12-2023

Using Retrieval Practice, Variability, and Spacing to Facilitate Understanding of Complex Information

Rachel Peirce
Washington University in St. Louis

WASHINGTON UNIVERSITY IN ST. LOUIS

Department of Education

Using Retrieval Practice, Variability, and Spacing to Facilitate Understanding of Complex Information

by

Rachel Nicole Smith Peirce

A thesis presented to
Washington University in St. Louis
in partial fulfillment of the
requirements for the degree
of Master of Arts

December 2023
St. Louis, Missouri

Table of Contents

List of Figures

List of Tables

Acknowledgments

I would like to thank my committee members, Dr. Andrew C. Butler, Dr. Christopher Rozek, and Dr. Rowhea Elmesky, for their guidance and support throughout the development and interpretation of this research project. I would also like to thank Julian Kim, Camila Dayan, and Stephanie Johnstone for their help coding the data. This research is supported by the NSF Graduate Research Fellowship Program (DGE-2139839).

Rachel Nicole Smith Peirce

Washington University in St. Louis

December 2023

Dedicated to R.E.S. and Z.R.P.

ABSTRACT OF THE THESIS

Using Retrieval Practice, Variability, and Spacing to Facilitate Understanding of Complex Information

by

Rachel Nicole Smith Peirce

Master of Arts in Education

Washington University in St. Louis, 2023

Professor Andrew C. Butler, Chair

Combining the learning strategies of retrieval practice and variability has been shown to be effective in student learning (Butler et al., 2017; Pan & Rickard, 2018), but the temporal structuring of these learning strategies (i.e., massing versus spacing of practice) may benefit or hinder learning. Study 1 investigated whether the benefits of variable retrieval practice relative to repeated retrieval practice that occurs with massed practice extends to spaced practice. Participants watched geology videos that contained a total of 12 concepts, and then either answered three questions or read three study points about each concept. Each of the three questions and study points were presented two days apart with the first presented immediately after the corresponding video. Overall, variable practice produced significantly greater transfer than repeated practice, but there was no significant difference between retrieval practice and study points. The lack of a retrieval practice effect and low level of performance during the initial learning and final test phases with spaced practice suggests that learners may have struggled to connect repetitions of the same concept, especially in the variable retrieval practice condition. Study 2 replicated the basic design of Study 1, but manipulated the temporal structure of the initial learning sessions. Each of the three questions were either presented two days apart (i.e., *spaced*) or in succession after the corresponding video (i.e., *massed*). On the final test,

spaced practice resulted in greater transfer relative to massed practice. However, unlike Study 1, there was no advantage of variability. Overall, the present two studies provide evidence that the knowledge acquired during initial learning depends on how learning strategies are implemented, as combining strategies known to be beneficial for learning does not necessarily result in the greatest final test performance.

Introduction

Imagine a typical college class where students are expected to learn copious amounts of information, and apply their knowledge on an exam. Each lecture contains many important concepts, and the information is constantly being built upon from previous lectures, requiring students to make connections across lectures. After multiple lectures spanning weeks, or even months, college classes typically measure learning through an exam assessing recall of the learned information (i.e., *retention*), and/or application of their knowledge to a novel scenario (i.e., *transfer*). In preparation for the exam, a student might answer practice questions to strengthen their learning of the information (i.e., *retrieval practice*). However, a student might also choose to simply re-read information from their notes, lecture slides, or textbook when preparing for the exam (i.e., *re-studying*). How might a student's learning strategies affect their performance on an exam? In addition to choosing the method in which they choose to study, students might repeat the exact same learning activities (i.e., *repetition*), or engage in different variations of the same activity (i.e., *variability*). For example, a student might repeat practice quizzes with the same questions or complete practice quizzes with different questions testing the same underlying concepts. How does implementing variability in learning affect performance on the exam? Lastly, a student can choose how structure their study methods for the exam. A student might choose to space their studying over multiple days (i.e., *spacing*), or cram their studying to the night before the exam (i.e., *massing*). How might the structure of the learning affect their performance on an exam?

The goal of the present study is to simulate different versions of the above scenario to understand how learning activities and structure could be combined to facilitate transfer of knowledge. The scenario of the college class described above contains many evidence-based

learning strategies that have been shown to facilitate learning in isolation: retrieval practice, variability in practice, and spaced practice (Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013). However, less is known about the optimal implementation of these learning techniques used in combination to facilitate deep understanding of complex information. Given the known benefits of retrieval practice and variability in practice, the combination of these two learning activities should produce greater learning. Additionally, spaced practice has been shown to lead to greater learning (Carpenter et al., 2012). Combining retrieval practice, variability in practice, and spaced practice might produce the greatest learning, as each individual component can be beneficial, and potentially produce additive effects. However, the combination of the learning strategies may hinder learning, as learners may fail to connect instances of the same concept when the practice is variable and spaced. In the following sections, I will (1) review research on transfer of knowledge, retrieval practice, variability, and spacing, (2) explain the theories pertaining to the present research, and what they could predict, and (3) introduce the rationale of Study 1.

Background

The purpose of the following sections is to provide a comprehensive overview of the key terms, and prior research pertaining to the present studies (see Table 1). A conceptualization of how learners transfer their knowledge will be introduced first, as learners encounter multiple instances in which they must transfer their knowledge throughout the general study paradigm. Next, the general paradigm of the two studies will be introduced, followed by an introduction of each strategy that learners will engage in during the practice. In terms of the three learning strategies, the type of practice that learners will perform in the paradigm (i.e., retrieval practice and re-studying) will be discussed first in reference to the general paradigm. Afterwards,

variability in practice will be introduced to show how retrieval practice and re-studying could be modified across multiple repetitions. Lastly, the structure of the practice repetitions will be discussed in combination with the type of practice activity and variability in the practice. The section for each learning strategy is designed to build upon the information presented in the previous sections to demonstrate how these learning strategies should be implemented.

Table 1. Reviewing Key Terms with Examples.

Key Term	Definition	Example
Transfer of Knowledge	The use of knowledge in a new and unfamiliar context.	A learner is asked to apply a math formula from a lecture to a word problem on an exam.
Re-studying	Practice that requires the learner to read or study the information.	A learner reads a word and the corresponding definition.
Retrieval Practice	Practice that requires the learner to retrieve information from memory.	A learner is given a word and they have to recall the corresponding definition from memory.
Testing Effect	The finding that engaging in retrieval practice leads to higher test performance compared to re-studying.	A learner does better on an exam that they answered practice questions rather than re-reading their notes.
Repeated Practice	Practice that is exactly the same in content.	A learner practices the same question to understand a concept before an exam.
Variable Practice	Practice that is varied with different content.	A learner practices a variety of questions to understand a concept before an exam.
Variability	The finding that variability in practice leads to greater performance on new questions compared to no variability.	A learner does better on an exam when they study a variety of material compared to studying the exact same material.
Massed Practice	Practice that is completed in close succession.	A learner masses their studying for an exam in one sitting.
Spaced Practice	Practice that is completed over an extended amount of time.	A learner spaces their studying for an exam over multiple days.
Spacing Effect	The finding that spaced practice before a test leads to greater long-term learning than massed practice.	A learner does better on an exam when they spaced out their studying across multiple days instead of cramming the night before.

Transfer of Knowledge

Learning information in one context and applying this knowledge to a novel context, or the transfer of knowledge, is often necessary in a learner's everyday life (Anderson et al., 2001; Barnett & Ceci, 2002). The process of transferring knowledge distinguishes between two distinct contexts: one context in which the information is originally learned (i.e., the original situation), and a subsequent context in which the information is used (i.e., the new situation). In order to achieve transfer of knowledge, the learner must (1) *recognize* the prior information from the original situation is applicable to the new situation, (2) *recall* the applicable prior information from the original situation, and (3) *apply* the recalled information to the context of the new situation (Barnett & Ceci, 2002). These three steps must be completed in order for the learner to successfully transfer their knowledge from the original situation to the new situation.

A common distinction made by researchers is whether the original situation is similar to the new situation (i.e., *near transfer*) or dissimilar to the new situation (i.e., *far transfer*). When considering transfer in educational practice, a long-term goal of higher education is to achieve *far transfer*, as college students are expected to take their knowledge learned from their courses, and apply this knowledge when working a full-time job after receiving their degree. However, this distinction between near and far transfer can be ambiguous on how one might define the original situation and new situation as similar or dissimilar. Barnett & Ceci (2002) proposed a framework in which transfer be characterized in terms of different dimensions occurring on a continuum. A few of these dimensions include the knowledge domain from the original context compared to the knowledge domain in the new context (e.g., psychology class versus history class), the environment of the original and new contexts (e.g., a classroom versus at home), and the amount of time between the original and new contexts (e.g., days versus months). Using

these dimensions, the degree or "farness" of transfer could be defined as the aggregation of dissimilarities across dimensions. An example of transferring knowledge to a rather dissimilar context could be a situation where a college student learns about effective learning strategies from an online psychology class, and then applies this knowledge to an in-person physics class two years later. In this example, the original learning situation and new situation are very dissimilar in the dimensions described above, demonstrating an instance of *far transfer*.

Retrieval Practice

Retrieval practice, or the act of bringing information to mind, has proven to be an effective learning strategy compared to more passive learning strategies (Roediger & Karpicke, 2006; Roediger & Butler, 2011; Dunlosky et al., 2013). The testing effect, or the advantage of retrieval practice over re-studying, has generalized across a variety of conditions (for systematic reviews, see Adesope, Trevisan & Sundararajan, 2017 and Rowland, 2014). The testing effect paradigm typically involves three distinct stages: (1) presenting the to-be-learned information, (2) practicing the information by engaging in learning strategies (e.g., retrieval practice, re-studying), and (3) taking a criterial test to assess the degree of which the information has been learned (see Figure 1). Learners can be presented with new information through a variety of forms: text passages, videos, word lists, lectures, etc. After the presentation of the to-be-learned information, learners engage in practice activities to strengthen the to-be-learned information in memory. These practice activities typically utilize some form of retrieval practice (e.g., answering practice questions, using flashcards, explaining information to a friend), and/or passively re-studying information (e.g., re-reading notes, highlighting key concepts from the textbook). How learners practice the information, either through retrieval practice or re-studying, is crucial to their future retrieval success on the criterial test. The criterial test is self-paced, and assesses the learner on the to-be-learned information. Henceforth, I will refer to the first two

stages of displaying the to-be-learned information and engaging in practice as the *initial learning* phase, and the criterial test stage as the *final test* phase.

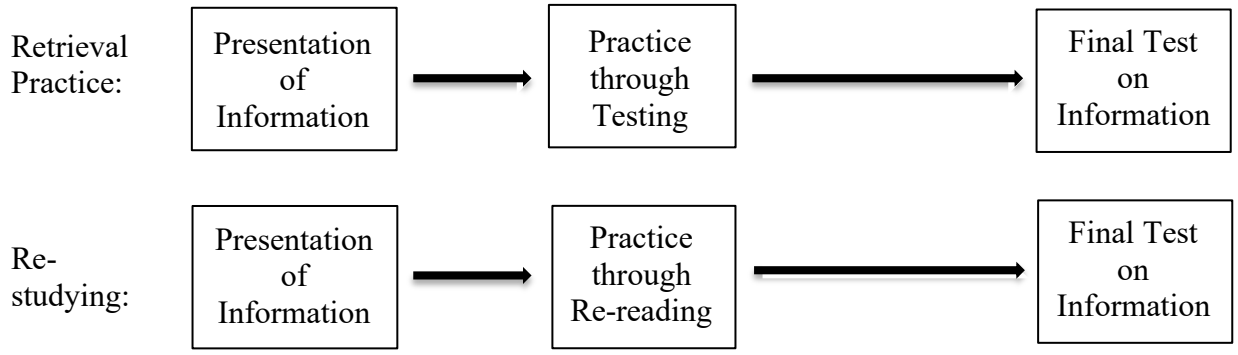

Figure 1. The Standard Testing Effect Paradigm.

