the main dependent variable was changed to be new inferential questions on the final transfer test.

Materials. The materials from Experiment 1 were used again with one exception: a new set of inferential questions was developed to assess transfer on the final test. 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 in the Method for Experiment 1 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 in the Method for Experiment 1 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 insect will reflect more of the sound wave as it gets closer*). Appendix B contains all of the inferential questions for both initially tested facts and concepts.

Procedure. The procedure was the same as in Experiment 1 except that the questions on the final test were changed from repeated questions to new inferential questions.

Results

Scoring. The author and a research assistant each scored 20% of the cued recall responses independently in the same manner as in Experiment 1. Inter-rater reliability was high ($\kappa = .87$), so the author resolved the few disagreements and then scored the remaining responses alone.

Initial Tests. Table 7 shows the proportion of correct responses on the initial cued recall tests as a function of question type and initial learning condition. As expected, the pattern of initial test performance was similar to that observed in Experiment 1. The proportion of correct responses increased across the successive tests, and greater gains occurred from Test 1 and Test 2 than from Test 2 to Test 3. Again, this pattern held for both types of question and both initial testing conditions.

Performance on the factual and conceptual questions was analyzed separately with a 3 (Test: 1, 2, 3) x 2 (Initial Learning Condition: Same Test, Variable Test) repeated measures ANOVAs. For factual questions, there was a significant main effect of test [$F(2,46) = 110.50$, $MSE = .03$, $\eta^2 = .66$]. A significant quadratic trend confirmed the observation of a curvilinear increase in performance across the three tests [$F(1,23) = 26.71$, $MSE = .02$, $\eta^2 = .26$]. 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, the ANOVA revealed 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).

Table 7

Mean proportion of correct responses on the three initial cued recall tests as a function of question type and initial learning condition for Experiment 2.

Question Type	Learning Condition	Test 1	Test 2	Test 3
Factual	Same Test	.34	.74	.88
	Variable Test	.39	.78	.86
Conceptual	Same Test	.38	.66	.75
	Variable Test	.39	.78	.92

Response Times. Table 8 shows the mean number of seconds that subjects spent on each item (i.e. both responding and reviewing feedback) as a function of question type and initial test condition (see Appendix C for full data). As in Experiment 1, subjects spent more time on average completing the conceptual questions than the factual questions, and this pattern held across both testing conditions. In addition, subjects spent slightly more time on each item in the same test condition relative to the variable test condition. A 2 (Question Type: Factual, Conceptual) x 2 (Initial Learning Condition: Same Test, Variable Test) repeated measures ANOVA revealed significant main effects of question type [$F(1,23) = 78.78$, $MSE = 86.53$, $\eta^2 = .61$] and initial learning condition [$F(1,23) = 6.22$, $MSE = 25.04$, $\eta^2 = .01$], but the interaction was not significant ($F < 1$).

The average total time spent on each passage was computed in the same manner as for Experiment 1. On average, subjects spent 310 seconds (5.2 minutes) and 290 seconds (4.8 minutes) completing the test for each passage in the same test and variable test conditions, respectively. Relative to the two testing conditions, subjects spent less time on each passage in the re-study passages condition, taking a total of 240 seconds (4.0 minutes) per passage. A one-way (Initial Learning Condition: Re-Study Passages, Same Test, Variable Test) repeated measures ANOVA confirmed that there was a significant difference among the means [$F(2,28) = 11.78$, $MSE = 4297.81$, $\eta^2 = .34$]. Pair-wise comparisons showed that subjects spent more time on task in the same test condition than in the variable test and the re-study passages conditions [310 sec vs. 290 sec: $t(23) = 2.92$, $SEM = 6.89$, $d = .24$; 310 sec vs. 240 sec: $t(23) = 3.95$, $SEM = 17.76$, $d = .99$, respectively]. Subjects also spent more time on task in the variable test condition than in the re-study passages condition [290 vs. 240: $t(23) = 2.88$, $SEM = 17.35$, $d = .77$].

Table 8

Mean number of seconds that subjects spent on each item (i.e. both responding and reviewing feedback) as a function of question type and initial learning condition for Experiment 2.

Question Type	Learning Condition	Mean Time Per Item
Factual	Same Test	30.1
	Variable Test	28.0
Conceptual	Same Test	47.3
	Variable Test	44.4

Note. Means represent the average of all three initial tests. See Appendix C for full results.

Final Test. Figure 4 shows the proportion of correct responses on the final cued recall test as a function of question type and initial learning condition. Despite the change in the questions on the final test (i.e. new inferential questions rather than repeated questions), the overall pattern of results was similar to Experiment 1. Performance was highest in the two initial testing conditions, both of which produced superior transfer relative to the re-study 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 the factual and conceptual transfer questions was analyzed separately by one-way (Initial Learning Condition: Re-Study Passages, Same Test, Variable Test) repeated measures ANOVAs. For the factual inferential questions, there was a main effect of initial learning condition [$F(2,46) = 16.73$, $MSE = .04$, $\eta^2 = .42$]. Planned pair-wise comparisons confirmed that both the same test and variable test conditions produced better transfer than the re-study passages condition [.60 vs. .30: $t(23) = 5.74$, $SEM = .05$, $d = 1.03$ and .57 vs. .30: $t(23) = 4.38$, $SEM = .06$, $d = .93$, respectively]. However, performance did not differ significantly between the two testing conditions ($t < 1$). For the conceptual inferential questions, there was also a main effect of initial learning condition [$F(2,46) = 15.63$, $MSE = .03$, $\eta^2 = .41$]. Planned pair-wise comparisons revealed that the same test and variable test conditions led to better performance on the final transfer test relative to the re-study passages condition [.60 vs. .36: $t(23) = 4.44$, $SEM = .05$, $d = .74$ and .64 vs. .36: $t(23) = 5.11$, $SEM = .05$, $d = .87$, respectively]. There was no significant difference between the testing conditions ($t < 1$).

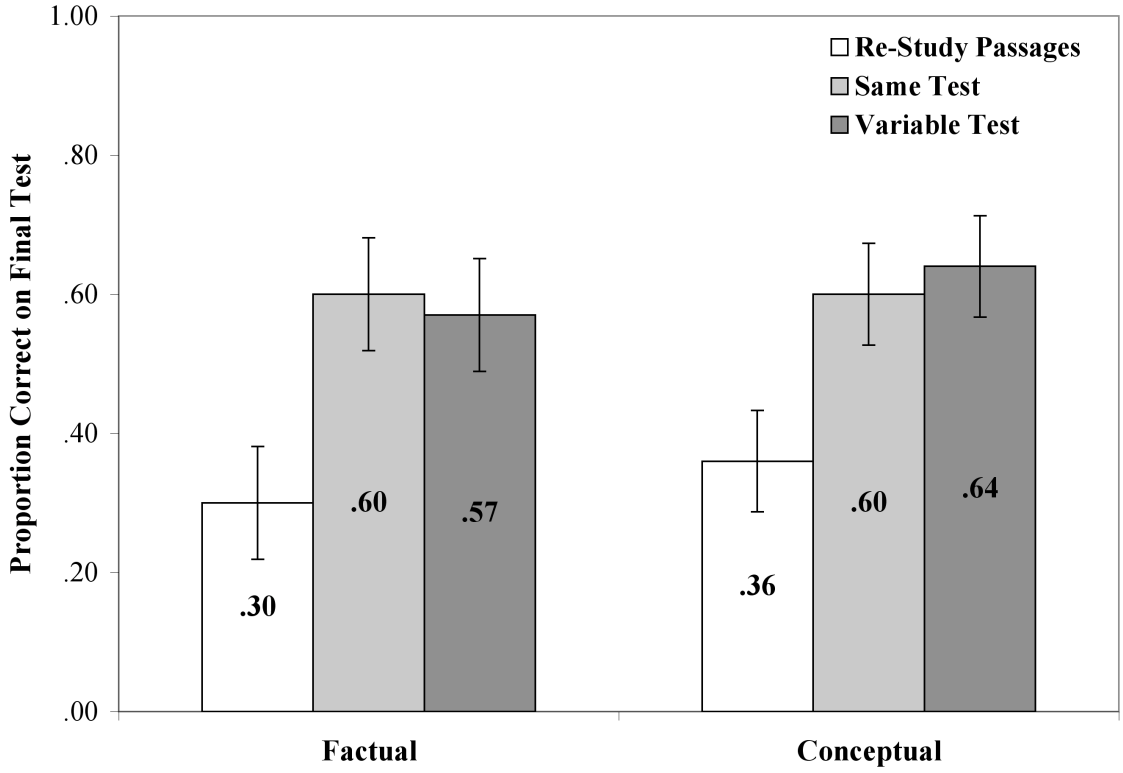

Figure 4. Mean proportion of correct responses on the final cued recall test as a function of question type and initial learning condition for Experiment 2. Error bars represent 95% confidence intervals.

Final test performance in the three initial learning conditions (collapsed across question type) was also analyzed with an ANCOVA that included total time on task as a covariate. The same test and variable test conditions led to significantly better performance on the final test than the re-study passages condition even after controlling for differences in total time on task.

Performance on the control questions was also analyzed. As in Experiment 1, the re-study 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: Re-Study 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 find a significant effect may be due to low power, and so a power analysis was conducted. In terms of estimating the expected effect size, no prior study could be identified that contained this exact comparison. The closest is a set of two experiments by Rawson and Kintsch (2005) in which they compared studying a text once versus studying it twice with an interval of one week between study opportunities. Both experiments contained this comparison and assessed retention with a final short answer test that was given after a retention interval of two days. The mean difference between the study twice condition and the study once condition was .13 (d = .76) and .12 (d = .66) in Experiments 1 and 2, respectively. In the present research, subjects studied the text four times in one of the conditions, but with much shorter inter-study intervals, so an expected mean difference of .15 was used in the power analysis. With an alpha of .05, a standard deviation of .23 (from the control question data), and power of .80, 53 subjects

would be needed to detect .15 mean difference between two or more of the conditions in a one-way ANOVA with three levels (Lenth, 2006-9). Since Experiment 2 had 24 subjects, the results of the power analysis indicate that there was insufficient power to detect the effect.

Confidence. As in Experiment 1, the confidence judgments were used to assess metacognitive monitoring through calibration and resolution. Table 9 shows the mean confidence judgments on the final cued recall tests as a function of question type and initial learning condition (as well as the mean proportion of correct responses and the difference between confidence and proportion correct). Much like in Experiment 1, subjects showed good calibration on items in the two testing conditions, but were slightly under-confident for both factual and conceptual questions. In the re-study passages condition, subjects were overconfident on conceptual questions, but relatively well-calibrated on factual questions (i.e. only slight overconfidence). The lack of overconfidence for factual questions in the re-study condition is odd given that subjects displayed similar degrees of overconfidence on the factual and conceptual questions in Experiment 1.

Table 9

Mean confidence judgment, mean proportion of correct responses, and mean difference between confidence and proportion correct on the final cued recall tests as a function of question type and initial learning condition for Experiment 2.

Question Type	Learning Condition	Confidence Judgment	Proportion Correct	Difference (CJ – PC)
Factual	Same Test	.57	.60	- .03
	Variable Test	.55	.57	- .02
	Re-Study Passages	.32	.30	+ .02
Conceptual	Same Test	.58	.60	- .02
	Variable Test	.58	.64	- .06
	Re-Study Passages	.46	.36	+ .10

Note. Mean confidence judgments were converted to proportions for ease of comparison with the mean proportions of correct responses and statistical analyses.

Statistical analyses were performed on the difference scores (confidence minus proportion correct). The data could not be collapsed across question type like in Experiment 1, so the factual and conceptual results were analyzed separately with one-way (Initial Learning Condition: Re-Study Passages, Same Test, Variable Test) repeated measures ANOVAs. For factual questions, there was no significant difference among the means ($F < 1$). However, for conceptual questions, there was a significant main effect of initial learning condition [$F(2,46) = 6.50$, $MSE = .02$, $\eta^2 = .22$]. Pair-wise comparisons indicated that subjects were more overconfident on conceptual questions in the re-study passages condition relative to the same test condition [.10 vs. -.03: $t(23) = 2.87$, $SEM = .04$, $d = .62$] and variable test condition [.10 vs. -.04: $t(23) = 3.93$, $SEM = .04$, $d = .74$].

To assess resolution, gamma correlations between accuracy (i.e. correct vs. incorrect) and confidence judgments on the final cued recall test were computed for each of the initial learning conditions. As in Experiment 1, the mean gamma correlations were positive and relatively large, which means that overall subjects were capable of accurately distinguishing between correct and incorrect responses. The same test condition produced a slightly bigger mean gamma correlation ($M = .55$) than the re-study passages condition ($M = .51$), which in turn was bigger than variable test condition ($M = .45$). However, there was no significant difference among the initial learning conditions ($F < 1$) as determined by a one-way (Initial Learning Condition: Re-Study Passages, Same Test, Variable Test) repeated measures ANOVA.

Conditional Analyses. As in Experiment 1, conditional analyses were conducted to explore the relationship between performance on the initial tests and performance on the final test. The main question of interest was whether transfer on the final test was

dependent on successful retrieval of the fact or concept on the initial tests. Table 10 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).

When subjects successfully retrieved the fact or concept at least once on the initial tests, they produced a much greater proportion of correct responses on the final transfer test than when they were unsuccessful on all the initial tests. However, some transfer did occur even when subjects failed to produce the correct response on the initial tests. This result may be due to learning from the feedback on the last test in the initial learning session. Alternatively, subjects could have gained some partial knowledge about the fact or concept from repeated testing (but not enough to constitute a correct response on the initial test), and this partial knowledge allowed them to work out the correct response to the associated transfer question on the final test.

A 2 (Retrieval Success: Successful, Unsuccessful) x 2 (Initial Learning Condition: Same Test, Variable Test) repeated measures ANOVA was conducted to analyze the results of the conditional analysis. 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 therefore did not produce a mean for unsuccessful retrieval on the initial tests. 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 the initial learning session led to superior transfer on the final test. Neither the main effect of initial learning condition nor the interaction was significant (Fs < 1).

Table 10

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 Experiment 2.

Learning Condition	Retrieval Success on Initial Tests	Proportion Correct on Final Test
Same Test	Successful	.65
	Unsuccessful	.32
Variable Test	Successful	.64
	Unsuccessful	.28

Discussion

Experiment 2 replicated most of the results from Experiment 1, but also produced important new findings. As expected, the proportion of correct responses increased on each successive test during the learning session, forming the typical negatively accelerated learning curve found in multi-trial learning experiments. This pattern held for both testing conditions and for both types of questions, replicating Experiment 1. Again, subjects took longer on average to complete the conceptual questions than the factual questions, and spent more time taking each test in the testing conditions than studying each passage in the re-study passages condition. 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, a finding that has not been previously reported in the testing effect literature. However, the variable test condition did not produce better transfer than the same test condition. Conditional analyses showed that retrieving a fact or concept at least once on the initial tests substantially increased the probability of correctly answering the related inferential question on the final transfer test. Repeated testing improved subjects' calibration on the final test, whereas repeated studying of the passages led them to be overconfident, but only for conceptual questions. Numerically, repeated studying of passages again produced better performance than repeated testing on the control questions, but the effect was not significant.

The finding that repeated testing produced better transfer than repeated studying is novel, and therefore highly interesting. The vast majority of the previous studies on the testing effect have used a final test with questions that are repeated verbatim from the

initial tests (e.g., Butler & Roediger, 2008; Carrier & Pashler, 1992; Karpicke & Roediger, 2008). One potential criticism of these prior studies is that the mnemonic benefit of retrieval practice is limited to the retention of a specific response. In other words, it leaves open the question of whether retrieval practice promotes the acquisition of knowledge that can be transferred to new contexts. Thus, the results of Experiment 2 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. In addition, repeated testing produced better transfer of both factual and conceptual information. If it can be replicated in Experiments 3 and 4, this finding would represent an important step forward in the effort to promote the use of testing as a learning tool.

As in Experiment 1, retrieval on the initial tests was critical in determining whether transfer occurred on the final test. When subjects retrieved a fact or concept at least once in the initial learning session, they were much more likely to produce the correct response to the related inferential question on the final test than if they had failed to retrieve it on the initial tests. Thus, the retrieval of information from memory may be the key mechanism that is responsible for the differences in performance on the final transfer test. Of course, also as in Experiment 1, total time on task cannot be ruled out as a possibility. Subjects spent more time completing the tests in the repeated testing conditions than they spent studying the passages in the re-study passages condition. However, the same arguments that were made in the discussion of Experiment 1 apply to Experiment 2 as well. First, subjects again spent more time on the conceptual questions

than they did on the factual questions, but similar levels of transfer occurred for both types of item. Second, spending more time processing materials does not always lead to better retention. Indeed, compared to the results of Experiment 1, performance on the control questions in Experiment 2 showed an even smaller benefit of studying a passage four times relative to studying it once (and the difference was not significant by conventional standards).

Although it was possible that the variable test condition would produce better transfer than the same test condition, there was no difference in performance between the two testing conditions on the final test in Experiment 2. As discussed above, one potential reason why the encoding variability induced by the variable test condition did not produce superior retention in Experiment 1 was the nature of the final test (i.e. verbatim repetition of the practice questions). However, the final test in Experiment 2 consisted of new inferential questions, and thus should have presented a situation in which encoding variability might be expected to help. Still, it is possible that the greater variety of features encoded in the variable test condition did not provide a better match for the features present in the retrieval cues than the features encoded in the same test condition. Alternatively, the way in which the questions were re-phrased 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 re-phrasing manipulation only added a small degree of variability. The failure to find support for the encoding variability hypothesis will be further 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 3 and 4.

