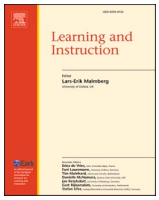
Learning and Instruction 100 (2025) 102219

Contents lists available at ScienceDirect

Learning and Instruction

journal homepage: www.elsevier.com/locate/learninstruc



Effects of retrieval practice on retention and application of complex educational concepts

Daniel Corral [a,*], Shana K. Carpenter [b]

[a] Department of Psychology, Syracuse University, USA
[b] School of Psychological Science, Oregon State University, Reed Lodge, 2950 SW Jefferson Way, Corvallis, OR, 97331, USA

ARTICLE INFO

Keywords:
Retrieval practice
Transfer of learning
Concept acquisition
Complex learning

ABSTRACT

Aims: Retrieval practice is effective for enhancing memory, but its effects on transfer are less clear. The current study compared the effects of retrieval versus non-retrieval-based strategies on retention and transfer of research methods concepts.

Sample and methods: In Experiment 1 (N = 309), participants completed one short-answer factual quiz and received correct-answer feedback (retrieval), one multiple-choice quiz with correct-answer feedback (recognition), restudied the original learning materials (restudy), or studied the short-answer quiz questions and answers (quiz study). Eight minutes later, participants received a final test over repeated questions (multiple-choice versions of the practice questions), and application questions (never-before-seen multiple-choice questions requiring application of the concepts). Experiments 2 (N = 158) and 3 (N = 255) involved the same retrieval, restudy, and quiz study conditions, but involved three rounds of retrieval practice and a one-week delayed final test.

Results: Retrieval enhanced performance compared to restudy, but not compared to quiz study or recognition, on repeated but not on application final test questions (Experiment 1). Retrieval produced better performance than restudy and quiz study on repeated final test questions (Experiment 2) and application final test questions (Experiment 3). Conditional analyses on application question performance given accurate repeated question performance revealed an advantage of retrieval, indicating that retrieval enhances the recognition component of transfer.

Conclusion: Retrieval practice benefits both retention and transfer of complex concepts. These benefits appear more likely to occur when a sufficient amount of retrieval practice is provided and learning is measured over a delay of several days.

1. Introduction

Research on the science of learning continuously reveals new insights about how to enhance the durability and efficiency of learning. Of particular interest are the strategies and techniques that promote long-term learning, as knowledge that is stable and reliable can benefit learners both during and after the formal education years. As such, an important goal of this research is to understand the conditions that produce meaningful and long-lasting learning.

Over a century of research has highlighted retrieval practice as one of the most effective learning strategies yet discovered (for a recent review, see Agarwal et al., 2021; Carpenter et al., 2022; McDermott, 2021). After initial study of some to-be-learned material (e.g., a textbook chapter, list of terms and definitions), practicing to retrieve that material from memory produces significant and often sizeable advantages on later memory, compared to a non-retrieval-based strategy, such as restudying. Such advantages have been observed in laboratory studies using a variety of materials from foreign language vocabulary (Kang et al., 2013), terms and definitions from different subject areas (Hui et al., 2021; Pan & Rickard, 2017), spelling (da Silva et al., 2023; Jones et al., 2016), texts (Agarwal, 2019; Endres et al., 2023), and in classroom studies using materials from the curriculum (Carpenter et al., 2018; Corral et al., 2020; McDaniel et al., 2011; Roediger et al., 2011). The benefits of retrieval are long-lasting, with studies documenting these

This article is part of a special issue entitled: Toward Lasting Learning published in Learning and Instruction.

* Corresponding author. Department of Psychology, Syracuse University, 346 Marley Education Building, Syracuse, NY, 13244, USA.
E-mail addresses: dcorral@syr.edu (D. Corral), shana.carpenter@oregonstate.edu (S.K. Carpenter).

https://doi.org/10.1016/j.learninstruc.2025.102219
Received 29 November 2023; Received in revised form 12 August 2025; Accepted 23 August 2025
Available online 20 September 2025

benefits over time intervals of several weeks and months (Carpenter et al., 2008, 2009; Kang et al., 2014; Lyle et al., 2020). Hundreds of demonstrations of retrieval practice—also referred to as the *testing effect*—have been documented in numerous meta-analyses, with effect sizes ranging from medium to large (e.g., Hedges' $g = .50$ in Rowland, 2014; Hedges' $g = .61$ in Adesope et al., 2017).

1.1. Retrieval practice and the transfer of learning

Despite the extensive literature on retrieval practice, a major limitation is that most studies are designed to measure direct memory retention of fairly simple materials, such as word lists and trivia facts. Fewer studies have looked at the effects of retrieval practice on more complex forms of learning, and in particular on the *transfer* of learning—the ability to use and apply learned information in a new context. Though transfer is less often explored in studies of retrieval practice, it is a critical component of learning, as many real-world situations depend on the flexible and adaptive use of knowledge in ways that are often unforeseen.

Studies looking at the effects of retrieval practice on transfer have revealed mixed findings. Some studies show that retrieval practice, relative to restudy, enhances a learner's ability to apply or generalize information, such as understanding how a given scientific concept applies in a new situation (Butler et al., 2017), applying a learned principle to a novel problem (Butler, 2010), or understanding how a scientific process would function when certain conditions are altered (Dobson et al., 2019). Other studies, however, have found that retrieval is not more effective than restudy for learning to apply rules and procedures, such as formulating deductive inferences from learned principles (Wissman et al., 2018; Tran et al., 2015) or applying a learned solution to a new problem that differs superficially from a previously-seen problem requiring that same solution (Corral et al., 2023; Peterson & Wissman, 2018).

1.1.1. Conceptualizing transfer

These inconsistent results might be understood by considering the different ways that transfer has been defined and measured (for a recent meta-analysis, see Pan & Rickard, 2018). A distinction is typically made between *near* and *far* transfer. Near forms of transfer involve applying knowledge from one context to another when the relevance of that knowledge is apparent in both contexts, such as when the two contexts involve the same topical domain. For example, a relatively near form of transfer would be using the same rule or concept from one physics problem in a given domain to another problem from the same domain (e.g., both problems are about objects revolving around a planet) requiring that rule or concept. In such cases, the shared features and topical domain across the two contexts serve as cues to the learner that the concept or solution that was applicable in a previous scenario is also applicable in the current scenario. In contrast, far forms of transfer involve applying a given concept from one context to another in situations where the two contexts differ in ways that the relevance of the knowledge needed in one context may not be apparent in the other context, such as when the contexts involve different topical domains. For instance, one could use a rule or concept that was originally learned within the context of a military problem for a medical problem that requires the same solution (e.g., see Gick & Holyoak, 1980, 1983; also see Barnett & Ceci, 2002). Scenarios that require far forms of transfer typically share very few, if any, surface features with previous, analogous scenarios that have been encountered.

1.1.2. Components of transfer

Considering transfer along this continuum (indeed, near and far transfer are more a matter of degree than clear categorical divisions) can help in identifying situations when transfers succeeds and when it fails. Memory is a critical component of knowledge transfer, but on its own it is not sufficient for transfer to occur (Butler et al., 2017). Transfer is posited to depend on (a) learners successfully remembering the learned information (memory component), (b) recognizing the relevance of that information in a new situation (recognition component), and (c) being able to apply that information successfully in the current situation (application component; Gick & Holyoak, 1987). Even though retrieval practice has consistent benefits on memory, it will not automatically facilitate transfer unless it also facilitates the ability to recognize the relevance of learned information and apply it in a new context (see Butler, 2010; Corral et al., 2023).

Consistent with the findings summarized above, therefore, it seems reasonable that the benefits of retrieval practice on transfer have more often been reported in cases that involve near forms of transfer (e.g., Butler, 2010; Butler et al., 2017; Dobson et al., 2017) than far forms of transfer (e.g., Corral et al., 2023; Peterson & Wissman, 2018; Tran et al., 2015; Wissman et al., 2018; see also Yeo & Fazio, 2019). Effects that involve far transfer have been observed, however, in situations where learners are given hints or assistance in recognizing the relevance of previous learning in a new situation (Butler, 2010; Gick & Holyoak, 1983).

In these cases, the recognition component of transfer is largely bypassed, which enables learners to focus on retrieving their corresponding knowledge and applying it to the current situation. Critically, recent work theorizes that recognition is the primary component that impedes the transfer of learning (Corral & Kurtz, 2025) and seems to lead to the *inert knowledge problem*, wherein learners have the necessary knowledge to solve a given problem, but do not apply it because they do not recognize that it is applicable (Whitehead, 1929). Support for this idea can be found in the problem-solving literature, in which learners can readily learn a given solution strategy and apply it successfully, but do not recognize when it should be applied (Mayer, 1998; Mayer et al., 1995).

Given the consistent benefits of retrieval practice on memory, situations where retrieval fails to benefit transfer may be more likely due to learners' failure to recognize the relevance of their prior learning in the transfer situation, such as whether and how a given learned concept applies in the new context. Though the near versus far transfer continuum is a useful framework for understanding how these recognition failures can occur (with such failures being more likely for far transfer), it is important to note that the key component of transfer is the recognition of the relevance of prior knowledge (which can be influenced by various factors) more so than the classification of a given transfer task as near or far on the transfer continuum.

1.2. Theoretical accounts of retrieval practice

Although theoretical mechanisms for the benefits of retrieval have been proposed, they have largely been developed to account for the benefits of retrieval on memory, and do not propose a clear mechanism for how retrieval might benefit the transfer of learning. One theory based on transfer-appropriate processing has proposed, for example, that the act of retrieval (more so than a non-retrieval-based task) is more likely to provide practice at the same aspects of the task that the final test requires (McDaniel & Fisher, 1991; Morris et al., 1977). Especially when the final test is the same as the initial test (as in many studies of retrieval practice), this account is based on the straightforward assumption that practicing to retrieve information aids later retrieval of that same information. According to the elaborative retrieval hypothesis (Carpenter, 2009, 2011; Carpenter & Yeung, 2017), retrieval involves recall of both the correct target information, along with information in semantic memory that might include one's prior knowledge or thoughts related to the retrieved information, which can serve as effective cues for recalling the target information again later.

Other theories have proposed that retrieval strengthens memory for the contextual details of the information being learned (Karpicke et al., 2014), or that retrieval creates a new memory for the information that is distinct from the memory of originally encoding it (Rickard & Pan,

2018), which provides additional retrieval cues for recalling that information again in the future. Retrieval has also been proposed to benefit learning in an indirect way through revealing to learners what they can and cannot recall, helping to improve metacognitive awareness and effective use of feedback (Roediger et al., 2011). Theories of retrieval practice are not mutually-exclusive, and it is quite likely that more than one of these mechanisms is at work simultaneously. Importantly, what all of these theories have in common is that they focus primarily on how retrieval benefits memory for the retrieved information, and less on how memory for that information is transferred to different contexts.

Given the benefits of retrieval practice on memory retention (Agarwal et al., 2021; Carpenter et al., 2022; McDermott, 2021), a plausible way that retrieval might aid the transfer of learning is through the memory component, as the content that is retrieved should be strengthened and therefore easier for learners to apply during transfer. Although this idea offers a specific account for how retrieval practice might facilitate the transfer of learning, it has not been directly tested. Thus, one of the goals of the current study was to test a memory-based transfer account of retrieval practice.

Assessing the memory-based contribution to transfer is important for both theoretical and practical reasons, because in real-world situations, students often learn information without knowing exactly where or when they will need to apply it. Hints or prompts, therefore, are unlikely to be present to highlight for the learner exactly what prior knowledge the situation calls for. Moreover, in education there are many situations involving transfer where students must learn things that have a connection between the basic concept and its application. In particular, scientific terms consist of both definitions and applications to real-world scenarios that students are expected to learn (see Corral et al., 2019; Corral et al., 2022). Given the educational relevance of this type of learning, the current study focused on the effects of retrieval practice on memory for the definitions of concepts, and the application of these concepts to new situations.

1.3. Effects of retrieval on application of learned concepts

Though some previous research has explored retrieval-enhanced transfer for these types of materials, there are few studies, and their results have been inconsistent. Particularly for studies looking at long-term effects of retrieval (assessing its effects several days or more after learning), results have so far been mixed. McDaniel et al. (2013) looked at exam performance in a middle school science class for information that had appeared on practice quizzes throughout the unit preceding the exam and found that exam performance was higher for quizzed than for un-quizzed information. However, this was only the case when exam questions were the same type of factual questions from the quiz, and not when they required application of the information. Application questions on the practice quizzes, however, did benefit later exam performance on both factual and application questions. At the University level, Thomas et al. (2018) had undergraduate students in an online Brain and Behavior course complete review quizzes with factual or application questions a week before the exam and found that exam performance was higher on both question types for information that had been quizzed versus information that was un-quizzed.

Though these studies show some benefits of quizzing relative to no quizzing on later application questions, neither included a restudy comparison group that learned the same material through a non-retrieval-based strategy. The potential benefits of retrieval thus could have been due to additional opportunities to engage with the material, and as such it is unknown whether restudying the same material without quizzing would have produced similar effects.

Other studies have included a restudy condition that involved additional opportunities to study the original learning material. McDaniel et al. (2015) had students read a textbook chapter on research methods, followed by either practice quizzes or additional time to study the chapter. Quizzes led to better performance on a five-day delayed final test, and this was true for both factual and application questions. Ebersbach et al. (2020) had undergraduate students attend a developmental psychology lecture, then answer factual questions about the lecture content or restudy the lecture slides. On a one-week delayed final test, retrieval produced significant benefits on the same factual questions, but only nominal (non-significant) benefits on never-before-seen application questions. Finally, Hinze and Rapp (2014) had students read scientific texts and then engage in either free recall or restudy of the texts. On a one-week delayed final test, performance on application questions (but not on factual questions) was higher following retrieval than restudy. However, unlike the other studies reviewed here, feedback was not provided after free recall of the text, and thus it is possible that the benefits of retrieval (on factual questions, and possibly on application questions as well) could have been underestimated.