Although there are many parameters that can be changed or manipulated within the testing effect paradigm, the testing effect has proven to have high generalizability. For example, providing feedback to learners after each retrieval attempt leads to a greater testing effect, even though not providing feedback to learners can produce a testing effect (Rowland, 2014). The retention interval, or the time between the end of the initial learning phase and the final test, has been measured in terms of minutes, days, weeks, and months. Retention intervals greater than one day led to a greater magnitude of the testing effect (Rowland, 2014), but shorter retention intervals produce reliable testing effects (Carpenter & DeLosh, 2005). The testing effect has been observed in simulated classroom environments (Butler & Roediger, 2007), and college classrooms beyond the typical lab setting (McDaniel, Roediger & McDermott, 2007; for reviews, see Agarwal, Nunes & Blunt, 2021; Yang et al., 2021). Numerous studies have shown that the testing effect is also present when transfer of knowledge is required on the final test (Butler, 2010; Carpenter, 2012; Hinze & Wiley, 2011), allowing the testing effect paradigm to generalize to educational settings where learners are often required to apply the to-be-learned information.

Pan and Rickard (2018) identified a three-factor framework to promote transfer in a testing effect paradigm: (1) response congruence, (2) elaborative retrieval practice, and (3) high practice performance. Response congruence refers to the match in format between the retrieval practice questions and final test questions (e.g., multiple-choice, free recall questions). If the retrieval practice questions match the format of the questions administered for the final test, transfer is more likely to occur than if the questions are different between retrieval practice and the final test. Second, elaborative retrieval practice can be achieved through broad encoding methods and elaborative feedback. Broad encoding methods instruct the learner to think of everything that comes to mind when answering the questions. Elaborative feedback, or extended and detailed feedback explaining *why* the answer to the question is correct, results in greater transfer of knowledge on the final test compared to simply presenting participants with the correct answer, or only providing the correct answer. Elaborative feedback facilitates understanding of the concept at a deeper level than only having the knowledge of the correct answer (Butler, Godbole & Marsh, 2013). Lastly, if a learner's performance on the retrieval practice questions is high (i.e., about 50% correct), then the learner demonstrates sufficient knowledge of the to-be-learned information. Creating connections within the learned information during retrieval practice promotes transfer to unlearned information present in the final test. When designing the present two studies, these three factors were heavily considered to facilitate transfer of knowledge.

Variability

Across the broader literature, variability in practice can be beneficial for transferring knowledge to a new context (see Raviv, Lupyan & Green, 2022 for a review). Providing learners with variability in the practice leads to lower initial performance during learning, but greater performance in instances that require transfer of knowledge. Prior studies have found benefits of

variability in the field of category learning (e.g., studying multiple animal and sound pairings versus the same pairing; Vukatana, Graham, Curtin & Zepeda; 2015), motor learning (e.g., performing a tennis shot in multiple locations on a court versus one location; Douvis, 2005), second language learning (e.g., being exposed to six speakers versus one speaker; Barcroft & Sommers, 2005), and problem solving (e.g., studying examples of geometrical problems with different values and problem formats versus examples with only different values; Paas & Merriënboer, 1994). However, research on the testing effect has primarily used the exact same initial learning items, repeated across multiple trials (e.g., Karpicke & Roediger, 2008; Roediger & Karpicke, 2006).

Figure 2 depicts how variability in the practice (i.e., *variable practice*) might be implemented in the testing effect paradigm as opposed to practice that is identical across repetitions (i.e., *repeated practice*). How variability can be implemented within initial learning has been done in few different ways. Empirical research on retrieval practice has incorporated variability by using the same sentence framing with different details (Glass, 2009), solving different arrangements of an anagram (Goode et al., 2008), re-phrasing the practice questions with identical answers (Butler, 2010), and providing different background information for face and name pairs (Smith & Handy, 2014; 2016). The results from these studies generally show a benefit of variability on the task requiring transfer of knowledge, but only implemented variability in the cues provided to learners (i.e., the wording or sentence structure of the practice questions), and rarely modified the target answer to the questions. To my knowledge, only two studies have investigated variability in initial learning by modifying both the content of the practice questions and the target answer, and this will be discussed further.

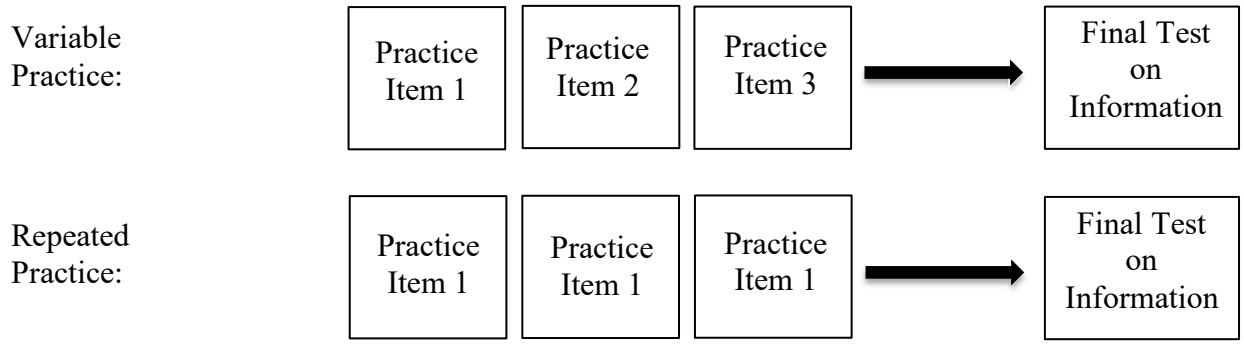

Figure 2. An Example of the Testing Effect Paradigm to Examine the Benefits of Variability. In both variable and repeated practice, the three practice items are related to the same concept.

Notably, Foss, Pirozzolo, and Kulesz (2023) found a benefit of variability in initial learning within an *Introduction to Methods in Psychology* college course. For the study design, four cumulative exams served as the initial learning questions for the class concepts, ending with a final cumulative exam of repeated and new application questions. Learners received half of the class concepts as repeated across the cumulative exams (i.e., *repeated* questions) and the other half of the class concepts were varied across the cumulative exams (i.e., *variable* questions). The time between the five exams ranged from 10 to 35 days, requiring learners to retain and apply information over long time periods. On the final cumulative exam, learners performed slightly better on the new application question when they received variable questions as opposed to repeated questions, but this difference was not statistically significant. This study provided descriptive evidence for the advantage of variable practice in a college course; however, the authors did not consistently change the answer to the questions, and thus, this study does not provide clear support for the benefit of variability in both the initial learning question and answer.

Perhaps the strongest evidence for benefits of variability in the initial learning questions and answers was a four-experiment research study conducted by Butler, Black-Maier, Raley, and

Marsh (2017). Butler and colleagues (2017) assigned learners to complete either repeated or variable practice for geological science concepts, and then learners answered new application questions for each concept two days later. In addition, learners engaged in either retrieval practice and re-reading study points equally during the initial learning phase. Engaging in variable retrieval practice produced superior transfer to new application questions compared to repeated retrieval practice. Therefore, variability in the initial learning led to greater transfer of knowledge across four lab-based experiments.

Spacing

Distributed learning, or practice repetitions separated by an amount of time, has been proven to be an effective strategy of structuring the initial learning sessions to produce greatest performance (for a review, see Carpenter, Cepeda, Rohrer, Kang, & Pashler, 2012). By spacing out the practice during the initial learning phase (e.g., answering questions, re-reading notes), information is encountered repeatedly at multiple time points as opposed to massing all of the repetitions of practice within one time period. Figure 3 illustrates how I will use different terminology for the time between initial learning sessions (i.e., the *spacing gap*), and the time between the practice sessions and final test, (i.e., the *retention interval*). Cepeda, Pashler, Wixted, and Rohrer (2006) reviewed the literature on distributed learning to identify potential factors for choosing the optimal length of the spacing gaps relative to the retention interval. When initial learning sessions were spaced more than two hours apart, final test performance increased compared to massing the initial learning sessions, regardless of retention interval length, providing evidence for an overall "spacing effect."

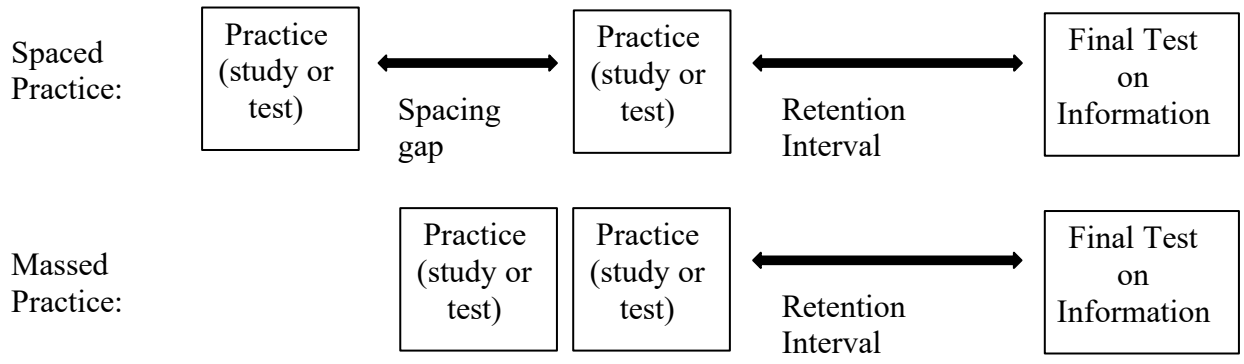

Figure 3. The Typical Paradigm for Investigating the Spacing Effect. Adapted from Kang (2016).

Despite the generalizability of the spacing effect, choosing the length of the spacing gaps and length of the retention interval has implications on the magnitude of the spacing effect. Cepeda, Vul, Rohrer, Wixted, and Pashler (2008) investigated long spacing gaps ranging from 0 to 105 days, and retention intervals between 7 and 350 days for retention of general knowledge facts. The findings showed that the optimal spacing gaps increase as the retention interval increases in the recall of facts, demonstrating that a shorter spacing gap of one day for a retention interval of seven days resulted in the greatest retention compared to longer spacing gaps (i.e., 7, 21, and 105 days). For a longer retention interval of 70 days, the optimal spacing gap between practice sessions was 21 days. This interaction for spacing gap and retention interval shows superior performance on increased spacing gaps for longer retention intervals, and superior performance on massed practice for shorter retention intervals (Maddox, 2016). Thus, I can conclude that the optimal relationship between the spacing gap and retention interval is a ratio of approximately 2:5.

Although there is generally a positive relationship between spacing gap and retention interval for retention of information with educationally relevant materials (e.g., Appleton-Knapp, Bjork & Wickens, 2005; Carpenter, Pashler & Cepeda, 2009; Rohrer & Taylor, 2006; Sobel,

Cepeda & Kapler, 2011), prior research is mixed about whether this relationship extends to the situation in which transfer of knowledge is required by the final test (Kang, 2016). There is some evidence that spaced retrieval practice can benefit transfer of knowledge to math concepts (e.g., Rohrer & Taylor, 2007), science concepts (e.g., Gluckman, Vlach, & Sandhofer, 2014; Kapler, Weston & Wiseheart, 2015), and categorization of artists' paintings (e.g., Kornell & Bjork, 2008; Kang & Pashler, 2012), but these studies did not necessarily address why the particular spacing gap and retention interval length were selected. Furthermore, the optimal ratio of spacing gap to retention interval length might not hold when variability in practice is added to spaced retrieval practice.

