

Building a Testing-Based Training Paradigm from Cognitive Psychology Principles

Daniel Corral, Alice F. Healy, Erica V. Rozbruch, & Matt Jones

University Colorado Boulder

Author Note

Daniel Corral, Alice F. Healy, Erica V. Rozbruch, and Matt Jones, Department of Psychology and Neuroscience, University of Colorado Boulder.

This research was supported by NSF Grant DRL1246588.

Correspondence about this article should be directed to Daniel Corral, Department of Psychology and Neuroscience, University of Colorado Boulder, 345 UCB, Boulder, CO, 80309-0345. Email: Daniel.Corral@Colorado.edu

Abstract

Cognitive psychology often produces findings that are relevant to educational instruction. However, many of these studies rely on artificial conditions, which often fail to transfer to realistic settings, resulting in a disconnection between cognitive psychology and education. This paper begins to address this issue by taking established principles from cognitive psychology and applying them to teach participants real academic concepts. We report a training paradigm that applies established principles from cognitive psychology: retrieval practice, feedback, self-paced studying, cognitive antidote, and levels of processing, in which participants are shown the correct response. This paradigm was used to teach undergraduates basic concepts of research design that are typically taught in university science courses. Participants studied PowerPoint-style slides that were divided into three sections. At the end of each section, participants were presented quiz questions. After each quiz response, the participant was shown the correct answer. This study also tested different forms of responding to quiz questions (between subjects): (a) fill-in-the-blank, (b) multiple-choice, and (c) fill-in-the-blank followed by a multiple-choice version of the same question. Participants completed two posttests, one immediately after training and another one week later. Both posttests consisted of items that tested retention and conceptual understanding. A control condition (wherein participants received no training) was used to assess the effectiveness of the training paradigm. Participants who used this paradigm outperformed control participants on both posttests. However, no differences in performance were found among participants who used different forms of responding.

Keywords: Retrieval Practice, Complex Concept Acquisition, Technology-Based Learning and Instruction, Translational Research

Teaching Scientific Principles using a Testing-Based Training Paradigm

One of the primary challenges in education is finding effective methods that increase students' retention and comprehension of course material. Factors that facilitate learning are thus of great interest to instructors. Over the past 70 years, many findings from cognitive psychology have shed light on this goal. This work has led to the discovery of various learning principles (e.g., *correct-answer feedback*: Benassi, Overson, & Hakala, 2014; *self-paced study*: Ariel, 2013; *cognitive antidote*: Healy, Jones, Lalchandani, & Tack, 2017; *levels of processing*: Craik & Lockhart, 1972). One of the most robust findings from cognitive psychology is that retrieving information from memory improves the retention of the information that was retrieved (formally known as *retrieval practice*; Carrier & Pashler, 1992; Kang, Gollan, & Pashler, 2013; Kang & Pashler, 2014; Karpicke & Roediger, 2008; Pan & Rickard, 2018; Pyc & Rawson, 2010; Roediger & Butler, 2011; Roediger & Karpicke, 2006a, 2006b). Specifically, work on the *testing effect* has demonstrated that testing learners on previously studied material (i.e., retrieval practice) often leads to better learning and retention than having them restudy that material (Butler, Black-Maier, Raley, & Marsh, 2017; Carpenter & Yeung, 2017; Eglington & Kang, 2018; Lehman & Karpicke, 2016; Pan & Rickard, 2018; Rickard & Pan, 2018). Retrieval practice has also been shown to aid learning in the classroom, as students who engage in retrieval practice, either through in-class clicker questions (Anderson, Healy, Kole, & Bourne, 2011, 2013; Mayer et al., 2009) or online practice quizzes (Carpenter et al., 2017; Corral, Carpenter, Perkins, & Gentile, 2019), often demonstrate better learning and retention than students who do not engage in these tasks.

On the other hand, many findings from cognitive psychology that appear to be relevant to education are often not translated to the classroom (Horvath, Lodge, & Hattie, 2017; Roediger, 2013). One reason for this lack of cross-fertilization may be that cognitive psychology studies

often use artificial learning tasks and materials (e.g., participants are asked to learn to distinguish among simple geometric figures; e.g., Corral, 2017; Corral & Jones, 2014, 2017; Corral, Kurtz, & Jones, 2018) that are not representative of the concepts that are taught in the classroom (e.g., a physics professor teaching the concept of buoyancy). The use of artificial conditions and simplified stimuli and concepts is fairly common in cognitive psychology and may lead instructors to view findings from such studies with skepticism, as it may seem unlikely that a given effect will hold under more ecologically valid conditions (Horvath et al. 2017; Oliver & Conole, 2003; Smeyers & Depaepe, 2013).

Adapting laboratory studies to real-world settings is a common issue in *translational science*—the application of laboratory findings to real-world settings—as researchers often struggle to apply findings from basic and theoretical research to real-world scenarios (Horvath et al., 2017; Oliver & Conole, 2003; Smeyers & Depaepe, 2013; Roediger, 2013; Woolf, 2008). One potential issue is that cognitive psychology studies typically use rigorous methodology to isolate the variable(s) of interest. Although this approach is appropriate for controlled scientific studies, it might not be conducive to translation in the classroom, which often involves many additional facets beyond what is required in a laboratory experiment (Hovrath et al., 2017; Oliver & Conole, 2003; Smeyers & Depaepe, 2013).

For example, although retrieval practice might aid learning and retention (Carrier & Pashler, 1992; Kang et al., 2013; Kang & Pashler, 2014; Pan & Rickard, 2018), an instructor might not know how to implement this principle in the classroom, as translation requires the instructor to make various decisions about how to implement numerous facets of retrieval practice. In particular, an instructor must decide what type of retrieval practice to provide students (e.g., recall vs. recognition), when to present retrieval practice during a lecture (e.g., beginning of lecture vs.

interspersed throughout lecture vs. end of lecture), as well as the type of feedback students should be presented after retrieval practice (e.g., no feedback vs. correct-answer feedback). As this example illustrates, each of these components offers the instructor an opportunity to translate different learning principles to the classroom, but this flexibility can produce uncertainty about when and how to apply these principles, and might thus make translation rather difficult.

Given these challenges, one way forward might be to develop a training paradigm that fully specifies each of its facets. The efficacy of this paradigm could then be tested in the laboratory with ecologically valid learning materials. With this aim in mind, the current paper takes well-established learning principles from cognitive psychology and integrates them with current instructional practices that are used in the classroom to develop a training paradigm that can be easily implemented by educators to supplement instruction. We therefore build a training paradigm around retrieval practice, one of the most reliable principles in cognitive psychology (Roediger, 2013), and specify and include additional learning principles for each of its facets. Specifically, this training paradigm incorporates the following four learning principles: (a) retrieval practice, (b) correct-answer feedback, (c) self-paced study, and (d) cognitive antidote. Correct-answer feedback involves showing participants the correct answer after they respond, and self-paced study allows them to control the time they spend studying. Cognitive antidote includes the idea that boredom or disengagement can be offset by alternating the tasks that learners complete, wherein two or more tasks are interspersed (as opposed to completing one task in its entirety and then the other).

These principles were selected for two reasons. The first reason is that each of these principles has been shown to aid learning and retention across numerous studies (e.g., *retrieval practice*: Brame & Biel, 2015; Carpenter, Pashler, & Cepeda, 2009; Carpenter & Yeung, 2017;

Dunlosky, Rawson, Marsh, Nathan, & Willingham, 2013; Roediger & Butler, 2011; Rowland, 2014; *correct-answer feedback*: Benassi et al., 2014; Butler, Karpicke, & Roediger, 2007; McDaniel, Anderson, Derbish, & Morrisette, 2007; Pashler, Cepeda, Wixted, & Rohrer, 2005; Vojdanoska, Cranney, & Newell, 2010; *self-paced study*: Ariel, 2013; de Jonge, Tabbers, Pecher, Jang, & Zeelenberg, 2015; Tullis & Benjamin, 2011; and *cognitive antidote*: Chapman, Healy, & Kole, 2016; Healy et al., 2017; Kole, Healy, & Bourne, 2008). The second reason is that by applying one of these principles for each facet that is involved in translating retrieval practice to a real-world paradigm we are able to fully specify this process (as discussed in detail in the Experiment and Method sections), which might greatly aid instructors in using or adapting this training paradigm.

It is important to note that none of these principles were manipulated between experimental groups, as our primary goal was to examine the efficacy of the paradigm as whole as compared to a control group that benefited from none of them. This approach highlights an important distinction between laboratory studies and translational research. Laboratory studies take a reductionist approach, as their typical goal is to isolate underlying mechanisms for a given phenomenon. In contrast, the goal of translational research is to produce a working, integrated system. As the example on translating retrieval practice demonstrates, translation is a complex process that involves multiple facets (Horvath et al., 2017; Oliver & Conole, 2003; Smeyers & Depaepe, 2013; Roediger, 2013; Woolf, 2008). Translation of a given learning principle is therefore likely to involve a multifaceted, complex training paradigm. Moreover, it is not clear that when a learning principle is translated and embedded within a larger system that any given component will aid learning, as it is possible that the different parts of the paradigm do not work well together and might offset or counteract the benefits of any single component. For these reasons, it is essential

for translational research to take a more holistic approach and examine whether a training paradigm as a whole improves learning.

However, as a secondary question, we examine whether manipulating mode of responding (i.e., type of retrieval practice), which is integral to implementing retrieval practice, produces differential learning and retention across experimental groups. One possibility is that how participants respond to a given question affects the extent to which they encode its information. Thus, a fifth principle we incorporate into our training paradigm (by manipulating form of responding) is *levels of processing*—the extent to which connections are formed between the information that is encoded and long-term memory (LTM; Craik & Lockhart, 1972; Craik & Tulving, 1975).

Recognition Versus Recall

When an instructor translates retrieval practice to a real-world setting, he or she must decide what type of retrieval practice to use. Many studies on retrieval practice (e.g., Butler et al., 2017; Carpenter, 2009; Carpenter & Yeung, 2017) involve recall, wherein participants must generate a response from memory. However, other forms of responding are possible, such as selecting an answer from a list of multiple-choice options, a process that often relies on recognition memory (Jacoby, 1991).