The studies reviewed above compared retrieval to additional time spent studying the material. This type of restudy opportunity has been used in other studies of retrieval practice and is certainly an educationally-relevant type of strategy that many students report engaging in (Corral et al., 2020; Geller et al., 2018; McAndrew et al., 2016; Yan et al., 2014). From a theoretical perspective, however, one benefit of retrieval practice in these cases could be the fact that retrieval provides students with practice questions over the concepts that will later appear on the final test. Especially with more complex materials like a textbook chapter or lecture, students who do not receive practice questions may have a difficult time identifying the specific information that will be tested.

A restudy opportunity that allows students to see the practice questions (with answers provided) provides the same exposure to the questions that retrieval practice does and would determine the degree to which retrieval per se, over and above exposure to the questions and answers, benefits retention and transfer. Thus, we included a new condition in the current study—the "*quiz study*" condition—that provided the same questions as in the retrieval practice condition, but with correct answers already provided so that participants simply studied the questions and answers together without the need to explicitly engage in retrieval.

1.4. Overview of the current experiments

The current experiments explored the effects of retrieval practice on application of complex concepts using a design that controlled for the issues present in previous studies, and that included additional conditions that were designed to test specific hypotheses about how retrieval affects retention and transfer of learning. In three experiments, participants learned concepts about research methods through either retrieval practice (i.e., answering factual, definition-based questions following an introductory tutorial) or through restudy. Participants received correct-answer feedback following retrieval practice. Unlike previous studies, we included two different types of study conditions, one involving restudy of the original learning materials (similar to previous studies), and the other involving study of the same questions and answers as in the retrieval practice condition.

We also included a control condition that only studied the original introductory tutorial and then did not engage further with the material. Such a condition has not been included in previous studies on this topic that compare retrieval to restudy. The primary advantage of a control condition is that it establishes a baseline learning level that can help interpret the effectiveness of retrieval versus restudy. This point is of particular importance in cases where two or more strategies might produce similar levels of final test performance, as the control group is the only way to know whether those strategies benefited performance to a similar degree or failed to benefit performance at all. The final test was administered after 8 min (Experiment 1) or one week (Experiments 2 and 3) and consisted of the same questions that were seen during retrieval practice, as well as never-before-seen application questions

over the same concepts.

2. Experiment 1

2.1. Overview

Experiment 1 investigated whether retrieval practice benefits retention and transfer of complex concepts from research methods (e.g., experimental control, threats to internal validity). The experiment consisted of three primary phases: (a) study, (b) training, and (c) posttest. All participants first studied a tutorial with PowerPoint-like slides illustrating several research methods concepts, which were taken from a subset of the materials used in a previous study (Corral et al., 2019). Some participants were then quizzed on this material through short-answer response questions (retrieval condition), whereas others were presented these same questions in multiple-choice format (recognition condition). After each response, participants in these two conditions were shown the correct answer (i.e., correct-answer feedback). Participants in a third group (quiz study condition) were presented these same quiz questions in short-answer format, which were identical to the quiz questions from the retrieval condition, but differed in that the correct answer was also included; these participants were asked to study each question and answer carefully. Participants in a fourth group were asked to restudy the slides from the tutorial (restudy condition). A control group was also included, which only completed the study phase over the tutorial, and did not receive any additional training afterward.

After the training phase (or after studying the tutorial for the control group), all participants completed a short 8-min filler task and were then given a multiple-choice posttest. To assess memory retention of the material, the posttest included *repeated questions*, which were identical to those from the quiz. To assess the transfer of learning, the posttest also included *application questions*, which tested the same concepts as those from the training phase but required participants to apply their knowledge about these concepts across novel scenarios. The scenarios in the application questions consisted of different domains from the definition-based questions that were used during the training phase (see Fig. 2 for examples), but both types of questions tested the same concepts.

Participants were informed that the final test was over concepts from the learning material, and the multiple-choice nature of the test made clear that the application questions pertained to the same material as the repeated questions. As such, participants were aware of the relevance of the learned concepts for the application questions. Given previous research showing that learners can successfully transfer knowledge to new situations when they are aware of the relevance of that knowledge (e.g., Butler, 2010), we expected that the conditions leading to enhanced memory would also lead to enhanced transfer. Experiment 1 tested the following specific a priori hypotheses.

2.1.1. Retrieval practice hypothesis

The short-answer quiz questions from the retrieval condition encourage participants to fully retrieve the corresponding material from the study phase, whereas this is not the case for the restudy and quiz study conditions. Although multiple-choice questions can engage some level of memory recall (Bjork & Whitten, 1974; Kintsch, 1970), short-answer questions depend more heavily on complete retrieval from memory that is unaided by cues or response alternatives and have also been shown to yield larger benefits of retrieval practice (Rowland, 2014). To the extent that such full and complete retrieval processes contribute to the benefits of retrieval practice (Carpenter & DeLosh, 2006; Glover, 1989), participants in the retrieval condition should learn the material better than participants in the recognition condition. The retrieval condition should thus outperform the restudy, quiz study, and recognition conditions on both repeated and application final test questions. Though our retention interval was short, various studies have shown that retrieval practice benefits do emerge over comparable retention intervals, typically under conditions where learners' initial retrieval success is high (Carpenter, 2009, 2011; Rowland et al., 2014; Sensenig et al., 2011) and when feedback is provided (Carpenter et al., 2008; Kornell et al., 2015).

2.1.2. Transfer-appropriate processing hypothesis

It is worth noting that the format of the quiz questions from the recognition condition matched the format of the posttest questions, as they were both multiple-choice. Based on principles of transfer appropriate processing (McDaniel & Fisher, 1991; Morris et al., 1977), participants in the recognition condition might benefit from this alignment. Indeed, in typical studies of retrieval practice, participants practice answering the same type of questions that they later see on the final test, raising the possibility that the match between practice conditions and final test conditions could contribute to the retrieval practice benefits. If so, the recognition condition would perform as well as the retrieval condition, and better than the restudy condition, on the repeated and application final test questions.

2.1.3. Question familiarity hypothesis

To the extent that familiarity with the test questions and answers contributes to the benefits of retrieval practice, these benefits may be attenuated or eliminated when the retrieval condition is compared to the quiz study condition. If simply seeing the material that will later be tested underlies the benefits of retrieval, then the quiz study condition would perform as well as the retrieval condition, and better than the restudy condition, on the repeated and application final test questions.

2.1.4. Retrieval-enhanced memory over transfer hypothesis

Retrieval has produced robust benefits on memory performance, but less consistent effects on transfer. Recent findings on problem solving suggest that the benefits of retrieval practice might be stronger for memory-based learning (i.e., remembering a solution to a particular problem) than for transfer (i.e., applying that solution to new problems; Corral et al., 2023; also see Yeo & Fazio, 2019). As such, we predicted that the benefits of retrieval would be stronger for repeated questions (which assess memory-based learning) than for application questions (which assess transfer) on the final test.

We also explored final test performance in each training condition relative to the control condition. Given the fact that each training condition involved more time spent learning the material than the control condition, we conducted planned comparisons to test the a priori hypotheses that each of the training conditions would outperform the control condition on the final test.

2.2. Method

2.2.1. Participants

Three hundred fifteen undergraduate students from Syracuse University (SU) participated in this experiment for course credit in an introductory psychology course. Approximately 53% of students who attend SU are White and approximately 52% of students identify as female. Approximately 61% of students are within the range of 18–21 years of age.

Previous work has shown medium-to-large effect sizes for retrieval practice (Adesope et al., 2017; Rowland, 2014), whereas other meta-analysis on the benefits of retrieval and the transfer of learning have reported small-to-medium effect sizes (Pan & Rickard, 2018; Yang et al., 2021). Critically, some of these latter estimates might underestimate the benefits of retrieval on transfer, as they have included studies on problem solving, which Pan and Rickard note are less likely to show benefits of retrieval practice, and in some cases show evidence against it. Pan and Rickard do note, however, that retrieval practice is more likely to benefit the transfer of learning when (a) learners are provided feedback during training, and (b) when learners are required to retrieve the to-be-learned information from memory (as opposed to other forms of testing; e.g., multiple-choice quizzing). Given that Experiment 1

incorporates both of these elements and does not involve any problem solving, we surmised that powering for a medium effect size seemed reasonable.

Our sample size was therefore based on the number of participants needed to detect at least a medium-sized effect at 80% power ($f = .25$; *alpha* = .05), which was 64 participants per condition. This estimate includes the smallest sample size required across all analyses that we conducted, including the omnibus ANOVA and all comparisons between conditions. We enrolled the maximum number of participants that could be collected within the semester that the experiment was run. Participants were randomly assigned to one of the five conditions: retrieval ($n = 65$), recognition ($n = 62$), quiz study ($n = 69$), restudy ($n = 57$), control ($n = 62$). Experiment 1 was approved by the Institutional Review Board (IRB) of human participants at SU. Experiment 1 was not preregistered. The materials and data for all three experiments are available here: https://osf.io/bcs86/files/osfstorage.

2.2.2. Design and materials

There was a total of five PowerPoint-like study slides for the introductory tutorial. The first slide covered *independent* and *dependent variables*; the second slide covered *confounds* and *experimental control*; the third slide covered *reverse causation*; the fourth slide covered the *third variable problem*; the fifth slide covered *self-selection*. Fig. 1 shows an example study slide from the tutorial.

The training quiz questions corresponded to material that was covered on slides 2–5. The posttest questions were in multiple-choice format (with options ranging from a–e, with only one correct answer). The content of the repeated questions on the posttest was identical to the content in the training quiz questions. Critically, these questions tested straightforward factual information that could be answered by memorizing the content in the tutorial study slides (e.g., *When it is unclear whether variable X causes variable Y or whether variable Y causes variable X, what kind of causal inference problem do we have?)*. Participants in the retrieval, recognition, and quiz study conditions could also answer these questions by memorizing the correct answers from the training phase quiz questions. The application questions, on the other hand, tested the same concepts as the repeated questions, but these concepts were tested across novel, hypothetical scenarios. Unlike with repeated questions, application questions could not be answered by simply memorizing the content from the tutorial study or training phases of the experiment. Fig. 2 shows an example repeated question (Panel A) and application question (Panel B).

The questions for the training phase were fairly straightforward definition-based questions, because these questions included explicit instruction on the to-be-learned concepts. Application-type questions do not explicitly define the to-be-learned concepts, but rather serve as more applied examples of the concepts, and thus we used these for assessing transfer to new situations.

2.2.3. Procedure

This was an in-person experiment conducted in a laboratory. All instructions and materials were presented at the center of a computer screen and all answers were entered using a computer keyboard.

Tutorial Study Phase. First, participants were instructed that they would be presented material about basic scientific research principles to study. Participants were notified that they would be given 8 min to study and that they would be tested on this material at the end of the experiment.

At the beginning of the tutorial study phase, participants were presented five PowerPoint-like slides for study. One slide was presented at a time and participants could toggle between each slide by pressing the "N" key to move to the next slide and the "B" key to move to the previous slide. A prompt was presented at the top of the screen which notified participants of how to move to the next or previous slide. A counter was displayed on the bottom right side of the screen, which showed participants which slide number they were on (e.g., Slide 2 of 5). If participants attempted to move past the fifth slide before 8 min had passed, the screen was cleared and they were instructed to continue to study for the remaining duration of the study time, which was displayed on the screen. These participants were also instructed to press the spacebar to return to the previous slide and continue to study.

Training Phase. After 8 min had passed, the screen was cleared, and participants were instructed that they would now receive some additional training on the material that they had just studied. These instructions served as a self-paced rest break, and participants were instructed to press the spacebar when they were ready to begin the next

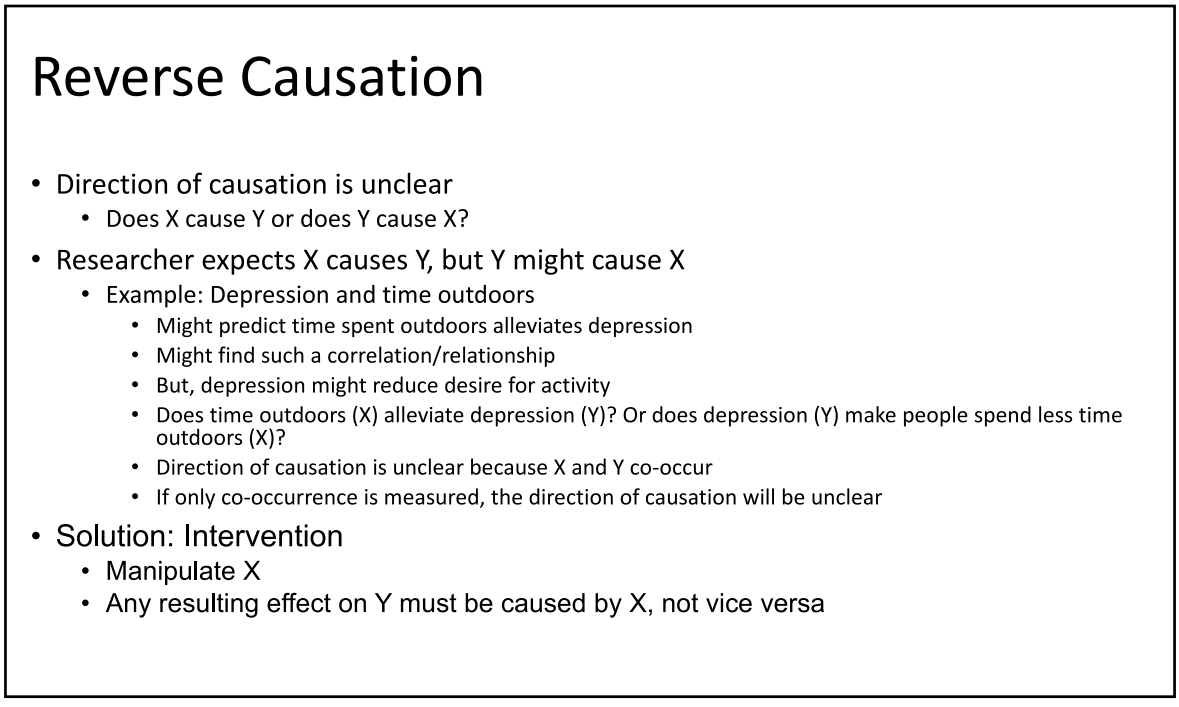


Fig. 1. *Example slide from the tutorial study phase of experiments 1-3*

A.