Theory

Although many theories for each of the three learning strategies exist, I will primarily use encoding variability theory to make predictions about findings when combining retrieval practice, variability, and spacing. Encoding variability theory is a dominant theoretical account for explaining the mnemonic benefits of retrieval practice, variability, and spacing, and thus, was chosen as the primary theoretical explanation in the present research. Encoding variability theory states that information is encoded differently for each presentation of the to-be-learned information due to the fluctuation of encoded features over time, leading to multiple retrieval routes for information. Thus, multiple retrieval routes allow for increased access to memory representations through cues, which increases the likelihood of retrieval (Estes, 1955; Melton, 1967; Bower, 1972; Bjork, 1975). Glenberg (1979) specified three categories of features (or components) that can be encoded with to-be-learned information: contextual features, structural features, and descriptive features. Contextual features relate to the general context in which the to-be-learned information is encoded, such as the physical environment, time of day, and learner's affective state. On the other hand, structural features pertain to the cognitive processing

that learners may engage in to structure the items of to-be-learned information during encoding. For example, learners may create visual imagery to connect multiple items of the to-be-learned information, and thus, strengthen the association across items to be later retrieved. Lastly, descriptive features are semantic memory representations of the information, including the meaning, pronunciation, and spelling of the individual words of the to-be-learned information. Unlike the contextual and structural features, descriptive features do not change across presentations of to-be-learned information, as these features cannot be changed by the learner. In the following paragraphs, I will discuss encoding variability theory as an explanation for retrieval practice, variability, and spacing individually, before hypothesizing how encoding variability theory could be used make predictions about combining these learning strategies.

Theoretical accounts within the retrieval practice literature have posited the benefit of encoding more contextual, structural, and descriptive features during the initial presentation of to-be-learned information. One theory in particular, the elaborative retrieval hypothesis, states that retrieval practice leads to greater semantic elaboration of the information, and activates relevant semantic knowledge, which results in better performance on the final test for information that has been retrieved. Semantic elaboration of items during retrieval practice can generate additional associations to the information, which promotes multiple retrieval routes, and thus, deeper processing of the information (Carpenter, 2009; Carpenter & DeLosh, 2006). Connecting back to ideas of encoding variability, learners must encode descriptive features of the to-be-learned information to create structural features through elaboration, and in turn, these structural features are strengthened through retrieval. Another theoretical explanation for the mnemonic benefits for retrieval practice is the episodic context hypothesis proposed by Karpicke, Lehman, and Aue (2014). The episodic context hypothesis states that retrieval practice

allows for the reinstatement of contextual features present in the initial encoding, leading to greater final test performance. During the initial encoding, learners tie various contextual features into the memory representation of the to-be-learned information. When learners attempt to retrieve an item from memory at a later time, they partially rely on these contextual features from the initial encoding context. According to the episodic context hypothesis, reinstating the temporal context, along with other contextual features, during retrieval will be more beneficial if the encoding and retrieval practice context are different, as the to-be-learned information has increased unique contextual features from multiple retrieval instances. By performing multiple retrievals to produce a variety of contextual features (i.e., variability in practice), the learned information might possess similar contextual features present on new application questions on a final test.

Encoding variability theory is just one of many theories explaining the benefits of variability across the domains of categorization, motor learning, and language acquisition (see Figure 3 in Raviv, Lupyan & Green, 2022). Almost all of the theories explaining the benefit of variability posit why lower initial performance in variable practice leads to greater performance on a test requiring transfer, and build upon similar ideas present in encoding variability theory. For example, Bayesian inference models of language learning suggest that even though there is low performance on initial learning instances, providing highly different practice items allow for learners to update their prior expectations during initial practice, promoting transfer of knowledge on a final test (Tenenbaum & Griffiths, 2001). Focusing on encoding variability theory in a testing effect paradigm, variability is likely to appear in the contextual features learners encode with the to-be-learned information to connect across repetitions (e.g., connecting to prior knowledge, using visual imagery), and descriptive features manipulated by the

experimenter (e.g., the content of the practice materials). Since prior research has primarily only implemented variability in cues provided to learners (e.g., changing the wording of a question) and not manipulated variability in the target response, the descriptive features could be the most salient to the learners, and the structural features would be less salient. Using different target responses across repetitions requires the utilization of structural features connect these instances. How learning strategies are implemented alters the contextual, structural, and descriptive features encoded with the to-be-learned information, which can influence their effectiveness when used in combination.

A key factor of encoding variability theory is the amount of time that passes between retrieval attempts, which allows for the change in descriptive, structural, and contextual features to change, and establish multiple retrieval routes for future retrieval attempts (Estes, 1955; Melton, 1967). In the spacing literature, encoding variability theory has served as the primary explanation for the benefits of spaced retrieval practice (see Maddox, 2016). When practice repetitions are massed, the contextual features of the to-be-learned information remain relatively unchanged. Incorporating spacing of repetitions increases the number of unique contextual features encoding with the to-be-learned information, as the learner's environment and mental state might be different across repetitions. Although massed practice would have fewer, but stronger contextual features to rely upon for subsequent retrieval, a final test that is days after the massed practice is unlikely to share the same contextual features encoded previously, resulting in lower final test performance. Compared to massed practice, final test performance is greater when preforming spaced practice due to the potential overlap in contextual features encoded during the initial retrieval attempts and the final test (Maddox, 2016).

To summarize, encoding variability theory highlights the role of various features that are encoded with the to-be-learned information to aid future retrieval. All of the theoretical accounts mentioned have hypothesized how different types of variability are present in contextual, structural, and descriptive features. Encoding variability theory could be used to predict and explain potential findings when combining retrieval practice, variability, and spacing. In one plausible scenario, combining retrieval practice, variability, and spacing allows for learners to encode a variety of contextual and descriptive features, and potentially use structural features to connect across variable repetitions. However, if learners do not recognize that variable repetitions should be connected together due to the increased spacing gaps, and fail to encode structural features of the to-be-learned information, then combining variability and spacing would be detrimental on a final test requiring transfer. Adding retrieval practice to variability and spacing could strengthen the structural features of the to-be-learned information through the act of retrieval. Although the three learning strategies may have additive benefits based on their individual theoretical predictions, it is possible that the same underlying mechanism boosts encoding for each learning strategy in isolation. Thus, combining learning strategies may result in sub-additive benefits, or lower combined performance than hypothesized, due to the redundancy of contextual, structural, and descriptive features (Begg & Green, 1988).

Study 1

The ability to utilize multiple learning strategies in combination is critical to achieve learning of complex material. Building on the findings from Butler et al. (2017), the present study investigated the impact of spaced learning on transfer of knowledge in a testing effect paradigm. Combining learning strategies of retrieval practice and variability in practice has been shown to be effective in student learning, but temporal structuring of the learning activities (i.e.,

massing versus *spacing*) may benefit or hinder learning. Thus, the goal of Study 1 was to examine these well-known learning activities and structures in combination during initial learning in the testing effect paradigm. Learners watched videos containing geological science concepts, and then either answered three application questions (i.e., *retrieval practice*) or read three study points (i.e., *re-studying*) for each concept. In addition, the three questions or study points presented to learners were either the exact same (i.e., *repeated practice*) or different (i.e., *variable practice*) across repetitions. On the final test, learners were presented with new application questions, assessing transfer of knowledge. Given the many variables and potential interactions in the present study, I proposed four hypotheses:

1. Engaging in retrieval practice will produce greater final test performance on new application questions compared to re-studying information based on the wealth of research on transfer in the testing effect paradigm (Pan & Rickard, 2018; Butler, 2010).
2. Engaging in variable practice will produce greater final test performance on new application questions compared to repeated practice, given the slight positive evidence for variable practice in prior research (Butler et al., 2017; Foss et al., 2023).
3. With the proposed advantage of retrieval practice and variable practice in the prior two hypotheses, learners that engage in variable, retrieval practice should result in the greatest performance (Butler et al., 2017).
4. Introducing spacing to variable retrieval practice could result in either a benefit or detriment to final test performance on the new application questions. On one hand, including spacing in variable, retrieval practice could benefit understanding of information spaced over multiple days through the principle of spacing introducing various contextual, structural, and descriptive features that facilitate transfer of

knowledge. Alternatively, participants might not make connections when the questions are different and spaced over multiple days, resulting in little to no transfer of knowledge.

Methods

Participants

Seventy-five students from the University of Texas, Austin participated in this study. Participants were recruited through the UT Austin Educational Psychology Subject pool, and received course credit for their participation. We planned to exclude participants if (1) their final test scores were outside of +/- 2 standard deviations, (2) they copied and pasted the feedback or their same answers across questions, and/or (3) the mean time to answer the questions was outside of +/- 2 standard deviations. None of the participants met the planned exclusion criteria, and therefore, they were all included in the final sample.

At the end of the study, participants self-reported their gender, race/ethnicity, and prior knowledge for geological science. There were slightly more female participants than male participants (57.3% female), and a variety of races/ethnicities: White (57.3%), Asian (25.3%), Black (6.7%), multiracial (4.0%), and Other (6.7%). When participants were asked if they had taken at least one geology course previously, only 14.7% reported taking one or two introductory geology courses at UT Austin.

Design

A 2 (Practice Activity: Retrieval Practice, Restudy) x 2 (Practice Type: Repeated, Variable) design was adopted for this study. Practice activity was manipulated within subjects, between concepts (see Table 2). Practice type was manipulated between-subjects. The presentation of the materials was counterbalanced across participants in three ways by rotating the three questions during the practice sessions through six order positions, creating six order

versions of the study (i.e., three order positions for the same condition and three order positions for the variable condition). A second counterbalancing method was necessary to assign concepts to retrieval practice questions and study points across participants, requiring two practice activity versions of the study. Participants were assigned retrieval practice questions for the odd-numbered concepts and study points for the even-numbered concepts, or vice versa. As a result of the two counterbalancing methods, there were a total of 12 versions of the study (see Appendix B). The main dependent variable was the accuracy on the final test questions.

Table 2. Design of Study 1.

	Initial Learning Sessions			Final Test
Condition	Day 1	Day 3	Day 5	Day 7
Retrieval – Repeated	V R_1	R_1	R_1	R_4
Retrieval – Variable	V R_1	R_2	R_3	R_4
Study – Repeated	V S_1	S_1	S_1	R_4
Study – Variable	V S_1	S_2	S_3	R_4

Note. V = videos on geological sciences. R = retrieval practice questions. S = study points. Subscripts denote whether the retrieval practice question or study points were the same (repeated condition) or different (variable condition).

Materials

The study materials, as used in Butler et al., (2017), consisted of five mini-lecture style videos on geological science concepts from a course titled "Nature of Earth: An Introduction to Geology" produced by The Great Courses. Each video clip was approximately eight minutes long, and contained 2-3 concepts with a concept defined as a piece of information that is integrated across multiple sentences in the video (see Table 3 for the 12 concepts).

Table 3. Concept Names and Primary Learning Objectives.