Recognition and recall are two distinct memory processes by which people access information from LTM (Kintsch, 1970). In a recognition task, participants are presented with a given item and are asked to determine whether or not it matches information that was previously encountered, as is often the case for multiple-choice questions. When students are presented with a multiple-choice question, they must select the correct response from a list of options. The option that is selected is often determined by whether the student recognizes the given option as correct

or finds it more familiar than other options (Marsh, Roediger, Bjork, & Bjork, 2007; see also Bjork, Little, & Storm, 2014; Little & Bjork, 2015). Likewise, in studies on recognition memory, participants are typically presented with a list of items and are asked to memorize them within an allotted period of time. At testing, participants are presented with a series of items and are asked to select which of those items were on the studied list. Similar to multiple-choice responding, the items that are selected in a recognition memory task are those that participants recognize or find most familiar (Kahana, 2012).

In contrast to this process, recall involves the retrieval of information from LTM, as is often the case for fill-in-the-blank questions. Fill-in-the-blank questions require students to provide a short response by recalling information from LTM. The process of generating a response is often more challenging than recognizing a previously encountered item (Anderson & Bower, 1972; Kintsch, 1970). As a result of this generation, recall has been posited to lead to *deeper encoding*—a greater number of connections are formed between the information that was probed in LTM and the information that was recalled—than does recognition, which can consequently improve retention (Hogan & Kintsch, 1971).

The differential effects of recall and recognition responding have been well documented in many memory experiments, but the manner in which they affect learning and transfer of knowledge is less clear. One prediction that follows directly from the experimental psychology literature is that, because questions that require recall processes (i.e., fill in the blank) lead to deeper encoding (Hogan & Kintsch, 1971; Kintsch, 1970), recall will produce superior learning. Another prediction is that engaging in both of the retrieval processes (recall followed by recognition) will produce cumulative effects, thus leading to better learning and retention than either recall or recognition alone.

We test these predictions by examining whether there are learning differences (measured through performance scores at testing) between participants who are trained by engaging in retrieval practice that invokes recognition versus recall. We also examine whether engaging in both recall and recognition, wherein participants first attempt recall followed by recognition, can aid learning and retention above and beyond engaging in only one of these retrieval processes.

The comparison among these three experimental groups (recall, recognition, and recall then recognition) complements the main question of this study, which is a comparison of all three of these groups to the control group. For this latter comparison, participants in the experimental groups were predicted to demonstrate better learning and retention than participants in the control group.

Experiment and Training Paradigm

We conducted a study to examine whether our training paradigm can be used to aid students in learning core scientific concepts that are typically taught in university-level statistics and research methods courses. These materials were chosen due to their direct relevance to all scientific fields (because these fields rely on sound research methodology), and thus wide applicability to education and instruction.

Three groups of undergraduate students, referred to as the experimental groups, were trained under our paradigm. These groups varied only in the type of retrieval practice participants were given (recall vs. recognition vs. recall-then-recognition). Participants were first asked to study PowerPoint-style slides (included in the supplementary materials) that were divided into three sections. Although the range of times participants were required or allowed to spend on each section was determined before the study (explained further in the Procedure), within these time restrictions participants could choose how long they studied each slide within a given section (a

form of self-paced studying). At the end of each section, participants were quizzed on the material for that section (thus we implement the cognitive antidote principle by alternating between study and retrieval practice) and after each response were shown the correct answer (correct-answer feedback). These participants completed two posttests, one immediately after training and another one week later. It is important to note that the first posttest affords participants in the experimental conditions additional retrieval practice, which might further benefit learning (Butler et al., 2017; Carpenter & Yeung, 2017; Eglington & Kang, 2018; Lehman & Karpicke, 2016; Pan & Rickard, 2018). For this reason, the first posttest can be viewed as another facet of the training paradigm.

Participants in a separate, control condition did not receive any training (i.e., were not shown any study materials or presented with any quiz questions) and were only asked to complete a single test, which was identical to the second posttest that participants who received the training paradigm completed. Posttest performance was compared between the control group and the trained groups.

The control condition was used to assess whether participants in the experimental conditions were indeed able to learn the concepts that were trained, as this condition provides a baseline measure of participants' knowledge of the material. Although extensive work has shown that retrieval practice can indeed aid retention (e.g., Carpenter et al., 2009; Carrier & Pashler, 1992; Dunlosky et al., 2013; Kang et al., 2013; Karpicke & Roediger, 2008; Pyc & Rawson, 2010; Roediger & Butler, 2011; Roediger & Karpicke, 2006a, 2006b), most of this literature is limited to direct memorization and does not typically involve true concept learning. Moreover, the limited work that has been conducted on this topic has yielded inconclusive results, as some of this work has shown a modest benefit of retrieval practice and testing on concept learning and transfer (Butler, 2010; Butler et al., 2017; Eglington & Kang, 2018; Pan & Rickard, 2018), but other studies

have failed to replicate this finding (Peterson & Wissman, 2018; Tran, Rohrer, & Pashler, 2015; van Gog & Kester, 2012; Wissman, Zamar, & Rawson, 2017). It is therefore an empirical question as to whether these principles can be used to help people learn ecologically valid, complex concepts.

The three trained groups were defined according to the format in which they were quizzed during training: recall, recognition, and recall-then-recognition. Participants in the recognition condition were provided with multiple-choice quiz questions, and participants in the recall condition were provided with fill-in-the-blank quiz questions. Participants in the recall-then-recognition condition responded to each quiz question twice, first with a fill-in-the-blank response and then with a multiple-choice response (multiple-choice options were shown only after the first response was given). This ordering was necessary to keep the multiple-choice options from contaminating the recall process for a given question, as the multiple-choice options might serve as memory cues for the correct response, and thereby trivialize the recall process.

Method

Participants. One hundred eighty-three undergraduate students participated for course credit in an introductory psychology course at the University of Colorado Boulder. This population consists primarily of freshmen and contains approximately 45% women and 71% White students, with an average age of 20 (5% of the students are 25 years of age or older); 17% of this population is classified as low-income students. One hundred fifty-four of these participants were randomly assigned to three experimental conditions (between subjects): recall only ($n = 51$), recognition only ($n = 51$), and recall-then-recognition ($n = 52$). The other participants ($n = 29$) were sampled concurrently from the same population and were assigned to the control condition. True random assignment was not possible because the online system participants use to sign up for studies

requires that one-part and two-part studies (such as our control and experimental conditions, respectively) be posted as separate sign-up options. However, this system randomizes the order of listed studies and provides prospective participants with no information other than time and location, which allows for a degree of random assignment. Thus there is a mild self-selection issue, because participants who chose to sign up for one- and two-part studies might differ from one another (although all students were subject to the same class requirement of six hours' total research participation that semester). However, we stress that the number of sessions participants signed up for was the only difference in sampling procedure between the experimental and control conditions.

Design and materials. All materials (instructions, study slides, and quiz and posttest questions) were presented on a computer monitor and were shown on a black background. All responses were entered using a computer keyboard. The training session consisted of PowerPoint-style slides that were modified from an undergraduate statistics lecture, which covered basic principles of research methods. These slides were adapted to exclude extraneous information, and each slide was carefully checked by the authors to ensure that each concept was fully explained. These slides covered 16 concepts, which were divided into three sections, and each section followed a conceptual progression (see Table 1). Section 1 (Slides 1-2) introduced the basic components related to scientific experiments. Section 2 (Slides 3-6) introduced issues related to causal inference and non-experimental studies. Section 3 (Slides 7-10) introduced methods that true experiments use to control for confounding variables and other related topics. Figure 1 shows an example study slide, Slide 9 from Section 3 in the training session.

Question types. Five question types were created for this study: (1) repeated, (2) definitional, (3) transfer, (4) analysis, and (5) application; all test items for each question type are

provided in the supplementary materials. These question types were divided into two subsets. We refer to question types 1-3 as *core questions* and question types 4-5 as *conceptual questions*. The immediate posttest comprised eight questions from each of types 1-3. The retention test comprised eight different questions from each of types 1-3, and 14 questions from each of types 5-6.

Core questions. To increase the chance that condition differences would be detected, the core questions were pilot tested with two different groups to ensure that no ceiling or floor effects were present; these participants were not trained on these materials. The first round of pilot testing was conducted with paid subjects from the university's paid subject pool and the second with undergraduate students (within the first few weeks of the semester) in an upper division psychology course on research methods. Given the course content, the participants in this latter group should have some background with these materials, and thus likely represent a more knowledgeable sample than the introductory psychology students who participated in the main experiment. The initial version of the core questions consisted of four multiple-choice options per question, but these materials proved too easy for students and were thus modified to be more challenging; one of these modifications was to switch from four multiple-choice options to five. This iterative process of pilot testing and revising these materials was concluded once an intermediate level of performance was found (between 50-60%).

The core question types tested the basic concepts that participants encountered during training. Repeated questions were identical in content to the recognition version of the quiz questions that were used during the training session (see Figure 4). Definitional questions were the inverse of repeated questions: Participants were shown a term and were asked to select the correct definition from the multiple-choice options, as shown in Figure 5. Transfer questions were similar

to repeated questions, but the description of the tested term was grounded in a hypothetical scenario (as shown in Figure 6).

Each of the core questions thus provides a different measure of retention. Repeated questions provide a direct measure of retention, as these questions can be answered correctly through rote memorization of the content that was trained and quizzed. Definitional questions measure whether participants can transfer their memory of the training material to the inverse of the concepts that were quizzed (i.e., matching a given term to the correct definition instead of matching a given definition to the correct term). In contrast, transfer questions provide a more robust measure of concept learning and transfer than definitional questions, in that they require participants to recognize the instantiation of a given concept in a superficially different scenario than what was encountered in training. Moreover, the transfer questions did not explicitly define the corresponding concept (as the repeated and definitional questions did), and thus recognizing these concepts required that the participant actually comprehends their meaning. Transfer questions therefore provide a measure of both retention and concept learning, as these questions require participants to remember a concept's definition and comprehend its meaning.