When it is unclear whether variable X causes variable Y or whether variable Y causes variable X, what kind of causal inference problem do we have?

a. Reverse causation
b. Reverse correlation
c. Third variable
d. Self-selection
e. Researcher expectancy

B.

Lisa just received her master's for her work in astronomy. Lisa found that planets that are closer to their moons rotate faster than planets that are farther away from their moons. However, Lisa is not sure how to interpret her data because it is unclear whether a planet that rotates faster draws its moons in closer to the planet or whether moons that are closer to a planet cause the planet to rotate faster. Ignoring the possibility that other factors may be involved, what kind of problem does Lisa have?

a. Third variable
b. Researcher expectation
c. Reverse causation
d. Reverse correlation
e. Reverse association

Fig. 2. *Example posttest repeated question (Panel A) and application question (Panel B)*

phase. Participants in the control condition skipped the training phase, and immediately after the 8-min tutorial went straight to the filler task (see below).

Restudy Condition. Participants in the restudy condition were asked to restudy the material from the tutorial study phase and were given 4 min to do so. Outside of reducing the study time from 8 min to 4 min, the tutorial study and training phases were identical for the restudy condition.

Retrieval Condition. At the beginning of the training phase, participants in the retrieval condition were notified that they would be quizzed on the material that they had just studied and that they would be shown the correct answer after each response. Participants were then given five short-answer quiz questions, each of which corresponded to a given concept from the tutorial. Each question was presented once, in randomized order. For each quiz question, participants were asked to type out their response into a textbox that was presented directly below the quiz question and to press the "Enter" key when they were ready to submit their response. Upon submitting their response, correct-answer feedback was presented in green text and was shown directly below the textbox. Feedback was self-paced, and participants were asked to study the question and answer carefully and press the "N" key when they were ready to move on to the next question.

Quiz Study Condition. At the beginning of the training phase, participants in the quiz study condition were notified that they would be presented quiz questions with the correct answers and that they would need to study the question and answer carefully. The quiz questions were identical to those in the retrieval condition, except that the correct answer was presented in green text below each question, identical to the feedback presentation in the retrieval condition. Questions were presented one at a time, in randomized order. Participants were asked to press the "N" key when they were ready to move on to the next question.

Recognition Condition. At the beginning of the training phase, participants in the recognition condition were given the same general instructions as participants in the retrieval condition, that they would be quizzed on the material and then shown the correct answer after each response. The quiz questions were the same as in the retrieval condition, except that they were in multiple-choice format with options ranging from a-e. Participants were asked to select the correct answer from the list of five options and to press the "Enter" key when they were ready to submit their response. Upon submitting their response, the correct option was displayed in green. As in the other conditions, participants were asked to study the question and correct answer carefully, and to press the "N" key when they were ready to move on to the next question. These questions were identical to the repeated questions on the posttest.

Filler Task. After completing the training phase (or directly after the tutorial for the control group), participants were asked to complete a

filler task, which consisted of a reading comprehension quiz over material unrelated to the research methods concepts. Specifically, five reading comprehension problems were taken from a GRE verbal practice test (Princeton Review, 2017); each problem was multiple choice and consisted of five response options (a–e). This task was designed to last for 8 min. If participants completed the reading comprehension task before 8 min passed, they were presented instructions on the screen that asked them to sit and wait quietly until it was time to move on.

Posttest. After the filler task, all participants were given the same multiple-choice posttest. The posttest consisted of five repeated questions (identical to those from the recognition condition) and five never-before-seen application questions. The application questions were presented first, followed by the repeated questions. Within each of these blocks, the presentation order of the questions was randomized for each participant, and no feedback was provided.

2.3. Results

2.3.1. Preliminary analyses

We first looked at completion times for Experiment 1 across each of the five conditions. Although the experiment was conducted in a laboratory, we carefully inspected the data for any evidence that participants could have been off task or distracted during the experiment. Data from six participants were excluded for taking more than three standard deviations longer than the group mean to complete the experiment. The following analyses are based on the remaining participants in each condition: retrieval (n = 64), restudy (n = 56), quiz study (n = 67), recognition (n = 61), and control (n = 61).

Analysis of quiz accuracy in the retrieval condition revealed that participants answered an average of 49.12% of questions correctly (SD = 29.96%). Participants in the recognition condition answered an average of 69.51% of questions correctly (SD = 23.27%). Completion time for the experiment varied across conditions, $F(4, 304) = 46.321$, $p < .001$, $MSE = 4.458$, $\eta p^2 = .379$. Post-hoc LSD comparisons showed that the retrieval condition ($M = 24.152$, $SD = 2.217$) took more overall time than the control condition (in minutes; $M = 21.653$, $SD = 2.135$), $p < .001$, 95% CI [1.755, 3.242], recognition condition ($M = 22.727$, $SD = 1.901$), $p < .001$, 95 % CI [.682, 2.169], and quiz study condition ($M = 22.081$, $SD = 1.609$), $p < .001$, 95% CI [1.345, 2.797], but took less time than the restudy condition ($M = 26.272$, $SD = 2.644$), $p < .001$, 95% CI [−2.880, −1.360].[1] Cronbach's alpha for the 10 final test questions was .651.

2.3.2. Overall final test performance

Fig. 3 shows mean performance for each condition partitioned by posttest question type (see also Table 1). To examine the results, we first conducted a mixed ANOVA, with training condition as a between-participants factor (retrieval vs. recognition vs. quiz study vs. restudy vs. control) and posttest question type as a within-participants factor (repeated vs. application). The results revealed a main effect of question type, $F(1, 304) = 85.516$, $p < .001$, $MSE = .033$, $\eta p^2 = .220$, as participants performed better on the repeated questions ($M = .814$, $SE = .012$) than on the application questions (M = .679, SE = .015). A non-significant effect of condition was also observed, $F(4, 304) = 1.920$, $p = .107$, $MSE = .086$, $\eta p^2 = .025$, and no interaction occurred between condition and question type, F (4, 304) = .922, $p = .451$, $MSE = .033$, $\eta p^2 = .012$.

2.3.3. Final test performance between training conditions

Next, to evaluate our a priori hypotheses, we conducted a series of planned comparisons among the training conditions to examine performance differences on the repeated and application questions. For completeness, the full set of pairwise comparisons is provided in Table A1 of the Appendix.

Repeated Questions. Participants in the retrieval condition numerically outperformed those in the restudy condition, however this difference was not significant, $t(118) = 1.909$, $p = .059$, $SE = .043$, $d = .349$, 95% CI [−.003, .166]. No performance differences were observed between the retrieval and the quiz study conditions, $t(129) = -.100$, $p = .920$, $SE = .036$, $d = -.017$, 95% CI [−.075, .067]), nor between the retrieval and recognition conditions, $t(123) = .711$, $p = .478$, $SE = .038$, $d = .127$, 95 % CI [−.048, .102]). Participants in the quiz study condition outperformed those in the restudy condition, $t(121) = 2.045$, $p = .043$, $SE = .042$, $d = .370$, 95% CI [.003, .168]). Participants in the recognition condition did not perform better than those in the restudy condition, $t(115) = 1.248$, $p = .215$, $SE = .044$, $d = .231$, 95% CI [−.032, .142]), or those in the quiz study condition, $t(126) = -.827$, $p = .410$, $SE = .037$, $d = -.146$, 95% CI [−.103, .042]).

Application Questions. No performance differences on the application questions were observed between the retrieval versus restudy, $t(118) = .851$, $p = .397$, $SE = .049$, $d = .156$, 95% CI [−.056, .140]), retrieval versus quiz study, $t(129) = -1.137$, $p = .258$, $SE = .047$, $d = -.199$, 95% CI [−.145, .039]), or retrieval versus recognition conditions, $t(123) = .249$, $p = .803$, $SE = .050$, $d = .045$, 95% CI [−.086, .111]). Participants in the quiz study condition outperformed those in the restudy condition, $t(121) = 2.054$, $p = .042$, $SE = .046$, $d = .372$, 95% CI [.003, .187]). Participants in the recognition condition did not perform better than those in the restudy condition, $t(115) = .599$, $p = .550$, $SE = .049$, $d = .111$, 95% CI [−.068, .127]), or those in the quiz study condition, $t(126) = -1.398$, $p = .164$, $SE = .047$, $d = -.247$, 95% CI [−.158, .027]).

2.3.4. Final test performance between training conditions versus control

Next, to assess whether the training conditions produced learning above and beyond the initial tutorial study phase, we compared the training conditions to the control group. Given our a priori hypothesis that each training condition would result in better learning than the control condition, we conducted planned comparisons of final test performance between the control condition and each of the other conditions. The full set of pairwise comparisons is provided in Table A2 of the Appendix.

On the repeated questions, participants in the retrieval condition performed better than control participants, $t(123) = 2.392$, $p = .018$, $SE = .037$, $d = .428$, 95% CI [.015, .163]), as did participants in the quiz study condition, $t(126) = 2.555$, $p = .012$, $SE = .036$, $d = .452$, 95% CI [.021, .165]). However, participants in the recognition condition did not outperform control participants, $t(120) = 1.628$, $p = .106$, $SE = .038$, $d = .295$, 95% CI [−.013, .138]), nor did participants in the restudy condition, $t(115) = .173$, $p = .863$, $SE = .043$, $d = .032$, 95% CI [−.078, .093]). On the application questions, none of the training conditions performed better than the control condition: retrieval, $t(123) = .186$, $p = .853$, $SE = .049$, $d = .033$, 95% CI [−.088, .106]), restudy: $t(115) = -.675$, $p = .501$, $SE = .049$, $d = -.125$, 95% CI [−.129, .064]), recognition: $t(120) = -.067$, $p = .947$, $SE = .049$, $d = -.012$, 95% CI [−.101, .094]), and quiz study: $t(126) = 1.346$, $p = .181$, $SE = .046$, $d = .238$, 95% CI [−.029, .154]).

2.4. Discussion

Results of Experiment 1 show that final test performance was better for repeated questions than for application questions, confirming the common finding in the literature that retention is typically easier than transfer (Carpenter et al., 2013; Corral et al., 2019, 2023). These results also indicate some positive effects of retrieval practice over restudy and control, but these effects were limited to repeated questions and did not occur for application questions. Moreover, the retrieval condition did not significantly outperform the quiz study condition. These results appear to support the question familiarity hypothesis, showing that

[1] The same pattern of results was observed for all reported analyses when time on training phase is included as a covariate.

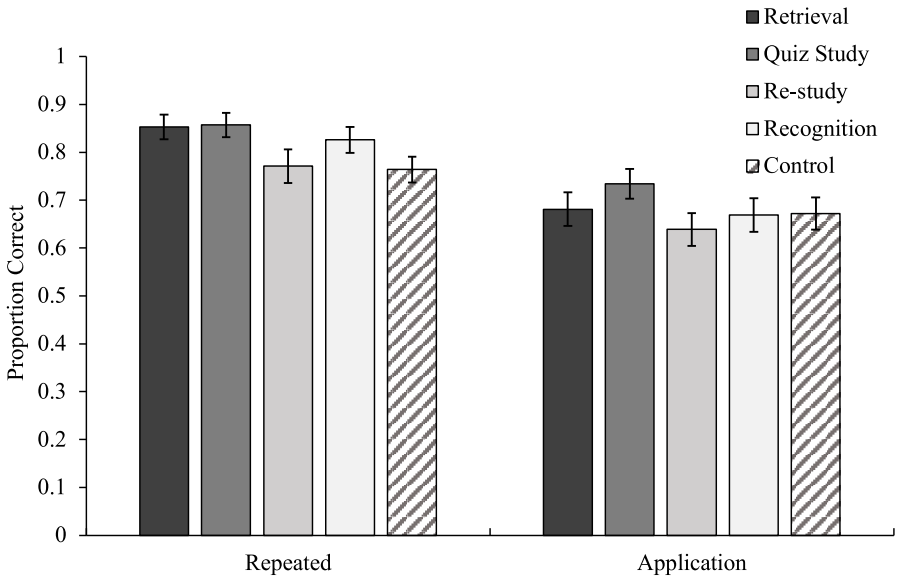


Fig. 3. *Mean performance for each condition and standard errors of the mean on the repeated and application questions in experiment 1*

Table 1
Unadjusted and adjusted[a] *mean performance on repeated and application questions, partitioned by condition, in experiments 1-3*

	Retrieval	Quiz Study	Restudy	Recognition	Control
Unadjusted means					
Experiment 1					
Repeated	.853 (.027)	.857 (.027)	.771 (.029)	.826 (.028)	.764 (.028)
Application	.681 (.033)	.734 (.033)	.639 (.036)	.669 (.034)	.672 (.034)
Experiment 2					
Repeated	.832 (.044)	.711 (.043)	.633 (.038)	–	.617 (.045)
Application	.681 (.051)	.553 (.050)	.521 (.045)	–	.526 (.052)
Experiment 3					
Repeated	.825 (.027)	.776 (.028)	.703 (.029)	–	.631 (.029)
Application	.654 (.030)	.570 (.031)	.503 (.032)	–	.469 (.033)
Adjusted means					
Experiment 1					
Repeated	.857 (.028)	.851 (.028)	.784 (.034)	.824 (.028)	.757 (.030)
Application	.691 (.034)	.720 (.034)	.673 (.042)	.662 (.034)	.653 (.036)
Experiment 2					
Repeated	.870 (.053)	.700 (.044)	.647 (.040)	–	.571 (.058)
Application	.700 (.062)	.547 (.051)	.528 (.046)	–	.502 (.068)
Experiment 3					
Repeated	.834 (.029)	.773 (.028)	.704 (.029)	–	.622 (.032)
Application	.668 (.033)	.566 (.031)	.505 (.032)	–	.454 (.036)

Note. Standard errors of the mean are shown in parentheses.
[a] Experiment 1: Adjusted for total experiment Time; Experiments 2–3: Adjusted for time on training phase. Performance in Experiment 1 was adjusted for total experiment time, because precise estimates for time on the training phase were not available.

exposure to questions during an initial quiz could be part of what underlies the benefits of retrieval practice.