Concept Number	Concept Name	Primary Learning Objective
1	Earth's structure	Recognize the relationship between the varying densities of the core, mantle, and crust
2	Layers beneath the surface	Understand the relationship between the asthenosphere and lithosphere (i.e., the lithosphere floats on the asthenosphere)
3	Convection cell systems	Apply the process of convection (i.e., hot air rises and cold air falls) to a similar process (e.g., the water cycle)
4	Rift zones vs. subduction zones	Compare how tectonic plates move for rift and subduction zones (i.e., rift = the plates diverge; subduction = the plates collide)
5	Locations of different types of magma	Recognize which type(s) of magma correspond to tectonic plate zones
6	Explosiveness depends on gas content of magma	Understand the relationship between amount of gas in magma and explosive power: more gas = bigger explosion
7	Viscosity of magma depends on silica content	Understand the relationship between viscosity of magma and silica content: higher silica content = greater viscosity
8	Explosiveness depends on consistency of magma	Understand the relationship between viscosity of magma and explosive power: greater viscosity = bigger explosion
9	Compressive vs. tensional forces	Apply compressive and tensional forces to structural impacts of these forces (e.g., specifying the forces present in arches vs. columns)
10	Elasticity and the release of energy	Recognize the amount of energy released differs between compressive and tensional forces: compressive forces release large amounts of energy while tensional forces release smaller amounts of energy
11	Compression vs. shear waves	Understand the difference in pathing for compression and shear waves: compression waves move horizontally while shear waves move both horizontally or vertically
12	Focus vs. epicenter	Compare the relative location between the focus and epicenter for an earthquake (i.e., the focus is underground and the epicenter is on the surface ground)

The initial learning and final test materials consisted of four retrieval practice questions with feedback, and four study points for each concept. Therefore, there were a total of 48 questions and 48 study points across the 12 concepts. Thirty-six questions and study points were used for the initial learning trials, reserving 12 questions to be presented only on the final test. The 36 questions and study points in the initial learning phase allowed for the 12 counterbalancing versions to be implemented. Each retrieval practice question required learners to apply knowledge from the video in a short-answer format, tapping into higher-order learning (see Appendix A for examples of questions and study points for the first concept). Elaborative feedback was presented after each question. The study points included identical information as the question and feedback, but rephrased to be a paragraph. Thus, every participant received the same information during the initial learning phase. All materials were presented to participants in a Qualtrics survey (www.qualtrics.com).

Procedure

Sessions 1-3: Initial Learning. The three sessions were spaced two days apart, lasting a total of two hours. Session 1 took place in a computer lab, where participants were presented with the geology videos, and then either read study points or answered questions. Participants were told that they would watch a series of five videos, and after each video, they would read a study points and answer questions requiring application of knowledge from the video. Participants were required to provide an answer for each question, and were instructed to give a complete response that consists of a few sentences. Participants were informed that some questions and study points might be repeated, and if so, then they should provide a complete response like before. Session 1 lasted approximately one hour.

Sessions 2 and 3 consisted of answering six questions and reading six study points. These two sessions were completed by participants remotely, each lasting approximately 30 minutes. An experimenter emailed participants a Qualtrics survey link with either same or different questions and study points according to the participant's counterbalancing version. Participants received the same instructions as in the previous session, and were required to complete the session within 12 hours of receiving the email to maintain the spacing of the sessions.

Session 4: Final Test. Two days after Session 3, participants completed a self-paced final test in the computer lab, lasting one hour. All participants answered 12 new inference questions in a random order. Like the previous sessions, participants were instructed to answer each question in a few sentences, but they did not receive feedback after each question. A second phase of the final test required participants to answer the same 12 questions with additional context provided; however, the analyses for these data will not be reported to limit the scope of the present study.

Analytic Approach

Data analysis was performed in R and RStudio (R Core Team, 2023). Descriptive statistics including the mean, standard deviation, and standard error were calculated for the initial learning trials and final test performance to assess the sample distribution. Inferential statistics of t-tests and mixed ANOVAs were computed to assess differences across the initial learning trials and final test performance. If the assumption of sphericity was violated in the mixed ANOVA, then the Greenhouse-Geisser correction was applied. For the third initial learning trial, 37 responses out of a total 450 responses were lost due to an error in data collection.

Results

Coding

All question responses were independently scored by two coders. Each response was marked as either correct (1) or incorrect (0), and interrater reliability was calculated using Cohen's Kappa (κ). The interrater reliability between the two coders was moderate for the final test (κ = .75), and the discrepancies were resolved by discussion between the two coders. Next, the coders graded the three initial learning responses, and each response was marked as either correct (1) or incorrect (0). The interrater reliability between two coders for the all three initial learning sessions were moderate (κ = .64). Discrepancies were discussed and resolved amongst coders.

Counterbalancing Analyses

Analyzing potential differences across the counterbalancing conditions was necessary to ensure consistency before the planned analyses. The counterbalancing analyses combined the two counterbalancing methods (i.e., order version and practice activity version) with the practice activity for each concept (retrieval practice or study). The order version combined the same practice question selected for the repeated condition, and the order of the practice questions for the variable condition to collapse across practice type. A 2 (Practice Activity: Retrieval Practice, Study) x 3 (Order Version: 1, 2, 3) x 2 (Practice Activity Version: 1, 2) mixed ANOVA examined if there were significant differences across the counterbalancing versions. Results revealed no significant differences across the counterbalancing versions nor significant interactions with practice activity (all Fs < 2.89, ps > .06). With the lack of main effect for the counterbalancing versions, and significant interactions between counterbalancing versions and

practice activity, the counterbalancing conditions were not included as a variable of interest in further analyses.

Initial Test Performance

Table 4 presents the mean accuracies for the retrieval practice conditions over the three practice trials. An independent samples t-test confirmed that performance did not differ significantly between the repeated and variable conditions for the first question, $t(72.86) = 0.33$, $p = .75$, $d = 0.08$. Comparing the trends across practice trials for repeated and variable conditions, participants increased in performance when shown the repeated question, whereas performance in the variable condition stayed relatively constant across the three questions. A 3 (Practice Trial: 1, 2, 3) x 2 (Practice Type: Repeated, Variable) mixed ANOVA with a Greenhouse-Geisser correction assessed differences in practice performance. There was a main effect of practice trial, $F(1.95, 142.62) = 5.32$, $p = .01$, $\eta_p^2 = 0.07$, as participants' performance increased across the practice trials. Post-hoc analyses revealed higher performance on the third question compared to the first question, $t(73) = 2.74$, $p = .02$, $d = 0.32$, and the second question, $t(73) = 3.10$, $p = .01$, $d = 0.36$. Both the main effect of practice type, and the interaction between question order and practice type were not significant (all Fs < 3.00, ps > .10).

Final Test Performance

Participants performed equally well on the final test when they engaged in retrieval practice ($M = .32$, $SE = .03$) compared to reading study points ($M = .30$, $SE = .03$), demonstrating no overall benefit of retrieval practice compared to re-studying. Collapsing across practice activity, participants given different questions and study points in the practice performed better on the final test ($M = .37$, $SE = .03$) than participants given the same question or study point ($M = .25$, $SE = .03$). A 2 (Practice Type: Repeated, Variable) x 2 (Practice Activity: Retrieval Practice,

Study) mixed ANOVA confirmed a significant main effect of practice type, $F(1, 73) = 5.95$, $p = .02$, $\eta_p^2 = 0.08$, as the variable condition produced superior performance than the repeated condition. Regardless of practice activity, variability in the practice sessions resulted in greater transfer in the final test. Both the main effect of practice activity and the interaction between practice activity and practice type were not significant (all Fs < 2.60, ps > .11). Table 4 lists the mean accuracies of the four conditions on the final test.

Table 4. Mean Accuracy on the Practice Trials and Final Test.

	Practice Trials			
Condition	First	Second	Third	Final Test
Retrieval – Repeated	.41	.45	.54	.29
Retrieval – Variable	.39	.34	.41	.36
Study – Repeated	–	–	–	.22
Study – Variable	–	–	–	.38

Note. Only retrieval practice conditions have accuracies for the practice sessions because the study conditions did not require a participant response.

Discussion

Study 1 investigated the effects of retrieval practice and variability in practice assessing transfer of knowledge. The results of Study 1 showed a benefit of variability in practice on final test sessions in that when participants answered different questions or re-read different study points, they performed better when answering new questions on the final test compared to repeated questions or study points. However, there was no overall benefit of testing compared to restudying information (i.e., no *testing effect*). Thus, learners were able to gain a deeper understanding of the geological science concepts regardless of whether they engaged in retrieval practice and read study points. This finding was surprising, given the presence of the testing

effect in Butler and colleagues (2017), and clear evidence for transfer to occur in a retrieval practice paradigm (Pan & Rickard, 2018). One explanation for the lack of the testing effect could be due to learners' low performance during retrieval practice. Clearly, learners struggled to apply information from the videos on the first question presented and answer subsequent questions regardless of variability condition. Although there are many factors in the current study, the challenge of answering application questions spaced over multiple days might have created a situation in which retrieval practice is equally as beneficial as re-studying the material.

Compared to Butler et al. (2017, Experiment 4), the general patterns of Study 1 seemed to mirror the massed practice, as variable practice consistently produced greater transfer of knowledge on the final test with new application questions. However, the final test performance in Study 1 was lower across all conditions relative to Butler and colleagues (2017), presumably due to the increased spacing of the practice sessions (see Figure 4). The low performance in the spaced practice reduced the degree to which learners could reach a high performance on the final test. Looking at performance on the third practice trial in Butler and colleagues (2017), participants reached above 90% accuracy in the repeated condition and 60% accuracy in the variable condition. The accuracy in both conditions was quite higher than the practice performance in the current study (see Table 4), providing evidence for how spaced practice could hinder transfer of knowledge.

Although Study 1 allows for an indirect comparison of massed and spaced practice, an experimental manipulation needs to be done to clearly assess the effects of spacing within the same study, as these findings seemingly contradict an abundance of research findings demonstrating the benefits of spacing. Furthermore, either the length of the spacing gap between practice trials, or length of the retention interval could be modified to better reflect the optimal

ratio to potentially produce greater performance (Cepeda et al., 2008). Study 2 was designed to address these gaps, and provide more clarity about the possible benefits of combining multiple learning strategies to facilitate transfer of knowledge.

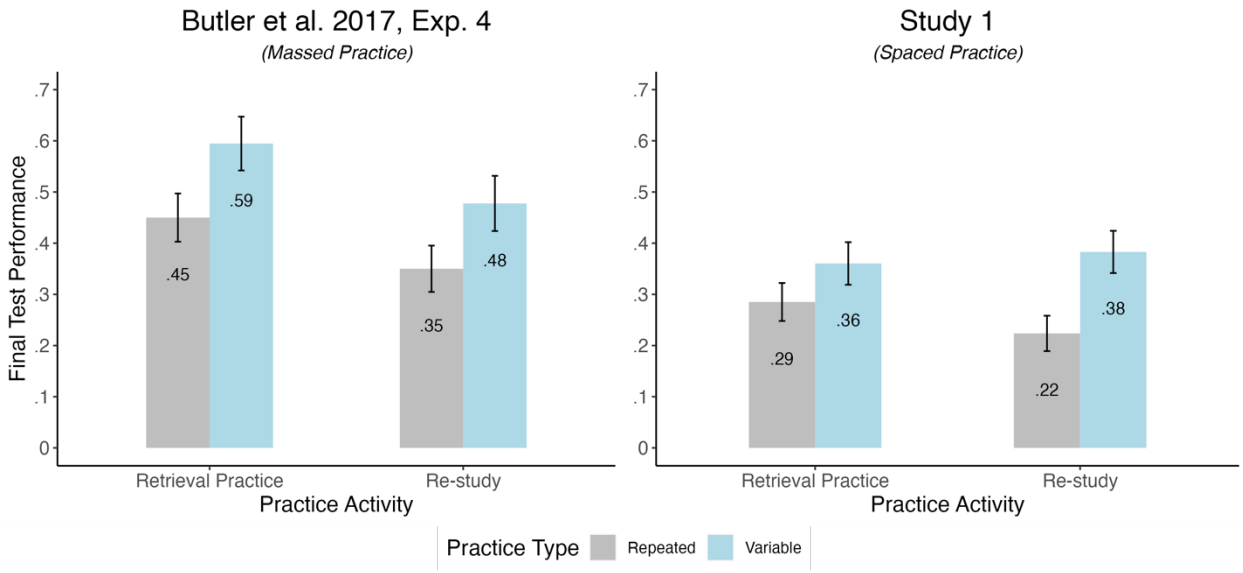

Figure 4. Final Test Performance for Butler et al. (2017) and Study 1. Massed practice results were taken from Butler et al. (2017, Experiment 4) that used the same materials and design as the current study. Error bars represent +/- one standard error.