As a final consideration, the confidence judgment data yielded an odd result. Whereas repeated studying led to overconfidence for both factual and conceptual questions in Experiment 1, subjects only exhibited overconfidence on conceptual questions in the re-study passages condition in Experiment 2. Of course, a key difference between Experiments 1 and 2 is the type of questions on the final test. Repeated studying may lead to overconfidence on repeated factual questions, but relatively good calibration on new inferential questions; however, this explanation seems unlikely. If this result is replicated in Experiment 3, then further consideration will be given to explaining it. The other finding of note was that both testing conditions produced good calibration on the final test, replicating the results of Experiment 1.

Experiment 3

There were two main goals in conducting Experiment 3. The first goal was to replicate the novel finding from Experiment 2 that repeated testing led to better transfer relative to repeated studying of the passages. The second goal was to compare repeating testing with a more stringent control condition: repeated studying of the isolated facts and concepts. In this new re-study isolated sentences control condition, subjects were presented with the individual facts and concepts and told to study them in anticipation of the final test (for a similar procedure see Butler & Roediger, 2008). Thus, the information processed in the re-study 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. The standard repeated re-study passages

condition of Experiments 1 and 2 was also included. The design, materials, and procedure were the same as in Experiment 2, except that the variable test condition was dropped in order to include the re-study isolated sentences condition.

Method

Subjects & Design. Twenty-four undergraduate psychology students at Washington University in St. Louis participated for course credit or pay. The design was a 3 (Type of Initial Learning: Re-Study Passages, Re-Study Isolated Sentences, Same Test) x 2 (Type of Initial Test Question: Factual, Conceptual) within-subjects design. Both variables were manipulated within-subjects, but between-materials. As in Experiment 2, the main dependent variable was new inferential questions on the final transfer test.

Materials & Counterbalancing. The materials from Experiment 2 were used.

Procedure. The procedure was the same as in Experiment 2 with the exception that the variable test condition was replaced by the re-study isolated sentences condition. In the re-study isolated sentences condition, subjects studied each fact and concept for 30 seconds. There were a total of four facts and four concepts per passage, so the re-study isolated sentences condition and re-study passages conditions were equated in terms of total time on task (4 minutes).

Results

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

Initial Tests. Table 11 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 1 and 2. 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 item was computed for the same test condition (see Appendix C for full data). As in the previous experiments, subjects spent more time on conceptual questions than on factual questions [42.1 vs. 26.2: $t(23) = 7.84$, $SEM = 2.03$, $d = 1.12$]. Total time on task was again calculated for each initial learning condition. On average, subjects spent 273 seconds (4.6 minutes) to complete a test on each passage, which was slightly more time than the 240 seconds (4.0 minutes) that they spent per passage in the re-study conditions. A one-way (Initial Learning Condition: Re-Study Passages, Re-Study Isolated Sentences, Same Test) repeated measures ANOVA showed a significant difference among conditions [$F(2,46) = 3.62$, $MSE = 2450.76$, $\eta^2 = .14$]. However, follow up pair-wise comparisons only yielded marginally significant differences between the same test condition and the re-study passages and re-study isolated sentences conditions [273 vs. 240: $t(23) = 1.90$, $SEM = 17.50$, $p = .07$; the results were the same for both comparisons].

Table 11

Mean proportion of correct responses on the three initial cued recall tests in the same test condition as a function of question type for Experiment 3.

Question Type	Test 1	Test 2	Test 3
Factual	.43	.73	.88
Conceptual	.39	.67	.77

Final Test. Figure 5 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 re-study conditions, and this pattern held for both factual and conceptual questions. Interestingly, re-studying the isolated facts and concepts did not lead to better transfer relative to re-studying the entire passage. This result provides additional evidence against the idea that differences in total time on task produced differential final test performance; subjects presumably spent more time processing each fact and concept in the re-study isolated sentences condition than in the re-study 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: Re-Study Passages, Re-Study Isolated Sentences, Same Test) repeated measures ANOVAs. There was a significant main effect of initial learning condition for factual inferential questions [$F(2,46) = 10.21$, $MSE = .03$, $\eta^2 = .31$]. Pair-wise comparisons showed that same test condition led to significantly higher final test performance than the re-study passages condition [.53 vs. .31: $t(23) = 5.74$, SEM = .05, $d = 1.03$] and the re-study isolated sentences condition [.53 vs. .33: $t(23) = 3.22$, $SEM = .06$, $d = .85$]. There was no significant difference between the two re-study conditions ($t < 1$). For the conceptual inferential questions, there was also a significant main effect of initial learning condition [$F(2,46) = 4.13$, $MSE = .05$, $\eta^2 = .15$]. Pair-wise comparisons confirmed that the same test condition produced significantly better transfer than the re-study passages condition [.58 vs. .41: $t(23) = 2.27$, $SEM = .06$, $d = .63$] and the re-study isolated sentences condition [.58 vs. .44: $t(23) = 2.61$, $SEM = .07$, $d = .54$]. Again, two re-study conditions did not differ significantly ($t < 1$).

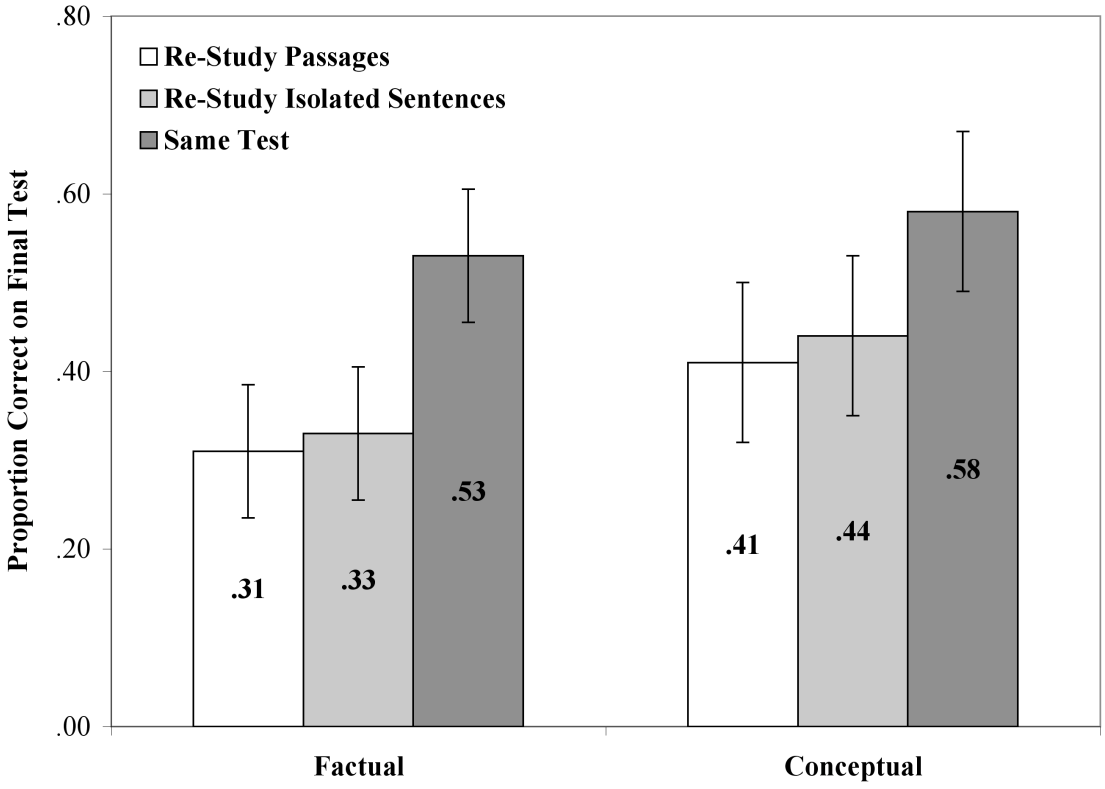

Figure 5. Mean proportion of correct responses on the final cued recall test as a function of question type and initial learning condition for Experiment 3. Error bars represent 95% confidence intervals.

Final test performance in the three initial learning conditions (collapsed across question type) was also analyzed with an ANCOVA that included total time on task as a covariate. The same test condition led to significantly better performance on the final test than the re-study passages and re-study isolated sentences conditions even after controlling for differences in total time on task.

The control questions on the final test were also analyzed to determine whether testing might benefit other (untested) material from the same passage. The re-study 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 1 and 2. The re-study passages condition also led to better performance than the re-study isolated sentences condition (M = .20); this result makes sense because in the latter condition subjects were only re-exposed 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 re-study passages condition, a one-way (Initial Learning Condition: Re-Study Passages, Re-Study Isolated Sentences, Same Test) repeated measures ANOVA did not show any significant differences among the conditions [$F(2,46) = 1.76$, MSE = .04, p = .19].

Much like Experiment 2, this null result may have been due to insufficient power. The results of the power analysis conducted for the control questions in Experiment 2 suggest that Experiment 3 also lacked sufficient power to detect an advantage of the re-studying passages condition over the other two conditions (see pages 62-63). To further address this issue, an additional analysis was performed on the data from Experiments 1-3 to compare performance on the control questions in the re-study passages and same test

conditions. This analysis yielded a significant result: re-studying the passages produced better performance relative to repeated testing [.35 vs. .23: $t(71) = 3.69$, $SEM = .03$, $d = .50$].

Confidence. The confidence judgments on the final test were used to assess subjects' success in metacognitive monitoring. Table 12 shows the mean confidence judgments and the mean proportions of correct responses on the final cued recall tests as a function of question type and initial learning condition. Subjects were overconfident on both types of questions in all three initial learning conditions. Given the results of Experiments 1 and 2, the overconfidence in the two re-study conditions was expected. However, the overconfidence in the same test condition is odd given that repeated testing led to relatively good calibration in Experiments 1 and 2.

Statistical analyses were performed on the difference scores (confidence minus proportion correct). Since the pattern of performance was the same for factual and conceptual questions, the data were collapsed across questions type for the purpose of statistical analysis (as in Experiment 1). A one-way (Initial Learning Condition: Re-Study Passages, Re-Study Isolated Sentences, Same Test) repeated measures ANOVA confirmed that the main effect was not significant [$F(2,46) = 1.29$, $MSE = .02$, $p = .28$], indicating that subjects were no more overconfident in the two re-study conditions than in the same test condition.

The confidence judgments for the control questions in Experiments 1-3 were also analyzed. However, none of the effects of encoding condition approached significance (all ps $> .10$), and there were no systematic patterns across the experiments. Due to the

low power in these analyses (only 2 observations per passage per participant), one must be cautious in interpreting any null effects.

Resolution was assessed via gamma correlations between accuracy (i.e. correct vs. incorrect) and confidence judgments on the final cued recall test. The re-study isolated sentences condition produced the highest gamma correlation (M = .54), followed by the same test condition (M = .50) and the re-study passages condition (M = .42), respectively. However, a one-way (Initial Learning Condition: Re-Study Passages, Re-Study Isolated Sentences, Same Test) repeated measures ANOVA did not reveal a significant difference among the means ($F < 1$).

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). As in Experiments 1 and 2, 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 vs. .31: $t(16) = 3.26$, SEM = .08, d = .97]. 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 therefore did not produce a mean for unsuccessful retrieval on the initial tests.

Table 12

Mean confidence judgment, mean proportion of correct responses, and mean difference between confidence and proportion correct on the final cued recall tests as a function of question type and initial learning condition for Experiment 3.

Question Type	Learning Condition	Confidence Judgment	Proportion Correct	Difference (CJ – PC)
Factual	Same Test	.64	.53	+ .11
	Re-Study Passages	.47	.31	+ .16
	Re-Study Isolated Sentences	.50	.33	+ .17
Conceptual	Same Test	.67	.58	+ .09
	Re-Study Passages	.53	.41	+ .12
	Re-Study Isolated Sentences	.59	.44	+ .15

Note. Mean confidence judgments were converted to proportions for ease of comparison with the mean proportions of correct responses and statistical analyses.

Discussion

Experiment 3 replicated and extended the key findings of Experiment 2 by incorporating a more stringent control condition. As in both the previous experiments, performance increased on each successive test in a curvilinear fashion for both factual and conceptual questions. The response time results also replicated both previous experiments: subjects took more time to complete the conceptual questions than the factual questions, even though they later recalled them no better. They also spent more total time on task in the same test condition than they did in the two re-study conditions. On the final test, repeated testing led to better performance than both repeated study of the passages (replicating Experiment 2) and repeated studying of the isolated facts and concepts. The latter two conditions did not differ. Conditional analyses again indicated that a much greater proportion of correct responses were produced on final transfer test when the related fact or concept had been successfully retrieved at least once on the initial tests. Unlike the previous experiments, all three initial learning conditions led to overconfidence on the final transfer test. 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 3 was that repeated testing produced better transfer than both repeated studying of the passages and repeated studying of the isolated facts and concepts. Any novel 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 2 could be obtained again. In addition, the

comparison of the same test and re-study isolated sentences conditions provided a more stringent assessment of whether retrieval might be the critical mechanism that produced the superior transfer in Experiment 2. The re-study 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 re-exposed 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 re-study 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.

The inclusion of the re-study isolated sentences control condition also helps to refute the idea that the differences in final test performance resulted from differences in the total time spent on task during the initial learning session. In Experiment 3, subjects did spend more time taking the tests than they did re-studying the passages or re-studying the facts and concepts. However, this difference was mainly due to the large amount of time spent completing the conceptual questions in the same test condition. Subjects spent approximately the same amount of time completing the factual questions in the same test condition (26.2 seconds) as they did studying the facts in the re-study isolated sentences condition (30.0 seconds). If the total time explanation is correct, then there should be no difference in final test performance between these two conditions. Of course, repeated testing produced substantially more transfer on the inferential questions related to the

facts than repeatedly studying the isolated facts. In addition, the re-study isolated sentences condition did not lead to better transfer than the re-study passages condition for either type of question, even though subjects presumably spent more time processing each fact and concept in the former condition. When the results of Experiment 3 are combined with the arguments put forth above in the discussion of Experiments 1 and 2, the total time explanation becomes untenable.

Experiment 3 yielded an odd result with respect to the confidence judgments given on the final test. Whereas repeated testing led to relatively good calibration in Experiments 1 and 2, subjects were overconfident for both types of question in the same test condition in Experiment 3. The method used for Experiment 3 was highly similar to Experiment 2; the major difference was that the variable test condition was replaced with the re-study isolated sentences condition. One possibility is that the switch from two testing conditions and one re-study condition to one testing condition and two re-study condition had some sort of global effect on subjects' metacognitive monitoring that increased confidence judgments. Indeed, collapsing across all conditions, the mean confidence judgment increased from Experiment 2 (M = .51) to Experiment 3 (M = .57). Nevertheless, this result needs to be replicated before any major conclusions are drawn. As a final note, subjects showed a high degree of overconfidence on factual questions in the re-study passages condition in Experiment 3, which indicates that the odd finding reported in Experiment 2 is probably an aberration.

Experiment 4

As discussed in the Introduction, far transfer is difficult to obtain in both laboratory and applied studies, but it is very important to understand (see Barnett & Ceci, 2002; Perkins & Grotzer, 1997). Indeed, Detterman (1993) has argued that experimental investigations of transfer should be considered trivial unless they demonstrate far transfer, and his criterion for far transfer essentially requires far transfer along multiple dimensions in the Barnett and Ceci's (2002) framework. 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 et al., 1986; Kosonen & Winne, 1995). In contrast with Detterman's (1993) criterion, the main goal of Experiment 4 was relatively modest: to explore whether retrieval practice could be used to promote far transfer along a single dimension in Barnett and Ceci's (2002) 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 Experiments 1-3, 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 is was difficult or impossible to find a related fact in a different domain for many items. In Experiments 1-3, 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 Experiment 1). The paragraphs in the passages that contained the critical facts were dropped, 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 & Design. Twenty undergraduate psychology students at Washington University in St. Louis participated for course credit or pay. The sole independent variable was type of initial learning (Re-Study Passages, Same Test), which was manipulated within-subjects, but between-materials. The main dependent variable was new inferential questions within different knowledge domains.

Materials & Counterbalancing. The materials from Experiments 1-3 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 1000 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 re-studied 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; Brown, 1989). 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 et al., 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 Experiments 1-3 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 re-studied 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 inter-rater reliability was high ($\kappa = .91$).

Initial Tests. As in Experiments 1-3, 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. 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 [$F(1,19) = 58.41$, $MSE = .02$, $\eta^2 = .76$].

Response Times. On average, subjects spent 52.5 seconds on each conceptual question in the same test condition (see Appendix C for full response time data). In terms of the average total time spent per passage, subject spent 210 seconds (3.5 minutes) completing each test for a passage, which was significantly more time than the 120 seconds (2.0 minutes) that they spent re-studying each passage in the re-study passages condition [210 vs. 120: $t(19) = 5.44$, $SEM = 16.53$, $d = 1.31$].

Final Test. Figure 6 shows the proportion of correct responses on the final cued recall test as a function of initial learning condition. The same test condition produced substantially better transfer relative to the re-study passages condition, and this observation was confirmed by a paired samples *t*-test [.68 vs. .44: $t(19) = 5.23$, $SEM = .05$, $d = .99$]. Final test performance in the two initial learning conditions was also analyzed with an ANCOVA that included total time on task as a covariate. The same test condition led to significantly better performance on the final test than the re-study passages condition even after controlling for differences in total time on task.