Sixteen items were constructed for each core question type (i.e., repeated, definitional, and transfer), one covering each of the 16 concepts that were introduced during training (as discussed in the first paragraph of the Design and materials). Thus, there was a one-to-one correspondence among the three core question types in terms of the concepts they tested. For purposes of explaining the experimental design, we refer to the questions of each core question type as numbered 1-16, following the numbering of training concepts (Table 1). For example, Question 8 tested the concept of the *third-variable problem* for all three question types. Core questions for each posttest were sampled using this numbering, as discussed in the following paragraph.

The core questions were divided into two equal subsets. Each experimental participant completed one of these subsets during the immediate posttest and the other subset during the delayed posttest, with this assignment counterbalanced across participants within each experimental condition. One subset covered even-numbered repeated questions and odd-numbered definitional and transfer questions; the other subset covered odd-numbered repeated questions and even-numbered definitional and transfer questions. Thus for each posttest, repeated questions covered different concepts than did transfer and definitional questions, whereas transfer and definitional questions covered the same concepts. Because definitional and repeated questions were the inverses of each other, this design avoided presenting participants highly similar questions on a given posttest.

Conceptual questions. Conceptual questions consisted of 14 analysis and 14 application questions.¹ The two conceptual question types (i.e., analysis and application) tested abstract principles that were not directly covered or quizzed during training, but which could be inferred with a sufficient conceptual grasp of the training material. Each of these questions contained a detailed description of a hypothetical experiment. Analysis questions required participants to determine which confounding variable(s), if any, were present (Figure 7 shows an example of an analysis question). Application questions required participants to determine how to eliminate confounding variables, if any were present (Figure 8 shows an example of an application question). Half of the analysis and application questions contained confounding variables, and half did not.

¹ Conceptual questions were not included on the first posttest so that we could administer them on the retention test (one week later) to assess participants' comprehension of the materials.

Conceptual questions thus tested the extent to which participants grasped the principles of sound research methodology, internal validity, and true experiments. These topics were chosen because they are of primary importance in research methods courses, and these questions examine the extent to which participants can transfer and apply the knowledge they acquired during training to complex study scenarios. For example, the question in Figure 7 examines whether participants can recognize the specific confounding in the hypothetical study scenario. Such recognition requires a strong grasp of the concepts of confounding, experimental manipulation and control, true experiments, and internal validity. Correctly answering these questions therefore involves more than simply memorizing a given definition. For these reasons, conceptual questions provide a strong measure of concept learning and far transfer.

Procedure. Participants in the experimental training groups participated in two sessions, each lasting a maximum of 55 min. At the start of the study, these participants were told they would be shown slides that contained information about basic scientific principles.

Training session. The training session was partitioned into three sections. Participants were instructed to study each slide carefully, as they would be tested on the material later in the experiment. The study slides were shown one at a time at that the center of the screen. A participant could view the next slide by pressing the right arrow key and the previous slide by pressing the left arrow key. Below each slide, a counter indicated which slide number the participant was viewing out of the total number of slides contained in the section, as well as which section the participant was working on (e.g., *Section 2, Slide 2 out of 4*). Participants could view slides only from the section they were studying and could not move ahead prematurely to the next section or return to a previous section once it was complete.

Navigating each section. Minimum and maximum time limits were implemented for each section, based on the number of slides contained in the section. These time constraints were meant to partially simulate real-world study conditions in which students are required to learn multiple concepts within a limited time frame. In such cases, students must devote a sufficient amount of study towards each concept in order to learn all the concepts, but must also balance the amount of time they allocate towards any single concept. Under such circumstances learners can control the amount of time they spend studying any given concept (as in the present study).

At the start of each section, a prompt showed the participant the number of slides that were contained in the section and the maximum study time that would be allowed. Time limits were set to allow an average study time of 2.5-3.5 min per slide. This range was intended to accommodate a wide spectrum of preferred pacing across different students. Section 1 contained two study slides and Sections 2-3 each contained four. Section 1 ran for 5-7 min, and Sections 2 and 3 ran for 10-14 min each.²

If participants attempted to move past the last slide in a section before the minimum study time had been reached, the screen was cleared and a prompt instructed them to return to the last slide by pressing the spacebar and to continue studying for at least the minimum duration of time that remained in the section. Once a section's minimum study time was reached, the screen was cleared and a prompt was presented that gave the participant the option of exiting the study phase by pressing the enter key or continuing to study and returning to the slide they were previously viewing by pressing the spacebar. If participants elected to continue studying, they could continue

² These time limits were used to accommodate the constraints of running a laboratory experiment.

navigating between slides by pressing the left- and right-arrow keys. Once a section's maximum time was exceeded, the screen was cleared and a prompt instructed the participant to press the spacebar to exit the section and continue to the quiz.

Quiz instructions. After studying each section, participants were given a self-paced rest break and were notified that they would be quizzed on the material that was covered in the section they had just completed. After completing their study of the slides in the first section, all participants were provided specific details on the format of quizzes they were going to be administered. Participants in the recall condition were instructed that they would need to type in a response for each quiz item. Participants in the recognition condition were instructed that they would be given a multiple-choice quiz and would be required to select a response for each quiz item. Participants in the recall-then-recognition condition were instructed that they would be shown two versions of the same question for each quiz item—a fill-in-the-blank version followed by a multiple-choice version—and would need to respond to each accordingly. Additionally, after responding to the first fill-in-the-blank question, these participants were shown a prompt reminding them they would be presented with two versions of each question throughout the quiz. All participants were asked to press the spacebar when they were ready to begin the quiz.

Quiz questions. Quiz questions were presented at the end of each section, which queried the material for that section. Sections 1-3 contained 3, 7, and 6 quiz questions, respectively (one per concept covered). The display for all quiz and posttest questions included a text box, located directly beneath the question, where participants were asked to enter their responses. Each quiz question consisted of a description of a given term, and participants were required to either type the correct term (recall-only, as shown in Figure 2), select the correct term from a list of five multiple-choice options (recognition-only, as shown in Figure 3), or complete both of these tasks

in succession (recall-then-recognition). For each quiz question in the recall-then-recognition condition, the participant was first provided with a fill-in-the-blank form of the question (as in Figure 2), followed by the same question in multiple-choice format (as in Figure 3).

Correct-answer feedback. After typing in a response, participants were required to press the enter key (this was also required for both posttests). Participants were then shown the correct answer at the bottom of the display. In all experimental conditions, only the correct answer was shown; the corresponding letter option was not displayed for multiple-choice items. Thus the feedback was identical in all conditions, matching verbatim the correct alternative from the multiple-choice version of the question. For the recall-then-recognition condition, participants were not shown the correct answer until after they entered their second response, on the multiple-choice version of the question. After being shown the correct answer, participants were asked to press the spacebar when they were ready to move on to the next question. There was a 300-ms interval following the feedback for each question on the quiz (as well as each question on both posttests).

Immediate posttest. All questions in both posttests were presented in multiple-choice format to explicitly test recognition learning, which is a common form of assessment in the classroom. The immediate posttest comprised 24 core questions, which were presented in a random order (different for each participant). After completing the immediate posttest, participants in the experimental conditions were thanked for their participation and reminded that they would be required to return in 7 days.

Delayed posttest. The delayed posttest consisted of 52 questions and followed the same procedure as the immediate posttest. Upon returning, participants in the experimental conditions were notified that they would be tested on the material that was covered in the first session of the

experiment (i.e., the previous week). The delayed posttest was partitioned into two sections. The first section consisted of 24 core questions (the subset not used in that participant's immediate posttest, as explained in the Design and materials section), and the second section consisted of all 28 conceptual questions. The order in which questions were presented within each section was randomized, separately for each participant.

Control group. Participants in the control condition were notified that they would be given a test on basic scientific principles. These participants were only asked to complete a single test, which was identical to the delayed posttest that participants in the experimental conditions completed. The rest of the procedure was identical to the second session that participants in the experimental conditions completed. Because there were two versions of this test, the version that was completed by each control participant was randomly selected, subject to the constraint that half of these participants completed one version and the other half completed the other version.

Results

Nine experimental participants were excluded from the analyses because they did not return for the second posttest (two from the recall-only condition, three from the recall-then-recognition condition, and four from the recognition-only condition), leaving 174 total participants.

It is important to note that core and conceptual questions assessed different aspects of participants' knowledge of the training material. It was possible for participants to correctly answer core questions by directly memorizing the training material. These questions hence provide a direct measure of retention. In contrast, conceptual questions tested participants' conceptual understanding, as they required participants to apply their knowledge of the training material to scenarios that tested these concepts' underlying principles. As a result, it was not possible for

participants to correctly answer conceptual questions just by memorizing the training material. Participants' performance on core and conceptual questions was therefore analyzed separately.

Experimental conditions vs. control condition. First, we examined whether participants in the experimental conditions were able to learn and retain the material they studied during the training session.³ Thus, performance on the core questions (i.e., repeated, definitional, and transfer) was compared between participants in the experimental and control conditions.

Performance on core questions. Figure 9 shows the mean performance on each type of core question for participants in the experimental and control conditions. Performance by participants in the experimental conditions on the immediate posttest exceeded control participants' performance, $M_{\text{experimental-immediate}} = .76$; $M_{\text{control}} = .49$, $t(172) = 8.47$, $p < .001$, $SE = .031$, $d = 1.74$. Experimental participants' delayed posttest performance also exceeded control participants' performance, $M_{\text{experimental-delayed}} = .69$, $t(172) = 6.19$, $p < .001$, $SE = .031$, $d = 1.28$.

Performance on conceptual questions. Furthermore, participants in the experimental conditions ($M = .30$) outperformed participants in the control condition ($M = .23$) on conceptual questions (analysis and application questions), $t(172) = 2.36$, $p = .020$, $SE = .029$, $d = .456$. It is also important to note that participants in the control condition did not perform reliably above chance (20%) on conceptual questions, $t(28) = .977$, $p = .337$, $SE = .029$, $d = .37$, whereas participants in the experimental conditions performed significantly above chance, $t(144) = 8.43$, $p < .001$, $SE = .012$, $d = 1.41$.