Study 2

Study 2 conceptually replicated the basic design of Study 1 with a few changes. Most notably, the temporal structure of the practice was manipulated to either be spaced out in two-day intervals (i.e., *spaced*) or completed in succession (i.e., *massed*). The massed condition followed the same procedure to Butler and colleagues (2017, Experiment 4). The low performance on the practice questions and final test in Study 1 could require a longer retention interval to adhere to the optimal ratio between the spacing gap length and retention interval length, and thus, observe the benefits of spacing on final test performance (Cepeda et al., 2008). Therefore, the retention interval was increased for the spaced condition from two days to four

days. To make comparisons between Study 1 and Study 2, the spacing gap remained the same length for the spaced condition (i.e., two days). To minimize the number of manipulated variables and isolate the conditions favorable for learning, only retrieval practice was implemented as the practice activity, dropping the re-studying conditions. A final change from Study 1 was allowing for participation to occur completely remote to ease data collection. To ensure participants were engaged during the sessions, Study 2 utilized attention checks after the videos, and recorded how long participants took to answer the questions.

In line with findings from Study 1, I hypothesized that variability in the practice questions would produce greater final test performance than repeated practice questions. Furthermore, the changes to the method performed in Study 2 were intended to produce a spacing effect. On one hand, the increased retention interval combined with spaced practice could lead to greater final test performance due to a more optimal ratio between the spacing gap length, and retention interval length. However, the increased retention interval could result in lower final test performance, as the combination of the conditions may not be ideal for observing the predicted effects. Given the hypothesized benefit of variability and spacing independently, combining the two favorable conditions for learning (i.e., spaced, variable practice) should produce the greatest final test performance.

Methods

Participants

The sample size was selected based on an a-priori power analysis using G*Power (Faul, Erdfelder, Lang & Buchner, 2007). The effect size was chosen based on the main effect for practice type (repeated or variable) from Experiment 4 in Butler et al. (2017). To achieve a Cohen's *f* of 0.24 with at least 90% power at a .05 significance level, at least 64 observations

should be obtained for analysis. Thus, it was determined a-priori that approximately 70 individuals should complete the entire research study to achieve sufficient power.

After recruiting students through the WUSTL Psychological and Brain Sciences Participant Pool during the 2023 spring semester, only 14.5% of the individuals that signed up to participate in the study finished all four sessions. This low completion rate could have been due to the challenging nature of the materials, and/or other priorities that disrupted participants' adherence to the spacing of the sessions. The closure of the participant pool at the end of the spring semester prompted the remainder of data collection to occur through CloudResearch, a third-party research recruitment platform. CloudResearch allows for researchers to reach individuals interested in research participation across the world (Litman, Robinson, & Abberbock, 2017). For individuals on CloudResearch to be eligible to participate, they had to be based in the United States, have completed at least 50 research studies previously on the platform, and have at least a 90% approval rate by researchers on prior studies. In addition to the eligibility criteria before beginning the study, participants were also excluded if failed to complete all four sessions of the study, failed more than two of quality control questions (i.e., "What is the second word in this sentence? Answer: *is*"), and/or failed more than two questions about the instructions.

In total, 68 participants between the ages of 18-50 completed their participation either through the WUSTL participant pool (n = 16), or CloudResearch (n = 52). Additional exclusion criteria were established due to data collection occurring completely remote. After data collection concluded, I excluded participants if (1) they did not follow directions (n = 3), (2) they did not type their ID number correctly across sessions (n = 2), (3) they copied the feedback and their previous answers entirely (n = 1), or (4) I assigned them to the wrong condition (n = 1).

Other exclusion criteria from Study 1 were assessed in Study 2, but didn't result in the exclusion of participants. The final sample included 61 participants: 12 participants from the WUSTL participant pool and 49 participants from the CloudResearch platform.

The same demographic information of gender, race/ethnicity, and prior knowledge was self-reported by participants at the end of the study. There were slightly more male participants than female participants and those that preferred not to specify (54.1% male, 44.3% female, and 1.6% preferred to not say). Participants also reported a variety of races/ethnicities: White (62.3%), Asian (14.8%), multiracial (9.8%), Black (8.2%), Hispanic (3.3%), and Pacific Islander (1.6%). When participants were asked if they had taken at least one geology course previously, 19.7% reported taking one or two introductory geology courses in high school or college.

Design

A 2 (Practice Structure: Spaced, Massed) x 2 (Practice Type: Repeated, Variable) design was used for this study. Practice type was manipulated between-subjects (see Table 5). Practice structure was manipulated within subjects, and counterbalanced across the 12 concepts. As done in Study 1, the presentation of the materials was counterbalanced across participants in three ways to rotate the three questions during the practice sessions through six order positions, creating six order versions of the study (i.e., three order positions for the same condition and three order positions for the variable condition). A second counterbalancing method assigned concepts to as massed and spaced across participants, requiring two practice structure versions of the study. Participants were assigned to massed practice for the odd-numbered concepts and spaced practice for the even-numbered concepts, or vice versa. As a result of the two counterbalancing methods, there were a total of 12 versions of the study (see Appendix B). The primary dependent variable was the accuracy on the final test questions.

Table 5. Design of Study 2.

Condition	Initial Learning Sessions			Final Test
	Day 1	Day 3	Day 5	Day 9
Massed – Repeated	V R_1 R_1 R_1			R_4
Massed – Variable	V R_1 R_2 R_3			R_4
Spaced – Repeated	V R_1	R_1	R_1	R_4
Spaced – Variable	V R_1	R_2	R_3	R_4

Note. V = videos on geological sciences. R = retrieval practice questions. Subscripts denote whether the retrieval practice question were the same (repeated condition) or different (variable condition).

Materials

The materials were identical to Study 1.

Procedure

The procedure was similar to Study 1 with a few differences. All sessions were conducted remotely, as opposed to both in-person and remote sessions for Study 1. Therefore, two attention check questions were included in the first session to confirm that participants are attending to the information presented in the videos (i.e., “Describe one of the images from the video in 1-2 sentences.”). In addition, participants will be asked about their compliance with the instructions at the end of each session (i.e., “Did you comply with the instructions throughout the study? Your answer will not affect your pay, but please be honest. We want to make sure the results of our study are valid.”). Compensation was adjusted to reflect the rates competitive for the CloudResearch platform.

Analytic Approach

The analytic approach for Study 2 was similar to Study 1, conducting the primary data analyses in R. Descriptive statistics was reported for the practice sessions and final test performance. A 2 (Practice Structure: Spaced, Massed) x 2 (Practice Type: Repeated, Variable) mixed ANOVA was performed to assess differences in final test performance. A 3 (Question Order: 1, 2, and 3) x 2 (Spacing: Spaced, Massed) x 2 (Variability: Repeated, Variable) mixed ANOVA will assess differences in initial test performance. Due to unequal sample sizes in the practice type conditions, the mixed ANOVAs used Type III sums of squares, and a Greenhouse-Geisser correction was applied when sphericity was violated.

Results

Coding

Given the overlap in materials across the two studies, the same coding procedure as Study 1 was used for the practice and final test responses. The interrater reliability between the two coders was moderate for the final test ($\kappa = .77$), and the discrepancies were resolved by discussion among the two coders. After, three coders graded the practice trial responses, and the interrater reliability between two coders for the three practice trials was substantial ($\kappa = .80$). Discrepancies were again discussed, and resolved amongst coders with guidance from the author.

Counterbalancing and Sample Analyses

As done in Study 1, analyzing potential differences across the counterbalancing conditions was necessary to ensure consistency before the planned analyses. The counterbalancing analyses combined the two counterbalancing methods (order version and practice structure version) with the practice structure for each concept (massed or spaced). The order version combined the same practice question selected for the repeated condition, and the

order of the practice questions for the variable condition to collapse across practice type. A 2 (Practice Structure: Massed, Spaced) x 3 (Order Version: 1, 2, 3) x 2 (Practice Structure Version: 1, 2) mixed ANOVA examined if there were significant differences across the counterbalancing versions. Results revealed no significant differences across the counterbalancing versions nor significant interactions with practice activity (all Fs < 2.65, ps > .10). With no significant differences across the counterbalancing versions, counterbalancing was not included as a variable in further analyses.

Given that the final sample of participants were recruited on two different platforms, an additional analysis tested whether there were significant differences between the two samples. A 2 (Practice Type: Repeated, Variable) x 2 (Practice Structure: Massed, Spaced) x 2 (Sample: CloudResearch, WUSTL) mixed ANOVA was conducted for this purpose. There was no significant main effect for sample, $F(1, 57) = 2.46$, $p = .12$, $\eta_p^2 = 0.04$, as well as no significant interactions between sample, variability, and spacing (all Fs < 0.02, ps > .68). Since the two samples were not significantly different in their final test performance, the two samples were combined in all further analyses.

Initial Test Performance

Table 6 displays the practice test performance across the three practice trials by practice structure and practice type condition. In the massed-repeated condition, practice performance substantially increased from the first trial ($M = .36$) to the second trial ($M = .80$), and maintained high performance in the third trial ($M = .85$). In contrast, practice performance steadily increased across the three practice trials ($M = .33$ vs. .43 vs. .64). in the spaced-repeated condition. The massed-variable condition followed a similar pattern to the massed-repeated condition, albeit to a lesser degree; practice performance increased from the first trial ($M = .33$) to the second trial (M

= .42), and maintained a similar performance in the third trial (M = .47). Surprisingly, there was also a slight increase in practice performance for the spaced-variable condition (M = .30 vs. .35 vs. .41). Overall, all practice conditions increased across the three practice trials.

Table 6. Mean Accuracy on the Practice Trials and Final Test.

	Practice Trials			
Condition	First	Second	Third	Final Test
Massed – Repeated	.36	.80	.85	.29
Massed – Variable	.33	.42	.47	.29
Spaced – Repeated	.33	.43	.64	.47
Spaced – Variable	.30	.35	.41	.35

A series of ANOVAs explored potential significant differences in the three practice trials. A 2 (Practice Type) x 2 (Practice Structure) mixed ANOVA demonstrated that performance on the first practice question did not differ significantly for neither of the main effects nor the interaction between practice type and practice structure (all Fs < 0.78, ps > .35). That is, all participants began at an equal performance before the practice type and practice structure conditions were implemented. A 3 (Question Order) x 2 (Practice Type) x 2 (Practice Structure) mixed ANOVA with a Greenhouse-Geisser correction assessed differences in practice trial performance. There was a main effect of question order, $F(1.93, 113.61) = 71.88$, $p < .001$, $\eta_p^2 = 0.55$, as participants' performance increased across the questions. Post-hoc analyses revealed higher performance on the third question compared to the first question, $t(59) = 11.08$, $p < .001$, d =1.44, and the second question, $t(59) = 4.64$, $p < .001$, $d = 0.60$. In addition, performance on the second question was higher than the first question, $t(59) = 7.46$, $p < .001$, d =0.97. Overall, there was a main effect of spacing, $F(1, 59) = 29.46$, $p < .001$, $\eta_p^2 = 0.33$, with spaced practice

resulting in greater performance than massed practice. There was also a main effect of variability, $F(1, 59) = 16.27$, $p < .001$, $\eta_p^2 = 0.22$, as participants in the same condition performed better than the participants in the variable condition. All two-way and three-way interactions were significant, given the differential improvement across the four conditions (Fs > 5.20, ps < .01).