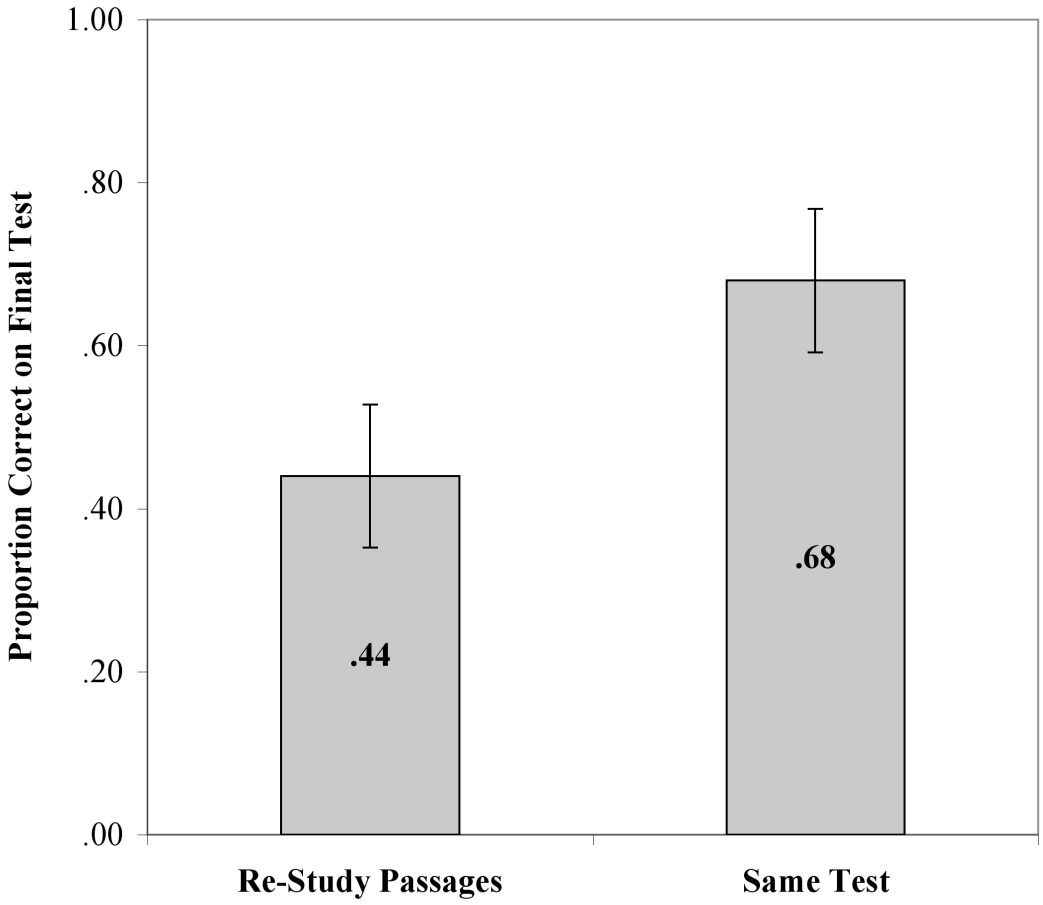

Figure 6. Mean proportion of correct responses on the final cued recall test as a function of initial learning condition for Experiment 4. Error bars represent 95% confidence intervals.

Confidence. Subjects' ability to subjectively assess the accuracy of their responses was examined through calibration and resolution. Table 13 shows the mean confidence judgment and mean proportion of correct responses on the final cued recall tests as a function of question type and initial learning condition. Statistical analyses were performed on the difference scores (confidence minus proportion correct). Subjects were well-calibrated for items in the same test condition, but overconfident in the re-study passages condition [.13 vs. -.03: $t(19) = 3.36$, $SEM = .05$, $d = .77$].

Gamma correlations between accuracy (i.e. correct vs. incorrect) and confidence judgments on the final cued recall test were computed to assess resolution. Unlike the previous three experiments in which there was no significant difference in resolution among the various initial learning conditions, the same test condition produced a higher mean gamma correlation than the re-study passages condition [.69 vs. .39: $t(20) = 3.61$, $SEM = .08$, $d = .98$]. One subject was excluded from this analysis because a gamma correlation could not be calculated for one or more of the initial learning conditions.

Conditional Analyses. Conditional analyses were conducted to examine whether final test performance was correlated with initial test performance. 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 at all [.72 vs. .49: $t(16) = 2.61$, $SEM = .09$, $d = .73$]. 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 therefore did not produce a mean for unsuccessful retrieval on the initial tests.

Table 13

Mean confidence judgment, mean proportion of correct responses, and mean difference between confidence and proportion correct on the final cued recall tests as a function of initial learning condition for Experiment 4.

Learning Condition	Confidence Judgment	Proportion Correct	Difference (CJ – PC)
Same Test	.66	.68	- .02
Re-Study Passages	.57	.44	+ .13

Note. Mean confidence judgments were converted to proportions for ease of comparison with the mean proportions of correct responses and statistical analyses.

Discussion

The results of Experiment 4 replicated many of the findings of Experiment 1-3, 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 4 will be 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 1 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 despite the possibility that inducing encoding variability during initial testing would produce better retention. Experiment 2 built upon Experiment 1 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 3 replicated Experiment 2 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 4 extended the findings of Experiments 2 and 3 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 and considering their significance within the broader memory literature. First, I will examine why the retrieval of information from memory produced superior transfer by discussing some possible theoretical explanations for this novel finding. Second, I will re-assess the encoding variability hypothesis in light of the results of Experiments 1 and 2, and discuss three possible explanations for the failure to find support for this idea. Finally, I will 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 considerable testing effect literature continues to expand with new studies that demonstrate the mnemonic benefits of retrieval practice (for review see Roediger & Karpicke, 2006a). Experiment 1 provides a conceptual replication of these studies, which generally show that testing produces superior retention relative to additional studying (or no activity at all) as measured by performance on a final test that consists of questions that are repeated verbatim from the initial test (e.g., Butler & Roediger, 2007; Carrier &

Pashler, 1992; Karpicke & Roediger, 2008). In addition, the finding that testing benefited the retention of concepts (as well as facts) in Experiment 1 is novel and important. Most previous testing effect studies have used relatively discrete factual information as the to-be-remembered materials, and so this finding demonstrates that the same mnemonic benefit of testing holds for more complex information like concepts. More broadly, this finding provides further evidence that retrieval is a mechanism for promoting retention of many types of information; if something can be successfully retrieved from memory, it will be better retained.

The most important finding that emerged from the present research was 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; Toppino & Cohen, in press) and establishing its generalizability to various materials (Carpenter & Pashler, 2007; Kang, in press) and applied contexts (e.g., Larsen et al., in press; 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.

In addition to the current research, there are two recent testing effect studies that have also included a final test with questions that are substantially different from those given on the initial test(s). Experiments 2 and 3 in the present research showed that

repeated testing produced better transfer to new inferential questions within the same knowledge domain than either repeatedly studying passages or repeatedly studying isolated facts and concepts. Two other studies have investigated whether retrieving information from memory promotes transfer within the same knowledge domain. McDaniel, Howard, and Einstein (2009, Experiment 2) had subjects use one of three study strategies while reading complex passages that described mechanical devices: 1) read the passage, attempt to recall it from memory, and then re-read the passage, 2) read the passage twice, or 3) read the passage twice and take notes while reading. On a final test one-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).

In an unpublished study, Marsh, Bjork, and Bjork (2009) had subjects study definitions of various scientific concepts (e.g., *ACCLIMATION is the slow adaptation of an organism to new conditions*). Next, subjects were given multiple-choice test for some of the concepts, while other concepts were not tested. On the multiple-choice test, the questions either presented the definition of the concept (e.g., *What is the name for the slow adjustment of an organism to new conditions?*) or an application of the concept (e.g., *What biological term describes fish slowly adjusting to water temperature in a new tank?*), but the concept name was the correct response for both types of question. Finally, subjects received a final cued recall test after a short delay on which each concept was tested with either a definition or application question. The results of the experiment showed successful transfer: taking the initial test with definition questions led to better

performance on the final test with application questions relative to not taking an initial test. However, unlike the present research, this experiment did not include a re-study control condition. Overall, initial testing with either type of question produced better performance on both types of final test question relative to not taking an initial test.

The present research included an experiment in which the difference between the initial learning and subsequent transfer contexts was extended farther than in any previous testing effect study. Experiments 4 showed that repeated testing produced better transfer to new inferential questions in different knowledge domains relative to repeatedly studying passages. 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; Perkins & Grotzer, 1997).

The finding that retrieval practice promotes superior transfer across knowledge domains is relatively novel because 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 1-3, 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 2 and 3. However, when the control question data were collapsed across Experiments 1-3, performance was significantly higher in the re-study 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 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 re-study the passages. The potential for lack of effort during re-study 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 re-studying 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 re-study 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 retention and transfer than repeated studying? 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 re-studying the material does not. Thus, this explanation for the testing effect incorporates the concept of encoding variability (Bower, 1972; Estes, 1955; Martin, 1968), which will be 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 et al., 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; Kang et al., 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 have 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 (e.g., Morris et al., 1977; Roediger, 1990). 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 re-studying 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 (Wheeler, Ewers, & Buonanno, 2003; Runquist, 1983).

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: 1) *storage strength*, which reflects how well the item is learned; and 2) *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' (1955) stimulus sampling theory. The "new theory of disuse" also incorporates his idea of stimulus fluctuation that motivated later encoding variability theories (J. R. Anderson & Bower, 1972; 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; Kang et al., 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 re-studying 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, in press; 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 re-studying the items, but only for the items that are successfully retrieved; re-studying produces smaller increases in storage and retrieval strength for all the items. If retention is assessed immediately, re-studying 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 re-studied items will have decreased faster than the retrieval strength of the tested items, resulting in testing producing superior performance relative to re-studying on the final test.

Although the aforementioned theories do not specifically address whether retrieval practice would be expected to promote superior transfer, they can be used to explain the results of Experiments 2-4 because retention is critical to the transfer process. As Barnett and Ceci (2002) have 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 re-studying, 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 re-studying it. For example, McDaniel et al. (2009) argued that retrieval practice promotes deep learning of the material more than re-studying 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.

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, 2001; Kulhavy, 1977; Kulhavy & Stock, 1989) and maintain correct responses (Butler et al., 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., Kane & Anderson, 1978; Slamecka & Fevreiski, 1983), a finding that is sometimes referred to as "test potentiation" (Izawa, 1967, 1970). For example, Kornell, Hays, and Bjork (2009, Experiment 4) had subjects either study weakly associated word pairs or try to guess the target word when given the cue word (which almost always resulted in unsuccessful "retrieval") and then receive feedback. On a subsequent cued recall test, subjects remembered more of the words in the test condition relative to the study condition despite their failure to produce the correct target during the initial learning test. 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 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 et al. (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, in press). 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, four findings argue against total time on task as an explanation. First, subjects generally spent more time on the conceptual questions than they did on the factual questions, but similar levels of retention and transfer occurred for both types of item. Second, 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. Third, subjects spent more time processing the critical facts and concepts in the re-study isolated sentences condition of Experiment 3 than they did in the re-study passages condition, and yet these two re-study conditions yielded equivalent performance on the final transfer test. Fourth, and

most importantly, subjects spent about the same amount of time completing the factual questions in the same test condition in Experiment 3 as they did studying the facts in the re-study isolated sentences condition, yet repeated testing produced substantially more transfer on the inferential questions related to the facts than repeated studying of the facts. 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

The second major goal of the present research was to explore whether repeated testing using re-phrased 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, this hypothesis was not supported: repeated testing with different versions of a question did not lead to better retention on a final test with repeated questions (Experiment 1) or transfer on a final test with new inferential questions (Experiment 2) than repeated testing with the same version of the question. Rather, the variable test condition and the same test condition produced almost identical final performance for both factual and conceptual questions in Experiments 1 and 2.

There are at least three possible explanations for the lack of support for the encoding variability hypothesis. First, the way in which the questions were re-phrased in the variable test condition may not have been sufficient to induce encoding variability. Despite efforts to pose questions in a different manner in each of the three versions, the

differences among the versions were largely superficial. That is, the differences among the versions were in the specific wording rather than some sort of deeper meaning. In addition, the decision to keep the correct response the same for each re-phrased version of the question may have also diminished the potential for variability in encoding.

Another possible explanation is that the re-phrasing of questions succeeded in inducing encoding variability, but the amount of variability produced by this manipulation was small relative to the amount of variability produced by other factors. The spaced presentation of the questions, the random ordering of questions within a given test (i.e. presentation in different contexts), and other aspects of the experimental procedure would be expected to induce substantial encoding variability. Thus, any effects of the re-phrasing manipulation may have been masked by the large degree of encoding variability induced in both testing conditions.

A third potential explanation is that the nature of the final tests used in Experiments 1 and 2 was such that any encoding variability induced during initial testing would not be expected to enhance performance. In Experiment 1, the final test consisted of a verbatim re-presentation of the version of the initial test questions that was presented three times in the same test condition, but only once in the variable test condition. Thus, these retrieval cues contained the same features, so any additional features encoded or retrieval routes created in variable test condition might not be expected to enhance the match between encoding and retrieval (but see Goode et al., 2007). In Experiment 2, the final test consisted of new inferential questions, and thus these retrieval cues contained a number of features that differed from those present in the initial test questions. Ostensibly, the presence of different features in the inferential questions presents a

situation in which encoding variability might be expected to help. However, it is possible that the additional features encoded in the variable test condition did not match the features present in the new retrieval cues any better than the features encoded in the same test condition.

Although the results of Experiments 1 and 2 did not support the encoding variability hypothesis, they did 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). Additional research will be needed to further investigate whether inducing encoding variability during repeated testing can help to promote retention and transfer. One improvement that could be made in future experiments is 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), and providing greater specification can be difficult for complex materials such as those used in the present research. However, it is possible to specify in greater detail the features involved in encoding variability (e.g., Glenberg, 1979), and thus researchers should make an effort to include such specification in future studies. On the whole, encoding variability theories potentially 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 and that is not the primary goal of education or vocational training. The results obtained in this study and other recent investigations (e.g., Marsh et al., 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. In addition, repeatedly retrieving information from memory and generating a response may help people to better understand material.

Clearly, retrieval practice holds great potential for promoting learning in the classroom and workplace that will transfer to new situations. Taken in the context of the broader testing effect literature (see Roediger & Karpicke, 2006a), the findings reported in the present research present a strong case for a fundamental re-evaluation of how tests are used in education and vocational training. All too often the use of testing seems to be reverse-engineered to produce the least amount of learning possible. Testing practices in higher education are a prime example of such misuse. At the college and graduate school levels, educators commonly give very few tests (e.g., a mid-term and a final), use recognition tests (e.g., multiple-choice and true / false), and withhold feedback to protect

their test banks. Giving tests more frequently (e.g., a quiz after every class; Leeming, 2002), using production tests (which produce better retention than recognition tests; e.g., Butler & Roediger, 2007), and providing feedback are examples of potential changes that could drastically improve students' learning from tests in higher education. If testing can be viewed as a learning tool first and as an assessment tool second, the potential benefit to long-term retention and transfer in education and vocational training could be substantial.

Nevertheless, retrieval practice should not be thought of as a panacea for education or vocational training. Rather, it is a tool that educators and trainers must decide how to utilize depending the context and the goals for learning, much like studying, lecture, group discussion, and other tools that can help people to learn. Although repeated testing produced better retention that repeated studying in the present research, there are conditions under which studying would be preferable. For example, if a student has an exam in an hour, re-studying all of the material would probably produce better performance than self-testing on a subset of it. Thus, educators and trainers must determine how to use testing and other tools so as to maximize their effectiveness. If only a subset of the material can be tested, then it may be best to focus on the core concepts in the to-be-learned material. If testing can enhance learning and retention of these core concepts, it may help students to learn and retain related or subsequent material that build upon these core concepts.

Concluding Remarks

In a set of four experiments, the present research showed that repeated testing produced better retention and transfer relative to repeated studying of the material. These

findings have implications for future research on both transfer of learning and the testing effect. As discussed in the introduction, 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). Future research should attempt to dissociate these two components of the transfer process in

order to determine whether retrieval practice influences one or both of them. Finally, the idea of introducing encoding variability during repeated testing should also be further examined because theoretically it should enhance the mnemonic benefits of retrieval practice. However, as discussed above, greater specification of the features involved in the encoding variability manipulation will be needed in order to thoroughly assess this idea.

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.

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 8*, 463-470.

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.

Ammons, R. B. (1956). Effects of knowledge of performance: A survey and tentative theoretical formulation. *Journal of General Psychology, 54*, 279-299.

Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review, 79*, 97-123.

Anderson, M. C. (2003). Rethinking interference theory: Executive control and the mechanisms of forgetting. *Journal of Memory and Language, 49*, 415-445.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 20*, 1063-1087.

Asch, S. E., & Ebenholtz, S. M. (1962). The principle of associative symmetry. *Proceedings of the American Philosophical Society, 106*, 135-163.

Azevedo, R., & Bernard, R. M (1995). A meta-analysis of the effects of feedback in computer-based instruction. *Journal of Educational Computing Research, 13*, 111-128.

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

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

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.

Barclay, J. R., Bransford, J. D., Franks, J. J., McCarrell, N. S., & Nitsch, K. (1974). Comprehension and semantic flexibility. *Journal of Verbal Learning and Verbal Behavior, 13*, 471-481.

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*. New York: 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: Wiley.

Brown, A. L. (1989). Analogical learning and transfer: What develops? In S. Vosniadou & A. Ortony (Eds.), *Similarity and analogical reasoning* (pp. 369-412). New York: Cambridge University Press.

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

Bugelski, B. R., & Cadwallader, T. C. (1956). A reappraisal of the transfer and retroaction surface. *Journal of Experimental Psychology, 52*, 360-366.

Butler, A. C., Karpicke, J. D., & Roediger, H. L., III (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied, 13,* 273-281.