Although the retrieval group performed numerically better than the recognition group on both repeated and application questions, counter to the transfer-appropriate processing hypothesis, these differences were small and non-significant, which might suggest that the completeness of retrieval, an effect found in previous studies using fairly simple materials (e.g., Carpenter & DeLosh, 2006; Glover, 1989), does not apply as readily to the materials used in the current study. It is worth noting that the overall learning effects were also quite small. Only the retrieval and quiz study groups performed better than the control group on repeated questions, and none of the training conditions outperformed the control condition on application questions.

Overall, Experiment 1 provides some preliminary support for the positive effects of retrieval practice on retention, but does not definitively support the retrieval practice hypothesis nor the retrieval-enhanced memory over transfer hypothesis. The effect sizes observed here may be influenced by the nature of the current materials, which are more complex than those used in many studies of retrieval practice. As such, it may have been challenging to recall the correct answers in the retrieval condition (indeed, initial accuracy on the short-answer quiz was only about 49 %).

Given the finding that retrieval practice benefits increase when more opportunities for retrieval are provided (e.g., Eriksson et al., 2011; Kornell & Bjork, 2008; McDaniel et al., 2012, 2013; Pyc & Rawson, 2009; Vaughn & Rawson, 2011), it is possible that these benefits were underestimated in Experiment 1 due to the fairly modest effects that resulted from just one retrieval opportunity. It is also possible that the short time interval between learning and final test (only 8 min) was insufficient for revealing stronger benefits of retrieval practice, which are more often observed after final test delays on the order of days (Rowland, 2014). Experiment 2 was designed to explore these possibilities.

3. Experiment 2

3.1. Overview

Experiment 2 was designed to follow up on the results from Experiment 1 under conditions in which the benefits of retrieval are more likely to occur. In particular, we included additional retrieval practice opportunities and administered the posttest after a longer retention interval than in Experiment 1. Using the same materials and basic procedure from Experiment 1, participants in Experiment 2 completed three

rounds of retrieving (retrieval condition) or restudying (quiz study condition) the material during training and then completed the posttest one week later.

These changes also serve to increase the ecological validity of the experiment, as students might typically quiz themselves on a to-be-learned concept repeatedly. Furthermore, students are required to retain the knowledge that they learn over extended periods (e.g., days, weeks, months). Thus, to meaningfully assess whether retrieval practice is a viable strategy for improving the transfer of learning in education, it is essential to determine whether its benefits hold over longer periods.

Experiment 2 included the retrieval, quiz study, restudy, and control conditions. We did not include the recognition condition from Experiment 1, because we sought to follow up specifically on the comparison between retrieval and quiz study conditions, both of which performed nominally better than the recognition condition on the final test. Including four rather than five conditions also helped to streamline data collection and retention over a two-part study, where data can be lost due to attrition and other factors. Aside from the extra retrieval practice and one-week final test delay, the design of Experiment 2 was largely the same as Experiment 1.

The specific a priori hypotheses we tested in Experiment 2 included the retrieval practice hypothesis (that retrieval would lead to better final test performance on both the repeated and application questions than the quiz study and restudy conditions), the question familiarity hypothesis (that participants in the quiz study condition would perform as well as those in the retrieval condition, and better than the restudy condition, on repeated and application final test questions), and the retrieval-enhanced memory over transfer hypothesis (that the benefits of retrieval would be stronger for repeated than for application final test questions). Additionally, we again tested the a priori hypothesis that each training condition would outperform the control condition.

3.2. Method

3.2.1. Participants

Seventy-one undergraduate students from Oregon State University (OSU) participated in Experiment 2 for course credit in an introductory psychology course. Approximately 61% of students enrolled at OSU are White and approximately 48% of students identify as female. Approximately 49% of students fall within the age range of 18–21. This experiment was approved by the IRB of human participants at OSU. In addition, we simultaneously collected a second sample of participants who were paid from Prolific (www.prolific.co). This sample consisted of 146 participants who were paid a total of $3 for their participation, which amounts to an hourly wage of $8 per hour. Participants were paid $2 for participating in part one of the experiment and $1 for participating in part two. Experiment 2 was not preregistered.

Based on estimated effect sizes for retrieval practice effects observed after a final test delay of longer than one day (Rowland, 2014, Hedges' $g = .69$), and from previous research implementing a similar design to Experiment 2 using three rounds of retrieval practice and a final test delay of five days (McDaniel et al., 2015, $d = .68$), our sample size was based on the number of participants needed to detect similar-sized effects at 80% power (*alpha* = .05), which was 35 participants per condition. This estimate includes the smallest sample size required across all analyses that we conducted, including the omnibus ANOVA and all comparisons between conditions. We enrolled the maximum number of participants that could be collected within the term that the experiment was run at OSU and enrolled enough participants through Prolific to account for some anticipated attrition across the two experimental sessions. As we report further in the results section, these different samples did not affect any of the results.

A total of 217 participants from both OSU and Prolific were randomly assigned to one of the four conditions: retrieval ($n = 55$), quiz study ($n = 49$), restudy ($n = 66$), and control ($n = 47$).

3.2.2. Design and procedure

Experiment 2 was conducted online. All participants were required to complete the experiment on a computer. All instructions and materials were presented on a computer screen and participants entered their responses using a computer keyboard and a mouse or trackpad.

The tutorial study phase and posttest were identical to Experiment 1, as was the procedure for the restudy and control conditions. For the training phase, participants in the retrieval and quiz study conditions were initially presented the same five quiz questions from Experiment 1, with the requirement to type in an answer followed by viewing correct answer feedback (in the retrieval condition) or to study the question and correct answer together (in the quiz study condition). After the first round of questions, these participants were instructed that they would be presented another round of questions, which was identical to the first one, and following this second round, participants were informed they would have a third and final round of the same questions. The same five quiz questions were used for all three rounds, amounting to 15 quiz questions in total. In each round, the order in which the quiz questions were presented was randomized for each participant.

After the training phase, all participants were thanked for completing the first part of the experiment and were reminded that they would receive an e-mail link to complete the second part of the study one week later, wherein they would be tested on the material that they had just learned about. Participants were also asked not to study or look up the material that they had just been tested on. On the morning of the posttest (i.e., one week after part one of the experiment), participants were sent an email that reminded them that they could now complete part two of the experiment, along with a link to access part two. Participants were given 24 h to complete part two after receiving this e-mail. The repeated questions were presented prior to the application questions on the final test.

After finishing the final test, participants were asked to report their level of prior knowledge of the material; participants were also asked if they had looked up or learned about any of the material from part one of the study during the interim week.

3.3. Results

3.3.1. Preliminary analyses

Out of the 217 participants who participated in part one of the experiment, 180 returned for part two. There was thus an attrition rate of approximately 17%. The attrition rate did not differ across conditions, $\chi^2 = 1.06$, $p = .787$. An additional 14 participants were excluded: two who accidently completed the first part of the experiment twice, one who reported already knowing 100% of the material (and achieving a perfect score on the posttest), four who reported learning the material in class in between parts one and two, and seven who reported looking up the material and studying it after the training phase but before the final test.

Due to the online nature of the study, we took extra precautions to check for any evidence of disengagement from the task and to verify that participants did not differ across conditions on any of the additional preliminary measures we collected. We carefully inspected completion times for both part one and part two. Eight participants took more than 3 standard deviations longer than the group mean to complete either part one or part two and were thus removed from the final analyses. This screening procedure was incorporated over concerns that participants who took too long to complete the experiment may not have been fully engaged and could have been carrying out other tasks during the experiment. The final sample thus consisted of 158 participants: retrieval ($n = 37$), quiz study ($n = 38$), restudy ($n = 48$), and control ($n = 35$). Cronbach's alpha for the 10 final test questions was .743.

Overall, participants reported little prior knowledge of the material, with an overall average self-reported rating of 1.32 ($SD = 1.035$) out of 4 (where 0 = no prior knowledge, 1 = some prior knowledge, 2 = knew about half the material, 3 = knew most of the material, and 4 = knew all

of the material). Degree of prior knowledge did not vary across the conditions, $F(3, 154) = .297$, $p = .827$, $MSE = 1.086$, $\eta p^2 = .006$. Because participants were allowed some flexibility of when they could complete part two (i.e., within 24 h of receiving the e-mail), we also examined the time delay in-between part one and part two. The average time delay was 7.27 days, and this did not vary across conditions, $F(3, 154) = .032$, $p = .992$, $MSE = .875$, $\eta p^2 = .001$.

A significant difference was observed between the conditions in total time to complete part one, $F(1, 154) = 65.228$, $p < .001$, $MSE = 3.858$, $\eta p^2 = .560$. Post-hoc LSD comparisons showed that the retrieval condition took more time (in minutes; $M = 14.406$, $SD = 3.145$) than the control condition ($M = 8.264$, $SD = .269$), $p < .001$, 95% CI [5.227, 7.057], the restudy condition ($M = 12.653$, $SD = .776$), $p < .001$, 95% CI [.904, 2.602], and the quiz study condition ($M = 10.850$, $SD = 2.367$), $p < .001$, 95% CI [2.660, 4.451].[2]

Finally, we looked at accuracy on the quiz questions during training for the retrieval group. As expected, a within-participants ANOVA revealed that quiz performance increased across the three rounds of retrieval practice, $F(2, 72) = 37.513$, $p < .001$, $MSE = .027$, $\eta p^2 = .510$. Specifically, post-hoc LSD comparisons showed that initial accuracy on the first round ($M = .416$, $SE = .045$) improved on the second round ($M = .660$, $SE = .050$), $p < .001$, 95% CI [.169, .317], and from the second round to the third round ($M = .735$, $SE = .050$), $p = .037$, 95% CI [.005, .147]. Thus, providing additional retrieval practice opportunities increased the accuracy of retrieval to a degree much higher than what was observed in Experiment 1 (only about 49%).

3.3.2. Overall final test performance

Fig. 4 shows mean final test performance for each condition, partitioned by question type (also see Table 1). To analyze our data, we first conducted a mixed ANOVA, with training condition (retrieval vs. quiz study vs. restudy vs. control) and sample (university students vs. Prolific) as between-participant factors and posttest question type as a within-participants factor (repeated vs. application). The results revealed a main effect of question type, $F(1, 150) = 24.284$, $p < .001$, $MSE = .043$, $\eta p^2 = .139$, 95% CI [.075, .174], as in Experiment 1, wherein participants performed better on the repeated questions ($M = .695$, $SE = .022$) than on the application questions ($M = .567$, $SE = .025$). More importantly, a main effect of condition was observed, $F(3, 150) = 4.983$, $p = .003$, $MSE = .122$, $\eta p^2 = .091$. No interaction occurred between condition and question type, $F(3, 150) = .200$, $p = .896$, $MSE = .043$, $\eta p^2 = .004$. No main effect of sample was observed, and sample did not interact with any of the factors (all $ps > .142$), indicating that the results did not vary according to where participants were sampled from (i.e., university students vs. Prolific).

Next, to evaluate our a priori hypotheses, we conducted a series of planned comparisons among the training conditions to examine performance differences on the repeated and application questions. For completeness, the full set of pairwise comparisons is provided in Table A3 of the Appendix.

3.3.3. Final test performance between training conditions

Repeated Questions. Participants in the retrieval condition performed significantly better than participants in both the restudy condition, $t(83) = 3.388$, $p = .001$, $SE = .059$, $d = .741$, 95% CI [.082, .316]) and the quiz study condition, $t(73) = 2.090$, $p = .040$, $SE = .058$, $d = .483$, 95% CI [.006, .238]). Participants in the quiz study condition did not perform significantly better than participants in the restudy condition, $t(84) = 1.201$, $p = .233$, $SE = .064$, $d = .261$, 95% CI [−.051, .205].

Application Questions. A similar pattern emerged for the application questions. Specifically, participants in the retrieval condition performed significantly better than participants in the restudy condition, $t(83) = 2.334$, $p = .022$, $SE = .069$, $d = .511$, 95% CI [.024, .297]), and nominally, but not significantly, better than participants in the quiz study condition, $t(73) = 1.762$, $p = .082$, $SE = .073$, $d = .407$, 95% CI [−.017, .274]). Participants in the quiz study condition did not perform significantly better than participants in the restudy condition, $t(84) = .461$, $p = .646$, $SE = .069$, $d = .100$, 95% CI [−.106, .169].

3.3.4. Final test performance between training conditions versus control

Next, to assess whether the training conditions produced learning above and beyond the initial tutorial study phase, we compared the training conditions to the control group. Given our a priori hypothesis that each training condition would result in better learning than the control condition, we conducted planned comparisons of final test performance between the control condition and each of the other conditions. The full set of pairwise comparisons is provided in Table A4 of the Appendix.

Participants in the retrieval condition reliably outperformed control participants on repeated questions, $t(70) = 4.077$, $p < .001$, $SE = .053$, $d = .961$, 95% CI [.110, .321] and application questions, $t(70) = 2.213$, $p = .030$, $SE = .070$, $d = .522$, 95% CI [.015, .295]. No performance differences were found between the quiz study and control conditions on either the repeated questions, $t(71) = 1.520$, $p = .133$, $SE = .061$, $d = .356$, 95% CI [−.029, .216] or the application questions, $t(71) = .379$, $p = .706$, $SE = .071$, $d = .089$, 95% CI [−.115, .168]. Lastly, no performance differences were found between the restudy and control conditions on either repeated questions, $t(81) = .263$, $p = .793$, $SE = .062$, $d = .058$, 95% CI [−.106, .139], or application questions, $t(81) = -.073$, $p = .942$, $SE = .067$, $d = -.016$, 95% CI [−.139, .129].