Final Test Performance

Table 6 lists the mean accuracies of the four conditions on the final test. Examining the Participants performed slightly better on the novel final test questions when they answered the same questions ($M = .38$, $SE = .04$) compared to different questions ($M = .32$, $SE = .03$). Collapsing across practice type conditions, participants performed better when concepts were spaced ($M = .40$, $SE = .03$), as opposed to massed ($M = .29$, $SE = .03$). To test whether these main effects of practice type and practice structure were significant, in addition to the interaction between the two variables, a 2 (Practice Type) x 2 (Practice Structure) mixed ANOVA was conducted. There was a significant main effect of practice structure, $F(1, 59) = 12.87$, $p < .001$, $\eta_p^2 = 0.18$, as spaced practice produced superior performance as opposed to massed practice. Although there seemed to be slight mean differences in practice type, the main effect for practice type was not significant, $F(1, 59) = 1.13$, $p = .30$, $\eta_p^2 = 0.02$. Lastly, there was no interaction between practice structure and practice type, $F(1, 59) = 3.30$, $p = .07$, $\eta_p^2 = 0.05$.

Discussion

Study 2 sought to confirm the benefit of variability found in Study 1, and experimentally manipulate the temporal structure of the initial learning sessions to assess transfer of knowledge. The findings from Study 2 show an advantage of spaced learning over massed learning, replicating the general spacing effect. By creating very short spacing gaps (i.e., seconds) relative

to the retention interval (i.e., nine days), the massed conditions resulted in the lowest final test performance. Despite the high performance reached on the initial learning questions for the massed, repeated condition (i.e., .85 proportion correct at the third practice trial), performance greatly decreased from initial learning to the final test with a retention interval of nine days. Even when initial learning questions were massed and variable, learners' performance slightly increased across the three practice trials, demonstrating that some connection across concepts. However, both of the massed conditions produced equivalent final test performance.

Out of the four conditions, learners that answered repeated questions spaced over two days apart performed the best on the final test. Perhaps repeating questions allowed for sufficient retention of the to-be-learned information, which led to a greater magnitude of transfer on new application questions. Learners in the spaced, repeated condition reached above 50% correct, representing adequate knowledge retrained during initial learning. In contrast, learners that received spaced, variable practice only slightly increased their performance across initial learning, suggesting that learners could not necessarily connect the contextual features of each concept across initial learning trials. Thus, requiring learners to transfer their knowledge across the three questions in the spaced, variable condition led to the smallest decrease in final test performance out of all four conditions.

General Discussion

The primary goal of the two present studies were to understand how combining retrieval practice, variability, and spacing during initial learning could influence transfer of knowledge on the final test. In both studies, learners watched mini-lecture style videos on 12 geological science concepts, and engaged in certain practice activities (i.e., retrieval practice versus study points), practice types (i.e., repeated practice versus variable practice), and practice structures (i.e.,

massed practice versus spaced practice), all occurring during the initial learning sessions. After a few days, learners answered new application questions, assessing transfer of knowledge. Although Study 1 found a benefit of variability, this finding was primarily driven by the re-study condition. When the re-study condition was dropped in Study 2, the advantage of variability in the initial practice disappeared, presumably due to high performance in the spaced, repeated condition. When purposefully manipulating spacing in Study 2, a spacing effect was found, replicating an abundance of prior work (see Maddox, 2016). Neither Study 1 nor Study 2 found the predicted advantage of spaced, variable retrieval practice, which could be due to the relatively low initial learning performance. Taken as a whole, both Study 1 and Study 2 results support the hypothesis that learners were primarily unable to connect contextual, structural, and descriptive features of the different questions and answers across the three initial learning trials.

Comparing to Massed Practice in Butler et al. (2017)

To better explain the potential patterns across the two studies in this thesis, Figure 5 presents the combined findings of practice type and practice structure across Butler et al. (2017; Experiment 4), Study 1, and Study 2. Butler et al. (2017) was included as a comparison because the authors used identical materials, and a similar design and procedure to the two present studies. Focusing on the massed practice, final test performance decreases substantially when the retention interval is increased by eight days (see Figure 4, left panel). When initial learning is massed (i.e., completed in succession), information is extremely susceptible to forgetting with longer retention intervals (Cepeda et al., 2005). Thus, massing practice is only optimal when information has to be transferred shortly after the initial learning phase (i.e., a shorter retention interval). Moreover, the benefit of variability from Butler and colleagues (2017) disappeared with the longer retention interval, demonstrating how variability in initial learning may depend on retention interval specified by the researchers. Turning to the spaced practice, repeated

retrieval practice in Study 2 outperformed repeated retrieval practice in Study 1 when the retention interval was increased by two days (see Figure 5, right panel). Just like in the massed practice, these results show the importance of choosing an optimal retention interval, as both Study 1 and Study 2 used a spacing gap of two days. Interestingly, when comparing the spaced, variable conditions for Study 1 and 2, final test performance was practically equivalent (i.e., .36 and .35, respectively), despite the few differences between Study 1 and Study 2.

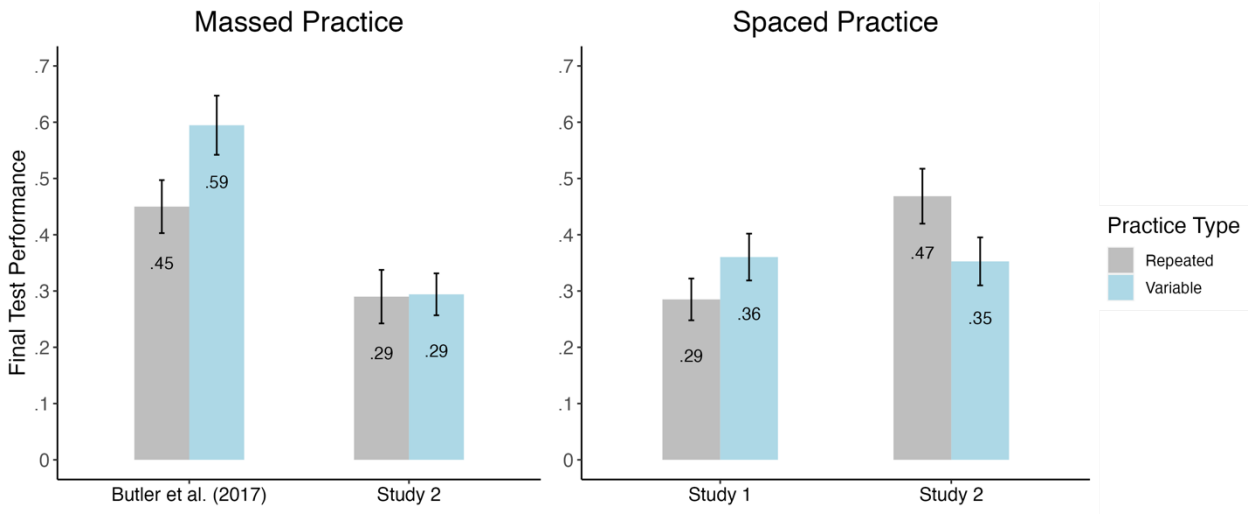

Figure 5. Final Test Performance for Massed and Spaced Practice. Massed practice results were taken from Butler et al. (2017, Experiment 4) that used the same materials and design as the current study. Error bars represent +/- one standard error.

Across Butler et al. (2017; Experiment 4), Study 1, and Study 2, similar patterns emerged in initial learning performance. In both Butler and colleagues (2017) and Study 2, learners in the massed, repeated condition reached the greatest initial learning performance due to the substantial increase in performance from the first practice trial to the second practice trial. In both instances, learners clearly recognized that the questions were repeated in succession, and did not have to transfer their knowledge across questions. When learners were required to apply their knowledge to different questions completed in succession (i.e., the massed, variable

condition), initial learning performance did not increase to the same degree. There were a few differences in initial learning performance between Butler et al. (2017), Study 1, and Study 2. Each of the three studies differed in their use of in-person and remote data collection: Butler et al. (2017) had all in-person sessions, Study 1 combined both in-person and remote sessions, and Study 2 was conducted completely remote. Across all four conditions, learners in Study 2 had a lower performance on the first practice trial of initial learning (i.e., < .37 proportion correct), presumably due to participants watching the videos and answering questions remotely. Secondly, the within-subjects manipulation of spacing for Study 2 resulted in only six questions presented in the second and third practice trial for the spaced conditions, instead of answering all 12 questions as done in Study 1. Having participants only answer six questions led to an increase in performance across the three initial learning trials, which could have allowed for learners to better connect the contextual, structural, and descriptive features during initial learning. Despite these few differences identified, achieving similar patterns across studies demonstrates replicability of findings, and the feasibility of using educationally relevant materials in completely remote research.

The Lack of a Testing Effect When Introducing Variability and Spacing

Despite an abundance of conceptual evidence for a testing effect (see Pan & Rickard, 2018), and multiple theoretical accounts predicting a testing effect (e.g., encoding variability theory, semantic elaboration hypothesis, episodic context hypothesis), a lack of a testing effect observed in Study 1 was a puzzling finding. Connecting back to the three-factor framework for transfer proposed by Pan & Rickard (2018), Study 1 implemented response congruence and elaborated retrieval practice beforehand, but initial learning performance could only be examined after the completion of data collection. Initial test performance could be a proxy for how complete the memory representation is for the to-be-learned information. If learners have high

initial learning performance, then leaners can rely on an increased number of contextual features to create multiple retrieval routes for the memory. Given the low initial learning performance reached in the spaced, variable condition (i.e., < 50% accuracy), the third factor of high initial learning performance was not achieved, which contributed to the lack of benefit in the retrieval practice condition compared to reading study points on final test performance.

Given the complexities of transfer occurring in the present two studies, scaffolding learning might be necessary to facilitate transfer of knowledge when combining learning strategies. In order to apply information, learners need to recognize that their prior knowledge is applicable, and recall the applicable prior knowledge (see Barnett & Ceci, 2002). Although it is possible that learners were unable to recognize that certain knowledge from the geological science videos was applicable to the questions, it seems more probable that learners were unable recall the information necessary before applying the information to the context of the question. Scaffolding could be executed in the present paradigm by having learners answer questions assessing retention before having them answer questions assessing transfer to ensure adequate retention of the to-be-learned information. In fact, Agarwal (2019) found that mixing retention and transfer questions during the practice, relative to only answering transfer questions, resulted in greater final test performance requiring transfer of knowledge. When using application questions, prior studies have typically given a final test assessing both retention and transfer of knowledge, which might help scaffold learning when the retention questions are answered before the transfer questions (e.g., Hinze, Wiley & Pellegrino, 2013; Johnson & Mayer, 2009; Woolridge, Bugg, McDaniel & Liu, 2014). Retention has been argued to be considered a building block for higher-order processes (e.g., Willingham, 2021), but this is not always the

case (see Agarwal, 2019). Future research should consider implementing scaffolding when combining multiple learning strategies to facilitate transfer of knowledge.

The Benefit of Variability Disappears with Spacing

The mixed findings of variability across Study 1 and Study 2 were surprising, given the results from Butler et al. (2017). The benefits of variability have primarily been found when there is variability in the descriptive features of the cues provided to learners (e.g., Butler, 2010; Glass, 2009; Goode et al., 2008; Smith & Handy, 2014). Less is known how variability can be beneficial for transfer of knowledge if both the cues and the response are manipulated, and the repetitions of the cues and responses are spaced. Even though Foss et al. (2023) did not consistently manipulate variability in the answers to the questions, the authors' approach to using long spacing gaps (i.e., 12-35 days) in a college course offer some insight into why the present two studies might have not found the benefit of variability. The long spacing gaps present in Foss et al. (2023) represent the days between cumulative exams, and students most likely studied the to-be-learned information multiple times in preparation for the exam. In the current two studies, it is reasonable to assume that learners did not encounter the geological science concepts outside of the experiment, limiting the learners' exposure to the material. When the practice is spaced, combining learning strategies may need more practice repetitions to help learners encode the features of the to-be-learned information, and thus, researchers might find a benefit of variability.