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.

Butler, D. L., & Winne, P. H. (2005). Feedback and self-regulated learning: A theoretical synthesis. *Review of Educational Research, 65,* 245-281.

Calkins, M. W. (1894). Association: I. *Psychological Review, 1*, 476-483.

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.

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

Chan, C. K., McDermott, K. B., & Roediger, H. L., III (2006). Retrieval induced facilitation: Initially nontested material can benefit from prior testing. *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: Lawrence Earlbaum Associates.

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.

Dallett, K. M. (1962). The transfer surface re-examined. *Journal of Verbal Learning & Verbal Behavior, 1*, 91-94.

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.

Dempster, F. N. (1989). Spacing effects and their implications for theory and practice. *Educational Psychology Review*, *1*, 309-330.

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

Ebbinghaus, H. (1964) *Memory: A contribution to experimental psychology* (H. A. Ruger & C. E. Bussenius, Trans.). New York: 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.

Gates, A. I. (1917). Recitation as a factor in memorizing. *Archives of Psychology, 6*(40).

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: Academic Press.

Gilman, D. A. (1969). Comparison of several feedback methods for correcting errors by computer-assisted instruction *Journal of Educational Psychology*, *60*, 503-508.

Glenberg, A. M. (1976). Monotonic and nonmonotonic lag effects in paired-associate and recognition memory paradigms. *Journal of Verbal Learning and Verbal Behavior, 15*, 1-16.

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.

Gulliksen, H. (1932). Transfer of response in human subjects. *Journal of Experimental Psychology, 15*, 496-516.

Hall, J. (1971). *Verbal learning and retention*. New York: J. B. Lippincott.

Halpern, D. F. (1998). Teaching critical thinking for transfer across domains. *American Psychologist, 53*, 449-455.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*, 81-112.

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). In R. L. Solso (Ed.), *Theories in cognitive psychology: The Loyola symposium* (pp. 77-99). Oxford: Lawrence Erlbaum.

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

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior, 10*, 562–567.

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

Izawa, C. (1967). Function of test trials in paired-associate learning. *Journal of Experimental Psychology, 75*, 194-209.

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. (1975). Physical features vs meaning: A difference in decay. *Memory & Cognition, 3*, 247-251.

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.

Jacoby, L. L. (1983). Remembering the data: Analyzing interactive processes in reading. *Journal of Verbal Learning & Verbal Behavior, 22*, 485-508.

James, W. (1890). *The principles of psychology*. New York: Holt.

Jones, H. E. (1923-1924). The effects of examination n the performance of learning. *Archives of Psychology*, *10*, 1-70.

Judd, C. H. (1908). The relation of special training to general intelligence. *Educational Review, 36*, 28-42.

Kahana, M. J. (2002). Associative symmetry and memory theory. *Memory & Cognition, 30*, 823-840.

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.

Kane, J. H., & Anderson, R. C. (1978). Depth of processing and interference effects in the learning and remembering of sentences. *Journal of Educational Psychology, 70*, 626-635.

Kang, S. H. K. (in press). Enhancing visuo-spatial learning: The benefit of retrieval practice.

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

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

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

Kline, L. W. (1914). Some experimental evidence in regard to formal discipline. *Journal of Educational Psychology, 5*, 259-266.

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

Koriat, A., & Goldsmith, M. (1996). Monitoring and control processes in strategic regulation of memory accuracy. *Psychological Review*, *103*, 490-517.

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.

Kulhavy, R. W. (1977). Feedback in written instruction. *Review of Educational Research, 47,* 211-232.

Kulhavy, R. W., & Anderson, R. C. (1972). Delay-retention effect with multiple-choice tests. *Journal of Educational Psychology, 63*, 505-512.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review, 1,* 279-308.

Kulik, J. A., & Kulik, C. C. (1988). Timing of feedback and verbal learning. *Review of Educational Research, 58*, 79-97.

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

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

Lenth, R. V. (2006-9). Java applets for power and sample size [Computer software]. Retrieved 7/16/09 from http://www.stat.uiowa.edu/~rlenth/power.

Leuba, J. H., & Hyde, W. (1905). An experiment in learning to make hand movements. *Psychological Review, 12*, 351-369.

MacLeod, M. D., & Macrae, C. N. (2001). Gone but not forgotten: The transient nature of retrieval-induced forgetting. *Psychological Science, 12*, 148-152.

Marsh, E. J., Roediger, H. L., III, Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review, 14*, 194-199.

Marsh, E. J., Bjork, R. A., & Bjork, E. L. (2009). Transfer of positive and negative effects of testing. Manuscript in preparation.

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

Martin, M. A. (1915). The transfer effects of practice on cancellation tests. *Archives of Psychology, 4*, No. 32.

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., Friedman, A., & Bourne, L. E. (1978). Remembering the levels of information in words. *Memory & Cognition, 6*, 156-164.

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.

McDaniel, M. A., & Sun, J. (submitted). The testing effect: Experimental evidence in a college course.

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

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

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

McKinney, F. (1933). Quantitative and qualitative essential elements of transfer. *Journal of Experimental Psychology, 16*, 854-864.

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior, 9*, 596–606.

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

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

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.

Mory, E. H. (2004). Feedback research review. In D. Jonassen (Ed.), *Handbook of research on educational communications and technology* (pp. 745-783). Mahwah, NJ: Erlbaum.

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.

Nelson, T. O. (1984). A comparison of current measures of the accuracy of feeling-of-knowing predictions. *Psychological Bulletin*, *95*, 109-133.

Nelson, T. O., & Dunlosky, J. (1991). When people's judgments of learning (JOLs) are extremely accurate at predicting subsequent recall: The "delayed-JOL-effect." *Psychological Science*, *5*, 207-213.

Osgood, C. E. (1949). The similarity paradox in human learning: a resolution. *Psychological Review, 56*, 132-143.

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

Pashler, H., Rohrer, D., Cepeda, N. J., & Carpenter, S. K. (2007). Enhancing learning and retarding forgetting: Choices and consequences. *Psychonomic Bulletin & Review, 14*, 187-193.

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

Perkins, D. N., & Grotzer, T. A. (1997). Teaching intelligence. *American Psychologist, 52*, 1125–1133.

Postman, L., & Knecht, K. (1983). Encoding variability and retention. *Journal of Verbal Learning & 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.

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

Reed, H. B. (1917). A repetition of Ebert and Meumann's practice experiment on memory. *Journal of Experimental Psychology, 2*, 315-346.

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

Roark, R. N. (1895). *Psychology in education*. New York: American Book.

Roediger, H. L., III (1985). Remembering Ebbinghaus. *Contemporary Psychology, 30,* 519-523.

Roediger, H. L., III (1990). Implicit memory: Retention without remembering. *American Psychologist, 45*, 1043-1056.

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.

Roediger, H. L., III, Weldon, M. S., & Challis, B. H. (1989). Explaining dissociations between implicit and explicit measures of retention: A processing account. In H. L. Roediger, III, & F. I. M. Craik (Eds.), *Varieties of memory and consciousness: Essays in honor of Endel Tulving* (pp. 3-41). Hillsdale, NJ: Erlbaum.

Roper, W. J. (1977). Feedback in computer assisted instruction. *Programmed Learning and Educational Technology, 14*, 43-49.

Ruediger, W. C. (1908). The indirect improvement of mental function thru ideals. *Education Review, 36*, 364-371.

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

Ryan, J. J. (1960). Comparison of verbal response transfer mediated by meaningfully similar and associated stimuli. *Journal of Experimental Psychology, 60*, 408-415.

Saunders, J., & MacLeod, M .D. (2002). New evidence on the suggestibility of memory: The role of retrieval-induced forgetting in misinformation effects. *Journal of Experimental Psychology: Applied, 8*, 127-142.

Schneider, W., Eschman, A., & Zuccolotto, A. (2002). *E-Prime Reference Guide*. Pittsburgh: Psychology Software Tools, Inc.

Skinner, B. F. (1954). The science of learning and the art of teaching. *Harvard Educational Review, 24*, 86-97.

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

Sleight, W. G. (1911). Memory and Formal Training. *British Journal of Psychology, 4*, 386-457.

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

Starch, D. (1911). Transfer of training in arithmetical operations. *Journal of Educational Psychology, 2*, 306-310.

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. (1913). *Educational psychology. Vol. 2: The psychology of learning*. New York: Teachers College Press.

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. (in press). The testing effect and the retention interval: Questions and answers. *Experimental Psychology*.

Tulving, E. (1967). The effects of presentation and recall of material in free-recall learning. *Journal of Verbal Learning and Verbal Behavior, 6*, 175-184.

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

Tulving, E., & Pearlstone, Z. (1966). Availability versus accessibility of information in memory for words. *Journal of Verbal Learning and Verbal Behavior, 5*, 381-391.

Tulving, E., & Thomson, D. M. (1973). Encoding specificity and retrieval processes in episodic memory. *Psychological Review, 80*, 352-373.

Webb, L. W. (1917). Transfer of training and retroaction: A comparative study. *Psychological Monographs, 24*, 1-90.

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

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.

Williams, W. M., Blythe, T., White, N., Li, J., Sternberg, R. J., & Gardner, H. (1996). *Practical intelligence for school*. New York: Harper Collins.

Wimer, R. (1964). Osgood's transfer surface: Extension and test. *Journal of Verbal Learning & Verbal Behavior, 3*, 274-279.

Winch, W. H. (1908). The Transfer of Improvement in Memory in School-Children. *British Journal of Psychology, 2*, 284-293.

Yum, K. (1931). An experimental test of the law of assimilation. *Journal of Experimental Psychology, 14*, 68-82.

Appendix A

Prose passages used in Experiments 1, 2, and 3.

BATS

Bats really stand out in the animal world. They are the only mammals that can fly, and they live much of their lives hanging upside down. Most species are active only from dusk until dawn, spending their days in dark caves. Many bats have developed adaptations that let them find their way (and their prey) in complete darkness. Bats have survived as a group for more than 50 million years, longer than most other modern animals. All bat species are part of an order called *Chiroptera*, which comes from the Greek words *cheir* ("hand") and *pteron* ("wing"). There are more than 1,000 bat species in the world, making them one of the most prevalent orders of mammals.

Traditionally, bat species are divided into two suborders: *Megachiroptera* (megabats) and *Microchiroptera* (microbats). Most megabat species are frugivores (fruit eaters) or nectavores (nectar drinkers) and look a lot like other mammals, with large eyes, small ears, and extended snouts. In contrast, most microbat species are insectivores and have a unique facial appearance, with large ears and peculiarly shaped, stubby snouts. While megabats have good eyesight, microbats use echolocation for navigation and finding prey. Also, the two suborders differ in terms of where they live: megabats are found only in Africa, Asia, and Australia, whereas microbats live all over the world. Although most scientists agree that the division of bat species into two suborders is a useful heuristic, the phylogenetic relationship among the different groups of bats has been the subject of much debate.

Although bats and birds both fly, a bat wing actually has more in common with a human arm than a bird wing. A bird's wing has fairly rigid bone structure, and the main flying muscles move the bones at the point where the wing connects to the body. In contrast, a bat has a much more flexible wing structure. It is similar to a human arm and hand, except it has a thin membrane of skin (called the *patagium*) extending between the "hand" and the body, and between each finger bone. Bats can use the wing like a hand, essentially moving through the air like a swimmer moves through water. The rigid bird wing is more efficient at providing lift, but the flexible bat wing allows for greater maneuverability.

To help them navigate and find their prey in the dark, microbat species have developed a remarkable system called echolocation. By emitting high-pitched sound waves and listening to the echoes, bats can determine with great precision the location of an object, how big it is, and the direction in which it is moving. Bats calculate the distance of the object by the amount of time it takes for the sound wave to return and the exact position of the object by comparing when the sound reaches its right ear to when the sound reaches its left ear. Similarly, a bat can tell how big an insect is based on the intensity of the echo: a smaller object will reflect less of the sound wave, and so will produce a less intense echo.

Although they hunt all night, bats will pass the daylight hours hanging upside down from a secluded spot, such as a cave or a hollowed-out tree. There are a couple of different reasons why bats roost this way. First of all, hanging upside down puts them in

position for takeoff, which is important because bats cannot launch themselves into the air from the ground. It is also a great way to hide from danger. During the hours when most predators are active, bats congregate where few animals look and most cannot reach. Although snakes, possums, and raccoons sometimes hunt bats, birds of prey are the main predator of bats. Most bat species roost in the same location every day, clustering with other bats for warmth and security.

Bats have a special physiological adaptation that enables them to hang upside down. A bat's talons work like human fingers, except that humans must contract muscles to grasp an object, whereas bats must do the opposite – relax their muscles. When humans grasp an object, they contract several arm muscles, which in turn pull tendons connected to their fingers, which pull the fingers closed. To hang upside down, a bat opens its talons to grab hold of the surface, and then simply lets its body relax. The weight of the upper body pulls down on the tendons connected to the talons, causing them to clench. Since it is gravity that keeps the talons closed, instead of a contracted muscle, the bat doesn't have to exert any energy to hang upside down.

Like all mammals, bats maintain their body temperature internally. However, unlike most mammals, bats allow their body temperature to sink to the ambient temperature whenever they are not active. As their temperature drops, they enter a torpor state, in which their metabolism slows down considerably. By reducing their biological activity and not maintaining a warm body temperature, bats conserve energy. This ability is important because flying all night is hard work. When the temperature is cold for long periods during the winter months, some bats enter a deeper torpor state called hibernation. Other bat species follow a yearly migration pattern, traveling to cooler climates in the warm months and warmer climates in the cool months. This is why some regions experience "bat seasons" every year.

Many people have a negative reaction to bats, and it's easy to see why. Just by virtue of their appearance and behavior, bats play into a number of human fears. However, bats play an important role in many ecosystems by keeping insect populations in check and pollinating plants. Insectivorous bats are the best bug killers on the planet. For example, a famous colony of more than 20 million Mexican free-tail bats that lives in Bracken Cave, Texas will eat up to 200 tons of insects in a night. Bats are also beneficial as plant pollinators. Many species feed on plant nectar, gathering pollen on their bodies as they feed and helping the plant to disperse its seed when they visit other plants.

TROPICAL CYCLONES

Few events on Earth rival the sheer power of a tropical cyclone. A tropical cyclone is a storm system characterized by a low-pressure center and numerous thunderstorms that produce strong winds and flooding rain. Depending on its location and strength, a tropical cyclone is sometimes called a hurricane or typhoon. Scientists at the National Center for Atmospheric Research estimate that the amount of heat energy released by a tropical cyclone in one hour is equivalent to 70 times the world energy consumption of humans per hour. Over the past two centuries, tropical cyclones have killed about 1.9 million people. Some researchers even theorize that the dinosaurs were wiped out by prehistoric hypercanes, a kind of super-sized tropical cyclone stirred to life by the heat of an asteroid strike.

Tropical cyclones often begin their lives as clusters of clouds and thunderstorms called tropical disturbances. In order to take the first step towards becoming a full-blown tropical cyclone, a disturbance must develop a pocket of low-pressure air at its center. This process, which can take anywhere from hours to days, begins with the thunderstorms in the disturbance releasing latent heat. This heat warms the air in the disturbance, causing oxygen molecules to expand and thereby lowering the density of the air. As the density of the air drops, so too does the air pressure. Once a low-pressure area exists, the first step is complete and the disturbance has the potential to take the next step in its development: beginning to rotate at high speeds.

The rotation of a tropical cyclone is a product of the Coriolis force, a natural phenomenon created by the rotation of the Earth that causes free-moving objects to veer to the right of their destination in the Northern Hemisphere and to the left in the Southern Hemisphere. This force causes cyclonic systems to rotate in the direction of the closest pole. As a result, tropical cyclones rotate counterclockwise in the Northern Hemisphere and clockwise in the Southern Hemisphere. The force also affects the path of the tropical cyclone, bending them towards the closest pole. Thus, tropical cyclones in the Northern Hemisphere usually turn north (before being blown east), and tropical cyclones in the Southern Hemisphere usually turn south (before being blown east).

Once rotation is initiated, a tropical cyclone builds in strength through rapidly rising air at the center of the storm. As it moves across the ocean, it sucks up warm, moist tropical air from the surface of the water and dispenses cooler air aloft. A tropical cyclone's primary energy source is the release of the heat of condensation from water vapor in this rising air. The release of heat creates a pattern of wind that circulates around a center, like water going down a drain, and brings the rotation of the tropical cyclone to high speeds. In addition to the warm air being sucked up into the center of the storm, converging winds at the surface and higher altitudes also push warm air upwards, increasing the rotation.

A tropical cyclone has two key parts. The low-pressure center of relative calm is called the eye. Weather in the eye is normally calm and free of clouds, although the sea may be extremely violent. Circular in shape, the eye may range in size from 5 to 120 miles in diameter, but most eyes are between 20 and 40 miles across. The area surrounding the eye is called the eye wall, and it consists of a dense wall of clouds and thunderstorms. The eye wall is the part of the storm where the greatest wind speeds are found, clouds reach the highest, and precipitation is the heaviest. Interestingly, the eye wall actually creates the eye by sucking out any clouds or rain in the area.

Tropical cyclone formation is still not fully understood, but five conditions appear to be critical. In most situations, water temperatures of at least 80° F are needed down to a depth of approximately 160 feet because it causes the overlying atmosphere to be unstable enough to sustain thunderstorms and facilitate convection (i.e. the transfer of heat). Another condition is rapid cooling with height, which allows the release of the heat that powers a tropical cyclone. High humidity is needed because moisture in the atmosphere helps disturbances to develop. Low amounts of wind shear are needed, as high shear disrupts the storm’s circulation. Lastly, tropical cyclones also need to form more than 345 miles away from the equator for the Coriolis force to initiate rotation.