3.3.5. Transfer contingent on retention

We conducted one additional exploratory analysis. Given the benefits of retrieval on memory retention and its partial benefits on transfer (significant compared to restudy, but not compared to quiz study), we explored whether any retrieval-enhanced transfer effects might reflect a specific form of memory-based transfer. Our learning materials were designed such that four of the five final test questions represented repeated-application question pairs. That is, for a repeated question over a given concept (e.g., third variable problem), an application question existed that tested that same concept. Thus, four pairs of questions existed for which the repeated and application version of the question tested the same concept and had the same correct answer (see test questions 1–4[3]; https://osf.io/h58bj.

This analysis allowed us to examine performance on application questions contingent on successful performance on repeated questions that assessed the same concepts. Thus, we can explore transfer across the four conditions under a situation where memory for each concept is correct. If memory is the key component contributing to retrieval-based transfer, then evaluating transfer only for items that were successfully remembered across all of the conditions—essentially, holding memory performance constant where it is perfectly accurate across all conditions—should eliminate any benefit of retrieval on transfer. Alternatively, if recognition of the relevance of a learned concept for answering the application question is the key component contributing to retrieval-based transfer, then this conditional analysis should still show a benefit of retrieval.

The conditional analysis did indeed show a benefit of retrieval practice ($M = .707$, $SE = .062$) over quiz study ($M = .580$, $SE = .064$, 95% CI [−.051, .303]), restudy ($M = .619$, $SE = .054$, 95% CI [−.075, .251]), and control ($M = .620$, $SE = .053$, 95% CI [−.076, .250]). However, a one-way ANOVA with condition as a between-participants

[2] The same pattern of results was observed for all reported analyses when time on training phase is included as a covariate.

[3] For Question 5, the application question assessed the same concepts as its corresponding repeated question, but we did not include this application question in our conditional analysis because the answers between the repeated and application question were different.

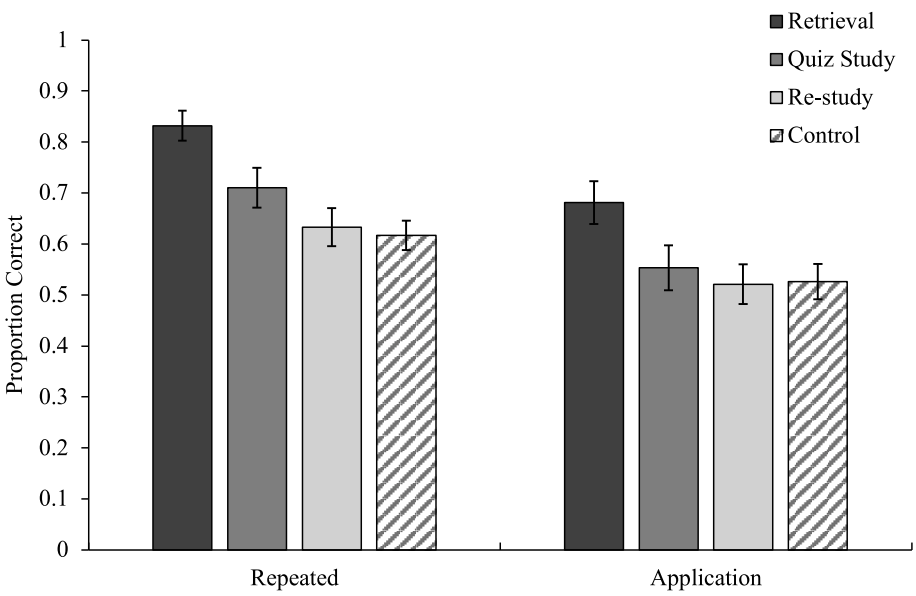


Fig. 4. *Mean performance for each condition and standard errors of the mean on repeated and application questions in experiment 2*

factor revealed that there were no significant differences between conditions, $F(3, 145) = .827$, $p = .481$, $MSE = .127$, $\eta p^2 = .017$.

3.4. Discussion

Results of Experiment 2 provide further insights into the effects of retrieval on retention and transfer. Relative to Experiment 1, in which only one retrieval opportunity was provided, three retrieval opportunities resulted in much higher performance on the initial quizzes. Combined with a one-week delayed final test, Experiment 2 resulted in a significant benefit of retrieval over restudy, and this benefit occurred for both repeated questions and application questions.

Experiment 2 also provides new evidence that retrieval practice appears to be more effective than studying the quiz questions with answers. Given the benefits of retrieval on both repeated and application questions, we did not observe a significant interaction whereby retrieval practice was more effective for repeated questions than for application questions, contrary to the retrieval-enhanced memory over transfer hypothesis.

Our results thus provide partial support for the retrieval practice hypothesis. However, the benefits of retrieval compared to quiz study were not as strong as they were compared to restudy. Although these results do not definitively support the question familiarity hypothesis by eliminating the retrieval practice benefit in the quiz study condition, they do suggest that there may be some learning benefit that occurs by having access to the questions and answers, even if learners merely study those materials and do not engage in retrieval.

Importantly, however, the effect sizes for retrieval in Experiment 2 were smaller than expected. Though we powered for an effect size similar to that found in previous research using a similar design with application final test questions (McDaniel et al., 2015, $d = .68$), the effect sizes we observed for application questions were $d = .511$ (for retrieval over restudy) and $d = .407$ (for retrieval over quiz study). Indeed, a sensitivity analysis (with 80 % power and *alpha* = .05) revealed that the sample size in Experiment 2 was sufficient to detect an effect size of $d = .656$, indicating that the study was underpowered. Though the results of the conditional analysis to assess memory-based transfer revealed some nominal (but not significant) benefit of retrieval, this analysis was very likely underpowered as well. To address these issues and provide additional data, we conducted Experiment 3, which was a close replication of Experiment 2 with a larger sample size.

4. Experiment 3

4.1. Overview

Experiment 3 was designed to replicate and extend the findings from Experiment 2. Using the same conditions as in Experiment 2—retrieval, restudy, quiz study, and control—Experiment 3 involved identical procedures during part one. During part two one week later, participants received the same repeated and application questions as in part two, but unlike in Experiment 2 in which the repeated questions preceded the application questions, in Experiment 3 the application questions preceded the repeated questions. This minor change was made in order to look more closely at the potential effects of retrieval on transfer. Though Experiment 2 revealed a benefit of retrieval on application questions (significant compared to restudy, but only nominal compared to quiz study), there is a possibility that answering the repeated questions first could have influenced participants' performance on the application questions.

More specifically, though feedback was not provided during final test questions, answering the repeated questions first could have reminded participants of the concepts or made these concepts more accessible, resulting in more effective performance on the application questions over the same concepts that immediately followed. If so, the benefits of retrieval on application questions could be due to this reminding or accessibility effect, as retrieval showed a clear benefit in memory performance on the repeated questions in Experiment 2. Though this is still certainly considered a form of transfer (and indeed studies on transfer sometimes intentionally provide a hint or reminder of how the information previously learned could be relevant in the transfer task, e.g., Butler, 2010), it does raise the question of whether benefits of retrieval would show up on application questions when there are no repeated questions beforehand, which might be considered a more "pure" measure of retrieval-enhanced transfer.

Experiment 3 was preregistered and designed to test specific a priori hypotheses. As in Experiment 2, we tested the retrieval practice hypothesis (that retrieval would lead to better final test performance on both the repeated and application questions than the quiz study and restudy conditions), and the question familiarity hypothesis (that participants in the quiz study condition would perform as well as those in the retrieval condition, and better than the restudy condition, on repeated and application final test questions). Although Experiment 2 did not provide strong evidence to support the retrieval-enhanced

memory over transfer hypothesis (that the benefits of retrieval would be stronger for repeated than for application final test questions), we tested this a priori hypothesis again in Experiment 3 given the switch in the order of the application versus repeated questions on the final test, on the premise that receiving the application questions first could make it more difficult to answer them, leading to the possibility that application questions were less likely than repeated questions to show benefits of retrieval.

Experiment 3 preregistered the additional a priori hypothesis that performance on application questions, given correct performance on repeated questions, would be greater for retrieval compared to the other conditions. The same conditional analysis from Experiment 2 was conducted to compare final test performance on the four application questions contingent on accurate performance on the four analog repeated questions that tested the same concepts.

Finally, as in Experiments 1 and 2, we again tested the a priori hypothesis that each training condition would outperform the control condition.

4.2. Method

4.2.1. Participants

A total of 398 participants from Prolific (www.prolific.co) participated in Experiment 3. The payrate and recruitment procedures were identical to those used in Experiment 2 with the Prolific sample. Participants were randomly assigned to one of four conditions: retrieval (n = 105), quiz study (n = 95), restudy (n = 99), and control (n = 99). Experiment 3 was approved by the IRB of human participants at SU. Experiment 3 was preregistered. The preregistration of this experiment can be found here: https://osf.io/bcs86/registrations?view_only=.

Based on the sample size rationale from Experiment 1, and the elements of Experiment 3 that rendered a retrieval practice benefit more likely (i.e., extra retrieval practice and a one-week delayed final test), our target sample size was based on the minimum number of participants needed to detect at least a medium-sized effect at 80% power (f = .25; *alpha* = .05), which was 64 participants per condition. This estimate includes the smallest sample size required across all analyses that we conducted, including the omnibus ANOVA and all comparisons between conditions. We oversampled, however, to account for potential attrition rates between part one and part two of the experiment.

4.2.2. Design, materials, and procedure

The design, materials, and procedure of Experiment 3 were identical to Experiment 2, except that the application questions preceded the repeated questions on the final test.

4.3. Results and discussion

4.3.1. Preliminary analyses

Three hundred nineteen participants returned for the second part of the experiment, which amounts to approximately a 20% attrition rate. No differences were found in the attrition rate across conditions, χ^2 = 5.28, p = .150. Following our preregistered exclusion criteria, an additional 64 participants were excluded from the final analyses: 57 who reported looking up the material or learning about it after the training phase but before the final test, one who reported knowing all of the material prior to participating in the experiment, and six for taking more than 3 standard deviations from the group mean to complete either part one or part two of the experiment. The final sample thus consisted of 255 participants: retrieval (n = 71), quiz study (n = 66), restudy (n = 60), and control (n = 58). Cronbach's alpha for the 10 final test questions was .622.

Participants reported having almost no prior knowledge of the material (M = .878, SD = .921; 0 = no prior knowledge, 1 = some prior knowledge, 2 = knew about half the material, 3 = knew most of the material, and 4 = knew all of the material). There were no reported differences in prior knowledge across the conditions, $F(3, 251) = 1.341$, p = .262, MSE = .844, ηp^2 = .016.

However, differences were observed between conditions in how long participants took to complete part one, $F(3, 251) = 32.826$, $p < .001$, MSE = 22.618, ηp^2 = .282. Post-hoc LSD comparisons showed that participants in the retrieval condition took more time (in minutes; M = 19.635, SD = 6.169) than participants in the control condition (M = 11.513, SD = 3.005), $p < .001$, 95% CI [6.464, 9.780], quiz study condition (M = 14.472, SD = 4.357), $p < .001$, 95% CI [3.561, 6.764], and restudy condition (M = 16.166, SD = 4.630), $p < .001$, 95% CI [1.827, 5.112].[4] Unlike in Experiment 2, all participants in Experiment 3 completed part two of the study exactly one week after part one, eliminating the need to analyze potential differences in retention interval time between the conditions.

Finally, we looked at accuracy on the quiz questions during training for the retrieval group. We conducted a within-participants ANOVA, which revealed that performance increased on each round of retrieval practice, $F(2, 140) = 64.708$, $p < .001$, MSE = .033, ηp^2 = .480. Specifically, post-hoc LSD comparisons showed that participants performed better on the second round (M = .670, SE = .035) than on the first round (M = .448, SE = .034), $p < .001$, 95% CI [.159, .286], and better on the third round (M = .789, SE = .032) than on the second round, $p < .001$, 95% CI [.071, .166]. Participants in the retrieval condition thus demonstrated clear evidence of learning during the training phase.

4.3.2. Overall final test performance

Fig. 5 shows mean performance for each condition on each final test question type (also see Table 1). To examine how participants performed on the posttest, we conducted a mixed ANOVA with condition as a between-participants factor (retrieval vs. quiz study vs. restudy vs. control) and question type as a within-participants factor (repeated vs. application). This analysis revealed a main effect of question type, $F(1, 251) = 135.542$, $p < .001$, MSE = .032, ηp^2 = .351, as participants performed better on the repeated questions (M = .740, SE = .015) than on the application questions (M = .555, SE = .016). As in Experiment 2, a main effect of condition was observed, $F(3, 251) = 10.851$, $p < .001$, MSE = .081, ηp^2 = .115. No interaction was observed between condition and question type, $F(3, 251) = .446$, p = .720, MSE = .032, ηp^2 = .005.

Next, to evaluate our a priori hypotheses, we conducted a series of planned comparisons among the training conditions to examine performance differences on the repeated and application questions. For completeness, the full set of pairwise comparisons is provided in Table A5 of the Appendix.

4.3.3. Final test performance between training conditions

Repeated Questions. These analyses revealed that participants in the retrieval condition performed better on the repeated questions than participants in the restudy condition, $t(129) = 3.283$, p = .001, SE = .037, d = .576, 95% CI [.048, .196]). Participants in the retrieval condition also numerically outperformed participants in the quiz study condition, but this difference was not significant, $t(135) = 1.345$, p = .181, SE = .037, d = .230, 95% CI [−.023, .123]). Participants in the quiz study condition nominally outperformed participants in the restudy condition, but this difference was not significant, $t(124) = 1.846$, p = .067, SE = .039, d = .329, 95% CI [−.005, .150].

Application Questions. Participants in the retrieval condition significantly outperformed participants in the restudy condition, $t(129) = 3.558$, $p < .001$, SE = .042, d = .624, 95% CI [.067, .234], and the quiz study condition, $t(135) = 2.032$, p = .044, SE = .041, d = .347, 95% CI [.002, .165]. Participants in the quiz study condition did not perform better than participants in the restudy condition, $t(124) = 1.467$, p = .145, SE = .045, d = .262, 95% CI [−.023, .156].