The Presence of a Spacing Effect When Transferring Knowledge

To my knowledge, the finding of a spacing effect in a testing effect paradigm assessing transfer of knowledge is a novel contribution to the broader literature. One of the few studies to specifically manipulate spacing gaps when assessing transfer of knowledge, Kapler and colleagues (2015), found that when learners were tested on meteorology concepts with longer

spacing gaps of eight days, final test performance on transfer questions was increased compared to concepts with shorter spacing gaps of one day. This study was conducted in a college course with a retention interval of five weeks, replicating the advantage of spaced practice for spacing gaps shorter than the retention interval (Cepeda et al., 2006). Although Kapler et al. (2015) used the same retention interval for both the 1-day and 8-day spacing gaps, the present two studies manipulated both the spacing gap (e.g., seconds or days) and retention interval (e.g., two days, four days, or nine days). Therefore, it is reasonable to assume that the benefit of spacing extends to transfer of knowledge when the spacing gap and retention interval are greater than one day. Future research can build upon this finding by examining other combinations of spacing gaps and retention intervals when measuring transfer or implement expanding spacing schedules, rather than uniform spacing schedules.

Implications for Encoding Variability Theory

According to encoding variability theory, combining multiple learning strategies should facilitate transfer due to a greater number of contextual, structural, and descriptive features present during encoding, which create multiple pathways for future retrieval, and thus, increasing the likelihood of future retrieval (Estes, 1955; Melton, 1967; Bjork, 1975; Glenberg, 1979). By comparing initial learning performance across the three repetitions, implementing spacing and/or variability in practice only gradually improved performance across the three practice trials, demonstrating that learners were not able use structural features to connect across retrieval attempts during initial learning. Furthermore, there was not the hypothesized additive benefit of combining retrieval practice, variability, and spacing in final test performance relative to the other conditions. Taken as a whole, these findings suggest that providing a greater variety of features do not necessarily result in higher final test performance, contrary to encoding variability theory.

Even though encoding variability theory has been used to explain the advantage of variability and retrieval practice for a final test requiring transfer of knowledge (Butler et al., 2017), encoding variability theory may not hold in situations that combine multiple learning strategies with longer spacing gaps and retention intervals. In terms of the temporal structure of practice repetitions, encoding variability theory does not necessarily make predictions about an optimal spacing gap and retention interval for the structure of repetitions (Maddox, 2016). Although encoding variability theory is a dominant theory for all three learning strategies, encoding variability theory does not explicitly predict how learning strategies should be implemented to facilitate transfer of knowledge. Currently, there is no comprehensive theory that explains how implementing multiple learning strategies in combination might be beneficial for performance requiring transfer, and future research should aim to understand in what situations results in super-additivity in performance when combining learning strategies.

Limitations & Future Directions

Given the challenges encountered during data collection for Study 2, a few limitations should be addressed in further detail. Based on data collection for Study 1, it was deemed feasible to conduct Study 2 completely remote, but only 14.5% of participants that signed up to participate in the study completed all four sessions, as previously mentioned. It is unclear exactly why the majority of the participants did not finish the study, but the stricter research parameters offered through CloudResearch allowed for Study 2 to be continue completely remote. The switch to CloudResearch resulted in low sample sizes for the repeated and variable conditions recruiting through the WashU participant pool (n = 8 and n = 4, respectively). Although there were no significant differences between the WashU sample and CloudResearch sample, a larger sample size is needed to assess potential differences between these two populations.

The final sample size for Study 2 did not reach the desired sample size in the a-priori power analysis, resulting in lower power than originally intended. In anticipation of a handful of participants to be excluded after data collection, more participants completed the study (n = 68) than the sample size proposed by the power analysis (n = 64). However, an in-depth evaluation of the data revealed some unexpected patterns that required a stricter exclusion criterion than specified in Study 1. Upon grading the participant responses, some participants did not follow the study instructions, resulting in a long time spent on questions, and many incorrect answers. Alternatively, conducting a post-hoc power analysis would not be recommended, since the p-value has been determined after statistical analyses which biases the power calculation (see Hoenig & Heisey, 2001; Lenth, 2000). The final sample size is comparable to similar lab-based work (Butler et al., 2017), and thus, it is reasonable to assume adequate power was achieved. However, future research should consider recruiting larger sample sizes to conduct individual-level analyses, as only certain learners might benefit from combining multiple learning strategies and activities.

Conclusion

Understanding how learning strategies are implemented in combination is critical to learners' engaging in transfer in educational contexts. The situation posed at the beginning of this paper highlights the many learning activities a student can choose to engage in when studying for a test that requires transfer of knowledge. Implementing variability in question and answer mirrors a more educationally-relevant scenario, as learners do not typically encounter the exact same answer when the question contains different descriptive features. However, if learners are unable to connect the structural features of the learning instances spaced days apart, then transfer is limited. Thus, educators should consider repeating application questions spaced a

few days apart to allow for the greatest performance on an exam requiring transfer of knowledge. Another possibility for facilitating greater transfer of knowledge could be introducing scaffolding when combining larger spacing gaps and variability in both the question and answers during initial learning. Despite the limitations and open questions, the present two studies illustrate that the performance on a test requiring transfer can depend how learning strategies are implemented in conjunction, as combining learning strategies is not always beneficial.

References

Adesope, O. O., Trevisan, D. A., & Sundararajan, N. (2017). Rethinking the use of tests: A meta-analysis of practice testing. *Review of Educational Research, 87*(3), 659-701. https://doi.org/10.3102/0034654316689306

Agarwal, P. K. (2019). Retrieval practice & Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology, 111*(2), 189–209. https://doi.org/10.1037/edu0000282

Agarwal, P.K., Nunes, L.D. & Blunt, J.R. (2021). Retrieval Practice Consistently Benefits Student Learning: a Systematic Review of Applied Research in Schools and Classrooms. *Educational Psychology Review, 33*(4), 1409–1453. https://doi.org/10.1007/s10648-021-09595-9

Appleton-Knapp, S. L., Bjork, R. A., & Wickens, T. D. (2005). Examining the spacing effect in advertising: Encoding variability, retrieval processes, and their interaction. *Journal of Consumer Research, 32*(2), 266-276. https://doi.org/10.1086/432236

Barcroft, J., & Sommers, M. S. (2005). Effects of acoustic variability on second language vocabulary learning. *Studies in Second Language Acquisition*, *27*(3), 387-414. https://doi.org/10.1017/S0272263105050175

Barnett, S. M., & Ceci, S. J. (2002). When and where do we apply what we learn?: A taxonomy for far transfer. *Psychological bulletin, 128*(4), 612. https://doi.org/10.1037/0033-2909.128.4.612

Begg, I., & Green, C. (1988). Repetition and trace interaction: Superadditivity. *Memory & cognition*, *16*(3), 232-242. https://doi.org/10.3758/BF03197756

Bjork, R. A. (1975). Retrieval as a memory modifier: An interpretation of negative recency and related phenomena. In Robert L. Solso (ed.), *Information Processing and Cognition: The Loyola Symposium.* Lawrence Erlbaum. pp. 123-144.

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

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 36*(5), 1118. https://doi.org/10.1037/a0019902

Butler, A. C., Black-Maier, A. C., Raley, N. D., & Marsh, E. J. (2017). Retrieving and applying knowledge to different examples promotes transfer of learning. *Journal of Experimental Psychology: Applied, 23*(4), 433. https://doi.org/10.1037/xap0000142

Butler, A. C., Godbole, N., & Marsh, E. J. (2013). Explanation feedback is better than correct answer feedback for promoting transfer of learning. *Journal of Educational Psychology, 105*(2), 290. https://doi.org/10.1037/a0031026

Butler, A. C., & Roediger III, H. L. (2007). Testing improves long-term retention in a simulated classroom setting. *European Journal of Cognitive Psychology*, *19*(4-5), 514-527. https://doi.org/10.1080/09541440701326097

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: the benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 35*(6), 1563. https://doi.org/10.1037/a0017021

Carpenter, S. K. (2012). Testing enhances the transfer of learning. *Current directions in psychological science, 21*(5), 279-283. https://doi.org/10.1177/0963721412452728

Carpenter, S. K., Cepeda, N. J., Rohrer, D., Kang, S. H., & Pashler, H. (2012). Using spacing to enhance diverse forms of learning: Review of recent research and implications for instruction. *Educational Psychology Review, 24*(3), 369-378. https://doi.org/10.1007/s10648-012-9205-z

Carpenter, S. K., & DeLosh, E. L. (2005). Application of the testing and spacing effects to name learning. *Applied Cognitive Psychology: The Official Journal of the Society for Applied Research in Memory and Cognition, 19*(5), 619-636. https://doi.org/10.1002/acp.1101

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. https://doi.org/10.3758/BF03193405

Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of US history facts. *Applied Cognitive Psychology, 23*(6), 760-771. https://doi.org/10.1002/acp.1507

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*(3), 354. https://doi.org/10.1037/0033-2909.132.3.354

Cepeda, N. J., Vul, E., Rohrer, D., Wixted, J. T., & Pashler, H. (2008). Spacing effects in learning: A temporal ridgeline of optimal retention. *Psychological science, 19*(11), 1095-1102. https://doi.org/10.1111/j.1467-9280.2008.02209.x

Douvis, S. J. (2005). Variable practice in learning the forehand drive in tennis. *Perceptual and motor skills*, *101*(2), 531-545. https://doi.org/10.2466/pms.101.2.531-545

Dunlosky, J., Rawson, K. A., Marsh, E. J., Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions

from cognitive and educational psychology. *Psychological Science in the public interest, 14*(1), 4-58. https://doi.org/10.1177/1529100612453266

Estes, W. K. (1955). Statistical theory of spontaneous recovery and regression. *Psychological review, 62*(3), 145. https://doi.org/10.1037/h0048509

Faul, F., Erdfelder, E., Lang, A.-G., & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods, 39*, 175-191. https://doi.org/10.3758/BF03193146

Foss, D. J., Pirozzolo, J. W., & Kulesz, P. A. (2023). Retrieving and transferring knowledge: Effects of test item variation on transfer in an authentic learning environment. *Learning and Instruction*, *88*, 101807. https://doi.org/10.1016/j.learninstruc.2023.101807

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

Glenberg, A. M. (1979). Component-levels theory of the effects of spacing of repetitions on recall and recognition. *Memory & Cognition, 7*(2), 95-112. https://doi.org/10.3758/BF03197590

Gluckman, M., Vlach, H. A., & Sandhofer, C. M. (2014). Spacing simultaneously promotes multiple forms of learning in children's science curriculum. *Applied Cognitive Psychology, 28*(2), 266-273. https://doi.org/10.1002/acp.2997

Goode, M. K., Geraci, L., & Roediger, H. L., III. (2008). Superiority of variable to repeated practice in transfer on anagram solution. *Psychonomic Bulletin & Review, 15*, 662– 666. http://dx.doi.org/10.3758/PBR.15.3.662

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

Hinze, S. R., Wiley, J., & Pellegrino, J. W. (2013). The importance of constructive comprehension processes in learning from tests. *Journal of Memory and Language*, *69*(2), 151-164. https://doi.org/10.1016/j.jml.2013.03.002

Hoenig, J. M., & Heisey, D. M. (2001). The abuse of power: the pervasive fallacy of power calculations for data analysis. *The American Statistician*, *55*(1), 19-24. https://doi.org/10.1198/000313001300339897

Johnson, C. I., & Mayer, R. E. (2009). A testing effect with multimedia learning. *Journal of Educational Psychology*, *101*(3), 621. https://doi.org/10.1037/a0015183

Kang, S. H. (2016). Spaced repetition promotes efficient and effective learning: Policy implications for instruction. *Policy Insights from the Behavioral and Brain Sciences, 3*(1), 12-19. https://doi.org/10.1177/2372732215624708

Kapler, I. V., Weston, T., & Wiseheart, M. (2015). Spacing in a simulated undergraduate classroom: Long-term benefits for factual and higher-level learning. *Learning and Instruction, 36*, 38-45. https://doi.org/10.1016/j.learninstruc.2014.11.001

Karpicke, J. D., & Aue, W. R. (2015). *The testing effect is alive and well with complex materials. Educational Psychology Review*, *27*(2), 317–326. https://doi.org/10.1007/s10648-015-9309-3

Karpicke, J. D., & Bauernschmidt, A. (2011). Spaced retrieval: absolute spacing enhances learning regardless of relative spacing. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 37*(5), 1250. https://doi.org/10.1037/a0023436

Karpicke, J. D., Lehman, M., & Aue, W. R. (2014). Retrieval-based learning: An episodic context account. In *Psychology of learning and motivation* (Vol. 61, pp. 237-284). Academic Press. https://doi.org/10.1016/B978-0-12-800283-4.00007-1

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

Kornell, N., & Bjork, R. A. (2008). Learning concepts and categories: Is spacing the "enemy of induction"?. *Psychological science*, *19*(6), 585-592. https://doi.org/10.1111/j.1467-9280.2008.02127.x

Lenth, R. V. (2000). Two sample-size practices that I don't recommend. In *Proceedings of the section on physical and engineering sciences* (pp. 8-11).