³ All reported analyses comparing the experimental and control groups meet the assumption of equal variance, as indicated by Levene's test.

Recall vs. recognition vs. recall-then-recognition. A separate analysis examined whether there were performance differences on core questions among the experimental conditions, and if so, whether such differences depended on the test and question types. This analysis was a mixed-model ANOVA with a between-subjects factor of training condition (recall only vs. recognition only vs. recall-then-recognition) and within-subject factors of question type (repeated questions vs. transfer questions vs. definitional questions) and test (immediate vs. delayed).

The analysis revealed a main effect of test, $F(1, 142) = 34.68, p < .001, MSE = .032, \eta_p^2 = .196$, such that participants performed better on the first posttest than on the second. There was also a main effect of question type, $F(2, 284) = 114.90, p < .001, MSE = .019, \eta_p^2 = .447$, as participants performed best on repeated questions. Additionally, there was an interaction between test and question type, $F(2, 284) = 3.29, p = .039, MSE = .02, \eta_p^2 = .023$, as there was a greater decrease in performance between the first and second posttest for repeated and definitional questions than for transfer questions (as shown in Figure 9). No differences in performance among the experimental conditions were found, and there were no interactions between condition and question or test type (all $ps > .216$, including all least-significant-difference post-hoc comparisons among the experimental conditions). Likewise, no performance differences were found among the experimental conditions on the conceptual questions ($p = .979$). Table 2 shows the mean performance of each experimental group on each of the core question types on the immediate and delayed posttests.

Exploratory analysis. One concern with the analyses contrasting the three experimental conditions is that they may not adequately capture true differences that might exist in conceptual understanding among these groups. Conceptual questions were meant to capture such differences, but the challenging nature of these questions might have obscured the effects of the experimental

manipulation. As noted in the second paragraph of the Results section, repeated and definitional questions could be correctly answered by memorizing the material presented during training, and thus they allowed for an adequate measure of retention but not of conceptual understanding. Although memorization could be used for transfer questions, doing so was more challenging because these questions were presented in novel contexts from what was encountered during training, and therefore required a deeper level of understanding. More specifically, it was necessary for participants to understand these concepts well enough to recognize them in unique scenarios. Transfer questions hence provide the best measure of conceptual understanding among the three core question types.

An exploratory analysis was thus conducted on transfer questions, to further examine whether participants who engaged in recall developed a better understanding and formed more durable memories of the concepts in the study material than did participants who did not engage in recall. Because participants in the recall-only and the recall-then-recognition conditions were asked to engage in recall during training, both groups were combined for this analysis. A mixed-model ANOVA was used to test for an interaction between type of training (between-subjects factor: recall conditions vs. recognition-only) and test type (within-subjects factor: immediate vs. delayed posttests). Comparing the immediate and delayed tests allows for an assessment of participants' retention and conceptual understanding of the study material. Figure 9B shows the mean performance on transfer questions by type of training and type of test. The analysis revealed a significant interaction between condition and test type, $F(1, 143) = 3.97, p = .048, MSE = .026, \eta_p^2 = .027$, as there was less of a decrease in performance between the first and second posttests for participants who engaged in recall ($M_{\text{immediate}} = .657; M_{\text{delayed}} = .647$) than for participants who engaged only in recognition ($M_{\text{immediate}} = .710; M_{\text{delayed}} = .620$). Thus this exploratory analysis

suggests that recall quizzing produced more durable knowledge that was less susceptible to forgetting, at least for the transfer questions, which required more conceptual understanding than the repeated or definitional questions.

Discussion

This paper presents a training paradigm that is built on the principle of retrieval practice. Translating this principle into a real-world paradigm requires addressing multiple facets, such as how much time to allow learners to study a given set of concepts, when to include retrieval practice, what type of retrieval practice to include, and whether to provide participants feedback on their responses. At each of these decision points, we fully specified the translation process by implementing findings from basic experimental psychology, regarding interspersed retrieval practice, different forms of responding, a restricted form of self-paced studying, and correct-answer feedback. To briefly summarize these facets: Participants were allowed to navigate the study slides within each section, permitting them to control which slides they spent more time studying (within the allotted time for each section). Concepts were divided into three sections, and interspersed retrieval practice was used, wherein participants were quizzed at the end of each section. After participants responded to a quiz question they were provided correct-answer feedback.

It is important to note that only form of responding was manipulated among the experimental groups (recall vs. recognition vs. recall-then-recognition), as the implementations of the other learning principles were held constant. Manipulating all of these principles as a unit enables a holistic test of their combined effect, which is more relevant to translation than is the reductionist approach of assessing each principle individually. Moreover, if left unspecified, each of the facets can lead to ambiguity in regards to the translation of retrieval practice to a real-world

paradigm. To avoid this ambiguity impeding translation (Horvath et al., 2013; Oliver & Conole, 2003; Smeyers & Depaepe, 2013; Roediger, 2013), we explicitly specify each facet of our training paradigm and base our decisions for each on the vast literatures on the learning sciences.

This training paradigm was developed with the goal that it might serve as a teaching tool that can be used to enhance student learning. Thus, we were not specifically interested in whether any one of these principles could enhance learning on its own, as each has been shown to do so in the context of the laboratory. Instead, our goal was to examine whether these principles could be translated into a realistic, complex learning system to aid learners in acquiring ecologically valid concepts, which could then be used by instructors in the classroom. Thus, we were interested in whether combining all of these principles into a single intervention would substantively impact performance in a realistic educational learning task. This holistic approach is often appropriate for translational research, because the translation of a given principle involves numerous facets beyond the variables that are manipulated in the laboratory (Horvath et al., 2013; Oliver & Conole, 2003; Smeyers & Depaepe, 2013; Roediger, 2013). With these issues in mind, the training paradigm was constructed in a manner that would allow for instructors to directly apply (in cases where the same concepts as those presented in this study are covered) or easily modify and adapt the paradigm accordingly (changing out the study slides and quiz questions), based on the course curriculum (discussed further below).

The training paradigm was effective in helping participants in the experimental conditions learn the concepts they were taught during training, and moreover these concepts were retained one week later. Importantly, the training paradigm also aided participants in correctly answering conceptual (application and analysis) questions, which required participants to have a thorough understanding of the study material. These question types tested complex scientific principles,

which, as many university professors who have taught a research methods course can affirm, can be extremely difficult for students to learn and retain (as indicated by the control group's chance performance on conceptual questions). Moreover, participants in the experimental conditions were not quizzed on these question types during training and were not tested on them until one week after they completed the training session. Thus, this finding seems to reflect the experimental participants' genuine conceptual understanding of the study material.

Perhaps more important is the extent to which such concepts were learned by participants who received training. On each of the posttests that participants in the experimental conditions completed, they outperformed control participants on core questions by approximately 20%, which amounts to a difference of two full letter grades. Notably, these learning gains were achieved with only a single training session, which consisted of less than an hour of actual training. Furthermore, participants who received training scored approximately 76% and 70% on the core questions in the first and second posttests, respectively, translating to passing letter grades of C and C-. This level of performance is noteworthy given that the amount of training participants in the experimental conditions were given is many orders of magnitude less than the instruction and study time that students in actual statistics and research methods courses receive. Taken together, these findings serve as a powerful demonstration of how the current training paradigm can aid students in acquiring and subsequently retaining complex concepts.

Type of Quizzing Format

Despite the evidence for the strong benefit of the training paradigm overall, performance appeared to be equivalent among the experimental conditions, suggesting that all three quizzing formats are equally effective. It is therefore unclear which format is ideal for presenting quiz questions for this training paradigm. One possibility is that the benefits of recall-based quizzing

were masked by the fact that the question format of the posttests matched that of the quiz questions that were presented to the recognition group. Research on transfer-appropriate processing has shown that test performance is superior when the training and testing conditions are similar (Balota & Neely, 1980). Thus, performance for participants in the recognition-only condition may have been inflated, reducing the performance advantage for the recall conditions. Future work will be required to more directly test this possibility.

An exploratory analysis, which examined whether the decline in performance between the two posttests on transfer questions differed between the recall conditions and the recognition-only condition, suggests that retention and transfer of concepts may have been stronger for participants who engaged in recall. Participants in the recall conditions performed equally well on the transfer questions on both posttests, suggesting that their memory for the concepts that were learned during training was not weakened by the one-week delay between the first and second posttest. In contrast, performance on the transfer questions for participants in the recognition-only condition decreased considerably between the first and second posttests (by approximately 9%), suggesting that their memory of the study material was somewhat tenuous in comparison to that of participants who engaged in recall during training. Thus, instructors who employ this training paradigm may wish to use a version that includes recall responding during quizzing. In the classroom, recall questions can be used during quizzing by asking students to write out their response to a given quiz question

and then showing students the correct response (similarly to the type of feedback used in our paradigm⁴).

Guide and Implications for Instructors

Instructors who wish to use this paradigm to train students on different content (e.g., physics, chemistry, mathematics) can do so by simply following our training procedure (discussed above in the Method section), and replacing our slides and quiz questions with those that correspond to the topic of interest. In this process, we recommend creating training slides that are concise and devoid of superfluous information, so that the slides fully and clearly explain all of the concepts that are introduced. Additionally, in cases where the training content builds on concepts that were introduced in earlier slides, we suggest presenting slides in a manner that follows a conceptual progression.

One area that instructors might wish to deviate from our training procedure is in the amount of time that students are permitted to study a given slide. Here, participants' study time was limited (although participants were given some autonomy in the amount of time they could spend studying) due to the time restrictions of the laboratory experiment. However, based on the principles of self-pacing (Ariel, 2013; de Jonge et al., 2015; Tullis & Benjamin, 2011), it might be more useful to allow participants full control over how much time they spend studying a given slide. On the other hand, one issue that this approach introduces is that some students might not spend a sufficient amount of time studying a given slide. Thus, it might be wise to keep a minimum

⁴ Instructors might also consider using more complex forms of feedback that encourage students to think carefully about the material, such as explanation feedback, wherein the correct answer is coupled with a detailed explanation (Butler, Godbole, & Marsh, 2013; Corral & Carpenter, 2019).

study time in place for any given set of slides, but provide participants the ability to advance to the next set of slides once the minimum time has been reached.