[4] The same pattern of results was observed for all reported analyses when time on training phase is included as a covariate.

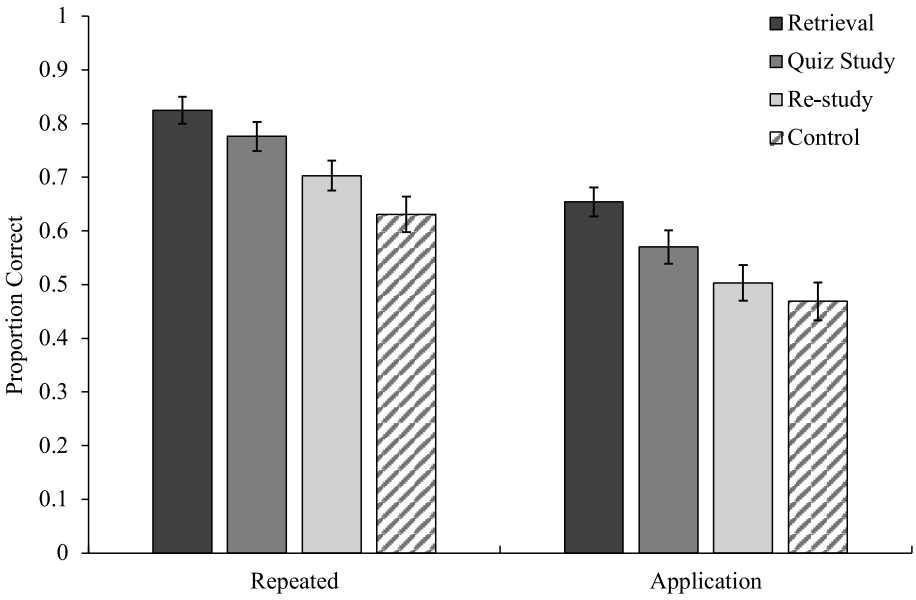


Fig. 5. *Mean performance for each condition and standard errors of the mean on repeated and application questions in experiment 3*

4.3.4. Final test performance between training conditions versus control

Next, to assess whether the training conditions produced learning above and beyond the initial tutorial study phase, we compared the training conditions to the control group. Given our a priori hypothesis that each training condition would result in better learning than the control condition, we conducted planned comparisons of final test performance between the control condition and each of the other conditions. The full set of pairwise comparisons is provided in Table A6 of the Appendix.

Participants in the retrieval condition significantly outperformed control participants on repeated questions, $t(127) = 4.839$, $p < .001$, $SE = .040$, $d = .856$, 95% CI [.115, .274]) and application questions, $t(127) = 4.229$, $p < .001$, $SE = .044$, $d = .749$, 95% CI [.098, .271]). Participants in the quiz study condition significantly outperformed control participants on repeated questions, $t(122) = 3.420$, $p < .001$, $SE = .042$, $d = .615$, 95% CI [.061, .229]) and application questions, $t(122) = 2.156$, $p = .033$, $SE = .047$, $d = .388$, 95% CI [.008, .193]). Lastly, no performance differences were found between the restudy and control conditions on either repeated questions, $t(116) = 1.690$, $p = .094$, $SE = .043$, $d = .311$, 95% CI [−.012, .157], or application questions, $t(116) = .718$, $p = .474$, $SE = .048$, $d = .132$, 95% CI [−.060, .129].

4.3.5. Transfer contingent on retention

We next examined our a priori, preregistered hypothesis that retrieval would show stronger transfer effects than the other conditions when application question performance was conditionalized on correct repeated question performance. We conducted the same conditional analysis as in Experiment 2, in which we examined application question performance on the final test contingent on correct performance on the analog repeated questions over those same concepts. More specifically, this analysis was based on four of the repeated-application question pairs (questions 1–4: https://osf.io/h58bj) that tested the same concept and had the same correct answer.

We conducted an ANOVA with condition as a between-participants factor and conditionalized performance as the dependent measure. The results showed a significant main effect of condition, $F(3, 246) = 3.319$, $p = .021$, $MSE = .116$, $\eta p^2 = .039$. Our planned comparisons showed that participants in the retrieval condition ($M = .773$, $SE = .031$) significantly outperformed participants in the quiz study condition ($M = .665$, $SE = .042$), $t(135) = 2.084$, $p = .039$, $SE = .052$, $d = .362$, 95% CI [.005, .210], restudy condition ($M = .609$, $SE = .047$), $t(129) = 2.963$, $p = .004$, $SE = .055$, $d = .525$, 95% CI [.055, .274], and control condition ($M = .611$, $SE = .053$), $t(127) = 2.729$, $p = .007$, $SE = .059$, $d = .486$, 95% CI [.045, .280]. The full set of pairwise comparisons is provided in Table A7 of the Appendix.

Thus, retrieval leads to the best transfer performance even under conditions where memory accuracy is held constant across all of the conditions. This finding suggests that memory is not the key component underlying retrieval-enhanced transfer. Though memory is certainly involved in transfer, it appears that over and above memory, recognition of the relevance of learned concepts to the application questions is necessary, and retrieval uniquely enhances the ability of learners to see where a learned concept is relevant and apply it correctly in a new situation. We provide further discussion in the General Discussion about how retrieval might enhance this process.

5. General Discussion

The current paper sheds new light on the effects of retrieval practice on the transfer of learning. In three experiments, we found that retrieval was more effective than restudy of the original learning materials on later performance over the same questions when the final test occurred after 8 min (though not significant in Experiment 1) and one week (Experiments 2 and 3). Though one retrieval practice opportunity did not produce significant benefits on new application questions in Experiment 1, three rounds of retrieval practice did produce significant benefits on application questions in Experiments 2 and 3. Though three rounds of retrieval practice produced only nominal benefits of retrieval over quiz study on application questions in Experiment 2, these benefits were larger and statistically significant in Experiment 3 with a larger sample size.

5.1. Mechanisms of retrieval-enhanced learning

Regarding specific hypotheses, these results support the retrieval practice hypothesis by showing that retrieval benefits performance more than restudy and quiz study on both repeated and application questions. The benefits of retrieval on memory retention are in line with the findings reported in many previous studies (for reviews, see Agarwal et al., 2021; Carpenter et al., 2022; McDermott, 2021). The benefit of retrieval over restudy of the original learning materials is consistent with previous studies using educationally-relevant materials that have

shown a similar benefit on a delayed posttest over repeated questions (Ebersbach et al., 2020; McDaniel et al., 2015) and application questions (Hinze & Rapp, 2014; McDaniel et al., 2015).

The current results contribute to our knowledge of retrieval-based learning in new ways by showing that retrieval is also more effective than quiz study for both memory- and transfer-based learning. Specifically, the current study evaluated a plausible but never-before-tested hypothesis (the question familiarity hypothesis) that the benefits of retrieval may be driven by the opportunity to see the questions and answers that would later be tested. The current results do not support the question familiarity hypothesis (Experiment 3), as we found that retrieval was beneficial for both repeated and application over and above the opportunity to study the questions and answers. It is worth noting, however, that the effect sizes associated with the retrieval advantage over quiz study ($d = .360$ for repeated questions, and $d = .375$ for application questions) were smaller than the effect sizes associated with the retrieval advantage over restudy ($d = .667$ for repeated questions, and $d = .558$ for application questions), so it does appear possible that seeing the test questions during training confers a learning benefit, and may be a more effective restudy activity than seeing the entire set of learning materials again.

In all three experiments, retrieval practice was always followed by feedback, raising the possibility that the benefits observed were due to some combination of retrieval itself in addition to the opportunity to review the correct answers. The quiz study condition provided the exact same questions and correct answers as the retrieval practice condition, and thus also allowed for the review of the correct answers, just without the act of retrieval beforehand. Thus, although the benefits of retrieval over restudy are likely due to both retrieval and feedback, the benefits of retrieval over quiz study are more likely due to the act of retrieval as both conditions included the correct answers.

The results of Experiment 1 do not support the transfer-appropriate processing hypothesis that the benefits of retrieval are driven by the match between the initial retrieval conditions and final test conditions. Indeed, other work has also pointed to mechanisms other than transfer-appropriate processing that appear to more consistently drive the benefits of retrieval, such as the elaboration or completeness of the retrieval process during initial learning (e.g., Carpenter, 2009, 2011; Carpenter & DeLosh, 2006). The current results are in line with and extend these same findings to conditions examining the transfer of learning due to retrieval practice.

Across Experiments 1–3, the benefits of retrieval did not depend on whether participants responded to repeated or application questions. Our results thus do not support the retrieval-enhanced retention over transfer hypothesis, whereby the benefits of retrieval would be stronger for retention than for transfer. The present results thus differ from recent work on problem solving, which has shown that the benefits of retrieval (via problem-solving practice) vary as a function of memory-versus transfer-based learning, wherein retrieval led to better memory-based learning than study, but both types of training led to comparable transfer-based learning (Corral et al., 2023; Yeo & Fazio, 2019).

Another novel feature of the current studies is the control condition. In contrast to retrieval and quiz study, we found no evidence that restudying the tutorial materials aids learning, as participants in the restudy condition achieved comparable performance to control participants on both repeated and application questions in all three experiments. These findings raise the possibility that restudying, at least how it is typically implemented in retrieval practice studies, is somewhat ineffective for learning. That the restudy and control conditions achieved similar performance across Experiments 1–3 is somewhat surprising given that participants in the restudy condition had considerably greater exposure to the learning material (approximately 50 % more exposure). One possible explanation for this finding is that participants in the restudy condition did not thoroughly engage with or study the materials in the second round of study. Indeed, these participants might have found the restudy portion to be somewhat boring or unnecessary, given that they had just spent 8 min studying the materials and might have thus thought that they did not need to carefully study them again. These speculations aside, it is difficult to interpret these null findings, because studies on retrieval practice do not typically include a control condition. For this reason, it is not clear to what extent the present findings are atypical.

Nevertheless, this null finding has critical implications for studies on retrieval practice, as retrieval has traditionally been compared to a restudy condition. Given that restudy seems somewhat ineffective, the present results raise the possibility that the purported benefits of retrieval practice in the literature might be overestimated. This point notwithstanding, the present paper shows clear evidence that retrieval practice indeed aids both memory- and transfer-based learning, but this benefit might not be as strong when compared to more comparable learning conditions (e.g., quiz study). Taken together, these issues highlight the critical need for studies on retrieval practice to incorporate more relevant comparison conditions and control conditions into their designs.

5.2. *Mechanisms underlying retrieval-based transfer*

A critical new finding from the current study is the evidence that retrieval practice appears to enhance transfer via improving learners' ability to see the relevance of the learned concepts in a new situation. Our study materials were uniquely designed to evaluate application question performance contingent upon correct retention performance at the individual concept level. This analysis showed that the conditional likelihood of answering an application question correctly, given a correct answer to the analog repeated question over the same concept, was significantly greater in the retrieval condition than in any of the other conditions (Experiment 3). Given the multitude of studies showing benefits of retrieval on memory, we hypothesized that retrieval may facilitate transfer via enhanced memory. We found instead that the conditionalized performance on application questions was greater for retrieval than the other conditions, indicating that retrieval-induced transfer operates not via the memory component, but via the recognition component, whereby learners recognize how a previously learned concept is relevant to a new application question.

How might retrieval enhance this recognition process? Our retrieval task involved short-answer questions where participants had to recall the concepts with no hints or clues. Early on during training, common errors on the retrieval task tended to involve intrusions (e.g., erroneously answering "third variable problem" when the correct answer was "reverse causation"). Performance improved over time with repeated retrieval practice, but the presence of these intrusion errors indicates that some of the concepts may have been easily confused with one another. Part of the learning from these materials, therefore, could involve understanding the difference between concepts like the third variable problem and reverse causation. The retrieval condition is the only condition that involves the chance for participants to expose their errors of confusion and receive feedback to correct these errors and clarify the differences between the concepts. Thus, in addition to facilitating memory for the correct concept itself, retrieval may facilitate understanding of how that concept is different from the other concepts, and this critical contrast is particularly beneficial for answering application questions that require learners to accurately distinguish between concepts that may be easily confused. The other conditions do not provide an opportunity to resolve this confusion during training, so although they may help participants learn the exact right answer to a repeated question, they do not facilitate the compare-and-contrast learning between the concepts that retrieval practice enhances, which is critical to successfully recognizing which concept is relevant to an application question.

The elaborative retrieval hypothesis proposes that information brought to mind during retrieval—even if that information is not the correct answer but is just meaningfully related to it in some way,

including error responses—benefits from retrieval (Carpenter, 2011; Carpenter & Yeung, 2017). The current results are consistent with this account by showing that retrieval benefits memory for the retrieved information (both correct and incorrect responses). Though the elaborative retrieval hypothesis has not been systematically explored in cases of retrieval-based transfer, the current results suggest that it may be a viable theoretical mechanism in cases where retrieval may prompt activation of related information that is relevant to the transfer task. The role of corrective feedback is critical as well, as it provides the opportunity to confront and correct errors that will likely arise when recalling concepts that are easily confused with one another. This idea accords with recent theory on knowledge revision, which holds that when learners are prompted to reconcile incongruencies between their own knowledge and new material, learners restructure and update their knowledge to better accord with that new material (Kendeou, 2024).

Exactly how retrieval might facilitate (or not facilitate) transfer is currently not well-understood, given the paucity of research on retrieval-induced transfer and the lack of clear theoretical guidance about how the act of retrieval might aid the components of transfer other than memory (i.e., recognition and application). The current results thus provide important new data on mechanisms of retrieval-based transfer and show that in particular, the recognition component of transfer appears to be enhanced by retrieval with the current learning materials. Given the novelty of this finding and extant theoretical reasoning, future research would benefit from further exploration of how retrieval practice aids the recognition component of transfer.