One measure of the size of a tropical cyclone is called the Radius of Outermost Closed Isobar (ROCI). The atmospheric pressure increases gradually as one moves away

from the center of the storm, and the outermost closed isobar is the point at which the pressure returns to normal. ROCI is determined by measuring the radii from the center of the storm to its outermost closed isobar in each of the four quadrants surrounding the storm. The distances of the radii are then averaged to come up with a single value. If the ROCI is between 2 and 3 degrees of latitude, then the cyclone is considered "small". A ROCI between 3 and 6 latitude degrees is considered "medium." A "large" tropical cyclone has a ROCI of between 6 and 8 degrees.

Coastal regions generally receive significant damage from a tropical cyclone because it produces high waves and damaging storm surge. In contrast, inland regions are relatively safe from the strong winds, but heavy rains can produce significant flooding. A recent example of such devastation is Hurricane Katrina, which claimed the lives of at least 1,836 people and caused an estimated $100 billion in damages when it hit Louisiana and Mississippi in August of 2005. Although their effects on human populations can be devastating, tropical cyclones are helpful in at least one way. They play an important role in the global atmospheric circulation mechanism by carrying heat away from the tropics and transporting it to more temperate latitudes, thereby helping to maintain a relatively stable and warm temperature worldwide.

VACCINES

A vaccine is a biological preparation that establishes or improves immunity to a particular disease. Most vaccines are prophylactic, which means that they prevent or ameliorate the effects of a future infection by any natural pathogen. The flu vaccine is an example of a prophylactic vaccine that is given annually to protect against the influenza virus. However, vaccines have also been used for therapeutic purposes, such as for alleviating the suffering of people who are already afflicted with a disease. An example of such a therapeutic use is the vaccines currently being developed for the treatment of various types of cancer. Until recently, most vaccines have been aimed at children, but the development of therapeutic vaccines has increased the number of treatments targeted at adults.

The early vaccines were inspired by the concept of variolation, which originated in Asia. Variolation is a technique in which a person is deliberately infected with a weak form of a disease through inhalation. Some historians claim that the earliest record of variolation can be found in an 8th century text from India called the *Nidana*. However, the first unequivocal reference to variolation comes from a Chinese text by Wan Quan called the *Douzhen Xinfa* written in 1549. The *Douzhen Xinfa* describes how dried smallpox scabs were blown into the nose of an individual who then contracted a mild form of the disease. Upon recovery, the individual was immune to smallpox. A small proportion of the people who were variolated died, but nowhere near the proportion that died when they contracted the disease naturally.

By 18th century, the practice of variolation had spread to Africa, India and the Ottoman Empire. In 1717, the wife of the British ambassador to the Ottoman Empire, Lady Mary Montagu, learned about variolation in Constantinople (which is known as Istanbul today) and advocated for the practice when she returned to England. At her behest, royal physicians conducted an experiment in which a number of prisoners and abandoned children were variolated by having smallpox inserted under their skin. When

the children and prisoners were deliberately exposed to smallpox several months later and none contracted the disease, the procedure was deemed safe. Neverthless, variolation carried a large degree of risk. Not only could the patient die from the procedure, but also the mild form of the disease could spread, causing an epidemic.

Over the following centuries, medical researchers like Edward Jenner and Louis Pasteur transformed the ancient technique of variolation into the modern day practice of inoculation with vaccines. Inoculation represented a major breakthrough because it reduced the risk of vaccination, while maintaining its effectiveness. Inoculation is the practice of deliberate infection through a skin wound. This new technique produces a smaller, more localized infection relative to variolation in which inhaled viral particles in droplets spread the infection more widely. The smaller infection works better because it is adequate to stimulate immunity to the virus, but it also keeps the virus from replicating enough to reach levels of infection likely to kill a patient.

Vaccines work because they prepare the immune system to deal with pathogens that it may encounter in the future. When a vaccine is given, the immune system recognizes the vaccine agents as foreign, destroys them, and then "remembers" them. When the real virulent version of an agent comes along, the body recognizes the protein coat on the virus and responds by destroying the infected cells before they can multiply. Of course, vaccines do not guarantee complete protection against developing the disease. Sometimes a person's immune system does not respond because of a lack of B-cells capable of generating antibodies to that antigen or a lowered immunity in general. Still, even when a vaccinated individual does develop the disease vaccinated against, the disease is likely to be milder than without vaccination.

Some vaccines are made from dead or inactivated virulent organisms that have been killed with chemicals or heat. Examples are vaccines against influenza, cholera, and hepatitis. Other vaccines contain live, attenuated virus organisms that are cultivated under conditions that disable their virulent properties. Examples include yellow fever, measles, rubella, and mumps. Aluminium-based adjuvants, such as squalene, are typically added to boost immune response. Vaccines can be monovalent or polyvalent. A monovalent vaccine is designed to immunize against a single antigen or single microorganism. A polyvalent vaccine is designed to immunize against two or more strains of the same organism, or against two or more organisms. In certain cases, a monovalent vaccine may be preferable for rapidly developing a strong immune response.

One challenge in vaccine development is economic: many of the diseases that could be eradicated with a vaccine, such as malaria, exist principally in poor countries. Although many vaccines have been highly cost effective and beneficial for public health, pharmaceutical firms and biotechnology companies have little incentive to develop vaccines for these diseases because there is little revenue potential. Even in more affluent countries, financial returns are usually minimal while the costs are great. The number of vaccines administered has actually risen dramatically in recent decades, but this rise is due to government mandates and support, rather than economic incentive. Thus, most vaccine development relies on "push" funding that is supplied by government, universities, and non-profit organizations.

Overall, the invention of vaccines has led to a marked decrease in the prevalence of certain diseases. For example, vaccines have contributed to the eradication of smallpox, one of the most contagious and deadly diseases known to man. Other diseases,

such as polio, measles, and typhoid, are nowhere near as common as they were a hundred years ago. As long as the vast majority of people are vaccinated, it is much more difficult for an outbreak of disease to occur and spread, an effect called herd immunity. Yet, critics have campaigned in opposition to vaccination for centuries. Disputes have arisen over the morality, effectiveness, ethics, and safety of vaccination. Still, the mainstream medical opinion is that the benefits of preventing suffering and death from serious infectious diseases greatly outweigh the risks of rare adverse effects following immunization.

BREAD

Bread is a prepared by baking dough made from two main ingredients: flour and water. Bakers call the inner, soft part of bread the crumb, which is not to be confused with small bits of bread that often fall off, called crumbs. The outer hard portion of bread is called the crust. Bread can either be leavened or unleavened. Leavening is the process of adding gas to the dough before or during baking to produce lighter, more chewable bread. Most of the bread consumed in contemporary cultures is leavened. However, unleavened bread has symbolic importance in many religions and, thus, nowadays it is primarily consumed in the context of religious rites and ceremonies. For example, Jews consume unleavened bread called *matza* during Passover.

Flour provides the primary structure to bread because it contains proteins – it is the quantity of these proteins that determines the quality of the finished bread. Wheat flour contains two non-water soluble protein groups (glutenin and gliadin), which form the structure of the dough. When worked by kneading, the glutenin forms long strands of chainlike molecules while the shorter gliadin forms bridges between the strands of glutenin, resulting in a network of strands called gluten. The network of strands, or gluten, is responsible for the softness of the bread because it traps tiny air bubbles as the dough is baked. If the network of strands is more cohesive or tightly linked, the bread will be softer. Gluten development improves if the dough is allowed to rest between mixing and kneading.

The amount of flour is the most significant measurement in a bread recipe. Professional bakers use a system known as Bakers' Percentage in their recipe formulations. They measure ingredients by weight rather than by volume because it is more accurate and consistent, especially for dry ingredients. Flour is always stated as 100%, and the rest of the ingredients are a percent of that amount by weight. For example, common table bread in the U.S. uses approximately 50% water, whereas most artisan bread formulas contain anywhere from 60 to 75% water. The water (or sometimes another liquid like milk or juice) is used to form the flour into a paste or dough.

Yeast is used in baking as a leavening agent. A single-cell microorganism (most commonly *Saccharomyces cerevisiae*), yeast help bread to rise because they convert the fermentable sugars present in the dough into carbon dioxide gas and alcohol. The alcohol, which burns off during baking, contributes to the bread's flavor. The carbon dioxide gas created by yeast causes the dough to expand or rise as the carbon dioxide forms bubbles. The stretchy, balloon-like consistency of the gluten in the bread dough traps the bubbles and keeps the carbon dioxide from escaping. When the dough is baked it "sets" and the bubbles remain, giving the baked product a soft and spongy texture. Most bakers in the

U.S. leaven their dough with commercially produced baker's yeast, which yields uniform, quick, and reliable results because it is obtained from a pure culture.

Gas-producing chemicals can also be used as a leavening agent. Whereas yeast takes two to three hours to produce its leavening action, a dry chemical leavening agent like baking powder is instantaneous. Many commerical bakeries use chemical additives to speed up mixing time and reduce necessary fermentation time, so that a batch of bread may be mixed and baked in less than 3 hours. "Quick bread" is the name that commercial bakers use for dough that does not require fermentation because of chemical additives. Often these chemicals are added to dough in the form of a prepackaged base, which also contains most or all of the dough's non-flour ingredients. Commercial bakeries also commonly add calcium propionate to retard the growth of molds.

The simplicity of bread is indicative of its history – it is one of the oldest prepared foods, dating back to the Neolithic era. The first breads produced were probably cooked versions of a grain-paste, made from ground cereal grains and water by hunter-gather tribes. The discovery of the first bread either occurred through accidental cooking or deliberate experimentation with water and grain flour. Descendants of these early breads are still commonly made from various grains worldwide, including the Middle Eastern *pita*, the Mexican *tortilla*, and the Indian *roti*. The basic flatbreads of this type also formed a staple in the diet of many early civilizations, including the Sumerians who ate a type of barley flat cake and the Egyptians who ate flat bread called *ta* in 12th century BC.

The development of leavened bread can probably be traced to prehistoric times as well. Yeast spores occur everywhere, so any dough left to rest will become naturally leavened. For example, an uncooked dough exposed to air for some time before cooking would probably contain airborne yeasts as well as yeasts that grow on the surface of cereal grains. Thus, the most common source of leavening was early bakers retaining a piece of dough from the previous day to utilize as a form of dough starter. Although leavening is likely of prehistoric origin, the earliest archaeological evidence comes from ancient Egypt. Scientific analysis using electron microscopy has detected yeast cells in some ancient Egyptian loaves.

Bread has been of great historical and contemporary importance in Western and Middle Eastern cultures, and it is commonly used in these cultures as a symbol of basic necessities, such as food and shelter. For example, the word bread is now commonly used in English speaking countries as a synonym for money (as is the case with the word "dough"). The political significance of bread is also considerable. In 19th century Britain, the inflated price of bread due to the Corn Laws caused major political and social divisions, prompting riots. The Assize of Bread and Ale, a 13th century law, showed the importance of bread in medieval times by setting heavy punishments for short-changing bakers. Today, bread remains a popular food in many societies, and the variety of breads enjoyed across these societies continues to expand.

THE RESPIRATORY SYSTEM

Humans breathe in and out anywhere from 15 to 25 times per minute. The main function of the respiratory system is gas exchange between the external environment and the circulatory system. A gas that the body needs to get rid of, carbon dioxide, is exchanged for a gas that the body can use, oxygen. Located within the chest cavity and

protected by the rib cage, the lungs are the most critical component of the respiratory system. The lungs are responsible for the oxygenation of the blood and the concomitant removal of carbon dioxide from the circulatory system. The other major function of the lungs is to manage the concentration of hydrogen ion in the blood, an important factor in regulating the acidity of blood (pH), which must be kept in a narrow range.

When a person inhales, the diaphragm and intercostal muscles (the muscles between the ribs) contract and expand the chest cavity. This expansion lowers the pressure in the lungs below the outside air pressure. Air then flows in through the airways (from high pressure to low pressure) and inflates the lungs. The lungs are made of spongy, elastic tissue that stretches and constricts during breathing. When a person exhales, the diaphragm and intercostal muscles relax and the chest cavity gets smaller. The decrease in volume of the cavity increases the pressure in the lungs above the outside air pressure. Air from the lungs (high pressure) then flows out of the airways to the outside air (low pressure). The cycle then repeats with each breath.

The respiratory system has many components. Air enters the body through the nose or mouth and goes past the epiglottis into the trachea, a rigid tube that connects the mouth with the bronchi. The epiglottis is a flap of tissue that closes over the trachea when a person swallows so that food and liquid do not enter the airway. The air continues down the trachea until it reaches the bronchi. From the bronchi, air passes into each lung and spreads out by following narrower and narrower bronchioles. The bronchioles are the numerous small tubes that branch from each bronchus into the lungs and get progressively smaller until they each end in an alveolus. Alveoli are tiny, thin-walled air sacs at the end of the bronchiole branches where gas exchange occurs.

Within the alveoli, gas exchange occurs through diffusion. Diffusion is the movement of particles from a region of high concentration to a region of low concentration. The oxygen concentration is high in the alveoli, so oxygen diffuses across the alveolar membrane into the pulmonary capillaries, which are small blood vessels that surround each alveolus. The hemoglobin in the red blood cells passing through the pulmonary capillaries has carbon dioxide bound to it and very little oxygen. The oxygen binds to hemoglobin and the carbon dioxide is released. Since the concentration of carbon dioxide is high in the pulmonary capillaries relative to the alveolus, carbon dioxide diffuses across the alveolar membrane in the opposite direction. The exchange of gases across the alveolar membrane occurs rapidly – usually in fractions of a second.

Humans do not have to think about breathing because the body's autonomic nervous system controls it. The respiratory centers that control the rate of breathing are located in the pons and medulla oblongata, which are both part of the brainstem. The neurons that live within these centers automatically send signals to the diaphragm and intercostal muscles to contract and relax at regular intervals. Neurons in the cerebral cortex can also voluntarily influence the activity of the respiratory centers. A region within the cerebral cortex, called motor cortex, controls all voluntary motor functions, including telling the respiratory center to speed up, slow down, or even stop. However, the influence of the nerve centers that control voluntary movements can be overridden by the autonomic nervous system.

Several factors can trigger such an override by the autonomic nervous system. One of these factors is the concentration of oxygen in the blood. Specialized nerve cells within the aorta and carotid arteries called peripheral chemoreceptors monitor the oxygen

concentration of the blood. If the oxygen concentration decreases, the chemoreceptors signal to the respiratory centers in the brain to increase the rate and depth of breathing. These peripheral chemoreceptors also monitor the carbon dioxide concentration in the blood. Another factor is chemical irritants. Nerve cells in the airways can sense the presence of unwanted substances like pollen, dust, water, or cigarette smoke. If chemical irritants are detected, these cells signal the respiratory centers to contract the respiratory muscles, and the coughing that results expels the irritant from the lungs.

Disorders of the respiratory system fall mainly into two classes. Some disorders make breathing harder, while other disorders damage the lungs' ability to exchange carbon dioxide for oxygen. Asthma is an example of a disease that influences the mechanics of breathing. During an asthma attack, the bronchioles constrict, narrowing the airways. This reduces the flow of air and makes the respiratory muscles work harder. In contrast, pulmonary edema is an example of a disease that minimizes or prevents gas exchange. Pulmonary edema occurs when fluid builds up in the area between the alveolus and pulmonary capillary, increasing the distance over which gases must exchange and slowing down the exchange. Various medical interventions are used treat disorders of the respiratory system, but coughing is the body’s main method of defense.

The respiratory systems of other animals differ from that of humans in varying degrees. Most other mammals have a similar respiratory system, but often have subtle differences. For example, horses do not have the option of breathing through their mouths and must take in air through their nose. The respiratory system of birds, which contains unique anatomical features such as air sacs, differs significantly from that found in mammals. Reptiles have a much simpler lung structure than mammals as they lack the extensive airway tree structure found in mammalian lungs. In amphibians, the skin is an important respiratory organ – it is highly vascularized and secretes mucus from specialized cells to facilitate rapid gas exchange. Overall, respiratory systems differ substantially across the animal kingdom.

THE INTERNET

The Internet is a global system of interconnected computer networks that interchange data using a set of standardized communications protocols. Essentially a "network of networks," the internet consists of millions of private and public networks of local to global scope that are linked by copper wires, fiber-optic cables, wireless connections, and other technologies. The Internet carries various information resources and services, such as electronic mail, online chat, and file sharing. Although the terms Internet and World Wide Web are often used synonymously, they are not same thing. The Internet is a global data communications system that uses hardware and software infrastructure to provide connectivity between computers, whereas the Web is a collection of interconnected documents and other resources that are communicated via the Internet.

The story of the Internet begins with the launch of the Soviet satellite Sputnik in 1957, which spurred the United States to establish the Advanced Research Projects Agency (ARPA) in order to regain a technological lead. A project leader at ARPA, Joseph Licklider, saw great potential in universal networking and initiated a project to build a network that relied on a new technology called packet switching. Packet

switching is a mode of data transmission in which data is broken into chunks, called packets, which are sent independently and then reassembled at the destination. Alternative modes of data transmission, such as circuit switching, require a fixed connection between terminals, so each circuit can handle only one user at a time. In contrast, packet switching can accommodate multiple users, optimizing network use and minimizing data transmission time.

After much work, the first two nodes of what would become the ARPANET were interconnected in Menlo Park, California in 1969. The next two decades featured rapid advances in the technology and infrastructure needed to create a global network. The language that modern-day computers use to communicate over the Internet, called the standardized Internet Protocol Suite (TCP/IP), arose from the experimental work of Vinton Cerf and Robert Kahn, who developed the first description of the protocol suite and published a paper on the subject in 1974. The first TCP/IP-based network was made operational in 1983 when the ARPANET was switched over from an older protocol. Two years later, the United States' National Science Foundation (NSF) commissioned the construction of the second TCP/IP-based network, called NSFNET.