Furthermore, as in our paradigm, we recommend that instructors quiz students on any concepts that are presented in the training slides. It is important to note that our quiz questions were presented in an abstract format so we could directly test participants' ability to transfer their knowledge to novel scenarios during testing. This aspect of the training paradigm was thus implemented for reasons of experiment design, and instructors may or may not wish to adopt a similar approach.

We also recommend that instructors implement an immediate posttest after training in order to assess how well participants are able to learn and retain the training material. This type of assessment can be particularly useful in helping both the student and instructor identify the aspects of the material that the student does not yet fully grasp. One noteworthy finding is that participants performed best on repeated questions, which were identical to the questions that were quizzed, and worst on transfer questions. However, performance decreased substantially on repeated (and definitional) questions between the first and second posttest, whereas performance was relatively stable for transfer questions. One reason for this finding might be that rote memorization could be used to answer repeated (and definitional) questions, but transfer questions required conceptual understanding. Thus, when participants were given the second test one week later, they may have forgotten the information that was memorized during training. In contrast, because performance on transfer questions might have been driven by conceptual understanding, as opposed to rote memorization, performance on these questions might have been more stable. These findings and explanation are in line with work on levels of processing (Craik & Lockhart, 1972; Craik & Tulving, 1975), wherein information that is processed in a deeper and more meaningful manner

(e.g., information that is comprehended by the learner) is more robust to decay than information that is learned through rote memorization (Symons & Thompson, 1997).

This explanation suggests that transfer items can better assess students' knowledge than items that can be answered through rote memorization. Moreover, training performance on the latter type of items might lead both students and instructors to form an inaccurate perception of the student's actual understanding of the tested content. This misperception can be problematic in cases where pretests are used to help prepare students for an upcoming exam, as students might develop a false sense of security due to their high performance on the items that were memorized during study or training. Consequently, students might reduce their study time, leaving them ill-prepared for an exam. The findings presented here therefore have direct implications for instructors who use clicker questions or pretests to assess their students' knowledge of course material. Our findings suggest that any such assessments should incorporate transfer-like questions, which are fairly similar to the type of test questions that instructors often use on exams.

Lastly, although instructors can use this training paradigm during lecture, it can also be applied outside of the classroom. For instance, our training paradigm can be implemented as an automated tutoring system that is made available to students. This option would allow students autonomy over when they study, and also provide them a structured and controlled training environment outside of the classroom. Our training paradigm might also be particularly well-suited for classroom laboratory courses (e.g., research methods, statistics), in which students are often required to complete assignments independently within a given time period (typically 1-3 hours). This context is highly similar to what participants in the experimental conditions encountered, and thus students in laboratory courses might greatly benefit from a training paradigm like the one used in the present study.

Limitations and Future Directions

From a translational and applied perspective, the implementation of multiple learning principles within a single training paradigm is a particular strength of this paper. However, a limitation of this approach from a theoretical perspective is that we did not isolate and test each of these principles. Thus, we do not know the extent to which each of these principles affected learning, as we examined only their combined impact. Nevertheless, a researcher or instructor might be interested in this question. Thus, a potential direction for future work is to methodically vary which facets are included in the paradigm and compare those conditions to the full paradigm (e.g., full paradigm vs. paradigm without correct-answer feedback or full paradigm vs. paradigm without retrieval practice).

One potential critique of the present study is that the control condition did not receive any instruction, and thus these results might be taken to demonstrate that the training paradigm merely leads to better learning than not receiving training at all. However, as we state above, the materials used in this study were highly complex (particularly the conceptual question types) and it is by no means a given that they can be readily acquired, even with extensive training. Indeed, as many research method instructors will likely attest, there are numerous students who fail to learn these exact concepts over an entire semester of rigorous instruction. Moreover, many training procedures fail to produce learning whatsoever, as is exemplified in studies where participants in some conditions perform at chance (e.g. Johnstone & Shanks, 2001; Quinn, Palmer, & Slater, 1999; Shanks, Johnstone, & Staggs, 1997). Thus, demonstrating that this training paradigm benefits complex learning is a critical first step to the present work.

Nevertheless, an instructor might certainly be interested in the extent to which this training paradigm benefits learning above and beyond simply studying the materials. One way to answer

this question in future work would be to provide one group of participants the full training paradigm and another group the PowerPoint-style slides for study. Another potential future direction is to examine how this paradigm might fare in comparison to how students typically study. Recent work suggests that students use suboptimal study strategies (Corral et al., 2019), and given that the training paradigm used here is premised on well-established learning principles, we would predict learning to be better for students who use the training paradigm than for those who receive the same study materials and are left to their own devices. To build on this idea, a particularly strong test of this paradigm's efficacy might be to select students in a course who are struggling (e.g., students with a letter grade of C- or lower) and randomly assign them to complete the training paradigm or to continue to study using their preferred method. These students' progress could also be monitored throughout the semester to examine whether the benefits of the training paradigm are observed over an extended period.

Conclusion

Translating basic and theoretical research towards real-world applications can be challenging (Woolf, 2008) and often fails to occur in the fields of cognitive psychology and education. One reason for this failure is that many cognitive psychology studies require participants to learn artificial concepts, which can make instructors skeptical of how well a given effect will transfer to the classroom. The current study lays out a blueprint for how principles from cognitive psychology, specifically the testing effect, form of responding, self-paced studying, and feedback, can be integrated in order to construct a valuable training paradigm. Furthermore, we have demonstrated the efficacy and applicability of this training paradigm with ecologically valid learning materials. These materials covered various core concepts of the scientific method, and quiz and posttest items were similar in structure and difficulty to exam questions that are typically

presented to students in a university-level research methods course. The findings for the current study are thus applicable to educators from a wide range of scientific domains. However, the current project takes only a small step towards utilizing cognitive psychology to aid students with the learning of real academic concepts. If the translation of cognitive psychology principles is to improve in the domain of education, future work must carefully demonstrate the efficacy of such principles with real academic concepts.

References

- Anderson, J. R., & Bower, G. H. (1972). Recognition and retrieval processes in free recall. *Psychological Review*, 79, 97-123. <https://doi.org/10.1037/h0033773>
- Anderson, L. S., Healy, A. F., Kole, J. A., & Bourne, L. E., Jr. (2011). Conserving time in the classroom: The clicker technique. *Quarterly Journal of Experimental Psychology*, 64, 1457-1462. <https://doi.org/10.1080/17470218.2011.593264>
- Anderson, L. S., Healy, A. F., Kole, J. A., & Bourne, L. E., Jr. (2013). The clicker technique: Cultivating efficient teaching and successful learning. *Applied Cognitive Psychology*, 27, 222-234. <https://doi.org/10.1002/acp.2899>
- Ariel, R. (2013). Learning what to learn: The effects of task experience on strategy shifts in the allocation of study time. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 39, 1697-1711. <https://doi.org/10.1037/a0033091>
- Balota, D. A., & Neely, J. H. (1980). Test-expectancy and word-frequency effects in recall and recognition. *Journal of Experimental Psychology: Human Learning and Memory*, 6, 576-587. <https://doi.org/10.1037/0278-7393.6.5.576>
- Benassi, V., Overson C., & Hakala, C. (Eds.). (2014). *Applying science of learning in education: Infusing psychological science into the curriculum*. Washington, DC: American Psychological Association.
- Bjork, E. L., Little, J. L., & Storm, B. C. (2014). Multiple-choice testing as a desirable difficulty in the classroom. *Journal of Applied Research in Memory and Cognition*, 3, 165-170. <https://doi.org/10.1016/j.jarmac.2014.03.002>

- Brame, C. J., & Biel, R. (2015). Test-enhanced learning: The potential for testing to promote greater learning in undergraduate science courses. *CBE—Life Sciences Education*, 14 (2, es4), 1–12. <https://doi.org/10.1187/cbe.14-11-0208>
- Butler, A. C., Black-Meier, 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, 433-446. <https://dx.doi.org/10.1037/xap0000142>
- Butler, A., Godbole, N., & Marsh, E. (2013). Explanation feedback is better than corrective feedback for promoting transfer of learning. *Journal of Educational Psychology*, 105, 290–298. <https://dx.doi.org/10.1037/a0031026>
- Butler, A. C., Karpicke, J. D., & Roediger, H. L. (2007). The effect of type and timing of feedback on learning from multiple-choice tests. *Journal of Experimental Psychology: Applied*, 13, 273-281. <https://doi.org/10.1037/1076-898X.13.4.273>
- Carpenter, S. K., Pashler, H., & Cepeda, N. J. (2009). Using tests to enhance 8th grade students' retention of U. S. history facts. *Applied Cognitive Psychology*, 23, 760–771. <https://doi.org/10.1002/acp.1507>
- Carpenter, S. K., Rahman, S., Lund, T. J. Armstrong, P. I., Lamm, M. H., Reason, R. D., Coffman, C. R., (2017). Students' use of optional online Reviews and its relationship to summative assessment outcomes in introductory biology. *CBE Life Sciences Education*, 16, 1–9. <https://doi.org/10.1187/cbe.16-06-0205>
- Carpenter, S. K., & Yeung, K. L. (2017). The role of mediator strength in learning from retrieval. *Journal of Memory & Language*, 92, 128-141. <https://dx.doi.org/10.1016/j.jml.2016.06.008>