5.3. Retention versus transfer

Across Experiments 1–3, we observed an overall benefit for repeated final test questions over application questions. This finding supports the results of previous studies showing general advantages of retention over transfer (Carpenter et al., 2013; Corral et al., 2019, 2023). Indeed, transfer is more challenging because it requires the extra steps of recognizing the relevance of learned information and knowing how to apply it in a new context (Barnett & Ceci, 2002). Depending on the type of transfer task, such a step could be quite difficult, especially if the new context bears little or no resemblance to the original learning context (Gick & Holyoak, 1980, 1983). Whereas it is fairly simple to assess pure memory retention, it is much more complex to assess transfer, as there are a multitude of different ways to set up a test situation that measures knowledge in a different way from how it was learned. Doing so is important, however, as real learning situations often call for the flexible and adaptive use of knowledge.

5.4. Limitations and future directions

The present learning materials were directly based off of lecture slides from a university-level research methods course. With greater complexity of material comes greater difficulty in retrieving the to-be-learned material. Even when feedback is provided, low rates of retrieval success means less engagement with the correct information and could also have less-than-optimal effects on motivation and interest. Thus, complex educational materials will likely require repeated engagement with retrieval practice, along with meaningful feedback that can help students not only check accuracy but also understand the lengthy and sometimes integrated nature of the information.

Although the learning materials used in the current experiments were fairly complex, they were presented in short PowerPoint-like slides, which did not require extensive reading. Moreover, the study phase in the experiments was only 8 min. Though this seems akin to the common practice of reading over PowerPoint slides, it remains an open question as to whether, and to what extent, the present findings extend to materials that require more involved and extensive reading. This question is important because students are often required to engage with text that can be quite extensive.

It is also worth noting that in the current experiments, the final test questions were in multiple-choice format. These multiple-choice questions increased the likelihood that participants would notice that the repeated and application questions tested the same concepts, eliminating the need for them to discover this on their own. Due to the complexity of the materials, multiple-choice questions might also decrease the chances of floor effects, especially after a one-week delay. It is possible, however, that multiple-choice posttest questions could underestimate the strength of the retrieval advantage. One direction for future research, therefore, is to explore whether the benefits of retrieval practice might be stronger when knowledge is assessed through more open-ended questions.

Another factor that might be of interest in future transfer studies is the domain of the application questions. In the current study, the application questions contained scenarios from domains outside of psychology, which were different from the questions practiced during training. Including application questions within the same domain as the practice questions could help to facilitate transfer. Future research might also explore how transfer is affected by application questions from domains that are familiar or of direct interest to learners.

Future research on transfer should also consider the type of questions used during retrieval practice. Long-term learning outcomes related to application may be enhanced by practicing not only factual questions, but also application questions during learning. McDaniel et al. (2015) incorporated both types of questions during initial retrieval practice and found benefits of retrieval (compared to restudy of learning materials) on both repeated and application questions several days later. Butler et al. (2017) showed the same benefits of application questions during retrieval practice (compared to studying statements describing the material) on a two-day delayed final test, while also showing that answering more application questions during initial retrieval more effectively benefited performance on new application questions later. The current paper used only factual questions during initial retrieval and observed significant yet smaller effects (in Experiments 2 and 3) on later application questions.

Although the present data demonstrate some long-term benefits of retrieval practice on the transfer of learning, the interval between when participants learned the material (part one) and when they were tested on it (part two) was only one week (Experiments 2 and 3). However, students must often retain the knowledge they learn for much longer periods (e.g., several months or years). For this reason, future research should investigate whether the present findings extend to longer retention intervals, which more closely align with students' typical learning experiences.

There are various other viable avenues for future research in this area, such as exploring the best ways to provide multiple retrieval opportunities, including the use of spaced retrieval practice (e.g., Lyle et al., 2020) and the idea of providing extra study opportunities or scaffolding prior to engaging in retrieval (Kalyuga, 2007). Another avenue to examine is how to best provide meaningful feedback after retrieval of complex materials (see Corral & Carpenter, 2020, 2024, for benefits of explanation feedback; also see Butler et al., 2013), as well as potential benefits of adaptive approaches that provide retrieval practice and extra study based on individual learners' levels of understanding (Greving et al., 2020).

Finally, it is worth exploring whether retrieval practice is actually the best method for promoting the learning and transfer of complex material. Retrieval practice is often compared to fairly simple procedures like re-reading, however there are many other non-retrieval-based approaches that students might use in their studying (Kuhbandner & Emmerdinger, 2019). A number of active learning approaches other than retrieval—such as generating examples from the learning materials, or from one's own life (Endres et al., 2017; Roelle & Nückles, 2019), summarizing (Ophuis-Cox et al., 2024), and generating one's own questions over the learning material (Ebersbach et al., 2020)—have been explored in lieu of, or in combination with, retrieval practice and

have been shown to be quite effective for learning complex educational materials. Future research can reveal important new insights by looking at retrieval practice as one tool in the box, among many others, and exploring which tools work best in which learning situations.

CRediT authorship contribution statement

Daniel Corral: Writing – review & editing, Writing – original draft, Software, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Shana K. Carpenter:** Writing – review & editing, Writing – original draft, Methodology, Investigation, Formal analysis, Conceptualization.

Acknowledgments

This material is based upon work supported by the James S. McDonnell Foundation 21st Century Science Initiative in Understanding Human Cognition, Collaborative Grant No. 220020483. Any opinions, findings, conclusions, or recommendations expressed in this material are those of the authors and do not necessarily reflect the views of the McDonnell Foundation.

Appendix

Table A1
Pairwise Comparisons Between the Training Conditions in Experiment 1

Experiment 1 Repeated Questions					95% CI		
		t	*p*	*SE*	*Lower*	*Upper*	Cohen's *d*
Retrieval ($n = 64$)	Recognition ($n = 61$)	.711	.478	.038	−.048	.102	.127
Retrieval ($n = 64$)	Restudy ($n = 56$)	1.909	.059	.043	−.003	.166	.349
Retrieval ($n = 64$)	Quiz Study ($n = 67$)	.100	.920	.036	−.075	.067	−.017
Restudy ($n = 56$)	Recognition ($n = 61$)	−1.248	.215	.044	−.142	.032	−.231
Restudy ($n = 56$)	Quiz Study ($n = 67$)	−2.045	.043	.042	−.168	−.003	−.370
Quiz Study ($n = 67$)	Recognition ($n = 61$)	.827	.410	.037	−.042	.103	.146
Experiment 1 Application Questions							
Retrieval ($n = 64$)	Recognition ($n = 61$)	.249	.803	.050	−.086	.111	.045
Retrieval ($n = 64$)	Restudy ($n = 56$)	.851	.397	.049	−.056	.140	.156
Retrieval ($n = 64$)	Quiz Study ($n = 67$)	−1.137	.258	.047	−.145	.039	−.199
Restudy ($n = 56$)	Recognition ($n = 61$)	−.599	.550	.049	−.127	.068	−.111
Restudy ($n = 56$)	Quiz Study ($n = 67$)	−2.054	.042	.046	−.187	−.003	−.372
Quiz Study ($n = 67$)	Recognition ($n = 61$)	1.398	.164	.047	−.027	.158	.247

Table A2
Pairwise Comparisons Between the Control Versus Training Conditions in Experiment 1

Experiment 1 Repeated Questions					95% CI		
		t	*p*	*SE*	*Lower*	*Upper*	Cohen's *d*
Control ($n = 61$)	Retrieval ($n = 64$)	−2.392	.018	.037	−.163	−.015	−.428
Control ($n = 61$)	Restudy ($n = 56$)	−.173	.863	.043	−.093	.078	−.032
Control ($n = 61$)	Quiz Study ($n = 67$)	−2.555	.012	.036	−.165	−.021	−.452
Control ($n = 61$)	Recognition ($n = 61$)	−1.628	.106	.038	−.138	.013	−.295
Experiment 1 Application Questions							
Control ($n = 61$)	Retrieval ($n = 64$)	−.186	.853	.049	−.106	.088	−.033
Control ($n = 61$)	Restudy ($n = 56$)	.675	.501	.049	−.064	.129	.125
Control ($n = 61$)	Quiz Study ($n = 67$)	−1.346	.181	.046	−.154	.029	−.238
Control ($n = 61$)	Recognition ($n = 61$)	.067	.947	.049	−.094	.101	.012

Table A3
Pairwise Comparisons Between the Training Conditions in Experiment 2

Experiment 2 Repeated Questions					95% CI		
		t	*p*	*SE*	*Lower*	*Upper*	Cohen's *d*
Retrieval ($n = 37$)	Restudy ($n = 48$)	3.388	.001	.059	.082	.316	.741
Retrieval ($n = 37$)	Quiz Study ($n = 38$)	2.090	.040	.058	.006	.238	.483
Quiz Study ($n = 38$)	Restudy ($n = 48$)	1.201	.233	.064	−.051	.205	.261
Experiment 2 Application Questions							
Retrieval ($n = 37$)	Restudy ($n = 48$)	2.334	.022	.069	.024	.297	.511
Retrieval ($n = 37$)	Quiz Study ($n = 38$)	1.762	.082	.073	−.017	.274	.407
Quiz Study ($n = 38$)	Restudy ($n = 48$)	.461	.646	.069	−.106	.169	.100

Table A4
Pairwise Comparisons Between the Control Versus Training Conditions in Experiment 2

Experiment 2 Repeated Questions					95% CI		
		t	*p*	*SE*	*Lower*	*Upper*	Cohen's *d*
Control ($n = 35$)	Retrieval ($n = 37$)	−4.077	< .001	.053	−.321	−.110	−.961
Control ($n = 35$)	Restudy ($n = 48$)	−.263	.793	.062	−.139	.106	−.058
Control ($n = 35$)	Quiz Study ($n = 38$)	−1.520	.133	.061	−.216	.029	−.356
Experiment 2 Application Questions							
Control ($n = 35$)	Retrieval ($n = 37$)	−2.213	.030	.070	−.295	−.015	−.522
Control ($n = 35$)	Restudy ($n = 48$)	.073	.942	.067	−.129	.139	.016
Control ($n = 35$)	Quiz Study ($n = 38$)	−.379	.706	.071	−.168	.115	−.089

Table A5
Pairwise Comparisons Between the Training Conditions in Experiment 3

Experiment 3 Repeated Questions					95% CI		
		t	*p*	*SE*	*Lower*	*Upper*	Cohen's *d*
Retrieval ($n = 71$)	Restudy ($n = 60$)	3.283	.001	.037	.048	.196	.576
Retrieval ($n = 71$)	Quiz Study ($n = 66$)	1.345	.181	.037	−.023	.123	.230
Quiz Study ($n = 66$)	Restudy ($n = 60$)	1.846	.067	.039	−.005	.150	.329
Experiment 3 Application Questions							
Retrieval ($n = 71$)	Restudy ($n = 60$)	3.558	< .001	.042	.067	.234	.624
Retrieval ($n = 71$)	Quiz Study ($n = 66$)	2.032	.044	.041	.002	.165	.347
Quiz Study ($n = 66$)	Restudy ($n = 60$)	1.467	.145	.045	−.023	.156	.262

Table A6
Pairwise Comparisons Between the Control Versus Training Conditions in Experiment 3

Experiment 3 Repeated Questions					95% CI		
		t	*p*	*SE*	*Lower*	*Upper*	Cohen's *d*
Control ($n = 58$)	Retrieval ($n = 71$)	−4.839	< .001	.040	−.274	−.115	−.856
Control ($n = 58$)	Restudy ($n = 60$)	−1.690	.094	.043	−.157	.012	−.311
Control ($n = 58$)	Quiz Study ($n = 66$)	−3.420	< .001	.042	−.229	−.061	−.615
Experiment 3 Application Questions							
Control ($n = 58$)	Retrieval ($n = 71$)	−4.229	< .001	.044	−.271	−.098	−.749
Control ($n = 58$)	Restudy ($n = 60$)	−.718	.474	.048	−.129	.060	−.132
Control ($n = 58$)	Quiz Study ($n = 66$)	−2.156	.033	.047	−.193	−.008	−.388

Table A7
Pairwise Comparisons Between Retrieval and the Other Conditions for Application Question Accuracy Conditionalized on Repeated Question Accuracy in Experiment 3

					95% CI		
		t	*p*	*SE*	*Lower*	*Upper*	Cohen's *d*
Retrieval ($n = 69$)	Restudy ($n = 59$)	2.963	.004	.055	.055	.274	.525
Retrieval ($n = 69$)	Quiz Study ($n = 64$)	2.084	.039	.052	.005	.210	.362
Retrieval ($n = 69$)	Control ($n = 58$)	2.729	.007	.059	.045	.280	.486

Data availability

All data are available online on OSF (see link in the manuscript)

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 and Bloom's taxonomy: Do students need fact knowledge before higher order learning? *Journal of Educational Psychology, 111*(2), 189. 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

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), 612–637. https://doi.org/10.1037/0033-2909.128.4.612

Bjork, R. A., & Whitten, W. B. (1974). Recency-sensitive retrieval processes in long-term free recall. *Cognitive Psychology, 6*(2), 173–189. https://doi.org/10.1016/0010-0285(74)90009-7

Butler, A. C. (2010). Repeated testing produces superior transfer of learning relative to repeated studying. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 36*(5), 1118–1133. 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), 433–446. 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), 290–298. https://doi.org/10.1037/a0031026

Carpenter, S. K. (2009). Cue strength as a moderator of the testing effect: The benefits of elaborative retrieval. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 35*(6), 1563–1569. https://doi.org/10.1037/a0017021

Carpenter, S. K. (2011). Semantic information activated during retrieval contributes to later retention: Support for the mediator effectiveness hypothesis of the testing effect. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 37*(6), 1547–1552. https://doi.org/10.1037/a0024140

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., Lohse, K. R., Healy, A. F., Bourne Jr, L. E., & Clegg, B. A. (2013). External focus of attention improves performance in a speeded aiming task. *Journal of Applied Research in Memory & Cognition, 2*(1), 14–19. https://doi.org/10.1016/j.jarmac.2012.11.002

Carpenter, S. K., Pan, S. C., & Butler, A. C. (2022). The science of effective learning with spacing and retrieval practice. *Nature Reviews Psychology, 1*(9), 496–511. https://doi.org/10.1038/s44159-022-00089-1