Until the late 1980s, the networks were used for governmental and scientific research purposes only. However, this restriction on the networks came to an end when the U.S. Federal Networking Council approved the interconnection of the NSFNET to the commercial MCI Mail system in 1988. The opening of the network to commercial interests greatly accelerated the expansion of what is now called the Internet. Motivated by potential profits, commercial companies aggressively pursued the connection of existing networks and the creation of new networks. Although the Internet had existed for almost a decade, the network did not gain a public face until the 1990s. In 1991, the European Organization for Nuclear Research publicized a new project called the World Wide Web. Over the following two decades, the Internet evolved into its present-day form.

One of the greatest things about the Internet is that nobody really owns it because it is a global collection of networks. Every computer that is connected to the Internet is part of a network. For example, people may use a modem and dial a local number to connect to an Internet service provider. When the computer connects to a provider, it becomes part of the provider’s network. The provider may then connect to a larger network and become part of that network. However, just because nobody owns the Internet, it does not mean that it is not monitored and maintained in different ways. The Internet Society, a non-profit group established in 1992, oversees the formation of the policies and protocols that define how people use and interact with the Internet.

Most large communications companies that provide Internet service have their own dedicated backbones connecting various regions. In each region, the company has a Point of Presence (POP). Each POP is a place for users to access the company's network, often through a local phone number or dedicated line. Interestingly, there is no overall controlling network. Instead, several high-level networks connect to each other through a Network Access Point (NAP). Each NAP is a physical infrastructure that allows different Internet service providers to exchange traffic between their networks. Dozens of large providers interconnect at NAPs in various cities, and trillions of bytes of data flow between the networks at these points. The Internet is largely a collection of huge corporate networks that all intercommunicate at the NAPs.

What is incredible about the Internet is that a message can leave one computer and travel halfway across the world through several different networks and arrive at another computer in a fraction of a second. To accomplish this feat, all of these networks rely on routers. Routers are specialized computers that have two main functions. First, routers ensure that information makes it to the intended destination by determining where to send it along thousands of pathways. Second, routers make sure that information doesn't go where it's not needed, which is crucial for keeping large volumes of data from clogging the connections of other users. Thus, the router joins the networks so they can communicate, but also protects them from one another.

Every machine on the Internet has a unique identifying number, called an IP Address. The IP stands for Internet Protocol, which is the language that computers use to communicate over the Internet. The IP address identifies both the individual computer and the network to which it belongs. Initially, the Internet consisted of a small network of computers and only the other computer's IP address was needed to establish a connection. However, this system became unwieldy as more computers came online. The Domain Name System (DNS), which maps human-friendly computer hostnames into IP addresses automatically, was created in 1983 to solve the problem of organizing the exponentially increasing number of IP addresses. With the DNS, a person only needs to remember www.google.com, for example, instead of Google's IP address.

Appendix B

Questions from Experiments 1, 2, 3, and 4. Questions are organized by passage.

BATS

Factual Questions

1. Bats are one of the most prevalent orders of mammals. Approximately how many bat species are there in the world?

More than 1,000 bat species have been identified.

2. Bats of the Microchiroptera suborder (microbats) live all over the world. Where do bats of the Megachiroptera suborder (megabats) live?

Megachiroptera bats (megabats) only live in Asia, Africa, and Australia.

3. Bats sleep hanging upside down in a high location to avoid predators. What is the main predator of bats?

Birds of prey are the main predator of bats.

4. A famous colony of Mexican free-tail bats lives in Bracken Cave, Texas. How many tons of insects per night does this colony of more than 20 million bats eat?

The bats in Bracken Cave will eat up to 200 tons of insects in a night.

Conceptual Questions

1. 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?

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.

2. Some bats use echolocation to navigate the environment and locate prey. How does echolocation help bats to determine the distance and size of objects?

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 of the object 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.

3. Bats have specially adapted talons that enable them to hang upside down. How do these talons function?

A bat must relax its muscles to grip an object, which is the opposite of how human fingers work. The weight of the upper body pulls down on the tendons connected to the talons, causing them to clench and gravity keeps the talons closed.

4. When bats sleep during the day, they enter a torpor state. What happens to bats physiologically when in a torpor state?

Bats allow their body temperature to sink to the ambient temperature whenever they are inactive. As their body temperature drops, they enter a torpor state. When in a torpor state, a bat's metabolism slows down, reducing biological activity and conserving energy.

Inferential Questions for Facts (same domain)

1. There are about 5,500 species of mammals in the world. Approximately what percent of all mammal species are species of bat?

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.

2. The brown bat (*Myotis lucifugus*) is one of the most common bats of North America. To what suborder of bat species does the brown bat likely belong?

The brown bat likely belongs to the Microchioptera suborder because it lives in North America – Megachiroptera bats (megabats) only live in Asia, Africa, and Australia.

3. Oddly, bats are not often attacked when they are sleeping upside down, but rather when they are flying at night. What specific animal is most likely to be responsible for these attacks on bats?

Owls. Birds of prey are the main predator of bats. Owls are a bird of prey and hunt at night.

4. The famous colony of 20 million Mexican free-tail bats, which lives in Bracken Cave, Texas, eats tons of insects every night. How many tons of insects does this colony eat in a week?

The bats in Bracken Cave will eat up to 200 tons of insects in a night, which would translate to 1400 tons of insects per week.

Inferential Questions for Concepts (same domain)

1. Bats are much better at catching mosquitoes than birds. Why are bats more proficient at hunting small, flying insects?

A bat has a much more flexible wing structure that allows for greater maneuverability, giving a bat a distinct advantage over a bird in hunting tiny, flying insects.

2. 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)?

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 insect will reflect more of the sound wave as it gets closer.

3. Sometimes bats die while they are sleeping. What will happen if a bat dies while it is hanging upside down?

Since it is gravity that keeps the talons closed, the bat will continue to hang upside down if it dies in that position.

4. Many zoologists believe that bats' ability to enter a temporary torpor state evolved in response to natural selection pressures related to food consumption. Why might the supply of the food have caused primitive bats to develop the ability to enter a temporary torpor state?

If the food supply for primitive bats fluctuated, then bats with the ability to enter a temporary torpor state would have an adaptive advantage in food shortages because they could conserve energy.

Control Questions (not tested during initial learning session)

1. Bat species are divided into two suborders: Megachiroptera (megabats) and Microchiroptera (microbats). What do the facial features of megabats look like?

The facial features of most megabat species look a lot like other mammals, with large eyes, small ears, and extended snouts.

2. Bats play an important role in many ecosystems by keeping insect populations in check. What other major role do they play in ecosystems?

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.

Re-Phrased Versions of a Factual Question

Version A. Bats are one of the most prevalent orders of mammals. Approximately how many bat species are there in the world?

Version B. Chiroptera is the name of the order that contains all bat species. What is the approximate number of bat species that exist?

Version C. Over millions of years, bats have evolved from a common ancestor into many species. When zoologists count up all the species of bat that have been identified, what is the total count?

Response to All: *More than 1,000 bat species have been identified.*

Re-Phrased Versions of a Conceptual Question

Version A. Some bats use echolocation to navigate the environment and locate prey. How does echolocation help bats to determine the distance and size of objects?

Version B. 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?

Version C. At night vision is not much use because there is little light, but some bats roam around without any problem by "seeing" with echolocation. How do bats "see" an object by using echolocation to calculate its distance and size?

Response to All: *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 of the object 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.*

Inferential Questions for Concepts (different domain)

1. 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?

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.

2. Submarines use SONAR to navigate underwater much like bats use echolocation to navigate at night. Using SONAR, how does a submarine determine that an object is moving towards it (i.e. rather than away from it)?

The submarine can tell the direction that an object is moving by calculating whether the time it takes for the sound waves to return changes over time. If the object is moving towards the submarine, the time it takes the sound wave to return will get steadily shorter. Also, the intensity of the sound wave will increase because object will reflect more of the sound wave as it gets closer.

3. An ascender is a mechanical device for rock climbing that functions in similar manner to a bat talon. If an ascender helps climbers move upwards on a rope, how do ascenders work?

A bat must relax its muscles to grip an object with its talons: once they are clenched, gravity keeps the talons closed. An ascender functions in a similar way by gripping the rope and preventing climbers from falling. When climbers want to move up, they must grasp the ascender to release its grip and move it up the rope.

4. Some scientists have argued that seasonal affective disorder, a type of depression, is similar to the torpor state that some animals (e.g., bats) enter when they are inactive. If loss of appetite is a symptom of seasonal affective disorder, why might this symptom support the hypothesis?

When in a torpor state, an animal's metabolism slows down, reducing biological activity and conserving energy. Thus, loss of appetite in seasonal affective disorder might indicate a reduced need for energy consumption due to reduced metabolism.

TROPICAL CYCLONES

Factual Questions

1. Tropical cyclones release an enormous amount of heat energy. Relative to the world energy consumption of humans, how much energy does a tropical cyclone release per hour?

Scientists at the National Center for Atmospheric Research estimate that the amount of heat energy released by a tropical cyclone in one hour is equivalent to 70 times the world energy consumption of humans per hour.

2. The Coriolis force initiates the rotation of a tropical cyclone during its formation. What is the Coriolis force?

The rotation of a tropical cyclone is a product of the Coriolis force, a natural phenomenon created by the rotation of the Earth that causes free-moving objects to veer to the right of their destination in the Northern Hemisphere and to the left in the Southern Hemisphere.

3. One of the necessary conditions for tropical cyclone formation is that the water temperature must be at least 80 degrees down to a depth of at least 160 feet. How does this condition affect the overlying atmosphere?

Water temperatures of at least 80 degrees are needed down to a depth of approximately 160 feet because it causes the overlying atmosphere to be unstable enough to sustain thunderstorms and facilitate convection (i.e. the transfer of heat).

4. Tropical cyclones are powerful storms that can devastate human populations, but they can also be helpful to humans. In what way are tropical cyclones helpful to humans?

Tropical cyclones play an important role in the global atmospheric circulation mechanism by carrying heat away from the tropics and transporting it to more temperate latitudes, thereby helping to maintain a relatively stable and warm temperature worldwide.

Conceptual Questions

1. In order to take the first step towards becoming a full-blown tropical cyclone, a disturbance must develop a pocket of low-pressure air at its center. How is this pocket of low-pressure air created?

A disturbance must develop a pocket of low-pressure air at its center before it can become a full-blown tropical cyclone. This process begins with the thunderstorms in the disturbance releasing latent heat. This heat warms the air in the disturbance, causing oxygen molecules to expand and thereby lowering the density of the air. As the density of the air drops, so too does the air pressure.

2. A tropical cyclone builds in strength as it moves across the ocean. What is the process through which it obtains energy?

As a tropical cyclone moves across the ocean, it sucks up warm, moist tropical air from the surface of the water and dispenses cooler air aloft. A tropical cyclone's primary energy source is the release of the heat of condensation from water vapor in this rapidly rising air. The release of heat creates a pattern of wind that circulates around a center, like water going down a drain, and brings the rotation of the tropical cyclone to high speeds.

3. The part of a tropical cyclone surrounding the eye is called the eye wall. What are the conditions in the eye wall like?

The area surrounding the eye is called the eye wall, and it consists of a dense wall of clouds and thunderstorms. The eye wall is the part of the storm where the greatest wind speeds are found, clouds reach the highest, and precipitation is the heaviest.

4. The Radius of Outermost Closed Isobar (ROCI) is a measure of the size of a tropical cyclone. How is ROCI determined?

The Radius of Outermost Closed Isobar (ROCI) is determined by measuring the radii from the center of the storm to its outermost closed isobar in the four quadrants surrounding the storm. The outermost closed isobar is the point at which the atmospheric pressure returns to normal as it gradually increases from the storm center. The distances of the radii are averaged to come up with a single value.

Inferential Questions for Facts (same domain)

1. On average, the world energy consumption of humans is about 2 gigawatts per hour. Based on this figure, what is amount of heat energy released per hour by a tropical cyclone?

140 gigawatts. The amount of heat energy released by a tropical cyclone in one hour is equivalent to 70 times the world energy consumption of humans per hour.

2. A plane is flying due south from the North Pole. How would the Coriolis force affect its path?

If the plane is flying south, the Coriolis force would push it slightly westward because the force causes free-moving objects to veer to the right of their destination in the Northern Hemisphere.

3. Scientists warn that the number of tropical cyclones may increase in the future as a result of global warming. Why might global warming cause more tropical cyclones to form?

Global warming will increase the temperature of the oceans and warm water temperatures are a necessary condition for the formation of tropical cyclones.

4. Located far away from the tropics, Sweden experiences only the indirect effects of tropical cyclones. If tropical cyclones ceased to exist, how would the climate of Sweden be affected?

The climate of Sweden would probably become colder. Tropical cyclones transport heat to more temperate latitudes, thereby helping to maintain a relatively stable and warm temperature worldwide.

Inferential Questions for Concepts (same domain)

1. The amount of time that it takes for a tropical disturbance to develop a low-pressure pocket of air at its center depends on the initial air temperature. Why does this process take longer to occur when the initial air temperature is low?

The development of a pocket of low-pressure air depends on latent heat warming the air in the disturbance. If the initial air temperatures are low, it will take longer to warm air to the point where it expands and lowers the air pressure.

2. Tropical cyclones weaken and eventually dissipate after they hit land. Why do they lose their power after making landfall?

A tropical cyclone's primary energy source is the release of the heat of condensation from water vapor in the warm, moist tropical air that it sucks up from the ocean. Once a tropical cyclone hits land, it loses this energy source.

3. A tropical cyclone is headed across the Gulf of Mexico towards Texas. Which part is most likely to do the most damage when it hits land?

The eye wall will do the most damage because it is the part of the storm where the greatest wind speeds are found and precipitation is the heaviest.

4. The distance from the center of a tropical cyclone to its outermost closed isobar measures 4, 5, 2, and 5 degrees of latitude in the four quadrants of the storm, respectively. What is the size of this tropical cyclone in terms of ROCI?

The ROCI for this tropical cyclone is 4. The radii are averaged to determine ROCI.

Control Questions (not tested during initial learning session)

1. Tropical cyclones can devastate human communities when they hit land, causing death and destruction. Approximately how many people have been killed by tropical cyclones over the past two centuries?

Over the past two centuries, tropical cyclones have killed about 1.9 million people.

2. Circular in shape, the eye of tropical cyclones ranges in size from 5 to 120 miles in diameter. What is the range of most eyes in terms of miles in diameter?

Most eyes range between 20 and 40 miles in diameter.

Re-Phrased Versions of a Factual Question

Version A. Tropical cyclones release an enormous amount of heat energy. Relative to the world energy consumption of humans, how much energy does a tropical cyclone release per hour?

Version B. On a constant basis, tropical cyclones emit energy in the form of heat. How does the amount of energy emitted by a tropical cyclone every hour compare to the hourly energy consumption of all the humans in the world?

Version C. The amount of energy consumed every hour by all the humans in the world is extremely large. How does the amount of heat energy generated every hour by a tropical cyclone compare?

> Response to All: *Scientists at the National Center for Atmospheric Research estimate that the amount of heat energy released by a tropical cyclone in one hour is equivalent to 70 times the world energy consumption of humans per hour.*

Re-Phrased Versions of a Conceptual Question

Version A. In order to take the first step towards becoming a full-blown tropical cyclone, a disturbance must develop a pocket of low-pressure air at its center. How is this pocket of low-pressure air created?

Version B. The creation of a pocket of low-pressure air at the center of a tropical disturbance represents the first step towards it reaching tropical cyclone status. What has to happen for this pocket of low-pressure air to develop?

Version C. Tropical cyclones begin their lives as smaller, more peaceful tropical disturbances. What is the process by which a low-pressure air pocket develops at the center of a tropical disturbance?

> Response to All: *A disturbance must develop a pocket of low-pressure air at its center before it can become a full-blown tropical cyclone. This process begins with the thunderstorms in the disturbance releasing latent heat. This heat warms the air in the disturbance, causing oxygen molecules to expand and thereby lowering the density of the air. As the density of the air drops, so too does the air pressure.*

Inferential Questions for Concepts (different domain)

1. Extreme summer heat affects the air in car tires in the same way that heat released by thunderstorms affects the atmospheric air in a tropical cyclone, except that the air in the car tires has nowhere to expand. How do hot summer temperatures affect the air in car tires?

> *The heat in a tropical cyclone causes oxygen molecules to expand, thereby lowering the density of the air and the air pressure. Similarly, heat in the summer*

causes oxygen molecules in car tires to expand. However, since there is nowhere for the air to expand, the air pressure increases.

2. Although cars are powered by a different energy source than tropical cyclones, the fundamental process that drives a car's pistons is essentially the same as that which powers the spinning vortex of a cyclone. What is the process that is responsible for spinning the engine components of a car?

A tropical cyclone's primary energy source is the release of the heat of condensation from water vapor in rapidly rising warm air from the surface of the ocean. In a car engine, gasoline is burned inside the cylinders, giving rise to a tremendous amount of heat, and this heat does the work of spinning the engine components.

3. Some economists who study financial crises make an analogy between economic activity during recessions and passing straight through the middle of a tropical cyclone. If a period of relative economic calm is reached in a recession, what would be expected in the near future if the analogy holds?

The eye is the calm part of a tropical cyclone and the eye wall that surrounds it is the part of the storm where the greatest wind speeds are found, clouds reach the highest, and precipitation is the heaviest. Thus, it would be expected that the some of the worst economic news would be expected in the near future.