- Carrier, M., & Pashler, H. (1992). The influence of retrieval on retention. *Memory and Cognition*, 20, 632-642. <https://doi.org/10.3758/bf03202713>
- Chapman, M. J., Healy, A. F., & Koe, J. A. (2016). Memory load as a cognitive antidote to performance decrements in data entry. *Memory*, 24, 1182-1196. <https://dx.doi.org/10.1080/09658211.2015.1086380>
- Corral, D. (2017). *A dual model of relational concept representation* (Unpublished doctoral dissertation). University of Colorado Boulder, Boulder, CO.
- Corral, D., & Carpenter, S. K. (2019). *Facilitating transfer through incorrect examples and explanatory feedback*. Under review.
- Corral, D., Carpenter, S. K., Perkins, K. M., & Gentile, D. A. (2019). *Assessing students' use of optional online reviews*. Under review.
- Corral, D., Kurtz, K. J. & Jones, M. (2018). Learning relational concepts from within- vs. between-category comparisons. *Journal of Experimental Psychology: General*, 147, 1571-1596. <https://dx.doi.org/10.1037/xge0000517>
- Corral, D., & Jones, M. (2017). Learning relational concepts through unitary-versus compositional-based representations. In G. Gunzelmann, A. Howes, T. Tenbrink, & E. Davelaar (Eds.), *In Proceedings of the 39th annual meeting of the cognitive science society* (pp. 1830–1835). Austin, TX: Cognitive Science Society.
- de Jonge, M., Tabbers, H. K., Pecher, D., Jang, Y., & Zeelenberg, R. (2015). The efficacy of self-paced study in multitrial learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 41, 851-858. <https://doi.org/10.1037/xlm0000046>
- Dunlosky, J., Rawson, K. A., Marsh, E. J, Nathan, M. J., & Willingham, D. T. (2013). Improving students' learning with effective learning techniques: Promising directions from cognitive

- and educational psychology. *Psychological Science in the Public Interest*, 14, 4–58.
<https://dx.doi.org/10.1177/1529100612453266>
- Eglington, L. G., & Kang, S. H. K. (2018). Retrieval practice benefits deductive inference. *Educational Psychology Review*, 30, 215-228.
<https://dx.doi.org/10.1007/s10648-016-9386-y>
- Healy, A. F., Jones, M., Lalchandani, L., & Tack, L. A. (2017). Timing of quizzes during learning: Effects on motivation and retention. *Journal of Experimental Psychology: Applied*, 23, 128-137. <https://dx.doi.org/10.1037/xap0000123>
- Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term recognition and recall. *Journal of Verbal Learning and Verbal Behavior*, 10, 562-567.
[https://doi.org/10.1016/s0022-5371\(71\)80029-4](https://doi.org/10.1016/s0022-5371(71)80029-4)
- Horvath, J. C., Lodge, J. M., & Hattie, J. (2017). *From the laboratory to the classroom: Translating science of learning for teachers*. New York, NY, US: Routledge/Taylor & Francis Group.
- Jacoby, L. L. (1991). A process dissociation framework: Separating automatic from intentional uses of memory. *Journal of Memory and Language*, 30, 513–541.
[https://dx.doi.org/10.1016/0749-596X\(91\)90025-F](https://dx.doi.org/10.1016/0749-596X(91)90025-F)
- Johnstone, T., & Shanks, D.R. (2001). Abstractionist and processing accounts of implicit learning. *Cognitive Psychology*, 42, 61–112. <http://dx.doi.org/10.1006/cogp.2000.0743>
- Kahana, M. J. (2012). *Foundations of human memory*. New York: Oxford University Press.
- Kang, S. H. K., Gollan, T. H., Pashler, H. (2013). Don't just repeat after me: Retrieval practice is better than imitation for foreign vocabulary learning. *Psychonomic Bulletin and Review*, 20, 1259-1265. <https://doi.org/10.3758/s13423-013-0450-z>

- Kang, S. H. K., & Pashler, H. (2014). Is the benefit of retrieval practice modulated by motivation? *Journal of Applied Research in Memory and Cognition*, 3, 183-188. <https://doi.org/10.1016/j.jarmac.2014.05.006>
- Karpicke, J. D., & Roediger, H. L. (2008). The critical importance of retrieval for learning. *Science*, 319, 966-968. <https://doi.org/10.1126/science.1152408>
- Kintsch, W. (1970). Models for free recall and recognition. In D. A. Norman (Ed.), *Models of human memory* (pp. 331-373). New York: Academic Press. <https://doi.org/10.1016/b978-0-12-521350-9.50016-4>
- Kole, J. A., Healy, A. F., & Bourne, L. E. Jr., (2008). Cognitive complications moderate the speed-accuracy tradeoff in data entry: A cognitive antidote to inhibition. *Applied Cognitive Psychology*, 22, 917-937. <https://dx.doi.org/10.1002/acp.1401>
- Lehman, M., & Karpicke, J. D. (2016). Elaborative retrieval: Do semantic mediators improve memory? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 42, 1573-1591. <https://dx.doi.org/10.1037/xlm0000267>
- Little, J. L., & Bjork, E. L. (2015). Optimizing multiple-choice tests as tools for learning. *Memory & Cognition*, 43, 14-26. <https://doi.org/10.3758/s13421-014-0452-8>
- Marsh, E. J., Roediger, H. L., Bjork, R. A., & Bjork, E. L. (2007). The memorial consequences of multiple-choice testing. *Psychonomic Bulletin & Review*, 6, 194-199. <https://doi.org/10.3758/bf03194051>
- Mayer, R. E., Stull, A., DeLeeuw, K., Almeroth, K., Bimber, B., Chun, D., ... Zhang, H. (2009). Clickers in college classrooms: Fostering learning with questioning methods in large lecture classes. *Contemporary Educational Psychology*, 34, 51-57. <https://doi.org/10.1016/j.cedpsych.2008.04.002>

- McDaniel, M. A., Anderson, J. L., Derbish, M. H., & Morrisette, N. (2007). Testing the testing effect in the classroom. *European Journal of Cognitive Psychology*, 19, 494-513. <https://doi.org/10.1080/09541440701326154>
- Oliver, M., & Conole, G. (2003). Evidence-based practice in e-learning and higher education: Can we and should we? *Research Papers in Education*, 18, 385-397.
- Pan, S. C., & Rickard, T. C. (2018). Transfer of test-enhanced learning: Meta-analytic review and synthesis. *Psychological Bulletin*, 144, 710-756. <https://dx.doi.org/10.1037/bul0000151>
- Pashler, H., Cepeda, N., Wixted, J., & Rohrer, D. (2005). When does feedback facilitate learning of words? *Journal of Experimental Psychology: Learning, Memory, and Cognition*, 31, 3-8. <https://doi.org/10.1037/0278-7393.31.1.3>
- Peterson, D. J., & Wissman, K. T. (2018). The testing effect and analogical problem-solving. *Memory*, 26, 1460-1466. <https://dx.doi.org/10.1080/09658211.2018.1491603>
- Pyc, M. A., & Rawson, K. A. (2010). Why testing improves memory: Mediator effectiveness hypothesis. *Science*, 333, 335. <https://doi.org/10.1126/science.1191465>
- Quinn, P. C., Palmer, V., & Slater, A. M. (1999). Identification of gender in domestic-cat faces with and without training: Perceptual learning of a natural categorization task. *Perception*, 28, 749-763. <https://doi.org/10.1068/p2884>
- Rickard, T. C., & Pan, S. C. (2018). A dual memory theory of the testing effect. *Psychonomic Bulletin & Review*, 25, 847-869. <https://dx.doi.org/10.3758/s13423-017-1298-4>
- Roediger, H. L. III. (2013). Applying cognitive psychology to education: Translational educational science. *Psychological Science in the Public Interest*, 14, 1-3. <https://dx.doi.org/10.1177/1529100612454415>

- Roediger, H. L., III, & Butler, A. C. (2011). The critical role of retrieval practice in long-term retention. *Trends in Cognitive Sciences*, 15, 20-27. <https://doi.org/10.1016/j.tics.2010.09.003>
- Roediger, H. L. & Karpicke, J. D. (2006a). Test-enhanced learning: Taking memory tests improves long-term retention. *Psychological Science*, 17, 249-255. <https://doi.org/10.1111/j.1467-9280.2006.01693.x>
- Roediger, H. L. & Karpicke, J. D. (2006b). The power of testing memory: Basic research and implications for educational practice. *Perspectives on Psychological Science*, 1, 181-210. <https://doi.org/10.1111/j.1745-6916.2006.00012.x>
- Rowland, C. A. (2014). The effect of testing versus restudy on retention: A meta-analytic review of the testing effect. *Psychological Bulletin*, 140, 1432–1463. <https://doi.org/10.1037/a0037559>
- Shanks, D. R., Johnstone, T., & Staggs, L. (1997). Abstraction processes in artificial grammar learning. *The Quarterly Journal of Experimental Psychology A: Human Experimental Psychology*, 50A, 216-252. <https://doi.org/10.1080/713755680>
- Smeyers, P., & Depaepe, M. (2013). Making sense of the attraction of psychology: On the strengths and weaknesses for education and educational research. In P. Smeyers & M. Depaepe (Eds.) *Educational research: The attraction of psychology* (pp. 1–10). Dordrecht, The Netherlands: Springer.
- Tullis, J. G., & Benjamin, A. S. (2011). On the effectiveness of self-paced learning. *Journal of Memory and Language*, 64, 109–118. <https://doi.org/10.1016/j.jml.2010.11.002>