Litman, L., Robinson, J., & Abberbock, T. (2017). TurkPrime. com: A versatile crowdsourcing data acquisition platform for the behavioral sciences. *Behavior research methods*, *49*(2), 433-442. https://doi.org/10.3758/s13428-016-0727-z

Maddox, G. B. (2016). Understanding the underlying mechanism of the spacing effect in verbal learning: A case for encoding variability and study-phase retrieval. *Journal of Cognitive Psychology, 28*(6), 684-706. https://doi.org/10.1080/20445911.2016.1181637

McDaniel, M. A., & Masson, M. E. (1985). Altering memory representations through retrieval. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, *11*(2), 371. https://doi.org/10.1037/0278-7393.11.2.371

McDaniel, M. A., Roediger, H. L., & McDermott, K. B. (2007). Generalizing test-enhanced learning from the laboratory to the classroom. *Psychonomic bulletin & review*, *14*(2), 200-206. https://doi.org/10.3758/BF03194052

Melton, A. W. (1967). Repetition and retrieval from memory. *Science (New York, NY), 158*(3800), 532.

Paas, F. G., & Van Merriënboer, J. J. (1994). Variability of worked examples and transfer of geometrical problem-solving skills: A cognitive-load approach. *Journal of educational psychology*, *86*(1), 122. https://doi.org/10.1037/0022-0663.86.1.122

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological bulletin, 144*(7), 710. https://doi.org/10.1037/bul0000151

R Core Team (2023). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna, Austria. https://www.R-project.org/.

Raviv, L., Lupyan, G., & Green, S. C. (2022). How variability shapes learning and generalization. *Trends in cognitive sciences*, *26*(6), 462-483. https://doi.org/10.1016/j.tics.2022.03.007

Roediger III, H. L., & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in cognitive sciences, 15*(1), 20-27. https://doi.org/10.1016/j.tics.2010.09.003

Roediger III, H. L., & Karpicke, J. D. (2006). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological science, 17*(3), 249-255. https://doi.org/10.1111/j.1467-9280.2006.01693.x

Rohrer, D., & Taylor, K. (2006). The effects of overlearning and distributed practice on the retention of mathematics knowledge. Applied Cognitive Psychology: The Official *Journal of the Society for Applied Research in Memory and Cognition, 20*(9), 1209-1224. https://doi.org/10.1002/acp.1266

Rohrer, D., & Taylor, K. (2007). The shuffling of mathematics problems improves learning. *Instructional Science*, *35*, 481-498. https://doi.org/10.1007/s11251-007-9015-8

Rowland, C. A. (2014). The effect of testing versus restudy on retention: a meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432. https://doi.org/10.1037/a0037559

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

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

Sobel, H. S., Cepeda, N. J., & Kapler, I. V. (2011). Spacing effects in real-world classroom vocabulary learning. *Applied Cognitive Psychology*, *25*(5), 763-767. https://doi.org/10.1002/acp.1747

Tenenbaum, J. B., & Griffiths, T. L. (2001). Generalization, similarity, and Bayesian inference. *Behavioral and brain sciences*, *24*(4), 629-640. https://doi.org/10.1017/S0140525X01000061

van Gog, T., & Sweller, J. (2015). Not new, but nearly forgotten: The testing effect decreases or even disappears as the complexity of learning materials increases. *Educational Psychology Review, 27*(2), 247–264. https://doi.org/10.1007/s10648-015-9310-x

Vukatana, E., Graham, S. A., Curtin, S., & Zepeda, M. S. (2015). One is not enough: Multiple exemplars facilitate infants' generalizations of novel properties. *Infancy, 20*(5), 548-575. https://doi.org/10.1111/infa.12092

Willingham, D. T. (2021). *Why don't students like school?: A cognitive scientist answers questions about how the mind works and what it means for the classroom*. John Wiley & Sons.

Wooldridge, C. L., Bugg, J. M., McDaniel, M. A., & Liu, Y. (2014). The testing effect with authentic educational materials: A cautionary note. *Journal of Applied Research in Memory and Cognition*, *3*(3), 214-221. https://doi.org/10.1016/j.jarmac.2014.07.001

Yang, C., Luo, L., Vadillo, M. A., Yu, R., & Shanks, D. R. (2021). Testing (quizzing) boosts classroom learning: A systematic and meta-analytic review. *Psychological Bulletin, 147*(4), 399–435. https://doi.org/10.1037/bul0000309

Appendix A

Example Questions, Answers, and Study Points for a Concept about the Earth's Structure

Item	Example
Question 1	In an article about the Earth's structure, a major U.S. newspaper reports the following densities for its three major layers: the core (2.2 g/cm³), the mantle (4.4 g/cm³), and the crust (11.5 g/cm³). What is wrong with this characterization of the Earth's structure?
Answer 1	The article reports that the core is the least dense and the crust is the densest. However, the core should be the densest, the mantle should be slightly less dense, and the crust should be the least dense. Earth's structure results from the upper layers "floating" on the lower layers because they are less dense.
Question 2	In 1692, Edmund Halley theorized that the Earth had a hollow center. However, his theory was considered logically impossible once the average density of the Earth (5.5 g/cm³) was calculated and the density of rock near the surface (2.2 g/cm³) was measured. Why did these findings make the "hollow earth" theory logically impossible?
Answer 2	If the rock near the surface (2.2 g/cm³) is less dense than the average density of the Earth (5.5 g/cm³), then the center must be denser than the average density of the Earth and thus cannot be hollow. Indeed, the Earth's core is the densest layer, and the upper layers "float" on the lower layers because of a difference in density.
Question 3	One theory suggests that planets form after the collapse of a nebula (an interstellar cloud of dust). The dust particles accumulate mass through gravitational attraction to form ever-larger bodies, and these concentrations differentiate by density to form the interior of a planet. How does the Earth's structure provide support for this theory?
Answer 3	The Earth's structure provides support for this theory because it is consistent with the planetary structure that would be produced through such a process of formation. Earth is composed of three major layers that vary in density with less dense upper layers "floating" on more dense lower layers.
Question 4	A new planet is discovered that is composed of following elements: silica dust (2.2 g/cm³), carbon dioxide (.0018 g/cm³), hermatite (4.5 g/cm³), iron and nickel mixture (7.6 g/cm³), water (.98 g/cm³), and amphibolite (2.9 g/cm³). What does your knowledge of the Earth's structure tell us about the structure of this new planet?
Answer 4	Based on Earth's structure, the most dense materials (iron and nickel) are probably at the core and then the other elements are layered on top in

	decreasing densities: hermatite, amphibolite, silica dust, water, and carbon dioxide.
Study Point 1	In an article about the Earth's structure, a major U.S. newspaper reports the following densities for its three major layers: the core (2.2 g/cm3), the mantle (4.4 g/cm3), and the crust (11.5 g/cm3). This characterization of the Earth's structure is wrong because it states that the core is the least dense and the crust is the densest. However, the core should be the densest, the mantle should be slightly less dense, and the crust should be the least dense. Earth's structure results from the upper layers "floating" on the lower layers because they are less dense.
Study Point 2	In 1692, Edmund Halley theorized that the Earth had a hollow center. However, his theory was considered logically impossible once the average density of the Earth (5.5 g/cm3) was calculated and the density of rock near the surface (2.2 g/cm3) was measured. These findings made the "hollow earth" theory logically impossible because if the rock near the surface (2.2 g/cm3) is less dense than the average density of the Earth (5.5 g/cm3), then the center must be denser than the average density of the Earth and thus cannot be hollow. Indeed, the Earth's core is the densest layer, and the upper layers "float" on the lower layers because of a difference in density.
Study Point 3	One theory suggests that planets form after the collapse of a nebula (an interstellar cloud of dust). The dust particles accumulate mass through gravitational attraction to form ever-larger bodies, and these concentrations differentiate by density to form the interior of a planet. The Earth's structure provides support for this theory because it is consistent with the planetary structure that would be produced through such a process of formation. Earth is composed of three major layers that vary in density with less dense upper layers "floating" on more dense lower layers.
Study Point 4	A new planet is discovered that is composed of following elements: silica dust (2.2 g/cm3), carbon dioxide (.0018 g/cm3), hematite (4.5 g/cm3), iron and nickel mixture (7.6 g/cm3), water (.98 g/cm3), and amphibolite (2.9 g/cm3). Our knowledge of the Earth's structure tells us about the structure of this new planet. Based on Earth's structure, the densest materials (iron and nickel) are probably at the core and then the other elements are layered on top in decreasing densities: hematite, amphibolite, silica dust, water, and carbon dioxide.

Appendix B

Table B1. Counterbalancing for Study 1.

Counterbalancing Condition	Time 1	Time 2	Time 3
Same – Version 1	R_1 or S_1	R_1 or S_1	R_1 or S_1
Same – Version 2	R_2 or S_2	R_2 or S_2	R_2 or S_2
Same – Version 3	R_3 or S_3	R_3 or S_3	R_3 or S_3
Variable – Version 1	R_1 or S_1	R_2 or S_2	R_3 or S_3
Variable – Version 2	R_2 or S_2	R_3 or S_3	R_1 or S_1
Variable – Version 3	R_3 or S_3	R_1 or S_1	R_2 or S_2

Note. R = Retrieval practice questions. S = Study points. Subscripts denote question or study point number for each concept.

Table B2. Counterbalancing for Study 2.

Counterbalancing Condition	Time 1	Time 2	Time 3
Same – Version 1	Sp_1 or M_1	Sp_1 or M_1	Sp_1 or M_1
Same – Version 2	Sp_2 or M_2	Sp_2 or M_2	Sp_2 or M_2
Same – Version 3	Sp_3 or M_3	Sp_3 or M_3	Sp_3 or M_3
Variable – Version 1	Sp_1 or M_1	Sp_2 or M_2	Sp_3 or M_3
Variable – Version 2	Sp_2 or M_2	Sp_3 or M_3	Sp_1 or M_1
Variable – Version 3	Sp_3 or M_3	Sp_1 or M_1	Sp_2 or M_2

Note. Sp = Spaced retrieval practice questions. M = Massed retrieval practice questions. Subscripts denote question number for each concept.