4. One way of measuring the size of sunspots, which are regions of intense magnetic activity on the Sun's surface, is through a method that is similar to the ROCI measure of tropical cyclones. How is this method used to measure the size of sunspots?

ROCI is determined by measuring the radii from the center of the storm to its outermost closed isobar in the four quadrants surrounding the storm, and then averaging them to come up with a single value. Thus, one way to measure sunspots is to measure the radii from the center of the sunspot to the edge of the region and then average the four radii.

VACCINES

Factual Questions

1. The practice of variolation, which led to the development of vaccines, began in Asia, but historians are unsure of the exact origin. What do historians agree was the first unequivocal reference to variolation?

The first unequivocal reference to variolation comes from a Chinese text by Wan Quan called the Douzhen Xinfa written in 1549.

2. The history of modern vaccines begins with the introduction of variolation to England. Who was the main person responsible for bringing variolation to England?

The wife of the British ambassador to the Ottoman Empire, Lady Mary Montagu, learned about variolation in Turkey and advocated for the practice when she returned to England.

3. A vaccine can be beneficial even if people develop the disease against which they have been vaccinated. What is the benefit of a vaccine if the illness it was supposed to prevent is developed?

Even if a vaccinated individual develops the disease that was vaccinated against, the disease is likely to be milder than without vaccination.

4. One argument for large-scale vaccination programs relies on the idea of herd immunity. How does herd immunity work?

The idea of herd immunity is that it is much more difficult for an outbreak of disease to occur and spread if the vast majority of people are vaccinated.

Conceptual Questions

1. Vaccines are biological preparations that commonly used in modern medicine. What are the two main ways in which vaccines are used today?

Most vaccines are used for prophylactic purposes, which means that they prevent or ameliorate the effects of a future infection by any natural pathogen. However, vaccines have also been used for therapeutic purposes, such as alleviating the suffering of people who are already afflicted with a disease.

2. A major breakthrough in the use of vaccines was the development of inoculation. Why is inoculation better than the older technique of variolation?

Inoculation is the practice of deliberate infection through a skin wound. This new technique produces a smaller, more localized infection relative to variolation in which inhaled viral particles in droplets spread the infection more widely. The smaller infection works better because it is adequate to stimulate immunity to the virus, but it also keeps the virus from replicating enough to reach levels of infection likely to kill a patient.

3. Vaccines vary in terms of their valence. What does the valence of a vaccine refer to?

The valence of the vaccine refers the number of different antigens contained in the vaccine. A monovalent vaccine is designed to immunize against a single antigen or single microorganism. A polyvalent vaccine is designed to immunize against two or more strains of the same organism, or against two or more organisms.

4. The number of vaccines administered has risen dramatically in recent decades. Where does the incentive for the development and administration of vaccines come from?

The rise in vaccinations is due to government mandates and support. Most vaccine development relies on "push" funding that is supplied by government, universities, and non-profit organizations. Pharmaceutical firms and biotechnology companies have little incentive to develop vaccines because many of the diseases that could be eradicated with a vaccine, such as malaria, exist principally in poor countries so there is little revenue potential.

Inferential Questions for Facts (same domain)

1. Smallpox was likely introduced in China around 100 AD from somewhere in modern-day India. Based on the first unequivocal reference to variolation, roughly how many years did it take for the Chinese to discover this method of protection against smallpox?

It took the Chinese roughly 1450 years to discover variolation. The first unequivocal reference to the technique comes from a Chinese text written in 1549.

2. A new history book about the British Empire argues that many of the great scientific advances made in England would not have occurred without the influx of new ideas that English citizens brought back home from foreign lands. How could the invention of the first modern vaccine by Englishman Edward Jenner be used to support this thesis?

Lady Mary Montagu learned about variolation in Turkey and brought it back to England, introducing an idea that was later used by Edward Jenner to develop the first modern vaccines.

3. A physician recommends that a child be vaccinated against a particular disease that permanently cripples people in its full-blown form. If clinical trials show that only 5% of people who get the vaccine develop immunity, what rationale might the physician provide for this decision other than the small chance of immunization?

Even if a vaccinated individual develops the disease that was vaccinated against, the disease is likely to be milder than without vaccination. It may be the even if the disease is contracted, it will not be strong enough to cripple the child.

4. Every year in the late fall, a university offers free flu shots to all its faculty, staff, and students. If a student decides not to get the shot, why might that student still be protected against the contracting the flu?

If everyone else is getting vaccinated, then fewer people will contract the flu, lessening the chance of others in the community contracting it. This rationale relies on the idea of herd immunity: it is much more difficult for an outbreak of disease to occur and spread if the vast majority of people are vaccinated.

Inferential Questions for Concepts (same domain)

1. Swedish researchers have developed a vaccine that may change the way the immune system responds in people who are newly diagnosed with Type 1 diabetes. What is the general purpose of this vaccine?

This vaccine is being used for a therapeutic purpose because it is being given to people who already have Type 1 diabetes.

2. The recently developed nasal spray flu vaccine, which is inhaled through the nose, contains weakened versions of the viruses that only cause infection at the cooler temperatures found within the nose. In what sense does this new method of vaccination combine the techniques of inoculation and variolation?

The nasal spray flu vaccine is similar to inoculation in that produces a smaller, more localized infection, but also like variolation in that the virus is inhaled.

3. Generally speaking, people given a monovalent vaccine develop immunity faster than people given a polyvalent vaccine. Why does immunity develop faster with a monovalent vaccine?

Immunity develops faster with a monovalent vaccine because the immune system only has to fight a single antigen, whereas it has to fight two or more antigens with a polyvalent vaccine, which takes longer.

4. Vaccines are being developed for the treatment of many types of cancer found in rich countries, but they have not yet been proven to work in clinical trials with humans. If the first round of cancer vaccines succeeds, how might it change the existing economic model of vaccine development?

Pharmaceutical firms and biotechnology companies will have more financial incentive to develop vaccines for diseases in rich countries, so development may not rely as heavily on "push" funding from government, universities, and non-profit organizations.

Control Questions (not tested during initial learning session)

1. Before the invention of modern vaccines, royal physicians in England conducted an initial experiment with variolation. Which two groups served as subjects in this experiment?

In the experiment, prisoners and abandoned children served as subjects. They were variolated by having smallpox blown into their noses.

2. Critics have campaigned in opposition to vaccination for centuries. What are two of the main issues that critics have raised about the use of vaccines?

> *Critics have argued about the morality, effectiveness, ethics, and safety of vaccination*.

Re-Phrased Versions of a Factual Question

Version A. The practice of variolation, which led to the development of vaccines, began in Asia, but historians are unsure of the exact origin. What do historians agree was the first unequivocal reference to variolation?

Version B. Historians argue about the exact origin of variolation because the evidence to support the earliest reference to the technique is incomplete. What is the earliest reference to variolation that historians agree is legitimate?

Version C. The legitimacy of some early references to variolation has been questioned by historians. What is the earliest reference to variolation that is considered by historians to be real?

> Response to All: *The first unequivocal reference to variolation comes from a Chinese text by Wan Quan called the Douzhen Xinfa written in 1549.*

Re-Phrased Versions of a Conceptual Question

Version A. Vaccines are biological preparations that commonly used in modern medicine. What are the two main ways in which vaccines are used today?

Version B. Modern medicine relies heavily on vaccines for two main purposes. How are vaccines are used in modern medicine?

Version C. The use of vaccines improves the lives of people all over the world every day. When vaccines are used in modern medicine, what two main purposes do they serve?

> Response to All: *Most vaccines are used for prophylactic purposes, which means that they prevent or ameliorate the effects of a future infection by any natural pathogen. However, vaccines have also been used for therapeutic purposes, such as alleviating the suffering of people who are already afflicted with a disease.*

Inferential Questions for Concepts (different domain)

1. Based on the same principle that helps humans inhale and exhale air, a bellows is a compressible container with an outlet nozzle that is used to deliver air in iron smelting. How does a bellows work?

Breathing in humans depends on air pressure. Similar to the lungs, when a bellows is expanded, it fills with air (high to low pressure). When a bellows is compressed, it increases the pressure in the bellows above the outside air pressure and the air flows out.

2. When a cube of sugar is placed into hot tea, it dissolves through the same process that makes gas exchange in the human respiratory system possible. Why does a sugar cube dissolve in hot tea?

Within the alveoli, gas exchange occurs through diffusion. Diffusion is the movement of particles from a region of high concentration to a region of low concentration. When a high concentration of sugar (the cube) is placed in hot tea, the sugar molecules with diffuse throughout the water because the concentration of sugar is lower.

3. Most cars that burn gasoline have an emissions control system that includes a component called an oxygen sensor and functions in a similar way to the system in the human body that can trigger involuntary breathing. How does emissions control system work?

Much like the system in the human body that can induce involuntary breathing, the emissions control system in a car monitors the oxygen concentration and can signal for increase in the concentration of oxygen if needed.

4. There are two main classes of disorders that can affect photosynthesis in plants and they are very similar to the two main classes of disorders that can affect the human respiratory system. If photosynthesis involves the conversion of carbon dioxide into sugars using the energy from sunlight, how does each class of disorder affect photosynthesis?

Disorders of the respiratory system can either make breathing harder or damage the lungs' ability to exchange carbon dioxide for oxygen. Likewise, disorders in photosynthesis in plants can either make obtaining carbon dioxide and sunlight more difficult or damage the plant's ability to convert of carbon dioxide into sugars.

BREAD

<u>Factual Questions</u>

1. Most of the bread consumed in contemporary cultures is leavened. What is unleavened bread primarily used for today?

Unleavened bread has symbolic importance in many religions – its primary use today is in various religious rites and ceremonies.

2. Yeast is often used as a leavening agent in baking bread. How does yeast help the bread to rise?

A single-cell microorganism, yeast help bread to rise because they convert the fermentable sugars present in the dough into carbon dioxide gas and alcohol.

3. Bread was discovered during the Neolithic era either through experimentation or by accident. What were the first breads made from?

The first breads produced were probably cooked versions of a grain-paste, made from ground cereal grains and water by hunter-gather tribes.

4. Bread has been of great historical and contemporary importance in Western and Middle Eastern cultures. What does bread symbolize in these cultures?

Bread is commonly used as a symbol of basic necessities, such as food and shelter, in Western and Middle Eastern cultures.

Conceptual Questions

1. Flour contains proteins. How do these proteins contribute to the consistency or texture of bread?

When worked by kneading, the non-water soluble proteins in flour form a network of strands called gluten, which is responsible for the softness of the bread because it traps tiny air bubbles as the dough is baked. If the network of strands is more cohesive or tightly linked, the bread will be softer.

2. Professional bread makers use a system called Bakers' Percentage. How does this system work?

Bakers' Percentage is a system in which ingredients are measured by weight instead of by volume. Measurement by weight is more accurate and consistent, especially for dry ingredients. Flour is always stated as 100%, and the rest of the ingredients are a percent of that amount by weight.

3. Many commercial bakeries use "quick bread." How is this bread different from traditional breads?

"Quick bread" is the name that commercial bakers use for dough that does not require fermentation because of chemical additives. Whereas yeast takes several hours to produce its leavening action, the chemical additives speed up mixing time, so that a batch of bread takes less than 3 hours.

4. Modern scientific analysis has detected yeast in ancient loaves, indicating that the process of leavening bread is quite old. What was the most common source of the yeast that early bakers used for leavening?

Yeast spores occur everywhere, so dough left to rest becomes naturally leavened by airborne yeasts and yeasts that grow on the surface of cereal grains. Thus, early bakers' most common source of yeast for leavening was a piece of dough from the previous day that was utilized as a form of dough starter.

Inferential Questions for Facts (same domain)

1. Roman Catholic Christians use bread when they celebrate the Eucharist, a rite derived from the narrative of the Last Supper. What type of bread is likely to be used in this religious ceremony?

Unleavened bread has symbolic importance in many religions, so Roman Catholic Christians probably use unleavened bread.

2. In addition to helping bread to rise, yeast are often used for fermentation in brewing both alcoholic beers and non-alcoholic beers, such as root beer. What are the two main ways in which the fermentation process can be adjusted to vary the alcohol content of beer?

Yeast convert sugar into alcohol and carbon dioxide. Either less sugar could be added, giving the yeast less sugar to convert, or the fermentation time could be reduced, giving the yeast less time to convert sugar to alcohol.

3. During the Neolithic era many hunter-gather tribes converted to farming and domesticated many wild plants. What is likely to be one of the first wild plants to be domesticated?

Cereal grains. Early hunter-gather tribes were the first to make bread from cooked versions of a grain-paste. They likely farmed cereal grains to make bread.

4. The Lord's Prayer, a popular prayer from the Bible, contains the line "Give us this day our daily bread." How might this line be interpreted metaphorically?

The line might be interpreted to be a request for God to provide the person saying the prayer with the basic necessities of life, such as food and shelter. Bread is commonly used as a symbol of basic necessities in Western and Middle Eastern cultures.

Inferential Questions for Concepts (same domain)

1. If bread is kneaded too much, the network of strands formed by the non-water soluble proteins will break down. How will over-kneading affect the consistency or texture of the bread?

The network of strands, or gluten, is responsible for the softness of the bread. If the network of strands is more cohesive or tightly linked, the bread will be softer. Thus, over-kneading will make the bread more dense and hard.

2. A recipe for bread calls for 30 pounds of flour and 20 pounds of water. How would this aspect of the recipe be expressed using the Bakers' Percentage system?

In the Bakers' Percentage system, flour is always stated as 100%, and the rest of the ingredients are a percent of that amount by weight. So, the recipe would call for 66% water.

3. Until the late 1950's, bakers usually had to get up at 2 A.M. to have fresh bread ready by 7 A.M., but nowadays they don't get up until 4 A.M. What change in bread making might have occurred in the 1950's?

Chemical additives speed up mixing time, so that a batch of bread may be made in less than 3 hours. The use of these chemical additives likely began in the 1950's.

4. On a camping trip, a group of people want to produce leavened bread. If they have flour and water, but forgot to bring yeast or any other leavening agent, how could they produce leavened bread?

Yeast spores occur everywhere, so they could leave a piece of dough out for a day and it could be utilized as a form of dough starter.

<u>Control Questions (not tested during initial learning session)</u>

1. Culinary professionals often have special terms to describe aspects of food. What terms do bakers use to refer to the inner and outer parts of bread?

Bakers call the inner, soft part of bread the crumb, while the outer hard portion is called the crust.

2. Bread was very important in medieval times, a fact that is illustrated by the Assize of Bread and Ale in the 13th century. What was the Assize of Bread and Ale?

The Assize of Bread and Ale was a law that set heavy punishments for short-changing bakers.

<u>Re-Phrased Versions of a Factual Question</u>

Version A. Most of the bread consumed in contemporary cultures is leavened. What is unleavened bread primarily used for today?

Version B. Bread can either be leavened or unleavened, but most people prefer leavened bread because it is lighter and more chewable. What primary purpose does unleavened bread serve nowadays?

Version C. Leavening is the process of adding gas to the dough before or during baking, but not all bread is leavened. In what context is most unleavened bread consumed in modern times?

> Response to All: *Unleavened bread has symbolic importance in many religions – its primary use today is in various religious rites and ceremonies.*

Re-Phrased Versions of a Conceptual Question

Version A. Flour contains proteins. How do these proteins contribute to the consistency or texture of bread?

Version B. Flour provides the primary structure to bread because it contains proteins. How is the texture of the bread determined by the proteins in the flour?

Version C. The quantity of the proteins in the flour used to make the dough determine the quality of the finished bread. What is the process through which proteins affect the texture of bread?

> Response to All: *When worked by kneading, the non-water soluble proteins in flour form a network of strands called gluten, which is responsible for the softness of the bread because it traps tiny air bubbles as the dough is baked. If the network of strands is more cohesive or tightly linked, the bread will be softer.*

Inferential Questions for Concepts (different domain)

1. The formula used for mixing concrete (which contains cement, sand, and water) is similar to the Bakers Percentage system used in making bread. If the primary ingredient in concrete is cement, how is concrete mixed according to the formula?

> *Like the Bakers Percentage system, the formula for concrete mixing requires the ingredients to be measured by weight instead of by volume. Cement is always stated as 100%, and the rest of the ingredients are a percent of that amount by weight.*

2. The polymer coating that is applied to nylon hot air balloons functions in a similar way to the gluten that is formed from proteins in bread making. Why is a polymer coating applied to nylon hot air balloons?

Gluten traps tiny air bubbles as the dough is baked. The polymer coating makes the nylon balloon impermeable to air, thus enabling the balloon to fly.

3. Much like the “quick bread” used by commercial bakers, many products that can be traced to ancient times have been enhanced recently by the use of chemical additives. Paint is one product that has been updated for a similar purpose to "quick bread" – what function do the chemical additives in paint serve?

"Quick bread" is the name that commercial bakers use for dough that does not require fermentation because of chemical additives, which speeds up mixing time. Similarly, chemical additives are used in paint to speed up the drying time.

4. Cladosporium is a type of mold that can induce asthmatic symptoms in people, and it can be found in many of the same places as the microorganism yeast that is used in bread making. While eliminating cladosporium would help to reduce asthma attacks, why would this task be difficult to achieve?

Yeast spores naturally occur everywhere, including in the air and many surfaces. Likewise, cladosporium occurs in the air and thus eliminating it would be very difficulty.

THE RESPIRATORY SYSTEM

Factual Questions

1. The lungs are the component of the human respiratory system responsible for gas exchange. What other major function do they serve?

The other major function of the lungs is to manage the concentration of hydrogen ion in the blood, an important factor in regulating the acidity of blood (pH), which must be kept in a narrow range.