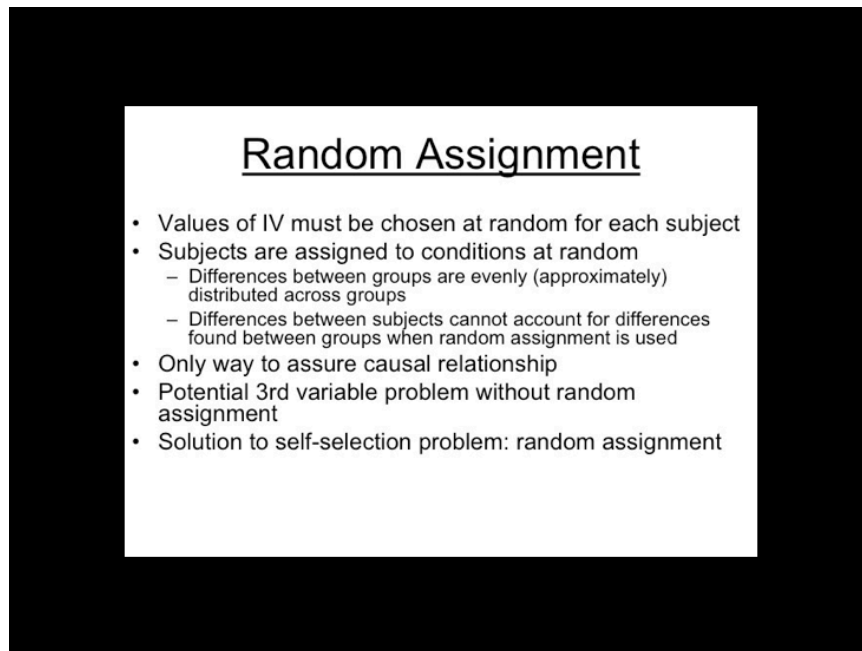
- Tran, R., Rohrer, D., & Pashler, H. (2015). Retrieval practice: The lack of transfer to deductive inferences. *Psychonomic Bulletin & Review*, 22, 135–140.
<https://dx.doi.org/10.3758/s13423-014-0646-x>
- Van Gog, T., & Kester, L. (2012). A test of the testing effect: Acquiring problem-solving skills from worked examples. *Cognitive Science*, 36, 1532-1541.
<https://dx.doi.org/10.1111/cogs.12002>
- Vojdanoska, M., Cranney, J., & Newell, B. R. (2010). The testing effect: The role of feedback and collaboration in a tertiary classroom setting. *Applied Cognitive Psychology*, 24, 1183–1195. <https://doi.org/10.1002/acp.1630>
- Wissman, K. T., Zang, A., & Rawson, K.A. (2018). When does practice testing promote transfer on deductive reasoning tasks? *Journal of Applied Research in Memory and Cognition*.
- Woolf, S. H. (2008). The meaning of translational research and why it matters. *The Journal of the American Medical Association*, 222, 211-213. <https://doi.org/10.1001/jama.2007.26>

Table 1. A complete list of the concepts and the order in which they were covered in training.

Section 1 (Slides 1-2)	Section 2 (Slides 3-6)	Section 3 (Slides 7-10)
1. Variables	4. Non-Experimental Study	11. Independent and Dependent Variables
2. Hypothesis	5. Causal Inference	12. Experimental Control
3. Experimental Study	6. Correlation	13. Confounds
	7. Reverse Causation	14. Random Assignment
	8. Third Variable Problem	15. Quasi-Independent Variables
	9. Self-Selection	16. Addressing Confounds
	10. Manipulation	

Table 2. Mean performance and standard deviations (displayed in the parentheses) for each experimental group on each of the core question types for each posttest.

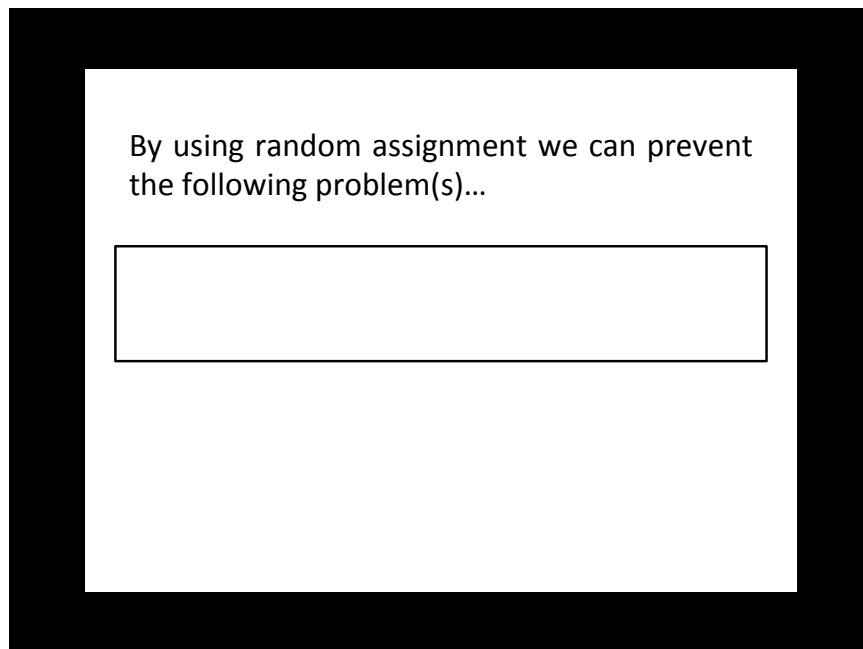
	Immediate Posttest			Delayed Posttest		
	Repeated	Definitional	Transfer	Repeated	Definitional	Transfer
Recall	.842 (.18)	.732 (.22)	.658 (.21)	.753 (.18)	.635 (.22)	.660 (.20)
Recognition	.899 (.16)	.747 (.19)	.710 (.21)	.790 (.17)	.650 (.19)	.620 (.23)
Recall-then- recognition	.872 (.13)	.747 (.19)	.656 (.22)	.789 (.17)	.694 (.18)	.635 (.23)



Random Assignment

- Values of IV must be chosen at random for each subject
- Subjects are assigned to conditions at random
 - Differences between groups are evenly (approximately) distributed across groups
 - Differences between subjects cannot account for differences found between groups when random assignment is used
- Only way to assure causal relationship
- Potential 3rd variable problem without random assignment
- Solution to self-selection problem: random assignment

Figure 1. Study Slide 9 from Section 3 of the training session.



By using random assignment we can prevent the following problem(s)...

Figure 2. An example fill-in-the-blank quiz question. The correct response is *self-selection*.

By using random assignment we can prevent the following problem(s)...

- a. Self-selection
- b. Researcher expectancy
- c. Reverse correlation
- d. Poor experimental control
- e. A and D

Figure 3. An example multiple-choice quiz question. The correct response is option *a*.

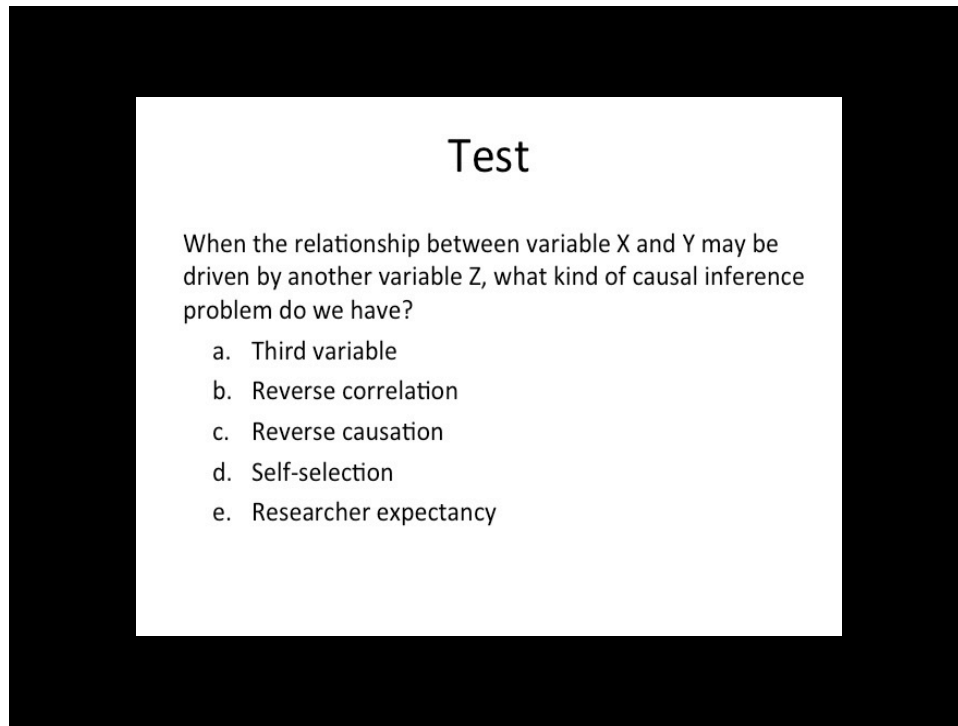


Figure 4. An example from the repeated question type (identical to the recognition version of questions given during training). The correct response is option *a*.