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/10.1038/s44159-022-00089-1

Carpenter, S. K., Pashler, H., Wixted, J. T., & Vul, E. (2008). The effects of tests on learning and forgetting. *Memory & Cognition, 36*, 438–448. https://doi.org/10.3758/MC.36.2.438

Carpenter, S. K., Rahman, S., & Perkins, K. (2018). The effects of prequestions on classroom learning. *Journal of Experimental Psychology: Applied, 24*(1), 34–42. https://doi.org/10.1037/xap0000145

Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory and Language, 92*, 128–141. https://doi.org/10.1016/j.jml.2016.06.008

Corral, D., & Carpenter, S. K. (2020). Facilitating transfer through incorrect examples and explanatory feedback. *Quarterly Journal of Experimental Psychology, 73*(9), 1340–1359. https://doi.org/10.1002/acp.3618

Corral, D., & Carpenter, S. K. (2024). Acquiring complex concepts through testing and explanatory feedback. *Cognitive Research: Principles and Implications, 9*, 81. https://doi.org/10.1186/s41235-024-00608-z

Corral, D., Carpenter, S. K., Perkins, K., & Gentile, D. A. (2020). Assessing students' use of optional online lecture reviews. *Applied Cognitive Psychology, 34*(2), 318–329. https://doi.org/10.1002/acp.3618

Corral, D., Carpenter, S. K., & St Hilaire, K. J. (2023). The effects of retrieval versus study on analogical problem solving. *Psychonomic Bulletin & Review, 30*(5), 1954–1965. https://doi.org/10.3758/s13423-023-02268-4

Corral, D., Healy, A. F., & Jones, M. (2022). The effects of testing the relationships among relational concepts. *Cognitive Research: Principles & Implications, 7*, 47. https://doi.org/10.1186/s41235-022-00398-2

Corral, D., Healy, A. F., Rozbruch, E. V., & Jones, M. (2019). Building a testing-based training paradigm from cognitive psychology principles. *Scholarship of Teaching & Learning in Psychology, 5*(3), 189–208. https://doi.org/10.1037/stl0000146

Corral, D., & Kurtz, K. J. (2025). *Improving the transfer of learning in education through category learning* [Manuscript submitted for publication]. Department of Psychology, Syracuse University.

da Silva, F. V., Ekuni, R., & Jaeger, A. (2023). Retrieval practice benefits for spelling performance in fifth-grade children. *Memory, 31*(9), 1197–1204. https://doi.org/10.1080/09658211.2023.2248420

Dobson, J. L., Linderholm, T., & Stroud, L. (2019). Retrieval practice and judgements of learning enhance transfer of physiology information. *Advances in Health Sciences Education, 24*, 525–537. https://doi.org/10.1007/s10459-019-09881-w

Ebersbach, M., Feierabend, M., & Nazari, K. B. B. (2020). Comparing the effects of generating questions, testing, and restudying on students' long-term recall in university learning. *Applied Cognitive Psychology, 34*(3), 724–736. https://doi.org/10.1002/acp.3639

Endres, T., Carpenter, S., Martin, A., & Renkl, A. (2017). Enhancing learning by retrieval: Enriching free recall with elaborative prompting. *Learning and Instruction, 49*, 13–20. https://doi.org/10.1016/j.learninstruc.2016.11.010

Endres, T., Kubik, V., Koslowski, K., Hahne, F., & Renkl, A. (2023). Immediate learning benefits of retrieval tasks. *Zeitschrift für Entwicklungspsychologie und Pädagogische Psychologie, 55*(2–3), 49–66. https://doi.org/10.1026/0049-8637/a000280

Eriksson, J., Kalpouzos, G., & Nyberg, L. (2011). Rewiring the brain with repeated retrieval: A parametric fMRI study of the testing effect. *Neuroscience Letters, 505*(1), 36–40. https://doi.org/10.1016/j.neulet.2011.08.061

Geller, J., Toftness, A. R., Armstrong, P. I., Carpenter, S. K., Manz, C. L., Coffman, C. R., & Lamm, M. H. (2018). Study strategies and beliefs about learning as a function of academic achievement and achievement goals. *Memory, 26*(5), 683–690. https://doi.org/10.1080/09658211.2017.1397175

Gick, M. L., & Holyoak, K. J. (1980). Analogical problem solving. *Cognitive Psychology, 12*(3), 306–355. https://doi.org/10.1016/0010-0285(80)90013-4

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

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

Glover, J. A. (1989). The" testing" phenomenon: Not gone but nearly forgotten. *Journal of Educational Psychology, 81*(3), 392–399. https://doi.org/10.1037/0022-0663.81.3.392

Greving, S., Lenhard, W., & Richter, T. (2020). Adaptive retrieval practice with multiple-choice questions in the university classroom. *Journal of Computer Assisted Learning, 36*(6), 799–809. https://doi.org/10.1111/jcal.12445

Hinze, S. R., & Rapp, D. N. (2014). Retrieval (sometimes) enhances learning: Performance pressure reduces the benefits of retrieval practice. *Applied Cognitive Psychology, 28*(4), 597–606. https://doi.org/10.1002/acp.3032

Hui, L., de Bruin, A. B., Donkers, J., & van Merriënboer, J. J. (2021). Does individual performance feedback increase the use of retrieval practice? *Educational Psychology Review, 33*, 1835–1857. https://doi.org/10.1007/s10648-021-09604-x

Jones, A. C., Wardlow, L., Pan, S. C., Zepeda, C., Heyman, G. D., Dunlosky, J., & Rickard, T. C. (2016). Beyond the rainbow: Retrieval practice leads to better spelling than does rainbow writing. *Educational Psychology Review, 28*, 385–400. https://doi.org/10.1007/s10648-015-9330-6

Kalyuga, S. (2007). Expertise reversal effect and its implications for learner-tailored instruction. *Educational Psychology Review, 19*, 509–539. https://doi.org/10.1007/s10648-007-9054-3

Kang, S. H., Gollan, T. H., & Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin & Review, 20*, 1259–1265. https://doi.org/10.3758/s13423-013-0450-z

Kang, S. H., Lindsey, R. V., Mozer, M. C., & Pashler, H. (2014). Retrieval practice over the long term: Should spacing be expanding or equal-interval? *Psychonomic Bulletin & Review, 21*, 1544–1550. https://doi.org/10.3758/s13423-014-0636-z

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

Kintsch, W. (1970). Models for free recall and recognition. In D. A. Normal (Ed.), *Models of human memory* (pp. 331–373). New York, NY: Academic Press.

Kornell, N., & Bjork, R. A. (2008). Optimising self-regulated study: The benefits—and costs—of dropping flashcards. *Memory, 16*(2), 125–136. https://doi.org/10.1080/09658210701763899

Kornell, N., Klein, P. J., & Rawson, K. A. (2015). Retrieval attempts enhance learning, but retrieval success (versus failure) does not matter. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 41*(1), 283–294. https://doi.org/10.1037/a0037850

Kuhbandner, C., & Emmerdinger, K. J. (2019). Do students really prefer repeated rereading over testing when studying textbooks? A re-examination. *Memory, 27*(7), 952–961. https://doi.org/10.1080/09658211.2019.1610177

Lyle, K. B., Bego, C. R., Hopkins, R. F., Hieb, J. L., & Ralston, P. A. (2020). How the amount and spacing of retrieval practice affect the short-and long-term retention of mathematics knowledge. *Educational Psychology Review, 32*, 277–295. https://doi.org/10.1007/s10648-019-09489-x

Mayer, R. (1998). Cognitive, metacognitive, and motivational aspects of problem solving. *Instructional Science, 26*, 49–63. https://doi.org/10.1023/A:1003088013286

Mayer, R. E., Sims, V., & Tajika, H. (1995). A comparison of how textbooks teach mathematical problem solving in Japan and the United States. *American Educational Research Journal, 32*(2), 443–460. https://doi.org/10.2307/1163438

McAndrew, M., Morrow, C. S., Atiyeh, L., & Pierre, G. C. (2016). Dental student study strategies: Are self-testing and scheduling related to academic performance? *Journal of Dental Education, 80*(5), 542–552. https://doi.org/10.1002/j.0022-0337.2016.80.5.tb06114.x

McDaniel, M. A., Agarwal, P. K., Huelser, B. J., McDermott, K. B., & Roediger, I. I. I. H. L. (2011). Test-enhanced learning in a middle school science classroom: The effects of quiz frequency and placement. *Journal of Educational Psychology, 103*(2), 399–414. https://doi.org/10.1037/a0021782

McDaniel, M. A., Bugg, J. M., Liu, Y., & Brick, J. (2015). When does the test-study-test sequence optimize learning and retention? *Journal of Experimental Psychology: Applied, 21*(4), 370–382. https://doi.org/10.1037/xap0000063

McDaniel, M. A., & Fisher, R. P. (1991). Tests and test feedback as learning sources. *Contemporary Educational Psychology, 16*(2), 192–201. https://doi.org/10.1016/0361-476X(91)90037-L

McDaniel, M. A., Thomas, R. C., Agarwal, P. K., McDermott, K. B., & Roediger, H. L. (2013). Quizzing in middle-school science: Successful transfer performance on classroom exams. *Applied Cognitive Psychology, 27*(3), 360–372. https://doi.org/10.1002/acp.2914

McDaniel, M. A., Wildman, K. M., & Anderson, J. L. (2012). Using quizzes to enhance summative-assessment performance in a web-based class: An experimental study. *Journal of Applied Research in Memory & Cognition, 1*(1), 18–26. https://doi.org/10.1016/j.jarmac.2011.10.001

McDermott, K. B. (2021). Practicing retrieval facilitates learning. *Annual Review of Psychology, 72*(1), 609–633. https://doi.org/10.1146/annurev-psych-010419-051019

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 519–534. https://doi.org/10.1016/S0022-5371(77)80016-9

Ophuis-Cox, F. H. A., Rozendal, L., Catrysse, L., Joosten-ten Brinke, D., & Camp, G. (2024). The effects of summarization and factual retrieval practice on text comprehension and text retention in elementary education. *Journal of Experimental Psychology: Applied, 30*(2), 258–267. https://doi.org/10.1037/xap0000507

Pan, S. C., & Rickard, T. C. (2017). Does retrieval practice enhance learning and transfer relative to restudy for term-definition facts? *Journal of Experimental Psychology: Applied, 23*(3), 278–292. https://doi.org/10.1037/xap0000124

Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin, 144*(7), 710–756. https://doi.org/10.1037/bul0000151

Peterson, D. J., & Wissman, K. T. (2018). The testing effect and analogical problem-solving. *Memory, 26*(10), 1460–1466. https://doi.org/10.1080/09658211.2018.1491603

Princeton Review. (2017). *Cracking the GRE Premium edition with 6 practice Tests, 2018*. New York, NY: Penguin Random House.

Pyc, M. A., & Rawson, K. A. (2009). Testing the retrieval effort hypothesis: Does greater difficulty correctly recalling information lead to higher levels of memory? *Journal of Memory and Language, 60*(4), 437–447. https://doi.org/10.1016/j.jml.2009.01.004

Rickard, T. C., & Pan, S. C. (2018). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review, 25*, 847–869. https://doi.org/10.3758/s13423-017-1298-4

Roediger, H. L., III, Agarwal, P. K., McDaniel, M. A., & McDermott, K. B. (2011). Test-enhanced learning in the classroom: Long-term improvements from quizzing. *Journal of Experimental Psychology: Applied, 17*(4), 382–395. https://doi.org/10.1037/a0026252

Roediger, H. L., III, Putnam, A. L., & Smith, M. A. (2011). Ten benefits of testing and their application to educational practice. In J. P. Mestre, & B. H. Ross (Eds.), *Psychology of learning and motivation* (pp. 1–36). Academic Press. https://doi.org/10.1016/B978-0-12-387691-1.00001-6.

Roelle, J., & Nückles, M. (2019). Generative learning versus retrieval practice in learning from text: The cohesion and elaboration of the text matters. *Journal of Educational Psychology, 111*(8), 1341–1361. https://doi.org/10.1037/edu0000345

Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin, 140*(6), 1432–1463. https://doi.org/10.1037/a0037559

Sensenig, A. E., Littrell-Baez, M. K., & DeLosh, E. L. (2011). Testing effects for common versus proper names. *Memory, 19*(6), 664–673. https://doi.org/10.1080/09658211.2011.599935

Thomas, R. C., Weywadt, C. R., Anderson, J. L., Martinez-Papponi, B., & McDaniel, M. A. (2018). Testing encourages transfer between factual and application questions in an online learning environment. *Journal of Applied Research in Memory & Cognition, 7*(2), 252–260. https://doi.org/10.1016/j.jarmac.2018.03.007

Tran, R., Rohrer, D., & Pashler, H. (2015). Retrieval practice: The lack of transfer to deductive inferences. *Psychonomic Bulletin & Review, 22*, 135–140. https://doi.org/10.3758/s13423-014-0646-x

Vaughn, K. E., & Rawson, K. A. (2011). Diagnosing criterion-level effects on memory: What aspects of memory are enhanced by repeated retrieval? *Psychological Science, 22*(9), 1127–1131. https://doi.org/10.1177/0956797611417724

Whitehead, A. N. (1929). *The aims of education and other essays*. New York, NY: Macmillan.

Wissman, K. T., Zamary, A., & Rawson, K. A. (2018). When does practice testing promote transfer on deductive reasoning tasks? *Journal of Applied Research in Memory & Cognition, 7*(3), 398–411. https://doi.org/10.1016/j.jarmac.2018.03.002

Yan, V. X., Thai, K. P., & Bjork, R. A. (2014). Habits and beliefs that guide self-regulated learning: Do they vary with mindset? *Journal of Applied Research in Memory & Cognition, 3*(3), 140–152. https://doi.org/10.1037/h0101799

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

Yeo, D. J., & Fazio, L. K. (2019). The optimal learning strategy depends on learning goals and processes: Retrieval practice versus worked examples. *Journal of Educational Psychology, 111*(1), 73. https://doi.org/10.1037/edu0000268