2. The human respiratory system has many components. What is the function of the epiglottis?

The epiglottis is a flap of tissue that closes over the trachea when a person swallows so that food and liquid do not enter the airway.

3. The human respiratory system is controlled by the brain. What part of the brain controls voluntary breathing?

A region within the cerebral cortex, called motor cortex, controls all voluntary motor functions, including telling the respiratory center to speed up, slow down or even stop.

4. The respiratory systems of other animals differ from that of humans. What is an example of a land-going class of animals that conducts gas exchange through an organ that is not the lungs?

Amphibians conduct gas exchange through their skin. The skin of an amphibian is an important respiratory organ – it is highly vascularized and secretes mucus from specialized cells to facilitate rapid gas exchange.

Conceptual Questions

1. The human respiratory system depends on air pressure to facilitate breathing. How does air pressure help people inhale and exhale?

When a person inhales, the chest cavity expands, lowering the pressure in the lungs below the outside air pressure. Air then flows in the lungs (from high to low pressure). When a person exhales, the chest cavity gets smaller, increasing the pressure in the lungs above the outside air pressure. Air from the lungs (high pressure) then flows out to the outside air (low pressure).

2. Gas exchange occurs in a part of the human respiratory system called the alveoli. How does the process of gas exchange work?

Within the alveoli, gas exchange occurs through diffusion. Diffusion is the movement of particles from a region of high concentration to a region of low concentration. The oxygen concentration is high in the alveoli, so oxygen diffuses across the alveolar membrane into blood in the pulmonary capillaries. The concentration of carbon dioxide is high in the pulmonary capillaries, so it diffuses in the other direction.

3. In the human respiratory system, a low concentration of oxygen in blood can trigger breathing automatically. How does this occur?

Low concentration of oxygen in the blood will trigger an override by the autonomic nervous system. Specialized nerve cells within the aorta and carotid arteries called peripheral chemoreceptors monitor the oxygen concentration. If the oxygen concentration decreases, the chemoreceptors signal the respiratory centers in the brain to increase the rate and depth of breathing.

4. There are two main classes of breathing disorders that can affect the human respiratory system. How does each class of disorder affect the respiratory system?

Disorders of the respiratory system fall mainly into two classes. Some disorders make breathing harder, while other disorders damage the lungs' ability to exchange carbon dioxide for oxygen.

Inferential Questions for Facts (same domain)

1. The human respiratory system manages the concentration of various gases and particles in the blood. If the acidity of the blood gets too high, what part of the respiratory system acts to reduce the pH levels?

The lungs manage the concentration of hydrogen ion in the blood, which regulates the acidity of blood.

2. Aspiration pneumonia is an infection that develops due to the entrance of food or liquid into the lungs. If someone contracts aspiration pneumonia, what part of the respiratory system has broken down?

The epiglottis has malfunctioned – it is the flap of tissue that closes over the trachea when a person swallows so that food and liquid do not enter the airway.

3. A patient with brain damage cannot hold his breath, but he is able to otherwise breathe normally. What part of his brain is likely to be damaged?

The damage is likely to be in the cerebral cortex, and more specifically in the part of the motor cortex that controls voluntary respiration.

4. Both fish and amphibians have the ability to breathe underwater. Whereas fish breathe through gills, how do amphibians breathe underwater?

Amphibians breathe underwater through their skin, an important respiratory organ that is vascularized and secretes mucus from specialized cells to facilitate rapid gas exchange.

Inferential Questions for Concepts (same domain)

1. The pressurization system of a submarine that is 10,000 feet below the surface suddenly begins to malfunction, increasing the air pressure in the cabin. Assuming there is still plenty of oxygen in the cabin, how would the respiration of the crew be affected?

Exhalation of air would be more difficult. The human respiratory system depends on air pressure to facilitate breathing. If the outside air pressure is high, people will have trouble contracting their chest cavities far enough to raise the pressure in the lungs above the outside air pressure and exhale air.

2. If people are having trouble breathing, they are often given pure oxygen to inhale. How does breathing pure oxygen facilitate gas exchange relative to regular air?

Breathing pure oxygen increases the oxygen concentration in the alveoli, so oxygen will diffuse more rapidly across the alveolar membrane into blood in the pulmonary capillaries.

3. People can voluntarily control their breathing. Why can't a person stop breathing completely?

Low oxygen concentration in the blood will trigger an override by the autonomic nervous system.

4. Emphysema is a chronic disease that is characterized by loss of elasticity of the lung tissue. Which of the two main classes of breathing disorders does emphysema fall under?

Emphysema is a disorder that makes breathing harder because it restricts the lungs ability to expand. It does not damage the lungs' ability to exchange carbon dioxide for oxygen.

Control Questions (not tested during initial learning session)

1. The muscles between the ribs are critical to the respiratory system because they contract and expand the chest cavity. What are the muscles between the ribs called?

The muscles between the ribs are called intercostal muscles.

2. Most other mammals have a similar respiratory system to humans, but often with subtle differences. How is a horse's respiratory system different from that of a human?

Horses do not have the option of breathing through their mouths and must take in air through their nose.

Re-Phrased Versions of a Factual Question

Version A. The lungs are the component of the human respiratory system responsible for gas exchange. What other major function do they serve?

Version B. The main purpose of the human respiratory system is gas exchange, which occurs in the lungs. In addition to gas exchange, what is the other major function of the lungs?

Version C. Respiration depends on the lungs, but gas exchange is not the only bodily process that the lungs manage. What is the other major function that the lungs perform?

Response to All: *The other major function of the lungs is to manage the concentration of hydrogen ion in the blood, an important factor in regulating the acidity of blood (pH), which must be kept in a narrow range.*

Re-Phrased Versions of a Conceptual Question

Version A. The human respiratory system depends on air pressure to facilitate breathing. How does air pressure help people inhale and exhale?

Version B. The pressure of the air surrounding people's bodies is critical to their ability to breathe. Why does inhalation and exhalation depend on air pressure?

Version C. Air pressure plays an important role in breathing. How does inhaling and exhaling involve air pressure?

> Response to All: *When a person inhales, the chest cavity expands, lowering the pressure in the lungs below the outside air pressure. Air then flows in the lungs (from high to low pressure). When a person exhales, the chest cavity gets smaller, increasing the pressure in the lungs above the outside air pressure. Air from the lungs (high pressure) then flows out to the outside air (low pressure).*

Inferential Questions for Concepts (different domain)

1. The company Symantec has developed software for computer viruses that can be used for the same two purposes as vaccines are used for humans. What two purposes can the Symantec software be used for?

> *Most vaccines are used for prophylactic purposes, but they can also be used for therapeutic purposes. The Symantec software can be used to prevent computer viruses from hurting a computer and also to help reduce the damage caused by a computer virus once it has infected a computer.*

2. Controlled burning is a forest management technique used to prevent wildfires that relies on the same principle as the practice of inoculation does in vaccinating people. How does controlled burning work?

> *Inoculation is the practice of deliberate infection that produces a small, localized infection. Similarly, controlled burning involves setting small fires under controlled conditions that eliminate the dry brush that fuels wildfires and limits the risk of the fire spreading out of control.*

3. Research on renewable energy technology is very similar to vaccine development in that they both depend on the same incentives. Where does the primary incentive for research on renewable energy technology come from?

> *Like vaccine development, research on renewable energy technology relies on "push" funding that is supplied by government, universities, and non-profit organizations.*

4. Much like valence is used to describe vaccines, the term is also used in linguistics to describe the relationship between the arguments contained in a sentence and a verbal predicate. What is the difference between a monovalent verb and a polyvalent verb?

Similar to its use with vaccines, the term valence refers to the number of arguments controlled by a verbal predicate. Thus, a monovalent verb takes one argument and a polyvalent verb takes more than one argument.

THE INTERNET

Factual Questions

1. Although the Internet and the World Wide Web are often used synonymously, they are not the same thing. What is the difference?

The Internet is a global data communications system that uses hardware and software infrastructure to provide connectivity between computers, whereas the Web is a collection of interconnected documents and other resources that are communicated via the Internet.

2. Vinton Cerf and Robert Kahn published a famous paper in 1974 that advanced the technology used in the Internet today. What aspect of the Internet did they develop?

Vinton Cerf and Robert Kahn developed the standardized Internet Protocol Suite (TCP/IP), which is the language that modern-day computers use to communicate over the Internet. They published a paper describing the protocol suite in 1974.

3. One of the greatest things about the Internet is that nobody really owns it. What organization monitors and maintains it?

The Internet Society, a non-profit group established in 1992, oversees the formation of the policies and protocols that define how people use and interact with the Internet.

4. The Internet relies on the Domain Name System (DNS), which was created in 1983. What problem was the DNS created to solve?

The Domain Name System (DNS), which maps human-friendly computer hostnames into IP addresses automatically, was created in 1983 to solve the problem of organizing the exponentially increasing number of IP addresses.

Conceptual Questions

1. A first step towards the creation of the Internet was the construction of an early network that relied on packet switching technology. How does packet switching facilitate more efficient use of networks?

Packet switching is a mode of data transmission in which data is broken into chunks, called packets, which are sent independently and then reassembled at the

destination. Unlike alternative modes of data transmission, which require a fixed connection between terminals, packet switching technology can accommodate multiple users, optimizing network use and minimizing data transmission time.

2. In 1988, the U.S. Federal Networking Council approved the connection of MCI Mail system to the NSFNET. Why was this event important to the growth of the Internet?

The interconnection of the NSFNET to the commercial MCI Mail system in 1988 signaled the opening of the network to commercial interests, which greatly accelerated the expansion of what is now called the Internet. Motivated by potential profits, commercial companies aggressively pursued the connection of existing networks and the creation of new networks.

3. Communications companies that provide Internet service to individuals depend on Points of Presence (POPs) and Network Access Points (NAPs). What is the difference between POPs and NAPs?

A Point of Presence is a place for users to access an Internet service provider's network, often through a local phone number or dedicated line. In contrast, a Network Access Point is a physical infrastructure that allows different Internet service providers to exchange traffic between their networks.

4. Routers are crucial to the workings of the Internet. What two main functions do they serve?

Routers are specialized computers that have two main functions. First, routers ensure that information makes it to the intended destination by determining where to send it along thousands of pathways. Second, routers make sure that information doesn't go where it's not needed, which is crucial for keeping large volumes of data from clogging the connections of other users.

Inferential Questions for Facts (same domain)

1. A U.S. Senator proclaims: "We are going to develop the World Wide Web so that everyone in America will have access to the Internet." What is wrong with this statement?

The Internet is the hardware and software infrastructure, whereas the Web is a collection of resources that are communicated via the Internet. The infrastructure of the Internet is what needs to be developed in order to offer great access to the World Wide Web.

2. When people cannot connect to the Internet, sometimes the support technician will advise changing the TCP/IP settings on the computer. Why might changing the TCP/IP settings help people to connect to the Internet?

The standardized Internet Protocol Suite (TCP/IP) is the language that modern-day computers use to communicate over the Internet. Changing the specific setting might enable the computer to communicate more effectively.

3. Due to the increasing threat from computer viruses, a international committee was recently formed to consider remaking the Internet. What organization has the power to decide whether to create an Internet 2.0?

The Internet Society has the power to decide because it is the organization that oversees the formation of the policies and protocols that define how people use and interact with the Internet.

4. Website names often contain a suffix specific to the country in which they are hosted, such as Italy (.it), Canada (.ca), or Japan (.jp). Since the Internet has no national boundaries, what main purpose do these country codes serve?

The country codes are part of the Domain Name System, which helps to organize all the IP addresses that exist around the world.

Inferential Questions for Concepts (same domain)

1. In the packet switching technology used in the Internet, the packets of data always contain information about their relative position in the sequence of original data. Why is this sequence information critical to the successful use of packet switching?

Packet switching is a mode of data transmission in which data is broken into chunks, called packets, which are sent independently and then reassembled at the destination. Sequence information is critical to the successful reassembly of the original data and ensuring no data packets were lost.

2. Many rural areas of the United States still do not have Internet access. Based on how the Internet was developed after 1988, what is likely to be the main reason that these areas have not yet been connected?

After 1988, commercial companies expanded the Internet because they were motivated by potential profits. There is probably not much of a financial incentive to connect rural areas to the Internet.

3. A person trying to connect to the Internet via the phone hears the modem dial the number, a few rings, and then a brief connection before the call is dropped. What part of the Internet Service Provider's system is likely to be out of order?

The provider's Point of Presence is likely to be out of order. The POP is a place for users to access a provider's network, often through a local phone number or dedicated line.

4. At a large company, people are erroneously receiving emails that were intended for their co-workers. What is likely to be the problem with the email system?

> *The problem is likely to be the router for the email system. Routers ensure that information makes it to the intended destination and that information doesn't go where it's not needed.*

Control Questions (not tested during initial learning session)

1. In 1958, the United States created the Advanced Research Projects Agency (ARPA), an institution that later helped to start the Internet. What event prompted the creation of ARPA?

> *The launch of the Soviet satellite Sputnik in 1957 spurred the United States to establish ARPA.*

2. The workings of the Internet rely on Internet Protocol (IP) addresses. What two pieces of information does an individual IP address contain?

> *An IP address identifies both the individual computer and the network to which it belongs.*

Re-Phrased Versions of a Factual Question

Version A. Although the Internet and the World Wide Web are often used synonymously, they are not the same thing. What is the difference?

Version B. People commonly interchange the terms "the Internet" and "the World Wide Web", but these terms really mean very different things. What does each term refer to?

Version C. In describing life "online", people use many different terms. What is the difference between the Internet and the World Wide Web?

> Response to All: *The Internet is a global data communications system that uses hardware and software infrastructure to provide connectivity between computers, whereas the Web is a collection of interconnected documents and other resources that are communicated via the Internet.*

Re-Phrased Versions of a Conceptual Question

Version A. A first step towards the creation of the Internet was the construction of an early network that relied on packet switching technology. How does packet switching facilitate more efficient use of networks?

Version B. A new technology called packet switching was used in an early network that ended up as a precursor of the modern Internet. Why is packet switching important to network communications?

Version C. Packet switching, a technology first used in an early precursor of the Internet, is important to the efficiency of network communications. How does packet switching work?

> Response to All: *Packet switching is a mode of data transmission in which data is broken into chunks, called packets, which are sent independently and then reassembled at the destination. Unlike alternative modes of data transmission, which require a fixed connection between terminals, packet switching technology can accommodate multiple users, optimizing network use and minimizing data transmission time.*

Inferential Questions for Concepts (different domain)

1. When engineers move historic buildings from one location to another, they use a method similar to the packet switching technology used in Internet. What is the method that engineers use to move historic buildings?

> *Packet switching is a mode of data transmission in which data is broken into chunks, called packets, which are sent independently and then reassembled at the destination. Engineers use a similar method in which they take apart the building, move the pieces of the building to the new location, and then reassemble them.*

2. The expiration of an old "radio telephone" patent in 1983 had the same effect on the mobile phone industry as the connection of MCI Mail system to the NSFNET in 1988 did on the Internet. How did the expiration of the "radio telephone" patent affect the mobile phone industry?

> *The interconnection of the NSFNET to the commercial MCI Mail system in 1988 signaled the opening of the network to commercial interests, which greatly accelerated the expansion of the Internet. Likewise, the expiration of the "radio telephone" patent opened the mobile phone industry to commercial interests and led to its expansion.*

3. Every stock exchange has a clearing house that serves essentially the same function as a Network Access Point does for the Internet. How does a clearing house facilitate the activities of the stock exchange?

> *A Network Access Point is a physical infrastructure that allows different Internet service providers to exchange traffic between their networks. Similarly, a clearing house is a system that allows the members of a stock exchange to exchange information.*

4. The U.S. Postal Service processes mail by using machines called Multiline Optical Character Readers that serve the same purpose as a router does for the Internet. What are the main two functions of the Multiline Optical Character Readers?

> *Similar to routers, Multiline Optical Character Readers ensure that mail makes it to the intended destination by determining where to send it and also keep mail from going where it's not needed.*

Appendix C

Mean number of seconds taken to answer a question, review the feedback message, and total time spent per item on the three initial cued recall tests as a function of question type and initial learning condition for Experiments 1, 2, 3, and 4.

Experiment	Question Type	Learning Condition	Test 1			Test 2			Test 3		
			Q	FB	Total	Q	FB	Total	Q	FB	Total
1	Factual	Same Test	30.2	6.6	36.8	22.4	3.3	25.7	17.5	1.4	18.9
		Variable Test	29.3	5.5	34.8	25.2	3.5	28.7	19.2	1.9	21.1
	Conceptual	Same Test	44.6	10.6	55.2	43.3	5.9	49.2	32.6	3.2	35.8
		Variable Test	43.2	9.3	52.5	39.8	5.3	45.1	30.8	2.6	33.5
2	Factual	Same Test	34.0	6.2	40.2	25.1	4.1	29.2	18.9	2.0	20.9
		Variable Test	29.0	5.1	34.1	24.5	3.4	27.9	20.4	1.5	21.9
	Conceptual	Same Test	45.8	10.5	56.3	42.3	7.6	49.9	33.0	3.0	36.0
		Variable Test	41.2	8.8	50.0	41.8	5.9	47.7	32.6	2.9	35.5
3	Factual	Same Test	27.1	6.0	33.1	22.6	3.5	26.1	17.7	1.8	19.5
	Conceptual	Same Test	40.3	9.2	49.5	38.1	5.3	43.4	30.9	2.5	33.4
4	Conceptual	Same Test	57.4	10.8	68.2	46.1	7.1	53.2	33.5	2.7	36.2

Note. Q = Question, FB = Feedback, Total = Total time spent per item.