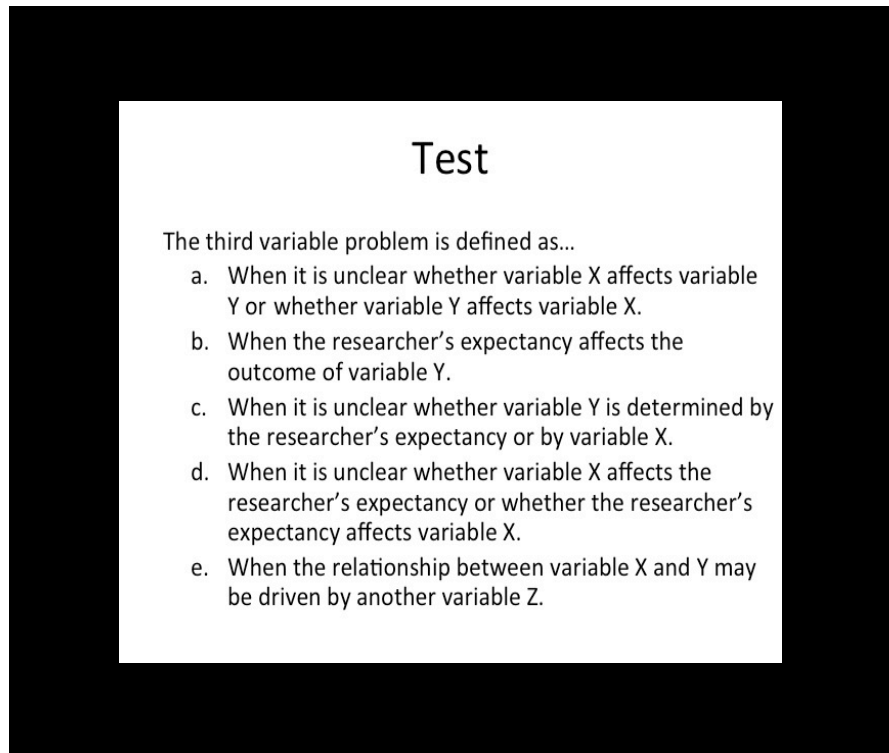
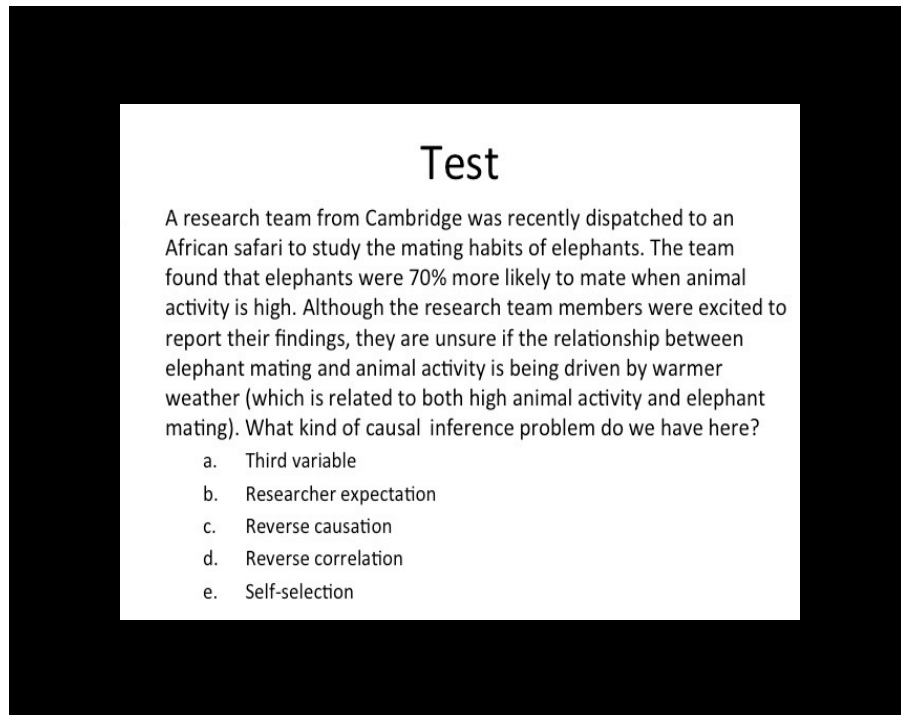


Figure 5. An example item from the definitional question type. The correct response is option *e*.



Test

A research team from Cambridge was recently dispatched to an African safari to study the mating habits of elephants. The team found that elephants were 70% more likely to mate when animal activity is high. Although the research team members were excited to report their findings, they are unsure if the relationship between elephant mating and animal activity is being driven by warmer weather (which is related to both high animal activity and elephant mating). What kind of causal inference problem do we have here?

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

Figure 6. An example item from the transfer question type. The correct response is option *a*.

Test

Arlene is testing how business attire affects generosity. Subjects were each paid \$100 for participating and were assigned to one of two conditions. In one condition subjects wore business attire; in the other condition subjects wore casual attire. Subjects were then asked to make donations to charity. The study ran for a month (30 days). Data for subjects in the business attire condition were collected during days 1-15. Data for subjects in the casual attire condition were collected in days 16-30. Subjects in the business attire condition were more charitable than subjects in the casual attire condition. Arlene concludes that wearing business attire makes people more charitable. Is there anything wrong with Arlene's conclusion?

- a. Yes, the type of attire worn by subjects is confounded by subjects' being paid.
- b. Yes, the type of attire worn by subjects is confounded with the time of the month that the data for subjects in each group was collected.
- c. Yes, it is unclear whether wearing business attire increases generosity or whether being more generous makes people wear more business attire.
- d. Yes, this is not a true experiment and we cannot make any causal inferences about the data.
- e. No, this is an experiment with no confounding variables, allowing us to draw a causal inference from the results.

Figure 7. An example item from the analysis question type. The correct response is option *b*.

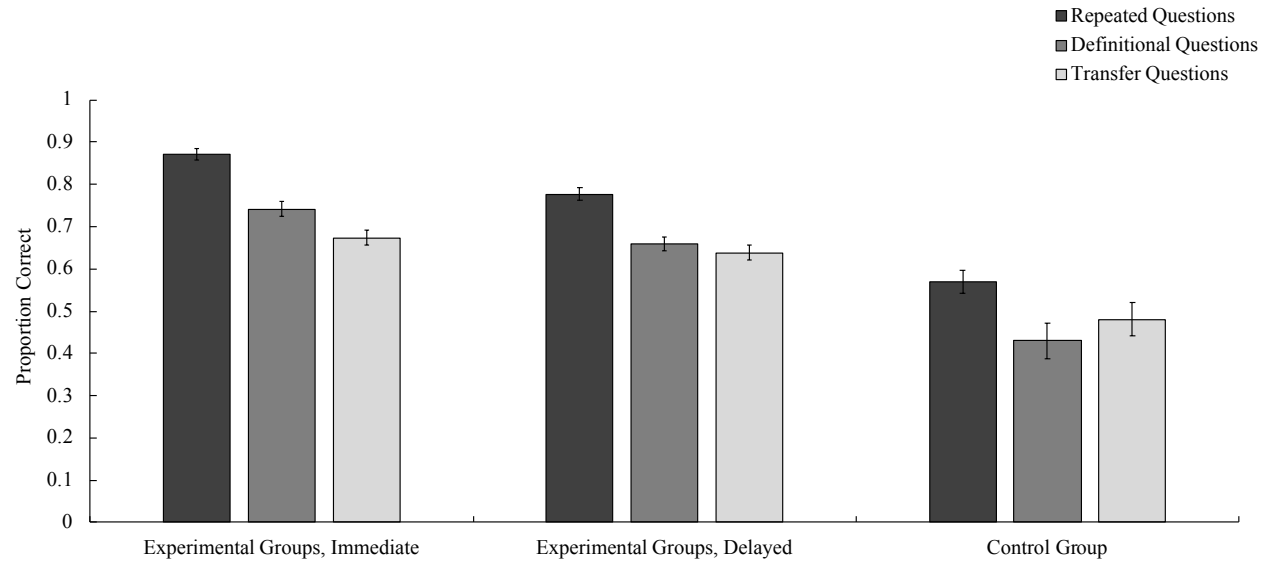
Test

Shaina is examining how memory is affected by a new memory pill. One group of subjects gets the memory pill; the other group gets a sugar pill. All subjects are then given a memory test. The first half of the subjects who sign up for the study are placed in the memory pill condition; the second half of subjects are placed in the sugar pill condition. How should Shaina change her study in order to avoid any confounding variables?

- a. A memory test needs to be administered the day before the experiment.
- b. A memory test needs to be administered the day after the experiment.
- c. Subjects must be randomly sampled from the population.
- d. Subjects must be randomly assigned to conditions.
- e. There are no confounding variables in this experiment and nothing needs to be changed.

Figure 8. An example item from the application question type. The correct response is option *d*.

A.



B.

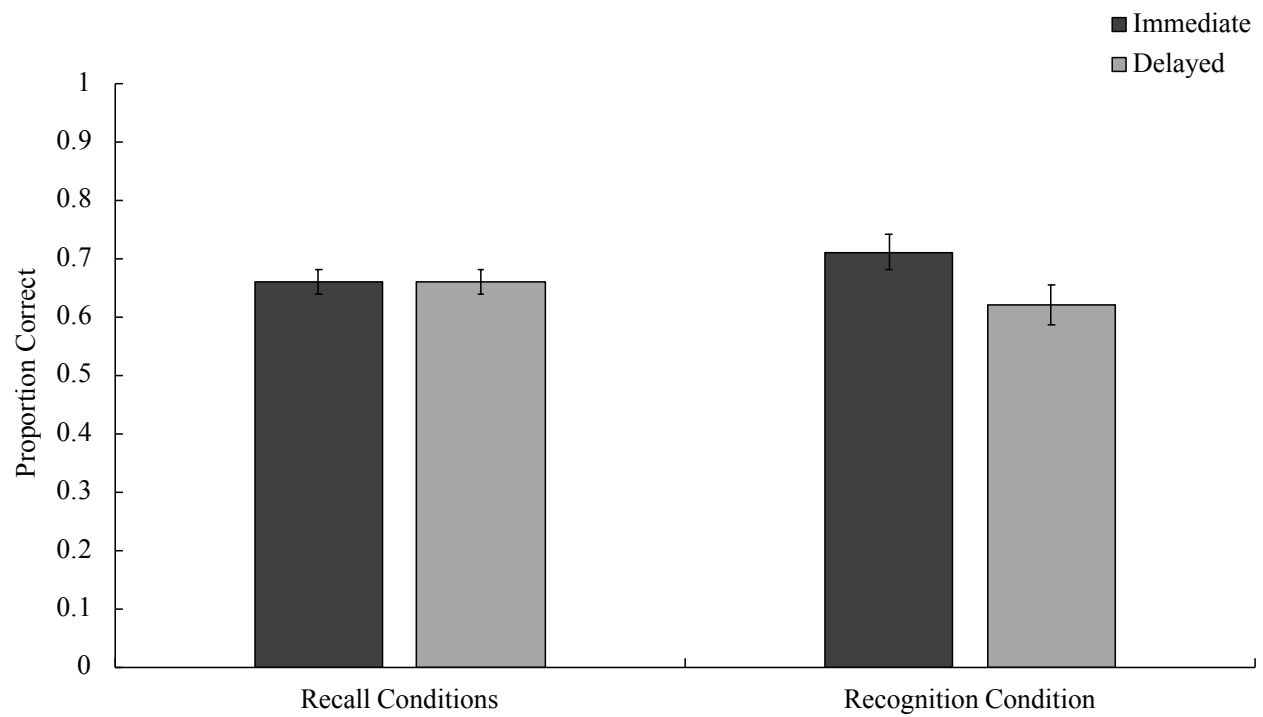


Figure 9. Panel A shows the experimental and control groups' mean performance on each type of core question (repeated, definitional, and transfer questions) for each posttest. Panel B shows mean performance for transfer questions on each posttest for the recall conditions (recall condition and

recall-then-recognition condition) and the recognition condition. Error bars indicate standard errors of the mean.